Stats300b Problem Set 4
Due: Thursday, February 11 at 5:00pm on Gradescope

**Question 7.6** (Smallest eigenvalue of a random, possibly heavy-tailed matrix)**:** Let $X_i$ be i.i.d. $\mathbb{R}^d$-valued random vectors, mean zero, where $\mathrm{Cov}(X_i) = \Sigma$ for a positive definite $\Sigma$. Assume also that $\mathbb{E}[|\langle v, X\rangle|] \geq \kappa\sqrt{v^T\Sigma v}$ for any vector $v \in \mathbb{R}^d$, where $\kappa > 0$ is a constant.

(a) Show that for any vector $v \in \mathbb{R}^d$,

$$\mathbb{P}\left(|\langle v, X\rangle| \geq \frac{\kappa}{2}\sqrt{v^T\Sigma v}\right) \geq \frac{\kappa^2}{4}.$$

(b) Let $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ denote the empirical second-moment matrix of the $X_i$, and for a symmetric matrix $A$, let

$$\lambda_{\min}(A) := \inf_v \left\{v^T A v \mid v \in \mathbb{S}^{d-1}\right\}$$

denote the minimum eigenvalue of $A$, where $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ denotes the sphere in $\mathbb{R}^d$. Show that there exist constants $C_1, C_2, C_3 \in (0, \infty)$, which may depend on $\kappa$, such that

$$\mathbb{P}\left(\lambda_{\min}(\widehat{\Sigma}_n) \geq \left(C_1 - C_2\sqrt{\frac{d}{n}} - C_3 t\right)_+ \lambda_{\min}(\Sigma)\right) \geq 1 - e^{-nt^2}$$

for all $t \geq 0$.

**Answer:**

(a) We use the Paley-Zygmund inequality (Question 1.11), which states that

$$\mathbb{P}(|\langle v, X\rangle| \geq \theta\mathbb{E}[|\langle v, X\rangle|]) \geq (1-\theta)^2 \frac{\mathbb{E}[|\langle v, x\rangle|]^2}{v^T\Sigma v - \theta(2-\theta)\mathbb{E}[|\langle v, X\rangle|]^2}.$$

Using that $\mathbb{E}[|\langle v, x\rangle|] \geq \kappa\sqrt{v^T\Sigma v}$ for a constant $\kappa > 0$ and setting $\theta = \frac{1}{2}$ above, we have

$$\mathbb{P}(|\langle v, X\rangle| \geq \frac{\kappa}{2}\|v\|_\Sigma) \geq \frac{1}{4}\frac{\kappa^2 v^T\Sigma v}{v^T\Sigma v - (3/4)\kappa v^T\Sigma v} = \frac{\kappa^2}{4 - 3\kappa}.$$

(b) Let $\|v\|_\Sigma^2 = v^T\Sigma v$ for shorthand, and recall that $\lambda_{\min}(\Sigma) = \inf_{v\in\mathbb{S}^{d-1}}\sqrt{v^T\Sigma v}$. The set of half-planes in $\mathbb{R}^d$ has VC-dimension at most $d+1$, while $\|v\|_\Sigma \geq \lambda_{\min}$ for all $v \in \mathbb{S}^{d-1}$. Thus if we define the random variable

$$B_i(v) = \mathbf{1}\left\{\langle v, X_i\rangle \geq \frac{\kappa}{2}\lambda_{\min}\right\} + \mathbf{1}\left\{\langle v, X_i\rangle \leq -\frac{\kappa}{2}\lambda_{\min}\right\},$$

then

$$\mathbb{E}[B_i(v)] \geq \frac{\kappa^2}{4 - 3\kappa}$$

by the first part of the question. Using the VC-dimension bounds from class, for a numerical constant $C$, we have

$$\mathbb{P}\left(\exists\, v \in \mathbb{S}^{d-1} \text{ s.t. } \frac{1}{n}\sum_{i=1}^n B_i(v) - \mathbb{E}[B_i(v)] \leq -C\sqrt{\frac{d}{n}} - t\right) \leq e^{-2nt^2}.$$

Written differently,

$$\mathbb{P}\left(\exists\, v \in \mathbb{S}^{d-1} \text{ s.t. } \frac{1}{n}\operatorname{card}(\{i \in [n] \mid \langle v, X_i\rangle^2 \ge \kappa^2 \lambda_{\min}(\Sigma)/4\}) \le \frac{\kappa^2}{4-3\kappa} - C\sqrt{\frac{d}{n}} - t\right) \le e^{-2nt^2}.$$

On the complement of the event within the probability above, we have

$$v^T \widehat{\Sigma}_n v \ge \left(\frac{\kappa^2}{4-3\kappa} - C\sqrt{\frac{d}{n}} - t\right)_+ \cdot \frac{\kappa^2}{4}\lambda_{\min}(\Sigma)$$

for all $v \in \mathbb{R}^d$.

