Statistics 210b Homework

Zixun Huang

Contents

1 Homework 1

1.1 Mills ratio (Wainwright 2.2)

Solution 1. *(a)*

$$\phi'(z) = -z\frac{1}{2\pi}e^{-z^2/2} = -z\phi(z).$$

(b) By the results of (a), we have

$$\mathbb{P}(Z \geq z) = \int_z^\infty \phi(t)\,\mathrm{d}t = -\int_z^\infty \frac{\phi'(t)}{t}\,\mathrm{d}t.$$

Notice that

$$\left(\left(\frac{1}{t} - \frac{1}{t^3}\right)\phi(t)\right)' = \frac{\phi'(t)}{t} + \frac{3}{t^4}\phi(t) \geq \frac{1}{t}\phi'(t)$$

and

$$\left(\left(\frac{1}{t} - \frac{1}{t^3} + \frac{3}{t^5}\right)\phi(t)\right)' = \frac{\phi'(t)}{t} - \frac{15}{t^6}\phi(t) \leq \frac{\phi'(t)}{t}.$$

and we derive the proof.

1.2 Sharp sub-Gaussian parameter for bounded random variable (Wainwright 2.4)

Solution 2. *(a)We have*

$$\phi(0) = \log \mathbb{E}1 = \log 1 = 0,$$

and

$$\phi'(0) = \frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}} \mid_{\lambda=0} = \mathbb{E}X = \mu.$$

(b)By derivation, we have

$$\phi''(\lambda) = \frac{\mathbb{E}X^2e^{\lambda X}}{\mathbb{E}e^{\lambda X}} - \left(\frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}}\right)^2 = \mathbb{E}_\lambda[X^2] - (\mathbb{E}_\lambda X)^2.$$

We denote $q_\lambda(X) = \frac{e^{\lambda X}}{\mathbb{E}e^{\lambda X}}$*, and we have*

$$\sup_{\lambda\in\mathbb{R}} \phi''(\lambda) \le \sup_{q \textit{ is a density function}} \operatorname{Var}_q(X) \le \frac{(b-a)^2}{4}.$$

(c) By taylor's expansion, we have

$$\log \mathbb{E}e^{\lambda X} = \phi(\lambda) = \phi(0) + \phi'(0)\lambda + \int_0^\lambda \phi''(s)(\lambda - s)\,\mathrm{d}s \le \mu\lambda + \frac{(b-a)^2}{4}\cdot\frac{\lambda^2}{2},$$

which implies that

$$\mathbb{E}e^{\lambda(X-\mu)} \le \frac{(b-a)^2\lambda^2}{8}.$$

And thus, $\sigma = \frac{b-a}{2}$*.*

1.3 Bennett's inequality (Wainwright 2.7)

Solution 3. *(a) By Taylor's expansion, we have*

$$\begin{aligned}\mathbb{E}e^{\lambda X_i} &= \mathbb{E}\sum_{k=0}^{\infty}\frac{(\lambda X_i)^k}{k!} \le 1+\sum_{k=2}^{\infty}\frac{\lambda^k b^{k-2}}{k!}\sigma_i^2 \le 1+\lambda^2\sigma_i^2\left(\frac{e^{\lambda b}-1-\lambda b}{\lambda^2 b^2}\right)\\ &\le \exp\left(\lambda^2\sigma_i^2\left(\frac{e^{\lambda b}-1-\lambda b}{\lambda^2 b^2}\right)\right),\end{aligned}$$

which gives the proof of (a).

Revise: we need to analyze $\lambda < 0$ separately. In fact, notice that we can replace $-X_i$ with X_i without affecting the condition or the result, allowing us to assume $\lambda > 0$.

(b) By Markov's inequality,

$$\begin{aligned}LHS &\le e^{-\lambda n\delta}\mathbb{E}e^{\lambda\sum_{i=1}^{N}X_i}\\ &\le e^{-\lambda n\delta}\prod_{i=1}^{n}\exp\left(\lambda^2\sigma_i^2\left(\frac{e^{\lambda b}-1-\lambda b}{\lambda^2 b^2}\right)\right)\\ &= e^{-\lambda n\delta}\exp\left(\lambda^2\sigma^2\left(\frac{e^{\lambda b}-1-\lambda b}{\lambda^2 b^2}\right)\right)\\ &= \exp\left(-\frac{n\delta}{b}\log\left(1+\frac{b\delta}{\sigma^2}\right)+\frac{n\sigma^2}{b^2}\left(\frac{b\delta}{\sigma^2}-\log\left(1+\frac{b\delta}{\sigma^2}\right)\right)\right).\end{aligned}$$

where the second line follows by (a), and the fourth line follows by choosing $\lambda = \frac{1}{b}\log\left(1+\frac{bn\delta}{\sigma^2}\right)$. Let $t = \frac{b\delta}{\sigma^2}$, and we have

$$LHS \le \exp\left(-\frac{n\sigma^2}{b^2}\left(\log(1+t)-t+t\log(1+t)\right)\right) = RHS.$$

(c) We only need to prove

$$\exp\left\{-\frac{n\sigma^2}{b^2}h\left(\frac{b\delta}{\sigma^2}\right)\right\} \le \exp\left(-n\frac{\delta^2}{2(\sigma^2+b\delta)}\right).$$

The right hand side is Bernstein's inequality, and the inequality is equivalent to

$$h(t) \ge \frac{t^2}{2(1+t)}.$$

Notice that

$$\frac{1}{1+t} \ge \frac{1}{(1+t)^3},$$

we have

$$h'(t) = \log(1+t) \ge \frac{t^2+2t}{2(1+t)^2},$$

which implies

$$h(t) \ge \frac{t^2}{2(1+t)}.$$

1.4 Sharp upper bounds on binomial tails (Wainwright 2.9)

Solution 4. *(a) By Markov's inequality, we have*

$$\mathbb{P}(Z_n \le \delta n) = \mathbb{P}(e^{-\lambda Z_n} \ge e^{-\lambda \delta n}) \le e^{\lambda \delta n} \mathbb{E} e^{-\lambda Z_n} = \exp(\lambda \delta n + n \log(1 - \alpha + \alpha e^{-\lambda}))$$

Let $\lambda = -\log\left(\frac{\delta}{\alpha}\right) - \log\left(\frac{1-\delta}{1-\alpha}\right)$*, and we have*

$$\mathbb{P}(Z_n \le \delta n) \le \exp(-nD(\delta||\alpha)).$$

(b) We only need to show that

$$D(\delta||\alpha) \ge 2(\delta - \alpha)^2.$$

For $\alpha > \delta$*, we have*

$$\frac{\alpha - \delta}{\alpha(1-\alpha)} \ge 4(\alpha - \delta),$$

and hence

$$D(\delta||\alpha) = \delta \log\left(\frac{\delta}{\alpha}\right) + (1-\delta)\log\left(\frac{1-\delta}{1-\alpha}\right) \ge 2(\delta - \alpha)^2.$$

1.5 Upper bounds for sub-Gaussian maxima (Wainwright 2.12)

Solution 5. *(a) We have*

$$\begin{aligned}
\mathbb{E} \max_{1 \le i \le n} X_i &= \frac{1}{\lambda} \mathbb{E} \log(\max_{1 \le i \le n} e^{\lambda X_i}) \\
&\le \frac{1}{\lambda} \log \max_{1 \le i \le n} \mathbb{E} e^{\lambda X_i} \\
&\le \frac{1}{\lambda} \log \left(\sum_{i=1}^{N} \mathbb{E} e^{\lambda X_i} \right) \\
&= \frac{1}{\lambda} (\log N + \frac{\lambda^2 \sigma^2}{2}).
\end{aligned}$$

where the second line follows from Jensen's inequality. Let $\lambda = \frac{\sqrt{2 \log N}}{\sigma}$*, we have*

$$\mathbb{E} \max_{1 \le i \le n} X_i \le \sqrt{2\sigma^2 \log n}.$$

(b) Consider $\{X_1, ..., X_n, -X_1, ..., -X_n\}$ *and apply the results in (a), we have*

$$\mathbb{E}[Z] \le \sqrt{2\sigma^2 \log(2n)} \le 2\sqrt{\sigma^2 \log n}$$

1.6 Operations on sub-Gaussian variables (Wainwright 2.13)

Solution 6. *(a) Since X_1 and X_2 are independent, we have*

$$\mathbb{E}e^{\lambda(X_1+X_2)} = \mathbb{E}e^{\lambda X_1}\mathbb{E}e^{\lambda X_2} \leq e^{\lambda^2\sigma_1^2/2}e^{\lambda^2\sigma_2^2/2} = e^{\lambda^2(\sigma_1^2+\sigma_2^2)/2}$$

So $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

(b)(c) Notice that

$$\sigma_1 + \sigma_2 \leq \sqrt{2}\sqrt{\sigma_1^2 + \sigma_2^2},$$

we only need to prove (c). By Hölder's inequality, we have

$$\mathbb{E}e^{\lambda(X_1+X_2)} \leq (\mathbb{E}e^{p\lambda X_1})^{1/p}(\mathbb{E}e^{q\lambda X_2})^{1/q} \leq \exp\left(\frac{p\lambda^2\sigma_1^2 + q\lambda^2\sigma_2^2}{2}\right)$$

choose $p = \frac{\sigma_1+\sigma_2}{\sigma_1}$ and $q = \frac{\sigma_1+\sigma_2}{\sigma_2}$,

$$\mathbb{E}e^{\lambda(X_1+X_2)} \leq \exp\left(\frac{\lambda^2(\sigma_1+\sigma_2)^2}{2}\right).$$

So $X_1 + X_2$ is sub-Gaussian with parameter $\sigma_1 + \sigma_2$.

(d) Notice that, by property 2 of the sub-Gaussian random variable,

$$\mathbb{E}|X_1|^p \leq C_1^p\sigma_1^p p^{\frac{p}{2}},$$

$$\mathbb{E}|X_2|^p \leq C_2^p\sigma_2^p p^{\frac{p}{2}},$$

for any $p \in \mathbb{Z}$. By Cauchy-Schwarz inequality,

$$(\mathbb{E}|X_1X_2|^p)^{\frac{1}{p}} \leq (\mathbb{E}|X_1|^{2p})^{\frac{1}{2p}}(\mathbb{E}|X_2|^{2p})^{\frac{1}{2p}} \leq 2C_1C_2\sigma_1\sigma_2 p$$

By property 2 and property 5 of sub-Exponential random variable (Vershynin Proposition 2.7.1), we have

$$\mathbb{E}\exp(\lambda X_1X_2) \leq \exp(C^2\sigma_1^2\sigma_2^2\lambda^2) \quad \textit{for} \quad |\lambda| \leq \frac{1}{C\sigma_1\sigma_2}.$$

We have X_1X_2 is sub-Exponential with parameter $(C\sigma_1\sigma_2, C\sigma_1\sigma_2)$

1.7 Robust estimation of the mean (Vershynin 2.2.9)

Solution 7. *(a)By Hoeffding's inequality, we have*

$$\mathbb{P}(|\hat{\mu} - \mu| > \epsilon) \leq 2\exp\left(-n\frac{\epsilon^2}{2\sigma^2}\right) =: \delta.$$

Hence, we have that

$$n = \frac{2\sigma^2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$$

which is the number of the required samples.

(b)By Markov's inequality

$$\mathbb{P}(|\hat{\mu} - \mu| > \epsilon) \leq 2\frac{\mathrm{Var}(\hat{\mu})}{\epsilon^2} = \frac{2\sigma^2}{n\epsilon^2} = \frac{1}{4}.$$

Hence, we have that

$$n = \frac{8\sigma^2}{\epsilon^2}.$$

(c)Let $n = mk$, where $m \in \mathbb{Z}$. By (b), when $m \geq \frac{32\sigma^2}{\epsilon^2}$, we have

$$\mathbb{P}(|Y_i - \mu| > \epsilon) \leq \frac{1}{16}$$

WLOG, assume k is an even number. Let $\hat{\mu} = medium(Y_1, ..., Y_k)$, we have

$$\mathbb{P}(|\hat{\mu} - \mu| > \epsilon) = \sum_{i=k/2}^{k}\binom{k}{i}\left(\frac{1}{16}\right)^i \leq \frac{1}{4^k}\sum_{i=0}^{k}\binom{k}{i} = \frac{1}{2^k} = \delta.$$

When $k = O(\log\frac{1}{\delta})$, then w.p $\geq 1 - \delta$

$$|\hat{\mu} - \mu| < \epsilon,$$

so the sample size $n = mk = O(\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta})$.

1.8 Survey question

problem 1: 20 min
problem 2: 20 min
problem 3: 50 min
problem 4: 10 min
problem 5: 15 min
problem 6: 40 min
problem 7: 40 min

2 Homework 2

2.1 Tail bound and almost sure convergence

Solution 8. *By Borel-Cantalli lemma and*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > t) \leq \sum_{n=1}^{\infty} \delta_{n,t} = \sum_{n=1}^{\infty} 2\exp(-\frac{nt^2}{2}) < \infty,$$

we know that

$$\mathbb{P}(\limsup_{n\to\infty} |X_n - X| \geq t) = 0,$$

which implies almost surely convergence.

For $\delta_{n,t} = \frac{1}{\sqrt{nt^2}}$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > t) \leq \sum_{n=1}^{\infty} \delta_{n,t} = \sum_{n=1}^{\infty} \frac{1}{\sqrt{nt^2}} = \infty,$$

and by Borel-Cantalli lemma

$$\mathbb{P}(\limsup_{n\to\infty} |X_n - X| \geq t) = 1,$$

which doesn't imply almost surely convergence.