□

**Question 7.8** (Covering numbers for low-rank matrices)**:** Let $\mathcal{M}_{r,d}$ be the collection of rank $r$ matrices $A \in \mathbb{R}^{d \times d}$ with $\|A\|_{\text{Fr}} = 1$, where we recall that the Frobenius norm $\|A\|_{\text{Fr}}^2 = \sum_{i,j} A_{ij}^2 = \operatorname{tr}(A^T A)$ is the usual Euclidean norm applied to the entries of $A$. Show that the covering numbers $N(\mathcal{M}_{r,d}, \|\cdot\|_{\text{Fr}}, \epsilon)$ of $\mathcal{M}_{r,d}$ in the Frobenius norm satisfy

$$\log N(\mathcal{M}_{r,d}, \|\cdot\|_{\text{Fr}}, \epsilon) \le 2rd \log\left(1 + \frac{4r}{\epsilon}\right).$$

*Hint:* Our solution uses the singular value decomposition that $A = U\Sigma V^T = \sum_{i=1}^r u_i \sigma_i v_i^T$, where $\Sigma \succeq 0$ is diagonal and $U = [u_1 \ \cdots u_r]$ and $V = [v_1 \ \cdots v_r] \in \mathbb{R}^{d \times r}$ are orthogonal, i.e., $U^T U = I_r$ and $V^T V = I_r$. *Note:* It is possible to get slightly sharper bounds than these, but we won't worry about that.

**Answer:** We use the hint, that is, that any matrix $A \in \mathcal{M}_{r,d}$ has a singular value decomposition $A = U\Sigma V^T$. Let $B = WSQ^T$ be the singular value decomposition of $B$, so that $A = \sum_{i=1}^r u_i \sigma_i v_i^T$ and $B = \sum_{i=1}^r w_i s_i q_i^T$, where we note that $\sum_{i=1}^r \sigma_i^2 \le 1$ and $\sum_{i=1}^r s_i^2 \le 1$ by construction. Then for any such $A, B$ we have

$$\begin{aligned}
\|A - B\|_{\text{Fr}} &\overset{(i)}{\le} \sum_{i=1}^r \left\|u_i \sigma_i v_i^T - w_i s_i q_i^T\right\|_{\text{Fr}} \\
&\overset{(ii)}{\le} \sum_{i=1}^r \left\|(u_i \sigma_i - w_i s_i) v_i^T\right\|_{\text{Fr}} + \left\|w_i s_i (v_i - q_i)^T\right\|_{\text{Fr}} \\
&= \sum_{i=1}^r \|u_i \sigma_i - w_i s_i\|_2 \|v_i\|_2 + \|w_i s_i\|_2 \|v_i - q_i\|_2 \\
&\overset{(iii)}{\le} \sum_{i=1}^r \|u_i \sigma_i - w_i s_i\|_2 + \|v_i - q_i\|_2 ,
\end{aligned} \tag{7.1}$$

where inequality $(i)$ is the triangle inequality, $(ii)$ is the triangle inequality after adding and subtracting $w_i s_i v_i^T$, and $(iii)$ follows because $\|w_i s_i\|_2 \le 1$ and $\|v_i\|_2 = 1$.

Each of the vectors $u_i \sigma_i, w_i s_i, v_i, q_i$ all belong to the $\ell_2$-ball $\mathbb{B}_2^d$. Let $\mathcal{N} = \{\tau_i\}_{i=1}^N$ be an $\epsilon$-cover of $\mathbb{B}_2^d$ in $\|\cdot\|_2$, which has cardinality at most $N \le (1 + 2/\epsilon)^d$ by our arguments in class. Then for every $2r$-tuple $\alpha = (i_1, \ldots, i_{2r}) \in \{1, \ldots, N\}^{2r}$, we define

$$A_\alpha := \sum_{j=1}^r \tau_{i_j} \tau_{r+i_j}^T.$$

Given any SVD of $U\Sigma V^T = A \in \mathcal{M}_{r,d}$, there is evidently a tuple $\alpha$ with

$$\left\|u_j\sigma_j - \tau_{i_j}\right\|_2 \le \epsilon \quad \text{and} \quad \left\|v_j - \tau_{r+i_j}\right\|_2 \le \epsilon$$

for $j = 1, \ldots, r$. For this tuple, we obtain

$$\|A - A_\alpha\|_{\text{Fr}} \le \sum_{j=1}^{r}(2\epsilon) = 2r\epsilon$$

by inequality (7.1).

The cardinality of such $2r$-tuples is at most $N^{2r} \le (1 + 2/\epsilon)^{2rd}$. Now we simply replace $\epsilon$ by $\epsilon/2r$ to get the claimed covering. □

**Question 7.13** (Moduli of continuity and high probability rates of convergence)**:** In this question, we show how convexity can be extremely helpful for many reasons in estimation and proving rates of convergence, including (more or less) free guarantees of consistency, as well as high-probability convergence possibilities. Let $\theta \in \mathbb{R}^d$ and define

$$f(\theta) := \mathbb{E}[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x) dP(x)$$

be a function, where $F(\cdot; x)$ is convex in its first argument (in $\theta$) for all $x \in \mathcal{X}$, and $P$ is a probability distribution. We assume $F(\theta; \cdot)$ is integrable for all $\theta$. Recall that a function $h$ is convex

$$h(t\theta + (1-t)\theta') \le th(\theta) + (1-t)h(\theta') \text{ for all } \theta, \theta' \in \mathbb{R}^d,\ t \in [0,1].$$

Let $\theta_0 \in \operatorname{argmin}_\theta f(\theta)$, and assume that $f$ satisfies the following $\nu$*-strong convexity* guarantee:

$$f(\theta) \ge f(\theta_0) + \frac{\nu}{2}\|\theta - \theta_0\|^2 \quad \text{for } \theta \text{ s.t. } \|\theta - \theta_0\| \le \beta,$$

where $\beta > 0$ is some constant. We also assume that the instantaneous functions $F(\cdot; x)$ are $L$-Lipschitz with respect to the norm $\|\cdot\|$.

Let $X_1, \ldots, X_n$ be an i.i.d. sample according to $P$, and define $f_n(\theta) := \frac{1}{n}\sum_{i=1}^n F(\theta; X_i)$ and let

$$\widehat{\theta}_n \in \operatorname*{argmin}_\theta f_n(\theta).$$

(a) Show that for *any* convex function $h$, if there is some $r > 0$ and a point $\theta_0$ such that $h(\theta) > h(\theta_0)$ for all $\theta$ such that $\|\theta - \theta_0\| = r$, then $h(\theta') > h(\theta_0)$ for all $\theta'$ with $\|\theta' - \theta_0\| > r$.

(b) Show that $f$ and $f_n$ are convex.

(c) Show that $\theta_0$ is unique.

(d) Let

$$\Delta(\theta, x) := [F(\theta; x) - f(\theta)] - [F(\theta_0; x) - f(\theta_0)].$$

Show that $\Delta(\theta, X)$ (i.e. the random version where $X \sim P$) is $4L^2\|\theta - \theta_0\|^2$-sub-Gaussian.

(e) Show that for some constant $\sigma < \infty$, which may depend on the parameters of the problem (you should specify this dependence in your solution)

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \ge \sigma \cdot \frac{1+t}{\sqrt{n}}\right) \le C \exp\left(-t^2\right)$$

for all $t \le \sigma'\sqrt{n}\beta$, where $\sigma' > 0$ is a constant depending on the parameters of the problem and $C < \infty$ is a numerical constant. *Hint:* The quantity $\Delta_n(\theta) := \frac{1}{n}\sum_{i=1}^n \Delta(\theta, X_i)$ may be helpful, as may be the bounded differences inequality in Question 6.4.

**Answer:**

(a) Fix $\theta'$ such that $\|\theta' - \theta_0\| > r$. Then there is some $\theta \in [\theta_0, \theta']$ such that $\|\theta - \theta_0\| = r$, that is, there is a $t \in (0, 1)$ with

$$\theta = t\theta_0 + (1-t)\theta', \quad \text{so} \quad h(\theta) \leq th(\theta_0) + (1-t)h(\theta').$$

Rearranging by subtracting $h(\theta_0)$ from both sides yields $h(\theta) - h(\theta_0) \leq (1-t)(h(\theta') - h(\theta_0))$. Noting that $h(\theta_0) < h(\theta)$ and that $t \in (0, 1)$, we thus obtain

$$0 < h(\theta) - h(\theta_0) \leq (1-t)[h(\theta') - h(\theta_0)], \text{ or } h(\theta') > h(\theta_0).$$

(b) This is immediate: for any (positive) measure $\mu$, including $P$ and $P_n$, we have

$$\int F(t\theta + (1-t)\theta'; x) d\mu(x) \leq \int tF(\theta; x) + (1-t)F(\theta'; x) d\mu(x).$$

(c) The uniqueness of $\theta_0$ is immediate by part (b) and (a), because $f(\theta) \geq f(\theta_0) + \frac{\nu\beta^2}{2} > f(\theta_0)$ for all $\theta$ with $\|\theta - \theta_0\| = \beta$.