For $\delta_{n,t} = \frac{1}{nt^2}$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > t) \leq \sum_{n=1}^{\infty} \delta_{n,t} = \sum_{n=1}^{\infty} \frac{1}{nt^2} = \infty,$$

and by Borel-Cantalli lemma

$$\mathbb{P}(\limsup_{n\to\infty} |X_n - X| \geq t) = 1,$$

which doesn't imply almost surely convergence.

For $\delta_{n,t} = \frac{1}{n^2t^4}$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > t) \leq \sum_{n=1}^{\infty} \delta_{n,t} = \sum_{n=1}^{\infty} \frac{1}{n^2t^4} < \infty,$$

and by Borel-Cantalli lemma

$$\mathbb{P}(\limsup_{n\to\infty} |X_n - X| \geq t) = 0,$$

which implies almost surely convergence.

2.2 Maximal version of Freedman's inequality

Solution 9. *Let* $X_K = \exp\left(\lambda \sum_{k=1}^{K} D_k - \lambda \sum_{k=1}^{K} \frac{\nu_k^2}{2}\right)$*, by the condition*

$$\mathbb{E}[X_K|\mathscr{F}_{K-1}] = X_{K-1} \cdot e^{-\lambda^2 \nu_K^2/2} \mathbb{E}[e^{\lambda D_K}|\mathscr{F}_{K-1}] \leq X_{K-1},$$

we have $\{(X_k, \mathscr{F}_k)\}$ *is a super-martingale. By Doob's maximal inequality (super-martingale version):*

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq K\leq T} \sum_{k=1}^{K} D_k - \sum_{k=1}^{K} \frac{\nu_k^2}{2} > t\right) &= \mathbb{P}\left(\max_{1\leq K\leq T} \exp\left(\lambda \sum_{k=1}^{K} D_k - \lambda \sum_{k=1}^{K} \frac{\nu_k^2}{2}\right) > e^{\lambda t}\right) \\ &\leq e^{-\lambda t}(\mathbb{E}[\max(X_0, 0)]) \\ &\leq e^{-\lambda t}\end{aligned}$$

Choose $t = \frac{\log(1/\delta)}{\lambda}$*, and we get the result.*

Then, we choose

$$\lambda = \min\left\{\frac{1}{\alpha}, \frac{\sqrt{2\log(1/\delta)}}{v_*}\right\},$$

and we have w.p. $\geq 1 - \delta$*,*

$$\sup_{1\leq K\leq T} \sum_{k=1}^{K} D_k < C \cdot \max\{\nu_* \sqrt{\log(1/\delta)}, \alpha \log(1/\delta)\}.$$

2.3 Concentration and kernel density estimation

Solution 10. *We prove that the function $g(X_1, ..., X_n) = \|\hat{f} - f\|(X_1, ..., X_n)$ is $(\frac{2}{n}, ..., \frac{2}{n})$-bounded differential.*

$$\begin{aligned}
\|g(X_1, ..., X_i, ..., X_n) - g(X_1, ..., X_i', ..., X_n)\| &\le \frac{1}{nh}\|K(\frac{x - X_i}{h}) - K(\frac{x - X_i'}{h})\|_1 \\
&\le \frac{1}{n}(\|K(x - hX_i)\| + \|K(x - hX_i')\|) \\
&= \frac{2}{n}
\end{aligned}$$

By the concentration inequality of bounded difference function, have $g(X_1, ..., X_n)$ is $sG(\frac{2}{\sqrt{n}})$, which implies

$$\mathbb{P}\left(\|\hat{f}_n - f\|_1 \ge \mathbb{E}\|\hat{f}_n - f\|_1 + \delta\right) \le e^{-\frac{n\delta^2}{8}}.$$

2.4 Concentration for spin glasses

Solution 11. *(a) Observe that* $F_d(\theta) = \log \sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}$. *We have*

$$\frac{\partial}{\partial\theta} F_d(\theta) = \frac{1}{\sqrt{d}} \frac{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}(\boldsymbol{x}\boldsymbol{x}^\top)}{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}}$$

and

$$\frac{\partial^2}{\partial\theta^2} F_d(\theta) = \frac{1}{d} \frac{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}(\boldsymbol{x}\boldsymbol{x}^\top)^{\otimes 2}}{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}}.$$

We have

$$U^\top \frac{\partial^2}{\partial\theta^2} F_d(\theta) U = \frac{1}{d} \frac{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}(\mathrm{tr}(U\boldsymbol{x}\boldsymbol{x}^\top)^2}{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}} \geq 0.$$

Thus, F_d *is a convex function.*

(b) Notice that

$$\|\frac{\partial}{\partial\theta} F_d(\theta)\|_2 \leq \left\| \frac{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}(\boldsymbol{x}\boldsymbol{x}^\top)}{\sum_{x\in\{\pm 1\}^d} \exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}} \right\|_2 \leq \|\boldsymbol{x}\boldsymbol{x}^\top\|_2 \leq d,$$

by mean value theorem, we have

$$\|F_d(\theta) - F_d(\theta')\|_2 \leq \sqrt{d}\|\theta - \theta'\|_2.$$

(c) First, we give the lower bound of the $\mathbb{E}[\frac{F_d(\theta)}{d}]$. *By Jensen's inequality,*

$$\begin{aligned}
\mathbb{E}\left[\frac{F_d(\theta)}{d}\right] &\geq \frac{1}{d}\log\left[\sum_{\boldsymbol{x}\in\{\pm 1\}^d} \mathbb{E}[\exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}]\right] \\
&= \log 2 + \frac{1}{d}\log \mathbb{E}[\exp\{\frac{1}{\sqrt{d}} \boldsymbol{x}^\top \theta \boldsymbol{x}\}] \\
&\geq \log 2 + \frac{1}{d}\sum_1^d \log \mathbb{E}[\exp\{\frac{1}{\sqrt{d}}\theta_{ii}\}] + \frac{1}{d}\sum_{i<j} \log \mathbb{E}[\exp\{\frac{1}{\sqrt{d}}\theta_{ij}\}] \\
&\geq \log 2 + \frac{d}{4}\beta^2.
\end{aligned}$$

Then, we prove the concentration inequality. Given that $\frac{F_d}{d}$ *is* $\frac{1}{\sqrt{d}}$*−Lipschitz, we have*

$$\mathbb{P}\left(\frac{F_d(\theta)}{d} - \mathbb{E}\left[\frac{F_d(\theta)}{d}\right] \geq t\right) \leq e^{-t^2 d/(2\beta^2)}, \quad \textit{for all } t > 0.$$

Combine them together and we get

$$\mathbb{P}\left(\frac{F_d(\theta)}{d} \geq \log 2 + \frac{\beta^2}{4} + t\right) \leq 2e^{-\frac{t^2 d}{2\beta^2}} \quad \textit{for all } t > 0.$$

> **Theorem 3.26** (Functional Hoeffding theorem) *For each $f \in \mathscr{F}$ and $i = 1, \ldots, n$, assume that there are real numbers $a_{i,f} \le b_{i,f}$ such that $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$. Then for all $\delta \ge 0$, we have*
>
> $$\mathbb{P}[Z \ge \mathbb{E}[Z] + \delta] \le \exp\left(-\frac{n\delta^2}{4L^2}\right), \tag{3.80}$$
>
> *where $L^2 := \sup_{f \in \mathscr{F}} \{\frac{1}{n} \sum_{i=1}^n (b_{i,f} - a_{i,f})^2\}$.*

Figure 1: Theorem 3.26

2.5 Rademacher chaos variables

Solution 12. *(a) We let $f(\epsilon) = \|Q^{1/2}\epsilon\|_2$ and we know $f(x) \in (-\sqrt{d}\|Q^{1/2}\|_{op}, \sqrt{d}\|Q^{1/2}\|_{op})$ as $x \in \{\pm 1\}^d$. By concentration inequality for Lipschitz function (Theorem 3.26 in the [2] as the figure 1 shows):*

$$\mathbb{P}(f(\epsilon) - \mathbb{E}[f(\epsilon)] \ge t) \le \exp(-\frac{dt^2}{16d\|Q\|_{op}^2}) = \exp(-\frac{t^2}{16\|Q\|_{op}^2}).$$

As $\mathbb{E}[f(\epsilon)] \le (\mathbb{E}[f(\epsilon)^2])^{1/2} = \sqrt{\mathrm{tr}(Q)}$, we derive the result directly!

(b) (I am inspired by [1] Lemma 6.2.2) We can write

$$Y = \langle \epsilon', M\epsilon \rangle.$$

Since

$$\mathbb{E}[e^{\lambda\langle \epsilon', M\epsilon\rangle}] \le \exp(\lambda^2\|M\epsilon\|^2/2),$$

So $\langle \epsilon', M\epsilon \rangle$ is $sG(\|M\epsilon\|)$ and

$$\mathbb{P}(Y \ge \delta \mid \epsilon) \le \exp(-\frac{\delta^2}{2\|M\epsilon\|_2^2}).$$

By (a), we have

$$\mathbb{P}(\|M\epsilon\|_2^2 \ge (\|M\|_F + t)^2) \le \exp(-\frac{t^2}{16\|M\|_{op}^2})$$

Thus

$$\begin{aligned}
\mathbb{P}(Y \ge \delta) &\le \exp(-\frac{t^2}{16\|M\|_{op}^2}) + \exp(-\frac{\delta^2}{2(\|M\|_F + t)^2}) \\
&\le \exp(-\frac{t^2}{16\|M\|_{op}^2}) + \exp(-\frac{\delta^2}{4(\|M\|_F^2 + t^2)}) \\
&\le 2\exp(-\frac{\delta^2}{4\|M\|_F^2 + 16\delta\|M\|_{op}})
\end{aligned}$$

The last line follows by choosing $t^2 = 2\delta\|M\|_{op}$.

2.6 Maximum likelihood and uniform laws

Solution 13. *(a) By definition, for Bernoulli,*

$$\begin{aligned}R(\theta,\theta_*) &= \frac{1}{1+e^{\theta_*}}\log\left(\frac{1/(1+e^{\theta_*})}{1/(1+e^{\theta})}\right)+\frac{e^{\theta_*}}{1+e^{\theta_*}}\log\left(\frac{e^{\theta_*}/(1+e^{\theta_*})}{e^{\theta}/(1+e^{\theta})}\right)\\ &= \frac{e^{\theta_*}}{1+e^{\theta_*}}(\theta^*-\theta)+\log(1+e^{\theta})-\log(1+e^{\theta_*}).\end{aligned}$$

For Poisson,

$$\begin{aligned}R(\theta,\theta_*) &= \sum_{x=0}^{\infty} p_{\theta_*}(x)(\theta_* x-\exp(\theta_*)-\theta x+\exp(\theta))\\ &= e^{\theta}-e^{\theta_*}+(\theta_*-\theta)e^{\theta_*}.\end{aligned}$$

For multivariate Gaussian,

$$\begin{aligned}R(\theta,\theta_*) = &\int\left[\frac{1}{2}(x-\theta)^\top\Sigma^{-1}(x-\theta)-\frac{1}{2}(x-\theta_*)^\top\Sigma^{-1}(x-\theta_*)\right]\\ &\cdot\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left\{-\frac{1}{2}(x-\theta_*)^\top\Sigma^{-1}(x-\theta_*)\right\}\mathrm{d}x\\ &= \frac{1}{2}(\theta-\theta_*)^\top\Sigma^{-1}(\theta-\theta_*)-\int(\theta-\theta_*)^\top\Sigma^{-1}y\cdot\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left\{-\frac{1}{2}y^\top\Sigma^{-1}y\right\}\mathrm{d}y\\ &= \frac{1}{2}(\theta-\theta_*)^\top\Sigma^{-1}(\theta-\theta_*)-(\theta-\theta_*)^\top\Sigma^{-1/2}\int z\frac{1}{\sqrt{(2\pi)^d}}\exp\left\{-\frac{1}{2}z^\top z\right\}\mathrm{d}z\\ &= \frac{1}{2}(\theta-\theta_*)^\top\Sigma^{-1}(\theta-\theta_*).\end{aligned}$$

(b) For Bernoulli,

$$\hat\theta = \log\frac{\sum_{i=1}^n X_i}{n-\sum_{i=1}^n X_i},$$

where we assume $\sum_i X_i \in (0,n)$.

$$E(\hat\theta,\theta_*) = R(\hat\theta,\theta_*) = \frac{e^{\theta_*}}{1+e^{\theta_*}}(\theta^*-\log\frac{\sum_{i=1}^n X_i}{n-\sum_{i=1}^n X_i})+\log\frac{n}{n-\sum_{i=1}^n X_i}-\log(1+e^{\theta_*}).$$

For Poisson,

$$\hat\theta = \log\sum_{i=1}^n X_i.$$

$$E(\hat\theta,\theta_*) = R(\hat\theta,\theta_*) = \sum_{i=1}^n X_i - e^{\theta_*}+(\theta_*-\log\sum_{i=1}^n X_i)e^{\theta_*}.$$

For multivariate Gaussian,

$$\hat\theta = \frac{1}{n}\sum_{i=1}^n X_i.$$

$$E(\hat\theta,\theta_*) = R(\hat\theta,\theta_*) = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^n X_i-\theta_*)^\top\Sigma^{-1}(\frac{1}{n}\sum_{i=1}^n X_i-\theta_*).$$

Now we give an upper bound on the excess risk. Let Empirical risk $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n [\log(p_{\theta^*}(X_i)) - \log(p_\theta(X_i))]$ *and population risk* $R(\theta) = R(\theta;\theta^*)$.