(d) We have that $\mathbb{E}[\Delta(\theta, X)] = 0$, and that

$$|\Delta(\theta, x)| \leq |F(\theta; x) - F(\theta_0; x)| + |f(\theta) - f(\theta_0)| \leq 2L \|\theta - \theta_0\|,$$

that is, $\Delta$ is bounded by $2L\|\theta - \theta_0\|$. Using the standard result that a variable $Z \in [a, b]$ is $\frac{(b-a)^2}{4}$-sub-Gaussian, we have that $\Delta$ is $16L^2 \|\theta - \theta_0\|^2 / 4 = 4L^2 \|\theta - \theta_0\|^2$ sub-Gaussian.

(e) Fix $\delta \leq \beta$ and let $\Theta_\delta = \{\theta : \|\theta - \theta_0\| \leq \delta\}$. Suppose that $\widehat{\theta}_n$ is not within $\delta$ of $\theta_0$, that is, $\|\widehat{\theta}_n - \theta_0\| \geq \delta$. Then by part (a), there must be some $\theta_\delta \in \Theta_\delta$ such that $f_n(\theta_\delta) \leq f_n(\theta_0)$. Then

$$\begin{aligned} f_n(\theta_\delta) \leq f_n(\theta_0) &= f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) + f(\theta_0) - f(\theta_\delta) \\ &\leq f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) - \frac{\nu}{2}\|\theta_\delta - \theta_0\|^2. \end{aligned}$$

Rearranging, we have

$$\frac{\nu}{2}\|\theta_\delta - \theta_0\|^2 \leq f_n(\theta_0) - f(\theta_0) + f(\theta_\delta) - f_n(\theta_\delta) \leq |\Delta_n(\theta_\delta)| \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|.$$

In particular, if we have that

$$\|\widehat{\theta}_n - \theta_0\| \geq \delta,$$

then it must be the case that

$$\frac{\nu}{2}\delta^2 \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|. \tag{7.2}$$

Now, let us understand this last event (7.2). Let $\Delta_n'$ be $\Delta_n$ with the point $x_i$ swapped for $x_i'$. Then

$$\begin{aligned} &\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| - \sup_{\theta \in \Theta_\delta} |\Delta_n'(\theta)| \leq \sup_{\theta \in \Theta_\delta} |\Delta_n(\theta) - \Delta_n'(\theta)| \\ &= \frac{1}{n} \sup_{\theta \in \Theta_\delta} \left|(F(\theta; x_i) - f(\theta)) - (F(\theta_0; x_i) - f(\theta_0) - (F(\theta; x_i') - f(\theta)) + (F(\theta_0); x_i' - f(\theta_0))\right| \\ &\leq \frac{1}{n} \sup_{\theta \in \Theta_\delta} \left\{|F(\theta; x_i) - F(\theta_0; x_i)| + |F(\theta; x_i') - F(\theta_0; x_i')|\right\} \leq \frac{2L}{n} \sup_{\theta \in \Theta_\delta} \|\theta - \theta_0\| = \frac{2L}{n}\delta. \end{aligned}$$

That is, $\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|$ satisfies bounded differences, and we have

$$\mathbb{P}\left(\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \geq \mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|\right] + t\right) \leq \exp\left(-\frac{nt^2}{2L^2\delta^2}\right).$$

Thus, we control the expected supremum of the errors $\Delta_n(\theta)$ over the neighborhood $\Theta_\delta$. We note by our standard symmetrization inequalities, and the fact that $\theta \mapsto \sqrt{n}\Delta_n(\theta)$ is $4L^2 \|\theta - \theta_0\|^2$-sub-Gaussian process, that