$$\begin{aligned} R(\hat{\theta},\theta_*) &= \mathbb{E}_{X_i}[R(\hat{\theta}) - R_n(\hat{\theta}) + R_n(\hat{\theta}) - R_n(\theta_*) + R_n(\theta_*) - R(\theta_*)] \\ &\leq 2\mathbb{E}_{X_i}[\sup_\theta |R(\theta) - R_n(\theta)|] \end{aligned}$$

We can write the expectation as

$$R(\theta) = \mathbb{E}_{X_i'}\left[\frac{1}{n}R_n'(\theta)\right],$$

and we have

$$\begin{aligned} R(\hat{\theta},\theta_*) &\leq 2\mathbb{E}_{X_i}[\sup_\theta \mathbb{E}_{X_i'}[|R_n'(\theta) - R_n(\theta)|]] \\ &\leq 2\mathbb{E}_{X_i,X_i'}[\sup_\theta |R_n'(\theta) - R_n(\theta)|] \\ &= 2\mathbb{E}_{X_i,X_i',\epsilon_i\sim Unif\{\pm1\}}[\sup_\theta |\epsilon_i(R_n'(\theta) - R_n(\theta))| \\ &= 2\mathbb{E}_{X_i,X_i',\epsilon_i\sim Unif\{\pm1\}}[\sup_\theta |\epsilon_i(\log p_\theta(X_i) - \log p_\theta(X_i'))|] \\ &\leq 4\mathbb{E}_{X_i\epsilon_i\sim Unif\{\pm1\}}[\sup_\theta |\epsilon_i \log p_\theta(X_i)|] \\ &= 4R_n(\mathscr{F}), \end{aligned}$$

where $\mathscr{F} = \{\log p_\theta(\cdot) : \theta \in \Omega\}$, *and* $R_n(\cdot)$ *is the Radamacher complexity, whose definition is*

$$R_n(\mathscr{F}) = \mathbb{E}_{X_i,\epsilon_i\in Unif\{\pm1\}} \sup_{f\in\mathscr{F}} \frac{1}{n}|\sum_{i=1}^n \epsilon_i f(X_i)|.$$

For Bernoulli:

$$\mathscr{F}_{Bernoulli} = \{f_\theta(x) = \theta x - \log(1+e^\theta) : x \in \{0,1\}, \theta \in \Omega\}.$$

For Poisson:

$$\mathscr{F}_{poisson} = \{f_\theta(x) = \theta x - \exp(\theta) - \sum_{i=1}^x \log i : x \in \mathbb{N}, \theta \in \Omega\}.$$

For multivariate Gaussian:

$$\mathscr{F}_{multi-gaussian} = \left\{f_\theta(x) = -\frac{1}{2}(x-\theta)^\top \Sigma^{-1}(x-\theta) - \frac{1}{2}\log(2\pi|\Sigma|) : x \in \mathbb{R}, \theta \in \Omega\right\}$$

2.7 Basic properties of Rademacher complexity

Solution 14. *(a) We have*

$$\begin{aligned}\mathscr{R}_n(\mathrm{conv}(\mathscr{F})) &= \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{f\in\mathrm{conv}(\mathscr{F})}\sum_{i=1}^n \epsilon_i f(X_i)] \\ &\leq \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{f\in\mathscr{F}}\sum_{i=1}^n \epsilon_i f(X_i)] = \mathscr{R}_n(\mathscr{F})) \leq \mathscr{R}_n(\mathrm{conv}(\mathscr{F})).\end{aligned}$$

The first inequality follows from the convexity of sup *and the second inequality follows from the fact that* $\mathscr{F}\subset\mathrm{conv}(\mathscr{F})$*.*

(b) We have

$$\begin{aligned}\mathscr{R}_n(\mathscr{F}+\mathscr{G}) &= \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{f\in\mathscr{F}}\sum_{i=1}^n \epsilon_i (f(X_i)+g(X_i))] \\ &\leq \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{f\in\mathscr{F}}\sum_{i=1}^n \epsilon_i f(X_i)] + \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{g\in\mathscr{F}}\sum_{i=1}^n \epsilon_i g(X_i)] = \mathscr{R}_n(\mathscr{F})+\mathscr{R}_n(\mathscr{G})\end{aligned}$$

If we choose $\mathscr{F}=\mathscr{G}$*, we have*

$$\mathscr{R}_n(\mathscr{F}+\mathscr{G}) = \mathscr{R}_n(2\mathscr{F}) = 2\mathscr{R}_n(\mathscr{F}),$$

the equation holds, which implies the bound is tight (cannot be improved in general).

(c) We have

$$\begin{aligned}\mathscr{R}_n(\mathscr{F}+g) &= \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{f\in\mathscr{F}}\sum_{i=1}^n \epsilon_i (f(X_i)+g(X_i))] \\ &\leq \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sup_{f\in\mathscr{F}}\sum_{i=1}^n \epsilon_i f(X_i)] + \mathbb{E}_{\epsilon,X_i}[\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)] \\ &\leq \mathscr{R}(\mathscr{F}) + \frac{1}{n}(\mathbb{E}[\sum_{i=1}^n \epsilon_i^2\ ])^{1/2}(\mathbb{E}[\sum_{i=1}^n g(X_i)^2\ ])^{1/2} \\ &= \mathscr{R}(\mathscr{F}) + \frac{1}{\sqrt{n}}\|g\|_2 \\ &\leq \mathscr{R}(\mathscr{F}) + \frac{1}{\sqrt{n}}\|g\|_\infty.\end{aligned}$$

The second inequality is Cauchy-Schwarz inequality. The last line follows from

$$\|g\|_2 \leq \|g\|_\infty.$$

3 Homework 3

3.1 Gaussian and Rademacher complexity (Wainwright 5.5)

Solution 15. *Recall that the definition of the Gaussian and Rademacher complexity of a set is*

$$\mathscr{G}(\mathbb{T}) = \frac{1}{d}\mathbb{E}_{w_i \sim \mathscr{N}(0,1)}\left[\sup_{a_i \in \mathbb{T}} \sum_{i=1}^{d} w_i a_i\right],$$

$$\mathscr{R}(\mathbb{T}) = \frac{1}{d}\mathbb{E}_{\sigma_i \sim Unif\{\pm 1\}}\left[\sup_{a_i \in \mathbb{T}} \sum_{i=1}^{d} \sigma_i a_i\right].$$

(a) By Jensen's inequality

$$\begin{aligned}
\mathscr{G}(\mathbb{T}) &= \frac{1}{d}\mathbb{E}_{\sigma_i \sim Unif\{\pm 1\}}\mathbb{E}_{w_i}\left[\sup_{a_i \in \mathbb{T}} \sum_{i=1}^{d} |w_i|\sigma_i a_i\right] \\
&\geq \frac{1}{d}\mathbb{E}_{\sigma_i \sim Unif\{\pm 1\}}\left[\sup_{a_i \in \mathbb{T}} \sum_{i=1}^{d} \sigma_i a_i\right] \cdot \mathbb{E}_{w_i}[|w_i|] \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{d}\mathbb{E}_{\sigma_i \sim Unif\{\pm 1\}}\left[\sup_{a_i \in \mathbb{T}} \sum_{i=1}^{d} \sigma_i a_i\right] \\
&= \sqrt{\frac{2}{\pi}}\mathscr{R}(\mathbb{T}).
\end{aligned}$$

(b) Consider contraction $\psi_r(x) = \frac{\langle r, x\rangle}{\max_{1 \leq i \leq d} r_i}$, *we have*

$$\mathscr{R}(\mathbb{T}) \geq \mathbb{E}[\mathscr{R}(\psi_w(\mathbb{T}))].$$

Thus,

$$\mathscr{G}(\mathbb{T}) = \frac{1}{d}\mathbb{E}_{w_i \sim \mathscr{N}(0,1)}\left[\sup_{a_i \in \mathbb{T}} \sum_{i=1}^{d} w_i a_i\right] \leq \mathbb{E}[\max_{1 \leq i \leq d} |w_i|] \cdot \mathscr{R}(\mathbb{T}) \leq \sqrt{2\log d} \cdot \mathscr{R}(\mathbb{T})$$

3.2 Maximal inequality for sub-exponential random variables

Solution 16. *We have*

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[n]}|X_i|\right] &\le \frac{1}{\lambda}\log\mathbb{E}\left[\max_{i\in[n]}[\exp(\lambda|X_i|)]\right]\\ &\le \frac{1}{\lambda}\log\mathbb{E}\left[\sum_{i\in[n]}\exp(\lambda|X_i|)\right]\\ &\le \frac{1}{\lambda}\left(\log(2n)+\log\mathbb{E}\left[\exp(\lambda X_i)\right]\right)\end{aligned}$$

Recall the properties of sub-Exponential random variables, we have

$$\log\exp(\lambda X_i)\le \frac{1}{2}\lambda^2\alpha_1^2,\ \textit{for}\ \lambda\le\frac{1}{\alpha_2}.$$

Thus,

$$\mathbb{E}\left[\max_{i\in[n]}|X_i|\right]\le \inf_{\lambda\le\frac{1}{\alpha_2}}\frac{1}{\lambda}\left(\log(2n)+\frac{1}{2}\lambda^2\alpha_1^2\right).$$

If $\frac{\sqrt{2\log(2n)}}{\alpha_1}\le\frac{1}{\alpha_2}$,

$$\mathbb{E}\left[\max_{i\in[n]}|X_i|\right]\le\sqrt{2}\alpha_1\sqrt{\log(2n)}.$$

If $\frac{\sqrt{2\log(2n)}}{\alpha_1}>\frac{1}{\alpha_2}$,

$$\mathbb{E}\left[\max_{i\in[n]}|X_i|\right]\le\alpha_2\log(2n)+\frac{\alpha_1^2}{2\alpha_2}\le\alpha_2\log(2n)+\frac{\sqrt{2}}{2}\alpha_1\sqrt{\log(2n)}$$

To conclude, we get

$$\mathbb{E}\left[\max_{i\in[n]}|X_i|\right]\le C\left[\alpha_1\sqrt{\log(2n)}+\alpha_2\log(2n)\right]$$

3.3 Covering numbers for low-rank matrices (Duchi 7.8)

Solution 17. *By singular decomposition, we can write*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

where $\sum_{i=1}^{r} \sigma_i^2 = 1$*. By triangle inequality, we have*

$$\|A - A'\|_F \leq \sum_{i=1}^{r} \|\sigma_i u_i - \sigma_i' u_i'\|_2 + \sum_{i=1}^{r} \|v_i - v_i'\|_2.$$

Thus, we can bound the covering number of $\epsilon-$*net of the right hand side* $\mathscr{M}_{r,d}$

$$N(\epsilon, \mathscr{M}_{r,d}, \|\cdot\|_F) \leq N(\epsilon/2r, B_d(1), \|\cdot\|_2)^{2r} \leq \left(1 + \frac{4r}{\epsilon}\right)^{2rd}$$

where the inequality follows from the proposition in the textbook:

$$N(\epsilon/2r, B_d(1), \|\cdot\|_2) \leq \frac{vol(B(1+\epsilon/4r))}{vol(B(\epsilon/4r))} \leq \left(1 + \frac{4r}{\epsilon}\right)^{d}$$

and both $\sigma_i u_i$ *and* v_i *are in the unit ball* $B_d(1)$*. Taking the logarithm on both sides, we obtain*

$$\log N(\epsilon, \mathscr{M}_{r,d}, \|\cdot\|_F) \leq 2rd \log(1 + 4r/\epsilon).$$

3.4 Rademacher complexity bound of Lipschitz functions on $[0,1]^d$

Solution 18. *(a) First, we denote $\mathscr{N}$ as the ϵ-covering of $\mathscr{F}_L^d$. We can prove that*

$$|\mathscr{N}| = N(\epsilon; \mathscr{F}_L^d, \|\cdot\|_\infty) \le \exp\left(\left(\frac{L}{\epsilon}\right)^d\right).$$

Second, we consider the random variable $\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)$.

$$\begin{aligned}\log \mathbb{E}[\exp(\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i))] &\le n \log \exp\left(\frac{L^2}{2n^2}\right) \\ &= \frac{L^2}{2n},\end{aligned}$$

thus $\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)$ is $sG(\frac{L}{\sqrt{n}})$ random variable.

By one step discretization,

$$\begin{aligned}\mathscr{R}_n(\mathscr{F}_L^d) &= \mathbb{E}_{x_i,\epsilon_i}\left[\sup_{f\in\mathscr{F}_L^d} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)|\right] \\ &\le \mathbb{E}_{x_i,\epsilon_i}\left[\sup_{\|f-f'\|_\infty\le\epsilon} |\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(x_i)-f'(x_i))|\right] + \mathbb{E}_{x_i,\epsilon_i}\left[\sup_{f\in\mathscr{N}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right] \\ &\le \epsilon + \sqrt{2\sigma^2 N(\epsilon; \mathscr{F}_L^d, \|\cdot\|_\infty)} \\ &\le \epsilon + \sqrt{2\frac{L^2}{n}\left(\frac{L}{\epsilon}\right)^d} \\ &\asymp \epsilon + \frac{L^{1+d/2}}{\sqrt{n}}\epsilon^{-d/2}.\end{aligned}$$

Choose $\epsilon = Ln^{-\frac{1}{d+2}}$, we have

$$\mathscr{R}_n(\mathscr{F}_L^d) \lesssim Ln^{-\frac{1}{d+2}}.$$

(b) Using Dudley's entropy integral method,

$$\begin{aligned}\mathscr{R}_n(\mathscr{F}_L^d) &\lesssim \epsilon + \frac{1}{\sqrt{n}}\int_\epsilon^{2L} \sqrt{\log N(u; \mathscr{F}_L^d, \|\cdot\|_\infty)}\,\mathrm{d}u \\ &\le \epsilon + \frac{1}{\sqrt{n}}\int_\epsilon^{2L}\left(\frac{L}{u}\right)^{d/2}\mathrm{d}u \\ &= \epsilon + \frac{L}{\sqrt{n}}\int_{\epsilon/L}^{2} u^{-d/2}\,\mathrm{d}u.\end{aligned}$$

If $d=1$,

$$\int_{\epsilon/L}^2 u^{-d/2}\,\mathrm{d}u \le 2\sqrt{2}.$$

Choose $\epsilon = 0$, and we have

$$\mathscr{R}_n(\mathscr{F}_L^d) \lesssim \frac{L}{\sqrt{n}}.$$

If $d=2$,

$$\int_{\epsilon/L}^2 u^{-d/2}\,\mathrm{d}u = \log 2 - \log\left(\frac{\epsilon}{L}\right) = \log 2 + \log L - \log\epsilon.$$

Choose $\epsilon = \frac{L}{\sqrt{n}}$*, and we have*

$$\mathscr{R}_n(\mathscr{F}_L^d) \lesssim \frac{L}{\sqrt{n}} \left(1 + \log 2 + \frac{1}{2} \log n\right) \lesssim \frac{L}{\sqrt{n}}(1 + \log n).$$

If $d \geq 3$*,*

$$\int_{\epsilon/L}^{2} u^{-d/2} \, \mathrm{d}u \leq \frac{2}{d-2} \left(\frac{L}{\epsilon}\right)^{\frac{d}{2}-1}.$$

Choose $\epsilon = Ln^{-1/d}$*, and we have*

$$\mathscr{R}_n(\mathscr{F}_L^d) \lesssim Ln^{-1/d}.$$

3.5 Upper bounds for l_0-"balls" (Wainwright 5.7)

Solution 19. *(a) We have*

$$\mathscr{G}(\mathbb{T}_d^s) = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\sup_{\theta\in\mathbb{T}_d^s}\langle\theta, w\rangle\right] = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\max_{|S|=s}\|w_S\|_2\right]$$

The second equality follows from Cauchy-Schwarz inequality

$$\langle\theta, w\rangle \le \|w_S\|_2\|\theta\|_2 \le \|w_S\|_2,$$

and the "=" holds when $\theta = w_S/\|w_S\|_2$.

(b) By Cauchy-Schwarz inequality,

$$\mathbb{E}[\|w_S\|_2] \le \sqrt{\mathbb{E}[\|w_S\|_2^2]} = \sqrt{s},$$

As $w \mapsto \|w_S\|_2$ is $1-$Lipchitz and w is a normal Gaussian random variable, we can prove that $\|w_S\|_2$ is $sG(1)$, which leads to the result.

(c) By maximal inequality.

$$\begin{aligned}
\mathbb{E}\left[\max_S\|w_S\|_2\right] &\le \sqrt{s} + \mathbb{E}\left[\max_S|w_S - \mathbb{E}\|w_S\|_2|\right] \\
&\le \sqrt{s} + \sqrt{2\sigma^2\log\left(\binom{d}{s}\right)} \\
&\le \sqrt{s} + \sqrt{2s\log\left(\frac{ed}{s}\right)} \\
&\lesssim \sqrt{s\log\left(\frac{ed}{s}\right)}.
\end{aligned}$$

The third line follows from

$$\binom{d}{s} \le \frac{d^s}{s!} \le \frac{d^s}{(s/e)^s} = \left(\frac{ed}{s}\right)^s.$$

The last line follows from

$$\sqrt{s} \le \sqrt{s\log e} \le \sqrt{s\log\left(\frac{ed}{s}\right)}.$$

3.6 Uniform laws and logistic loss (Duchi 7.1)

Solution 20. *WLOG, we can assume* $n \geq 2$. *Denote* $X_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i, Y_i) - M_\theta(X, Y)$, $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\| \leq r\}$, $\mathscr{X} = \{x \in \mathbb{R}^d : \|x\|_* \leq M\}$, $\mathscr{F} = \{m_\theta(z) : x \in \mathscr{X}, y \in \{\pm 1\}, \theta \in \Theta\}$ *Notice that*

$$
\begin{aligned}
m_\theta(X_i, Y_i) - M_\theta(X, Y) &\leq \sup_{x \in \mathscr{X}, y \in \{\pm 1\}, \theta \in \Theta} \log(1 + \exp(-y\theta^\top x)) - \inf_{X \in \mathscr{X}, Y \in \{\pm 1\}, \theta \in \Theta} \log(1 + \exp(-y\theta^\top x)) \\
&\leq \log(1 + e^{Mr}) - \log(1 + e^{-Mr}) \leq Mr.
\end{aligned}
$$

From this, we have

- X_θ *is* $sG(\frac{Mr}{\sqrt{n}})$ *on* θ.
- $\sup_\theta X_\theta$ *is* $(\frac{Mr}{n}, ..., \frac{Mr}{n})$*-bounded, which implies* $\sup_\theta(X_\theta)$ *is* $sG(\frac{Mr}{\sqrt{n}})$ *on* $z_1, ..., z_n$.

As a result,

$$
\mathbb{P}(\sup_\theta X_\theta - \mathbb{E}\,[\sup_\theta X_\theta] \geq \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2M^2r^2}\right).
$$

By one-step discretization, we have

$$
\mathbb{E}\,[\sup_\theta X_\theta] \leq \inf_{\epsilon'} \left(\epsilon' + \sqrt{2\frac{M^2r^2}{n}\log N(\epsilon'; \mathscr{F}, L^\infty)}\right). \tag{1}
$$

To bound the covering number of $\mathscr{F}$, *we next prove the Lipschitz condition of* m_θ.

$$
\begin{aligned}
\|m_{\theta_1}(z) - m_{\theta_2}(z)\|_\infty &= \sup_{x \in \mathscr{X}, y \in \{\pm 1\}} \langle \partial_\theta(x, y), \theta_1 - \theta_2 \rangle \\
&\leq \sup_{x \in \mathscr{X}, y \in \{\pm 1\}} \langle \frac{\exp(-y\theta^\top x)}{1 + \exp(-y\theta^\top x)}(-yx), \theta_1 - \theta_2 \rangle \\
&\leq \sup_{x \in \mathscr{X}} \|x\|_* \|\theta_1 - \theta_2\| \\
&\leq M\|\theta_1 - \theta_2\|.
\end{aligned}
$$

So $m_\theta(z)$ *is an* M*-Lipschitz function and we can bound the covering number now.*

$$
N(\epsilon'; \mathscr{F}, L^\infty) \leq N\left(\frac{\epsilon'}{M}; \Theta, \|\cdot\|\right) \leq N\left(\frac{\epsilon'}{Mr}; B_d(1), \|\cdot\|\right) \leq \left(\frac{Mr}{\epsilon'}\right)^d.
$$

Substitute the covering number into the equation (1) and we have

$$
\begin{aligned}
\mathbb{E}\,[\sup_\theta X_\theta] &\leq \inf_{\epsilon'} \left(\epsilon' + \sqrt{2\frac{M^2r^2}{n} d \log\left(\frac{Mr}{\epsilon'}\right)}\right) \\
&\leq \frac{Mr}{\sqrt{n}} + \sqrt{\frac{M^2r^2}{n} d \log n} \\
&\lesssim \sqrt{\frac{M^2r^2}{n} d \log n}.
\end{aligned}
$$

The second line follows from choosing $\epsilon' = \frac{Mr}{\sqrt{n}}$, *and the third line holds if* $n \geq 2$. *Then we choose* $\epsilon_n(\delta) = C\sqrt{\frac{r^2M^2}{n}\left(d\log n + \log\frac{1}{\delta}\right)} \geq \mathbb{E}\,[\sup_\theta X_\theta] + \sqrt{\frac{2r^2M^2}{n}\log\frac{1}{\delta}}$, *thus*

$$
\mathbb{P}(\sup_\theta X_\theta \geq \epsilon(\delta)) \leq \mathbb{P}\left(\sup_\theta X_\theta - \mathbb{E}\,[\sup_\theta X_\theta] \geq \sqrt{\frac{2r^2M^2}{n}\log\frac{1}{\delta}}\right) \leq \delta.
$$

4 Homework 4

4.1 Lower bounds for l_0-"balls"(Wainwright 5.8)

Solution 21. *(a) Denote that*

$$T = \Big\{\theta \in \mathbb{R}^d : \theta_i \in \{-s^{-1/2}, 0, s^{1/2}\}, \|\theta\|_0 \le s\Big\} \subset T^d(s),$$

and we have the inequlity of the covering number

$$N\left(1/\sqrt{2}; T^d(s), \rho_{Euc}\right) \ge N\left(1/\sqrt{2}; T, \rho_{Euc}\right).$$

Observe that for fixed point $A \in T$, we have

$$\left|\{x \in T : \rho_{Euc}(x, A) \le 1/\sqrt{2}\}\right| = |\{x \in T : d_{Hamming}(x, A) \le s/2\}| \le \sum_{j=1}^{s/2-1} \binom{d}{j} 2^j.$$

And we can get

$$\begin{aligned} N\left(1/\sqrt{2}; T, \rho_{Euc}\right) &\ge \frac{3^d}{\sum_{j=1}^{s/2-1} \binom{d}{j} 2^j} \\ &\ge \frac{3^d}{3^s \binom{d}{s}} \\ &\ge \frac{3^d}{(3ed/s)^s} \\ &\ge \left(\frac{ed}{s}\right)^s. \end{aligned}$$

The last line is equivalent to

$$3^{d/2s} \ge \sqrt{3}\frac{ed}{s} \Longleftrightarrow 3^k \ge 3e^2k^2,$$

which can be proved by

$$3^k \ge (3^{k/6})^6 \ge (k/2)^6 = k^2\frac{k^4}{64} \ge k^2\frac{21^4}{64} \ge 3e^2k^2.$$

(b) By (a) and the properties of packing number

$$M(1/\sqrt{2}; T^d(s), \rho_{Euc}) \ge N(1/\sqrt{2}; T^d(s), \rho_{Euc}) \ge \left(\frac{ed}{s}\right)^s.$$

Define Guassian Process $X_\theta = \langle w, \theta\rangle$, $\mathscr{G}(T^d(s)) = \mathbb{E}_w\left[\sup_{\theta \in T^d(s)} X_\theta\right]$. By Sudakov Minority,

$$\begin{aligned} \mathbb{E}_w\left[\sup_{\theta \in T^d(s)} X_\theta\right] &\ge \frac{1/\sqrt{2}}{2}\sqrt{\log M(1/\sqrt{2}; T^d(s), \rho_{Euc})} \\ &\gtrsim \sqrt{s\log\left(\frac{ed}{s}\right)}. \end{aligned}$$

4.2 Sub-Gaussian matrices and mean bounds (Wainwright 6.8)

Solution 22. *(a) First we prove that*

$$\mathbb{E}\left[e^{\lambda \cdot \lambda_{\max}(S_n)}\right] \leq d e^{\frac{\lambda^2 \sigma^2}{2n}}. \tag{2}$$

Proof. We have

$$\mathbb{E}\left[e^{\lambda \cdot \lambda_{\max}(S_n)}\right] \leq \mathbb{E}\left[\operatorname{tr}\left(e^{\lambda S_n}\right)\right] \leq \operatorname{tr}\left(\prod_i \left(\mathbb{E}[e^{\frac{\lambda Q_i}{n}}]\right)\right).$$

Notice that

$$\mathbb{E}[e^{\frac{\lambda Q_i}{n}}] \preceq e^{\frac{\lambda^2 V_i}{2n^2}},$$

and we can get

$$\operatorname{tr}\left(\prod_i \left(\mathbb{E}[e^{\frac{\lambda Q_i}{n}}]\right)\right) \leq \operatorname{tr}\left(\exp\left(\sum_{i=1}^n \frac{\lambda^2 V_i}{2n^2}\right)\right) \leq d \exp\left(\left\|\sum_{i=1}^n \frac{\lambda^2 V_i}{2n^2}\right\|_{op}\right) \leq d e^{\frac{\lambda^2 \sigma^2}{2n}}.$$

□

Then, by Jensen's inequality,

$$\mathbb{E}[\lambda_{\max}(S_n)] = \frac{1}{\lambda}\mathbb{E}\left[\log e^{\lambda \cdot \lambda_{\max}(S_n)}\right] \leq \frac{1}{\lambda}\log \mathbb{E}\left[e^{\lambda \cdot \lambda_{\max}(S_n)}\right] \leq \frac{\log d}{\lambda} + \frac{\lambda \sigma^2}{2n}.$$

Choose $\lambda = \frac{\sqrt{2n \log d}}{\sigma}$ *and we derive the result.*

(b) Define

$$\tilde{Q}_i = \begin{pmatrix} Q_i & 0 \\ 0 & -Q_i \end{pmatrix},$$

$\tilde{S}_n = \sum_{i=1}^n \frac{1}{n}\tilde{Q}_i$, *and use the result in (a) for* $\tilde{S}_n$,

$$\mathbb{E}[\lambda_{\max}(\tilde{S}_n)] \leq \sqrt{\frac{2\sigma^2 \log(2d)}{n}}.$$

Notice that

$$\lambda_{\max}(\tilde{S}_n) = \left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_{op},$$

and we derive the result.

4.3 Tail bounds for non-symmetric matrices (Wainwright 6.10)

Solution 23. *(a) We have*

$$\|Q_i\|_{op}^2 = \lambda_{\max}[Q_i^\top Q_i] = \lambda_{\max}\left(\begin{pmatrix} A_iA_i^\top & 0 \\ 0 & A_i^\top A_i \end{pmatrix}\right) = \lambda_{\max}\left(A_iA_i^\top\right) = \|A_i\|_{op}^2$$

which implies $\|Q_i\|_{op} = \|A_i\|_{op}$.