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|\right] \leq \frac{CL}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\Theta_\delta, \|\cdot\|, \epsilon)} d\epsilon,$$

where $N$ denotes the covering numbers of $\Theta_\delta$ in norm $\|\cdot\|$ at radius $\epsilon$ as usual. But then we have $\log N(\Theta_\delta, \|\cdot\|, \epsilon) \leq d \log(1 + \frac{\delta}{\epsilon})$ for $\epsilon < \delta$, and 0 otherwise, so that

$$\mathbb{E}\left[\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)|\right] \leq \frac{CL\sqrt{d}}{\sqrt{n}} \int_0^\delta \sqrt{\log\left(1 + \frac{\delta}{\epsilon}\right)} d\epsilon \leq C\frac{L\sqrt{d}\delta}{\sqrt{n}}.$$

That is, for some numerical constant $C$, we have

$$\mathbb{P}\left(\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \geq C\frac{L\delta}{\sqrt{n}}(\sqrt{d} + t)\right) \leq e^{-t^2} \tag{7.3}$$

for all $t \geq 0$.

On the event that $\sup_{\theta \in \Theta_\delta} |\Delta_n(\theta)| \leq \frac{L\sqrt{d}}{\sqrt{n}}\delta + \frac{\sqrt{2}L}{\sqrt{n}}\delta t$, which occurs with probability at least $1 - e^{-t^2}$ by inequality (7.3), we have by inequality (7.2)

$$\delta^2 \leq C\frac{L}{\nu\sqrt{n}}\left(\sqrt{d} + t\right)\delta,$$

where $C < \infty$ is an absolute constant, as long as $\delta \leq \beta$ (where $\beta$ is the radius of strong convexity). Setting $\sigma = CL\sqrt{d}/\nu\sqrt{n}$, that

$$\delta \leq \sigma(1 + t).$$

That is,

$$\mathbb{P}\left(\|\widehat{\theta}_n - \theta_0\| \leq C\frac{L}{\nu\sqrt{n}}\left(\sqrt{d} + t\right)\right) \geq 1 - e^{-t^2}$$

so long as $\frac{L}{\nu\sqrt{n}}(\sqrt{d} + t) \leq c\beta$, where $c > 0$ is a numerical constant.

□

**Question 7.15:** We consider a few different contraction inequalities and complexities, relating Gaussian to Rademacher complexities. For this problem, define the Rademacher and Gaussian complexities of a set $T \subset \mathbb{R}^n$ by

$$R_n(T) := \mathbb{E}[\sup_{t \in T} |\langle \varepsilon, t \rangle|] \quad \text{and} \quad G_n(T) := \mathbb{E}[\sup_{t \in T} \langle g, t \rangle]$$

where $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{Uni}\{\pm 1\}$ and $g \sim \mathsf{N}(0, I_n)$. Note the lack of an absolute value in the Gaussian complexity.

(a) Let $X \sim \mathsf{N}(0, \Sigma)$ be a Gaussian vector. Argue that for any index $i_0$,

$$\mathbb{E}[\max_{i,j} |X_i - X_j|] = 2\mathbb{E}[\max_i X_i] \;\; \text{and} \;\; \mathbb{E}[\max_i |X_i|] \le 2\mathbb{E}[\max_i X_i] + \mathbb{E}[|X_{i_0}|].$$

(b) Show that for any[1] set $T \subset \mathbb{R}^n$,

$$R_n(T) \le \sqrt{2\pi} G_n(T) + \sqrt{\frac{\pi}{2}} \inf_{t_0 \in T} \mathbb{E}[|\langle g, t\rangle|].$$

If $T$ is symmetric (so $T = -T$) show that $R_n(T) \le \sqrt{\frac{\pi}{2}} G_n(T)$.

(c) Let $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$, be a $M$-Lipschitz functions, meaning $|\phi_i(x) - \phi_i(y)| \le M|x - y|$ for $x, y \in \mathbb{R}$, and define $\phi(t) = (\phi_i(t_i))_{i=1}^n$ to be the elementwise application of $\phi$. Using the result of part (b), show that

$$R_n(\phi(T)) \le M\sqrt{2\pi} G_n(T) + \sqrt{\frac{\pi}{2}} \inf_{t \in T} \mathbb{E}[|\langle g, \phi(t)\rangle|].$$

(d) For a function class $\mathcal{F} \subset \{\mathbb{R}^d \to \mathbb{R}\}$, define the Rademacher and Gaussian complexities

$$R_n(\mathcal{F} \mid x_1^n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \;\; \text{and} \;\; G_n(\mathcal{F} \mid x_1^n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n g_i f(x_i) \right| \right]$$

for any collection $x_1^n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$. Let the function class $\mathcal{F} = \{f(x) = \langle \theta, x\rangle \mid \|\theta\|_1 \le 1\}$, and let $\phi$ be 1-Lipschitz with $\phi(0) = 0$. Show that for $\sigma_{n,j}^2 = \sum_{i=1}^n x_{i,j}^2$ (the sum of squares of the $j$th component of the vectors $x_i \in \mathbb{R}^d$),

$$R_n(\phi \circ \mathcal{F} \mid x_1^n) \le C\sqrt{\max_{j \le d} \sigma_{n,j}^2 \log(2d)}$$

for a numerical constant $C$.