(b) We have

$$\mathrm{Var}(Q_i) = \mathbb{E}[Q_i^2] = \mathbb{E}\left[\begin{pmatrix} A_iA_i^\top & 0 \\ 0 & A_i^\top A_i \end{pmatrix}\right],$$

thus

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Q_i)\right\|_{op} &= \left\|\mathbb{E}\left[\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n A_iA_i^\top & 0 \\ 0 & \frac{1}{n}\sum_{i=1}^n A_i^\top A_i \end{pmatrix}\right]\right\|_{op} \\ &= \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[A_iA_i^\top]\right\|_{op}, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[A_i^\top A_i]\right\|_{op}\right\} = \sigma^2.\end{aligned}$$

(c) Denote that $S_n = \sum_{i=1}^n Q_i$, *by conditions and (b),*

$$\|Q_i\|_{op} \le b \quad \textit{and} \quad \|\mathrm{Var}(S_n)\|_{op} \le n\sigma^2,$$

So by matrix Bernstein's inequality, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n A_i\right\|_{op} \ge n\delta\right) \le 2(d_1+d_2)e^{-\frac{n\delta^2}{2(\sigma^2+b\delta)}}.$$

4.4 Pairwise incoherence (Wainwright 7.3)

Solution 24. *(a) Assume v is the eigenvector according to the $\lambda_{\min}$, and $|v_1| = \max_i |v_i|$. Assume*

$$\frac{1}{n} X_S^\top X_S = \begin{pmatrix} a_{11} & \cdots & a_{1S} \\ \cdot & \cdots & \cdot \\ a_{S1} & \cdots & a_{SS} \end{pmatrix}.$$

By pairwise incoherence, we have$a_{ii} \in [1 - \frac{\gamma}{s}, 1 + \frac{\gamma}{s}]$ and $a_{ij} \in [-\frac{\gamma}{s}, \frac{\gamma}{s}]$. We compare the first coordinate of the eigenvector

$$\lambda_{\min} v_1 = \sum_{i=1}^{s} a_{1i} v_i \Longrightarrow a_1 - \lambda_{\min} \le \sum_{i=2}^{s} \frac{|v_i|}{|v_1|} |a_{1i}| \Longrightarrow \lambda_{\min} \ge 1 - s\frac{\gamma}{s} = 1 - \gamma =: c(\gamma).$$

(b) Assume $\Delta \in Null \bigcap \mathbb{C}(S)$, we have

$$X\Delta_S = -X\Delta_{S^c} \text{ and } \|\Delta_S\|_1 \ge \|\Delta\|_{S^c}.$$

We know

$$\left\| \frac{X\Delta_S}{\sqrt{n}} \right\|^2 \ge \lambda_{\min} \|\Delta_S\|_2^2 \ge (1-\gamma)\|\Delta_S\|_2^2.$$

On the other hand,

$$\left\| \frac{X\Delta_S}{\sqrt{n}} \right\|^2 = \left\langle \frac{X\Delta_S}{\sqrt{n}}, -\frac{X\Delta_{S^c}}{\sqrt{n}} \right\rangle = \left| \Delta_S \left(\frac{X^\top X}{n} - I \right) \Delta_{S^c} \right| \le \frac{\gamma}{s} \|\Delta_S\|_1 \|\Delta_{S^c}\|_1 \le \sqrt{s}\frac{\gamma}{s} \|\Delta_S\|_2 \|\Delta_{S^c}\|_1.$$

Thus,

$$(1-\gamma)\|\Delta_S\|_2^2 \le \frac{\gamma}{\sqrt{s}} \|\Delta_S\|_2 \|\Delta_{S^c}\|_1 \Longrightarrow (1-\gamma)\|\Delta_S\|_1 \le \sqrt{s}(1-\gamma)\|\Delta_S\|_2 \le \gamma \|\Delta_{S^c}\|_1 \le \gamma \|\Delta_S\|_1.$$

If $\gamma < \frac{1}{3}$, then $\Delta_S = 0 = \Delta_{S^c}$, which implies that

$$Null \bigcap \mathbb{C}(S) = \emptyset.$$

4.5 Pairwise incoherence and RIP for isotropic ensembles (Wain- wright 7.7)

Solution 25. *(a) Apply Bernstein's inequality:*

$$\mathbb{P}\left(\max_{ij}|X^\top X/n - I_d| \geq \frac{1}{3s}\right) \leq d\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_{1i}^2 - 1\right| \geq \frac{1}{3s}\right) + (d^2-d)\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_{1i}X_{2i}\right| \geq \frac{1}{3s}\right)$$
$$\lesssim d^2 \exp\left(-Cn\min\left\{\frac{1}{s^2}, \frac{1}{s}\right\}\right) =: \delta.$$

Solve the equation and we only need:

$$n \gtrsim s^2 \log\left(\frac{d^2}{\delta}\right) \gtrsim s^2 \log\left(\frac{d}{\delta}\right)$$

(b) First, we fix $|S| = s$ and bound the $\|X_S^\top X_S/n - I_S\|_{op}$. By concentration of sample covariance of sub-Gaussian, we have

$$\|X_S^\top X_S/n - I_S\|_{op} \lesssim \frac{s + \log(1/\delta_1)}{n} + \sqrt{\frac{s+\log(1/\delta_1)}{n}} \leq \frac{1}{3}.$$

Choose

$$\delta_1 \sum_{i=1}^{2s}\binom{d}{k} \leq \delta,$$

we have

$$\delta_1 \leq \frac{\delta}{\binom{d}{2s}} \leq \delta\left(\frac{2s}{ed}\right)^{2s}.$$

And we only need

$$n \gtrsim s + s\log\left(\frac{ed}{2\delta s}\right) \gtrsim s\log\left(\frac{ed}{s}\cdot\frac{1}{\delta}\right).$$

4.6 Weighted l_1-norms (Wainwright 7.6)

Solution 26. *(a)* ***Statement:*** *Let* $\tilde{X} = (\frac{1}{w_1}X^{(1)}, \cdots, \frac{1}{w_d}X^{(d)})$*,* $Null(\tilde{X}) = \{\Delta : \tilde{X}\Delta = 0\}$*. For* $\Delta \in Null(\tilde{X})$*, we have*

$$\max_{|S|=k} \|\Delta_S\|_1 \leq \frac{1}{2}\|\Delta\|_1.$$

Proof. Let $\tilde{\theta_i} = w_i|\theta_i|$, we can rewrite the minimization problem as

$$\min_{\tilde{\theta}\in\mathbb{R}^d} \|\tilde{\theta}\|_1 \text{ such that } \tilde{X}\tilde{\theta} = y.$$

The necessary and sufficient conditions for uniquely recovery are

$$Null(\tilde{X}) \bigcap \mathbb{C}(S) = \{0\}$$

for all $|S| = k$. The condition can be written as: for $\Delta \in Null(\tilde{X})$, we have

$$\max_{|S|=k} \|\Delta_S\|_1 \leq \frac{1}{2}\|\Delta\|_1.$$

□

(b) ***Statement:*** *The condition for recovery is:*

$$c_{\min} > 2\frac{s}{t}\delta_{PW}(X).$$

Proof. Let $\theta \in Null(\tilde{X})$, $|S| = s$, we'll show that

$$\|\theta_{S^c}\|_1 \geq \|\theta_S\|_1.$$

Notice that

$$0 = \|\tilde{X}\theta\|_2^2 = \left\|X_S\theta_S + \frac{1}{t}X_{S^c}\theta_{S^c}\right\|_2^2 \geq \theta_S^\top X_S^\top X_S\theta_S + 2\frac{1}{t}\theta_{S^c}X_{S^c}X_S\theta_S.$$

So we derive

$$\begin{aligned} &\frac{1}{s}c_{\min}\|\theta_S\|_1^2 \leq c_{\min}\|\theta_S\|_2^2 = \lambda_{\min}\left(\frac{X_S^\top X_S}{n}\right)\|\theta_S\|_2^2 \\ \leq &\theta_S^\top\left(\frac{X_S^\top X_S}{n}\right)\theta_S \leq 2\frac{1}{t}|\theta_{S^c}X_{S^c}X_S\theta_S| \leq 2\frac{1}{t}\|\theta_S\|_1\|\theta_{S^c}\|_1\left\|\frac{X_{S^c}^\top X_S}{n}\right\|_{\max}. \end{aligned}$$

Thus

$$\frac{\|\theta_{S^c}\|_1}{\|\theta_S\|_1} \geq \frac{tc_{\min}}{2st\left\|\frac{X_{S^c}^\top X_S}{n}\right\|_{\max}} \geq \frac{c_{\min}}{2s\delta_{PW}(X)} > 1.$$

□

When $t \to \infty$*, the conditions will be naturally satisfied.*

4.7 Sharper bounds for Lasso (Wainwright 7.15)

Solution 27. *(a) By exercise 7.9*

$$L_1(s) = \{\|\Delta\|_2 \le 1, \|\Delta\|_1 \le \sqrt{s}\} \subset 2\,\mathrm{conv}\{\|\Delta\|_2 \le 1, \|\Delta\|_0 \le s\} = 2\,\mathrm{conv}\, L_0(s).$$

Thus,

$$\sup_{\Delta \in L_1(s)} \left| \left\langle \Delta, \frac{1}{n} X^\top w \right\rangle \right| \le 2 \sup_{\Delta \in \mathrm{conv}\, L_0(s)} \left| \left\langle \Delta, \frac{1}{n} X^\top w \right\rangle \right| = 2 \sup_{\Delta \in L_0(s)} \left| \left\langle \Delta, \frac{1}{n} X^\top w \right\rangle \right| := Z'.$$

Next, we prove that, w.p. $\ge 1 - c_2 e^{-c_3 n \delta^2}$,

$$\frac{Z'}{C\sigma} \le c_1 \sqrt{\frac{s \log(ed/s)}{n}} + \delta.$$

Proof. Let $f(w) = \sup_\Delta \langle \Delta, \frac{1}{n} X^\top w \rangle$ and $Z' = f(w)$. Notice that

$$f(w_1) - f(w_2) \le \frac{1}{n} \|X\|_{op} \|w_1 - w_2\|_2 \le \frac{C}{\sqrt{n}} \|w_1 - w_2\|_2.$$

Hence, f is $\frac{C}{\sqrt{n}}$-Lipschitz and by concentration we have

$$\mathbb{P}(Z' - \mathbb{E}[Z']) \le \exp\left(-\frac{n\delta^2}{2C^2} \right).$$

Then, we bound the expectation.

$$\mathbb{E}[Z'] \le \mathbb{E}\left[\sup_{|S| \le s, \|\Delta\|_2 = 1} \frac{2}{n} \|X_S^\top w\|_2 \right].$$

Using chaining method and $\|X_S^\top w\|_2$ is $C\sqrt{n}$-Lipschitz, we have

$$\mathbb{E}[\|X_S^\top w\|_2] \le \mathbb{E}[\sup_{\theta \in \mathbb{S}^{s-1}} \langle \theta, X_s^\top w \rangle] \le \inf_\epsilon \epsilon \mathbb{E}[\|X_S^\top w\|_2] + C\sigma\sqrt{n} \int_0^{\sqrt{2}} \sqrt{s \log(1 + \frac{2}{u})}\, \mathrm{d}u.$$

$$\implies \sup_{|S| \le s} \mathbb{E}[\|X_S^\top w\|_2] \le C\sigma\sqrt{n}\sqrt{s} \lesssim C\sigma\sqrt{n}\sqrt{s \log \frac{ed}{s}}.$$

Combine them together, we have

$$\mathbb{E}[Z'] \lesssim C\sigma \sqrt{\frac{s \log(ed/s)}{n}}.$$

Combine the concentration inequality and we prove the result. □

(b) By the LASSO condition, we have

$$\kappa \|\Delta\|_2^2 \le \frac{1}{n} \|X\Delta\|_2^2 \le \frac{2w^\top X \Delta}{n},$$

where $\Delta = \hat{\theta} - \theta^*$. *We have*

$$\kappa \|\Delta\|_2 \le 2 \left\langle \frac{\Delta}{\|\Delta\|_2}, \frac{1}{n} X^\top w \right\rangle.$$

Notice that $\|\Delta\|_1 \le 2\|\hat{\theta}_S - \theta^*_S\|_1 \le 2\sqrt{s}$, *we have*

$$\kappa\|\Delta\|_2 \le 2 \sup_{\Delta \in L_1(2s)} \left| \left\langle \Delta, \frac{1}{n} X^\top w \right\rangle \right| \le 2C\sigma \sqrt{\frac{2s\log(ed/2s)}{n}},$$

$$\Longrightarrow \|\Delta\|_2 \lesssim \frac{C\sigma}{\kappa} \sqrt{\frac{s\log(ed/s)}{n}}.$$

w.p. $\ge 1 - c_2' e^{-c_3' n\delta^2}$.

5 Homework 5

5.1 Restricted isometry property (RIP) implies restricted eigen- value (RE) (Wainwright 7.10)

Solution 28. *(a) Again, we use the result of Exercise 7.9.*

$$L_1(s) = \{\|\Delta\|_2 \leq 1, \|\Delta\|_1 \leq \sqrt{s}\} \subset 2\,\mathrm{conv}\{\|\Delta\|_2 \leq 1, \|\Delta\|_0 \leq s\} = 2\,\mathrm{conv}\, L_0(s).$$

Thus, if $\|\theta\|_1 \leq \sqrt{s}\|\theta\|_2$, $\tilde{\theta} = \frac{\theta}{\|\theta\|_2} \in 2\,\mathrm{conv}(L_0(s))$. *Let* $\tilde{\theta} = \sum_i \alpha_i \theta_i$

$$\left(\sum_i \alpha_i \theta_i\right)^\top \Gamma \left(\sum_i \alpha_i \theta_i\right) = \sum_i \alpha_i^2 \theta_i^\top \Gamma \theta_i + \sum_{i,j} \alpha_{ij} \theta_i^\top \Gamma \theta_j.$$

By condition

$$\left|\theta_i^\top \Gamma \theta_i\right| \leq 4\delta.$$

$$4\left|\theta_i^\top \Gamma \theta_j\right| \leq (\theta_i + \theta_j)^\top \Gamma (\theta_i + \theta_j) - (\theta_i - \theta_j)^\top \Gamma (\theta_i - \theta_j) \leq 32\delta \Longrightarrow \left|\theta_i^\top \Gamma \theta_j\right| \leq 8\delta.$$

Combine them together, we have

$$\left(\sum_i \alpha_i \theta_i\right)^\top \Gamma \left(\sum_i \alpha_i \theta_i\right) \leq 4\delta \sum_i \alpha_i^2 + 16\delta \sum_{i<j} \alpha_i \alpha_j \leq 8\delta.$$

And by rescaling,

$$\left|\theta^\top \Gamma \theta\right| \leq 8\delta \|\theta\|_2^2.$$

Otherwise,

$$\|\theta\|_1 > \sqrt{s}\|\theta\|_2.$$

Consider $\tilde{\theta} = \frac{\theta}{\|\theta\|_1/\sqrt{s}}$, *we have* $\|\tilde{\theta}\|_2 \leq 1$ *and* $\|\tilde{\theta}\|_1 \leq \sqrt{s}$. *Then similarly to the above proof, we can prove result in this case.*

In conclusion, we have

$$\left|\theta^\top \Gamma \theta\right| \leq \begin{cases} 8\delta\|\theta\|_2^2 & , \|\theta\|_1 \leq \sqrt{s}\|\theta\|_2; \\ \frac{8\delta}{s}\|\theta\|_1^2 & , \|\theta\|_1 > \sqrt{s}\|\theta\|_2. \end{cases}$$

(b) Let X *satisfy* $\mathrm{RIP}(2s)$, *then for any* $d \times 2s$ *submatrix* X_{2s},

$$(1 - \delta_{2s})\|\Delta\|_2^2 \leq \frac{1}{d}\|X_{2s}\Delta\|_2^2 \leq (1 + \delta_{2s})\|\Delta\|_2^2.$$

$$\Longrightarrow \left|\left\langle \theta, \left(\frac{1}{d}X^\top X - \mathrm{Id}\right)\theta \right\rangle\right| \leq \delta_{2s}, \quad \textit{for all } \theta \in L_0(2s).$$

Then, we prove the result for all $\theta \in C_\alpha(S)$, $|S| \leq s$. *By part (a), if we have*

$$\|\theta\|_1 \leq \sqrt{s}\|\theta\|_2.$$

Then

$$\left|\left\langle \theta, \left(\frac{1}{d}X^\top X - \mathrm{Id}\right)\theta \right\rangle\right| \leq 12\delta_{2s}\|\theta\|_2^2,$$

and $\kappa = 1 - 12\delta_{2s}$. *Otherwise*

$$\left|\left\langle \theta, \left(\frac{1}{d}X^\top X - \mathrm{Id}\right)\theta \right\rangle\right| \leq 12\frac{\delta_{2s}(1+\alpha)^2}{s}\|\theta\|_1^2 \leq 12\delta_{2s}(1+\alpha)^2\|\theta\|_2^2,$$

and $\kappa = 1 - 12\delta_{2s}(1+\alpha)^2$.

(c) Hard Problem.

5.2 Random matrices satisfy RIP (Exercise in Lecture 18)

Solution 29. *We only need to prove w.p.* $\geq 1-\delta$

$$\left\|\frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}\right\|_{op} \leq \epsilon \quad \textit{for any } |S| = 2s.$$

We use covering number to prove the result.

$$\left\|\frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}\right\|_{op} = \sup_{\|u\|_2=1} \left\langle u, \left(\frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}\right) u \right\rangle = \sup_{\|u\|_2=1} \frac{1}{n}\sum_{i=1}^n (x_i^\top u)^2 - 1.$$

Using $\epsilon-$*net* $\mathscr{N}$ *to cover* $\mathbb{S}^{2s-1}$, $A = \frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}$,

$$\|A\|_{op} = \langle Ax, x\rangle \leq \langle Ax_0, x_0\rangle + \langle Ax_0, x - x_0\rangle + \langle A(x-x_0), x\rangle \leq \sup_{x_0\in\mathscr{N}} \langle Ax_0, x_0\rangle + 2\epsilon\|A\|_{op}.$$

Choose $\epsilon = \frac{1}{4}$ *and we get*

$$\left\|\frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}\right\|_{op} \leq 2 \sup_{u\in\mathscr{N}} \frac{1}{n}\sum_{i=1}^n (x_i^\top u)^2 - 1.$$

Notice that the covering number can be bounded

$$|\mathscr{N}| \leq \left(\frac{2}{\epsilon} + 1\right)^{2s} = 9^{2s}.$$

Fix u, $x_i^\top u \sim \mathscr{N}(0,1)$. *By Hoeffding's inequality,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (x_i^\top u)^2 - 1 \geq \epsilon/2\right) \leq \exp\left(-\frac{n\epsilon^2}{8}\right).$$

By union bound, we get

$$\mathbb{P}\left(\sup_{u\in\mathscr{N}} \frac{1}{n}\sum_{i=1}^n (x_i^\top u)^2 - 1 \geq \epsilon/2\right) \leq 9^{2s}\exp\left(-\frac{n\epsilon^2}{8}\right),$$

hence

$$\mathbb{P}\left(\left\|\frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}\right\|_{op} \geq \epsilon\right) \leq 9^{2s}\exp\left(-\frac{n\epsilon^2}{8}\right).$$

By union bound, we get

$$\mathbb{P}\left(\left\|\frac{X_S^\top X_S}{n} - \mathrm{I}_{2s}\right\|_{op} \geq \epsilon \quad \textit{for any } |S| = 2s\right) \leq \binom{d}{2s} 9^{2s}\exp\left(-\frac{n\epsilon^2}{8}\right) \leq \left(\frac{9ed}{2s}\right)^{2s}\exp\left(-\frac{n\epsilon^2}{8}\right) := \delta.$$

Hence

$$n \geq 8\frac{\log(1/\delta) + 2s\log(9ed/2s)}{\epsilon^2} \gtrsim \frac{\log(1/\delta) + s\log(ed/s)}{\epsilon^2}.$$

5.3 Analysis of square-root Lasso (Wainwright 7.17)

Solution 30. *Consider 4 problems*

$$\arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{2n}\|y-X\theta\|_2^2+\lambda_n\|\theta\|_1\right\}. \tag{3}$$

$$\arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{\sqrt{n}}\|y-X\theta\|_2+\gamma_n\|\theta\|_1\right\}. \tag{4}$$

$$\arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{2n}\|y-X\theta\|_2^2\right\}, \quad s.t.\|\theta\|_1\le R_1. \tag{5}$$

$$\arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{\sqrt{n}}\|y-X\theta\|_2\right\}, \quad s.t.\|\theta\|_1\le R_2. \tag{6}$$

In the lecture, we proved the 3 and 5 are equivalent. For the same reason 4 and 6 are equivalent. Obviously, 5 and 6 are equivalent. Thus 3 and 4 are equivalent.

(b) Take the derivative of the objective function, we have

$$0=\frac{\partial}{\partial\theta}\left\{\frac{1}{\sqrt{n}}\|y-X\theta\|_2+\gamma_n\|\theta\|_1\right\}=\frac{\frac{1}{n}X^T(X\hat{\theta}-\mathbf{y})}{\frac{1}{\sqrt{n}}\|\mathbf{y}-X\hat{\theta}\|_2}+\gamma_n\hat{z}.$$

(c) By definition,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|y-X\hat{\theta}\|_2+\gamma_n\|\hat{\theta}\|_1 &\le \frac{1}{\sqrt{n}}\|y-X\theta^*\|_2+\gamma_n\|\theta^*\|_1\\
\Longrightarrow \frac{1}{\sqrt{n}}\|w-X\hat{\Delta}\|_2-\frac{1}{\sqrt{n}}\|w\|_2 &\le \gamma_n\|\theta^*\|_1-\gamma_n\|\hat{\theta}\|_1\\
\Longrightarrow \left(\frac{1}{\sqrt{n}}\|w-X\hat{\Delta}\|_2-\gamma_n\|\theta^*\|_1+\gamma_n\|\hat{\theta}\|_1\right)^2 &\le \frac{1}{n}\|w\|_2^2
\end{aligned}$$

$$\Longrightarrow \frac{1}{n}\|X\hat{\Delta}\|_2^2-2\left\langle\hat{\Delta},\frac{1}{n}X^\top w\right\rangle+\gamma_n^2(\|\hat{\theta}\|_1-\|\theta^*\|_1)^2-2\gamma_n\frac{\|y-X\hat{\theta}\|_2}{\sqrt{n}}(\|\hat{\theta}\|_1-\|\theta^*\|_1)\le 0.$$

Notice $\|\hat{\theta}\|_1-\|\theta^*\|_1\le\|\hat{\Delta}_S\|_1-\|\hat{\Delta}_{S^c}\|_1$, *we have*

$$\frac{1}{n}\|X\hat{\Delta}\|_2^2\le 2\left\langle\hat{\Delta},\frac{1}{n}X^\top w\right\rangle+2\gamma_n\frac{\|y-X\hat{\theta}\|_2}{\sqrt{n}}(\|\hat{\Delta}_S\|_1-\|\hat{\Delta}_{S^c}\|_1).$$

That's not enough, we need to improve the result using (b).

$$\begin{aligned}
\frac{1}{n}\|X\hat{\Delta}\|_2^2 &= \frac{1}{n}\left(\hat{\theta}-\theta^*\right)^\top X^\top X\left(\hat{\theta}-\theta^*\right)\\
&= \frac{1}{n}\hat{\Delta}\left(X^\top X\hat{\theta}-X^\top y\right)+\frac{1}{n}\hat{\Delta}\left(X^\top y-X^\top X\theta^*\right)\\
&= \frac{1}{n}\langle\hat{\Delta},X^\top w\rangle-\frac{\gamma_n}{\sqrt{n}}\|y-X\theta\|_2\langle\hat{\Delta},\hat{z}\rangle.
\end{aligned}$$

The last line comes from (b). To bound $\langle\hat{\Delta},\hat{z}\rangle$, *we have*

$$\langle\hat{\Delta},-\hat{z}\rangle=\langle\hat{\Delta}_S,-\hat{z}_S\rangle-\langle\hat{\Delta}_{S^c},\hat{z}_{S^c}\rangle\le\|\hat{\Delta}_S\|_1-\|\hat{\Delta}_{S^c}\|_1.$$

Combine them together, we prove the result.

(d) By (c)

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq 2\left\langle \hat{\Delta}, \frac{1}{n}X^\top w \right\rangle + 2\gamma_n \frac{\|y - X\hat{\theta}\|_2}{\sqrt{n}}(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)$$
$$\leq 2\frac{1}{n}\|\hat{\Delta}\|_1\|X^\top w\|_\infty + 2\gamma_n \frac{\|y - X\hat{\theta}\|_2}{\sqrt{n}}(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

If $\|\hat{\Delta}_{S^c}\|_1 \leq \|\Delta_S\|_1$, then the result holds. Otherwise,

$$2\frac{\|X^\top w\|_\infty \|y - X\hat{\theta}\|_2}{n\|w\|_2}(\|\hat{\Delta}_{S^c}\|_1 - \|\hat{\Delta}_S\|_1) \leq \gamma_n \frac{\|y - X\hat{\theta}\|_2}{\sqrt{n}}(\|\hat{\Delta}_{S^c}\|_1 - \|\hat{\Delta}_S\|_1) \leq \frac{1}{n}\|\hat{\Delta}\|_1\|X^\top w\|_\infty$$

We can choose

$$R_2 = \|\theta^*\|_1,$$

n problem 6 and $\|y - X\hat{\theta}\|_2 \geq \|y - X\theta^*\|_2 = \|w\|_2$. (We still need to verify that this choice of R_2 is valid by some optimization theory.)

$$\Longrightarrow 2(\|\hat{\Delta}_{S^c}\|_1 - \|\hat{\Delta}_S\|_1) \leq \|\hat{\Delta}\|_1 = (\|\hat{\Delta}_{S^c}\|_1 + \|\hat{\Delta}_S\|_1),$$

which derives the result.

(e) By RE condition, we have

$$\frac{1}{n}\|X\hat{\Delta}||_2^2 \geq \kappa\|\hat{\Delta}\|_2^2.$$

By the inequality in (d), we have

$$\kappa\|\Delta\|_2^2 \leq \frac{1}{n}\|X\hat{\Delta}\|_2^2 \lesssim \frac{\gamma_n}{\sqrt{n}}\|w\|_2 \left(3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\right) \lesssim \frac{\gamma_n}{\sqrt{n}}\|w\|_2\sqrt{s}\|\hat{\Delta}\|_2.$$

Thus, we have

$$\|\hat{\Delta}\|_2 \lesssim \frac{\|w\|_2}{\sqrt{n}}\gamma_n\sqrt{s}.$$

5.4 Unitarily invariant matrix norms (Wainwright 8.2)

Solution 31. *(a) (i), (ii) and (iii) are unitarily invariant.*

As$\sigma(VMM^\top V^\top) = \sigma(MM^\top)$*, the singular values are invariant under orthogonal transform.*

$$\|M\|_F^2 = \sum_i \sigma_i^2.$$

$$\|M\|_{nuc} = \sum_i |\sigma_i|.$$

$$\|M\|_{op} = \sup_i \sigma_i.$$

Hence (i)-(iii) are unitarily invariant.