**Answer:**

(a) For the first question, we use the symmetry of the differences to see that $\max_{i,j} |X_i - X_j| = \max_{i,j}(X_i - X_j)$, and so by symmetry

$$\mathbb{E}[\max_{i,j} |X_i - X_j|] = \mathbb{E}[\max_{i,j}(X_i - X_j)] = \mathbb{E}[\max_i X_i + \max_j(-X_j)] = 2\mathbb{E}[\max_i X_i].$$

For the second, we use the triangle inequality to obtain

$$\mathbb{E}[\max_i |X_i|] \le \mathbb{E}[\max_i |X_i - X_{i_0}|] + \mathbb{E}[|X_{i_0}|] \le \mathbb{E}[\max_{i,j} |X_i - X_j|] + \mathbb{E}[|X_{i_0}|],$$

as claimed.

[1] ignoring measurability issues, and assuming that for any random vector $X$ and function $f$ we require that $\mathbb{E}[\sup_{t \in T} f(t, X)] = \sup_{k \in \mathbb{N}} \sup_{|T_0| \le k, T_0 \subset T} \mathbb{E}[\max_{t \in T_0} f(t, X)]$

(b) Note that $\mathbb{E}[|g_i|] = \sqrt{2/\pi}$ for $g_i \sim \mathsf{N}(0,1)$. Then the Rademacher complexity satisfies

$$R_n(T) = \mathbb{E}[\sup_{t \in T} |\langle \varepsilon, t\rangle|] = \mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n \sqrt{\frac{\pi}{2}}\varepsilon_i \mathbb{E}[|g_i|] t_i\right|\right] \overset{(\star)}{\le} \sqrt{\frac{\pi}{2}}\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n \varepsilon_i |g_i| t_i\right|\right],$$

where $(\star)$ uses Jensen's inequality. As $\varepsilon_i |g_i| \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, whenever $T$ is symmetric this gives the result. When $T$ may be non-symmetric, we apply part (a) to give the claimed result.

(c) We develop a comparison inequality. Define the processes

$$Z_t := \sum_{i=1}^n \phi(t_i) g_i \quad \text{and} \quad Y_t := M\langle t, g\rangle$$

on $T$, where $g \sim \mathsf{N}(0, I_n)$. Then

$$\mathbb{E}[(Z_t - Z_u)^2] = \sum_{i=1}^n (\phi(t_i) - \phi(u_i))^2 \le M^2 \|u - t\|_2^2 = \mathbb{E}[(Y_t - Y_u)^2].$$

Applying the Sudakov-Fernique comparison inequality gives

$$G_n(\phi \circ T) \le M G_n(T),$$

and so by applying part (b) we have

$$R_n(\phi(T)) \le \sqrt{2\pi} G_n(\phi(T)) + \sqrt{\frac{\pi}{2}} \inf_{t\in T} \mathbb{E}[|\langle g, \phi(t)\rangle|]$$

Applying the comparison inequality gives the result.

(d) We use that $t = 0 \in T$, and apply the previous parts; we immediately obtain

$$R_n(\phi \circ \mathcal{F} \mid x_1^n) \le O(1) G_n(\phi\circ\mathcal{F} \mid x_1^n) \le O(1) G_n(\mathcal{F} \mid x_1^n) = \mathbb{E}\left[\sup_{\|t\|_1 \le 1} \sum_{i=1}^n g_i x_i^T t\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n g_i x_i\right\|_\infty\right].$$

Each coordinate $j$ of $\sum_{i=1}^n g_i x_{i,j}$ is $\mathsf{N}(0, \sum_{i=1}^n x_{i,j}^2)$, and so standard Gaussian maxima (Question 1.7) give

$$G_n(\mathcal{F} \mid x_1^n) \le \sqrt{2 \max_{j \le d} \sum_{i=1}^n x_{i,j}^2} \sqrt{\log(2d)}.$$

□