However, for (iv), we can choose $M = \operatorname{diag}\{1, 0 \cdots, 0\}$ *and* $V(1) = \frac{1}{\sqrt{d}}(1, \cdots, 1)$*. Then* $\|VMU\|_\infty = \frac{1}{\sqrt{d}} = \|M\|_\infty$.

(b) To prove it is a matrix norm, we need to check the four conditions. As ρ *is a norm on* $\mathbb{R}^{d_1}$*, we have*

$$\|M\|_\rho \geq 0.$$

$$\|M\|_\rho = 0 \text{ if and only if } \sigma_i(M) = 0 \Longleftrightarrow M = 0.$$

$$\|\alpha M\|_\rho = \rho(\alpha\sigma_1(M), \cdots, \alpha\sigma_{d_1}(M)) = |\alpha|\rho(\sigma_1(M), \cdots, \sigma_{d_1}(M)) = |\alpha|\|M\|_\rho.$$

At last, we will prove the triangle inequality.

$$\|M + N\|_\rho = \rho(\sigma_1(M+N), \cdots, \sigma_{d_1}(M+N))$$

First, we define

$$x \prec_w y \Longleftrightarrow \sum_{i=1}^k x_i \leq \sum_{i=1}^k y_i.$$

By Ky Fan's inequality,

$$\sigma(M+N) \prec \sigma(M) + \sigma(N).$$

Proof.

$$\sum_{i=1}^k \sigma_i(M) = \max_{\operatorname{rank}(U)=k} \operatorname{tr}(MU)$$

$$\max_{\operatorname{rank}(U)=k} \operatorname{tr}((M+N)U) \leq \max_{\operatorname{rank}(U)=k} \operatorname{tr}(MU) + \max_{\operatorname{rank}(U)=k} \operatorname{tr}(NU).$$

□

We only need to prove

$$\rho(x) \leq \rho(y) \quad \text{as } x \prec_w y.(Hard)$$

(c) We have $\sigma(VMU) = \sigma(M)$*, as* ρ *is invariant to permutations and sign changes,* $\|\cdot\|_\rho$ *are unitarily invariant.*

5.5 PCA for Gaussian mixture models (Wainwright 8.6)

Solution 32. *(a) Our model is*

$$x_i = \epsilon_i \theta^* + \sigma w_i,$$

where $\epsilon_i \in \mathrm{Unif}\{\pm 1\}$ *and* $w_i \in N(0, I_d)$.

First, we restate the PCA model. The empirical covariance matrix is

$$\hat{\Sigma} = \frac{1}{n}\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i^\top,$$

and

$$\hat{\theta} = \arg\max_{\|\theta\|_2=1} \theta^\top \hat{\Sigma}\theta.$$

Compared to the textbook theorem $v = \frac{1}{\sigma^2}$ *and we have*

$$\|\hat{\theta} - \theta\|_2 \lesssim \sqrt{\frac{v+1}{v^2}}\sqrt{\frac{d}{n}} \lesssim \sigma\sqrt{1+\sigma^2}\sqrt{\frac{d}{n}} \quad w.h.p.$$

(b) The classification rule is

$$\psi(x) = \begin{cases} 1 & , \textit{if } \|x - \hat{\theta}\| \le \|x + \hat{\theta}\|. \\ -1 & , \textit{if } \|x - \hat{\theta}\| > \|x + \hat{\theta}\|. \end{cases}$$

(c) We can rescale the model by $\Sigma^{-\frac{1}{2}}$*:*

$$\hat{x}_i = \Sigma^{-\frac{1}{2}} x_i = \epsilon_i \Sigma^{-\frac{1}{2}}\theta^* + \Sigma^{-\frac{1}{2}} w_i.$$

Similarly, we use the largest eigenvalue of empirical covariance matrix:

$$\hat{\theta} = \arg\max_{\|\theta\|_2=1} \theta^\top \left(\frac{1}{n}\hat{x}_i\hat{x}_i^\top\right)\theta.$$

But now, our estimator doesn't work because as $n \to \infty$,

$$\hat{\theta} \to \frac{\Sigma^{-\frac{1}{2}}\theta^*}{\|\Sigma^{-\frac{1}{2}}\theta^*\|_2} \neq \theta^*.$$

5.6 PCA for retrieval from absolute values (Wainwright 8.7)

Solution 33. *By definition,*

$$\left(\mathbb{E}[y_i^2 x_i x_i^\top]\right)_{kl} = \sum_{ij} \theta_i^* \theta_j^* x_i x_j x_k x_l = \begin{cases} 2\theta_k^* \theta_l^* & , \textit{if } k \neq l. \\ 2\theta_k^{*2} & , \textit{if } k = l. \end{cases}$$

And thus,

$$\mathbb{E}[y_i^2 x_i x_i^\top] = 2{\theta^*}^\top \theta^* + I_d.$$

Obviously,

$$\lambda_{\max}(2{\theta^*}^\top \theta^* + I_d) = 3.$$

Then we can construct our PCA estimator

$$\hat{\theta} = \arg \max_{\|\theta\|_2 = 1} \theta^\top \left(\frac{1}{n} y_i^2 x_i x_i^\top \right) \theta.$$

5.7 Lower bounds on the critical inequality (Wainwright 13.5)

Solution 34. *(a) By critical inequality,*

$$\mathscr{G}(\delta, \mathscr{F}) = \mathbb{E}_w \left[\sup_{\|g\| \leq \delta, g \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} w_i g_i \right] \leq \delta \mathbb{E} \|w\|_2 \leq \delta = \frac{\delta^2}{2\sigma},$$

as $\delta^2 = 4\sigma^2$.

(b) By the lower bound of $\mathscr{G}(\delta, \mathscr{F})$.

$$\frac{\delta^2}{2\sigma} \geq \mathscr{G}(\delta, \mathscr{F}) \geq \frac{\delta}{\sqrt{n}} \sqrt{\frac{2}{\pi}}.$$

Thus, we get

$$\delta^2 \geq \frac{8}{\pi} \frac{\sigma^2}{n}.$$

As $\delta \in (0, 1]$, *we have*

$$\delta^2 \geq \min \left\{ 1, \frac{8}{\pi} \frac{\sigma^2}{n} \right\}.$$

6 Homework 6

6.1 Local Gaussian complexity and adaptivity (Wainwright 13.6)

Solution 35. *(a) By definition, we have*

$$\mathscr{G}(\mathscr{F}_{l_1}(1),\delta) = \mathbb{E}\left[\sup_{\theta\in\mathscr{F}_{l_1}(1),\|f\|_n\le\delta}\sum_{i=1}^n w_i\theta_i\right] \le \mathbb{E}\left[\sup_{\theta\in\mathscr{F}_{l_1}(1)}\sum_{i=1}^n w_i\theta_i\right],$$

where $w_i \in \mathscr{N}(0,\frac{1}{n})$. By contraction inequality,

$$\mathbb{E}\left[\sup_{\theta\in\mathscr{F}_{l_1}(1)}\frac{1}{n}\sum_{i=1}^n w_i\theta_i\right] \lesssim \frac{1}{n}\mathbb{E}\left[\sup_{1\le i\le n} w_i\right]$$

By maximal inequality,

$$\mathbb{E}\left[\sup_{1\le i\le n}\tilde{w}_i\right] \lesssim \sqrt{\log n}.$$

Combine them together, we get

$$\mathscr{G}(\mathscr{F}_{l_1}(1),\delta) \le c_1\frac{\sqrt{\log n}}{n}.$$

By the definition of $\hat{\theta}$,

$$\frac{1}{n}\sum_{i=1}^n(y_i - f_{\hat{\theta}}(e_i))^2 \le \frac{1}{n}\sum_{i=1}^n(y_i - f_{\theta^*}(e_i))^2$$

$$\Longrightarrow \sum_{i=1}^n\left(\hat{\theta}_i - \theta_i^* + \frac{\sigma}{\sqrt{n}}w_i\right)^2 \le \sum_{i=1}^n\left(\frac{\sigma}{\sqrt{n}}w_i\right)^2$$

$$\Longrightarrow \left\|\hat{\theta}-\theta^*\right\|_2^2 \le 2\frac{\sigma}{\sqrt{n}}\sum_{i=1}^n\left(\hat{\theta}_i - \theta_i^*\right)w_i \le 2\sqrt{n}\sigma\mathscr{G}(\mathscr{F}_{l_1}(1),\delta)\|\hat{\theta}-\theta^*\|_1 \le c_1'\sigma\sqrt{\frac{\log n}{n}}.$$

The last line follows from $\|\hat{\theta}-\theta^*\|_1 \le \|\hat{\theta}\|_1 + \|\theta^*\|_1 \le 2$.

(b) WLOG, assume $\theta^* = e_1$. We have

$$1 \ge \|v+e_1\|_1 \ge 1 - |v_1| + \sum_{j\ge 2}|v_j|$$

$$\Longrightarrow \|v\|_1 \le 2|v_1| \le 2\delta\sqrt{n}.$$

Then we write the critical equality

$$\frac{\mathscr{G}_n(\delta,\mathscr{F}^*)}{\delta} \le \frac{\delta}{2\sigma/\sqrt{n}}.$$

Now we analyze the Empirical Gaussian process.

$$\mathscr{G}_n(\delta,\mathscr{F}^*) \le \frac{\sqrt{\log n}}{n}\sup_{\|\theta\|_1\le 1,\|\theta-\theta^*\|_1\le\delta\sqrt{n}}\|\theta-\theta^*\|_1 \le \frac{\sqrt{\log n}}{n}\sup_{\|v+e_1\|\le 1,\|v\|_1\le\delta\sqrt{n}}\|v\|_1 \le 2\delta\frac{\sqrt{\log n}}{\sqrt{n}}.$$

Solve the critical equality and we can choose

$$\delta_n \asymp \sigma\frac{\sqrt{\log n}}{n}.$$

We have

$$\|\hat{\theta} - \theta^*\|_n^2 \lesssim \delta_n^2 \asymp \sigma^2 \frac{\log n}{n^2}$$

$$\Longrightarrow \|\hat{\theta} - \theta^*\|_2^2 = n\|\hat{\theta} - \theta^*\|_n^2 \leq c_2' \sigma^2 \frac{\log n}{n}.$$

Remark 6.1. *I use the definition*

$$\mathscr{G}(\mathscr{F}, \delta) = \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathscr{F}} \sum_{i=1}^n w_i f(x_i)\right],$$

which is consistent with the textbook. The result may be a little different.

6.2 Rates for twice-differentiable functions (Wainwright 13.8)

Solution 36. *We apply Corollary 13.7 of the textbook*

$$\frac{16}{\sqrt{n}} \int_{\delta^2/4\sigma}^{\delta} \sqrt{\log N_n(t, B_n(\delta, \mathscr{F}_C - \mathscr{F}_C)} \leq \frac{\delta^2}{4\sigma}.$$

As $\mathscr{F}_C - \mathscr{F}_C \subset \mathscr{F}_{2C}$. By the result of Example 5.11 and let $\alpha = \gamma = 1$, we get

$$\log N_n(t, B_n(\delta, \mathscr{F}_C - \mathscr{F}_C) \lesssim \left(\frac{1}{t}\right)^{\frac{1}{2}}.$$

Thus, we combine the results and identify that

$$\frac{1}{\sqrt{n}} \delta^{\frac{3}{4}} \lesssim \delta^2 \Longrightarrow \delta^2 \asymp \left(\frac{\sigma^2}{n}\right)^{\frac{4}{5}}.$$

And by Example 13.11 of the textbook, we get

$$\mathbb{P}\left(\|\hat{f} - f^*\|_n^2 \geq c_0 \left(\frac{\sigma^2}{n}\right)^{4/5}\right) \leq c_1 e^{-c_2 (n/\sigma^2)^{1/5}}.$$

6.3 Rates for additive nonparametric models (Wainwright 13.9)

Solution 37. *(a) We have*

$$\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i \hat{\Delta}(X_i)\right| \leq \frac{\sigma}{n}\sum_{j=1}^{d}\left|\sum_{i=1}^{n} w_i \hat{\Delta}_j(X_{ij})\right|.$$

For fixed j, when $t > \delta_{n,j}$, $\|\hat{\Delta}_j\|_n \geq \sqrt{t\delta_{n,j}}$. Then, w.h.p, we have

$$\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i \hat{\Delta}_j(X_{ij})\right| \leq \sqrt{t\delta_{n,j}}\|\hat{\Delta}_j\|_n,$$

by Lemma 13.12 in the textbook. And when $\|\hat{\Delta}_j\|_n \leq \sqrt{t\delta_{n,j}}$,

$$\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i \hat{\Delta}_j(X_{ij})\right| \leq \sup_{\|g_j\|_n \leq \sqrt{t\delta_{n,j}}} \frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i g_j(X_{ij})\right|.$$

Then we consider the union probability,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_j \hat{\Delta}_j(X_{ij})\right| \geq t\delta_{ij} + \sqrt{t\delta_{ij}}\|\hat{\Delta}_j\|_n\right) && (7)\\
\leq &\underbrace{\mathbb{P}\left(\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_j \hat{\Delta}_j(X_{ij})\right| \geq t\delta_{ij}, \hat{\Delta}_j(X_{ij}) \geq \sqrt{t\delta_{n,j}}\right)}_{\mathrm{I}} && (8)\\
+ &\underbrace{\mathbb{P}\left(\frac{\sigma}{n}\left|\sum_{i=1}^{n} w_j \hat{\Delta}_j(X_{ij})\right| \geq \sqrt{t\delta_{ij}}\|\hat{\Delta}_j\|, \hat{\Delta}_j(X_{ij}) \leq \sqrt{t\delta_{n,j}}\right)}_{\mathrm{II}}. && (9)
\end{aligned}$$

By Lemma 13.12, we have

$$\mathrm{I} \leq \exp\left(-\frac{nt\delta_{n,j}}{2\sigma^2}\right).$$

To bound the II*, we prove is a sub-Gaussian first. As*

$$\left|\sup_{\|g\|_n \leq \sqrt{t\delta_{n,j}}} \frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i g_j(X_{ij})\right| - \sup_{\|g\|_n \leq \sqrt{t\delta_{n,j}}} \frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i' g_j(X_{ij})\right|\right| \leq \frac{\sigma}{\sqrt{n}}\sqrt{t\delta_{n,j}}\|w - w'\|_2,$$

we have $\sup_{\|g_j\|_n \leq \sqrt{t\delta_{n,j}}} \frac{\sigma}{n}|w_j g_j(X_{ij})|$ *is* $\frac{\sigma}{\sqrt{n}}\sqrt{t\delta_{n,j}}$*−Lipschitz. Thus,* $\sup_{\|g_j\|_n \leq \sqrt{t\delta_{n,j}}} \frac{\sigma}{n}|w_j g_j(X_{ij})| \sim \mathrm{sG}(\frac{\sigma}{\sqrt{n}}\sqrt{t\delta_{n,j}})$*. Notice that* $\mathscr{G}(\delta_{n,j}) \leq \delta_{n,j}^2 \leq t\delta_{n,j}$*.*

$$\begin{aligned}
\Longrightarrow \mathrm{I} &\leq \mathbb{P}\left(\sup_{\|g_j\|_n \leq \sqrt{t\delta_{n,j}}} \frac{\sigma}{n}|w_j g_j(X_{ij})| \geq 2t\delta_{n,j}\right)\\
&\leq \exp\left(-\frac{4t^2\delta_{n,j}^2}{(\frac{\sigma}{\sqrt{n}}\sqrt{t\delta_{n,j}})^2}\right)\\
&= \exp(-cnt\delta_{n,j}).
\end{aligned}$$

Combine them together we prove the inequality 7. Consider union bound over d dimension, we get

$$\mathbb{P}\left(\frac{\sigma}{n}\sum_{j=1}^{d}\left|\sum_{i=1}^{n} w_j\hat{\Delta}_j(X_{ij})\right| \geq dt\delta_{n,\max} + 2\sqrt{t\delta_{n,\max}}\left(\sum_{j=1}^{d}\|\hat{\Delta}_j\|_n\right)\right) \leq c_1 d e^{-c_2 nt\delta_{n,\max}}.$$

(b) First, by Cauchy-Schwarz inequality, we have

$$\sum_{j=1}^{d}\|\hat{\Delta}_j\|_n \leq \sqrt{d}\sqrt{\sum_{j=1}^{d}\|\hat{\Delta}_j\|_n^2} \leq \sqrt{d}\sqrt{K}\|\hat{\Delta}\|_n.$$

Notice that

$$\hat{\Delta} = \hat{f} - f^* = \sum_{j=1}^{d}\hat{g}_j - \sum_{j=1}^{d} g_j^*.$$

By basic inequality,

$$\|\hat{\Delta}\|_n^2 \leq dt\delta_{n,\max} + \sqrt{t\delta_{n,\max}}\sqrt{dK}\|\hat{\Delta}\|_n \leq dt\delta_{n,\max} + t\delta_{n,\max}dK + \frac{\|\hat{\Delta}\|_n^2}{4}$$

$$\Longrightarrow \|\hat{\Delta}\|_n^2 \leq c_3 K dt\delta_{n,\max} = c_3 K d\delta_{n,\max}^2.$$

The last equal sign follows by choosing $t = \delta_{n,\max}$.

6.4 Properties of Kullback-Leibler divergence (Wainwright 15.3)

Solution 38. *(a) By Jensen's inequality and* $\log x$ *is a concave function of* x*.*

$$-D(P||Q) = -\int P(x) \log \frac{P(x)}{Q(x)} \,\mathrm{d}x = \int P(x) \log \frac{Q(x)}{P(x)} \,\mathrm{d}x \le \log \int P(x) \frac{Q(x)}{P(x)} \,\mathrm{d}x = \log 1 = 0.$$

Thus $D(P||Q) \ge 0$ *and the equality holds if and only if* $P(x) = Q(x)$ *a.s., which means* $p(x) = q(x)$ *a.s..*

(b) As $D(P||Q) = D(Q||P)$*, we only need to prove the first equation.*

$$D(\sum_{j=1}^m \lambda_j P_j || Q) = \int \sum_{j=1}^m \lambda_j P_j(x) \log \frac{\sum_{j=1}^m \lambda_j P_j(x)}{Q(x)} \,\mathrm{d}x$$

We only need to prove that

$$\int \sum_{j=1}^m \lambda_j P_j(x) \log \left(\sum_{j=1}^m \lambda_j P_j(x) \right) \mathrm{d}x \le \sum_{j=1}^m \lambda_j \int P_j(x) \log P_j(x) \,\mathrm{d}x,$$

which follows the Jensen's inequality and that $x \log x$ *is a convex function of* x*.*

(c) Notice that

$$\log \frac{\mathrm{d}P_1 \otimes P_2}{\mathrm{d}Q_1 \otimes Q_2} = \log \frac{P_1}{Q_1} + \log \frac{P_2}{Q_2}.$$

Thus we have

$$D(P_1 \otimes P_2 || Q_1 \otimes Q_2) = \mathbb{E}_{P_1 \otimes P_2} \left[\log \frac{\mathrm{d}P_1 \otimes P_2}{\mathrm{d}Q_1 \otimes Q_2} \right] = \mathbb{E}_{P_1} \left[\log \frac{P_1}{Q_1} \right] + \mathbb{E}_{P_2} \left[\log \frac{P_2}{Q_2} \right] = D(P_1||Q_1) + D(P_2||Q_2).$$

By mathematical induction, we get the result immediately.

6.5 Bounds on the TV distance (Wainwright 15.10)

Solution 39. *(a) By definition,*

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\int |p(x) - q(x)|\nu(\mathrm{d}x).$$

So, we only need to check

$$\left(\int |p(x) - q(x)|\nu(\mathrm{d}x)\right)^2 \leq \int \frac{p(x)^2}{q(x)}\nu(\mathrm{d}x) - 1.$$

By Cauchy-Schwarz inequality, we have

$$\left(\int |p(x) - q(x)|\nu(\mathrm{d}x)\right)^2 \leq \left[\int \frac{(p(x) - q(x))^2}{q(x)}\nu(\mathrm{d}x)\right]\left[\int q(x)\nu(\mathrm{d}x)\right] = \int \frac{p(x)^2}{q(x)}\nu(\mathrm{d}x) - 1.$$

(b) We have

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \quad \textit{and} \quad q(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Denote $p_n(x) = p(x)^{\otimes n}$ *and* $q_n(x) = q(x)^{\otimes n}$ *We have*

$$\int \frac{p_n(x)^2}{q_n(x)}\nu(\mathrm{d}x) = \left(\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2} + \frac{2\theta x}{\sigma^2} - \frac{\theta^2}{\sigma^2}\right)\mathrm{d}x\right)^n = e^{\left(\frac{\sqrt{n}\theta}{\sigma}\right)^2},$$

which leads to the result.

(c) We have

$$p(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) + \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x+\theta)^2}{2\sigma^2}\right),\ q(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Denote $p_n(x) = p(x)^{\otimes n}$ *and* $q_n(x) = q(x)^{\otimes n}$ *We have*

$$\int \frac{p_n(x)^2}{q_n(x)}\nu(\mathrm{d}x) = \left(\frac{1}{4}A + \frac{1}{2}B + \frac{1}{4}C\right)^n.$$

where

$$A = \int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2} + \frac{2\theta x}{\sigma^2} - \frac{\theta^2}{\sigma^2}\right)\mathrm{d}x = e^{\left(\frac{\theta}{\sigma}\right)^2},$$

$$B = \int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2} - \frac{\theta^2}{\sigma^2}\right)\mathrm{d}x = e^{-\left(\frac{\theta}{\sigma}\right)^2},$$

$$C = \int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2} - \frac{2\theta x}{\sigma^2} - \frac{\theta^2}{\sigma^2}\right)\mathrm{d}x = e^{\left(\frac{\theta}{\sigma}\right)^2}.$$

Combine them together, we get the result,

$$\left(\frac{1}{4}A + \frac{1}{2}B + \frac{1}{4}C\right)^n \leq e^{\left(\frac{\sqrt{n}\theta}{\sigma}\right)^2}.$$

6.6 KL divergence for multivariate Gaussian (Wainwright 15.13)

Solution 40. *(a) We prove (b) first, and (a) directly follows from (b).*

(b) We have

$$
\begin{aligned}
\mathbb{E}_{q_1}\left[\log\left(\frac{q_1}{q_2}\right)\right] &= \mathbb{E}_{q_1}\left[\frac{1}{2}\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) + \frac{1}{2}(x-\mu_2)^\top\Sigma_2^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\right] \\
&= \frac{1}{2}\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) + \frac{1}{2}\langle\mu_1-\mu_2, \Sigma_2^{-1}(\mu_1-\mu_2)\rangle + \frac{1}{2}\mathbb{E}_{q_1}\operatorname{tr}(\Sigma_2^{-1}xx^\top) - \frac{1}{2}\mathbb{E}_{q_1}\operatorname{tr}(\Sigma_1^{-1}xx^\top) \\
&= \frac{1}{2}\left\{\langle\mu_1-\mu_2, \Sigma_2^{-1}(\mu_1-\mu_2)\rangle + \log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) + \operatorname{tr}(\Sigma_2^{-1}\Sigma_1) - d\right\}.
\end{aligned}
$$

6.7 Sharper bounds for Gaussian location family (Wainwright 15.8)

Solution 41. *Denote the minimax lower bound as $\mathscr{M}$.*

(a) Use Le Cam's two point method and Pinsker inequality, we have

$$\mathscr{M} \geq \frac{\delta^2}{2}\left(1 - \|P_{2\delta}^n - P_0^n\|_{\mathrm{TV}}\right) \geq \frac{\delta^2}{2}\left(1 - \sqrt{\frac{1}{2}D(P_{2\delta}^n || P_0^n)}\right) = \frac{\delta^2}{2}\left(1 - \frac{2n\delta^2}{\sigma^2}\right) = \frac{\sigma^2}{16n},$$

where we choose $\delta^2 = \frac{\sigma^2}{4n}$.

(b) Use Le Cam's two point method and Le Cam's inequality, we have

$$\mathscr{M} \geq \frac{\delta^2}{2}\left(1 - \|P_{2\delta}^n - P_0^n\|_{\mathrm{TV}}\right) \geq \frac{\delta^2}{2}\left(1 - H(P_{2\delta}^n || P_0^n)\sqrt{1 - \frac{H^2(P_{2\delta}^n || P_0^n)}{4}}\right).$$

First, we calculate the squared Hellinger distance

$$H^2(P_{2\delta} || P_0) = 2 - 2\exp\left(-\frac{\delta^2}{2\sigma^2}\right).$$

Thus, we have

$$H^2(P_{2\delta}^n || P_0^n) \leq n\left(2 - 2\exp\left(-\frac{\delta^2}{2\sigma^2}\right)\right) \leq \frac{n\delta^2}{\sigma^2}.$$

As a result, we can choose $\delta^2 = \frac{\sigma^2}{4n}$ and get

$$\mathscr{M} \geq \frac{\delta^2}{8n}\left(1 - \frac{\sqrt{15}}{8}\right).$$

(c) Use Le Cam's two point method and Problem 6.5, we can choose $\delta^2 = \frac{\sigma^2}{4n}$ and get

$$\mathscr{M} \geq \frac{\delta^2}{2}\left(1 - \|P_{2\delta}^n - P_0^n\|_{\mathrm{TV}}\right) \geq \frac{\delta^2}{2}\left(1 - \frac{1}{2}\sqrt{e^{\frac{4n\delta^2}{\sigma^2}} - 1}\right) = \frac{\sigma^2}{8n}\left(1 - \frac{1}{2}\sqrt{e - 1}\right).$$

6.8 Lower bounds for generalized linear models (Wainwright 15.17)

Solution 42. *(a) By tensorization properties of KL divergence*

$$
\begin{aligned}
D(P_\theta||P_{\theta'}) &= \sum_{i=1}^n \int P_\theta(y_i)\log\left(\frac{P_\theta(y_i)}{P_{\theta'}(y_i)}\right)\\
&= \sum_{i=1}^n \int P_\theta(y_i)\frac{1}{s(\sigma)}\left(y_i(\langle x_i,\theta-\theta'\rangle - \Phi(x_i,\theta)+\Phi(x_i,\theta'))\right)\\
&= \sum_{i=1}^n \frac{1}{s(\sigma)}\mathbb{E}_\theta[Y_i]\left(y_i(\langle x_i,\theta-\theta'\rangle - \Phi(x_i,\theta)+\Phi(x_i,\theta'))\right).
\end{aligned}
$$

By taking derivatives on both sides

$$
\begin{aligned}
1 &= \int P_\theta(y)\\
\Longrightarrow 0 &= \frac{\mathrm{d}}{\mathrm{d}\langle x,\theta\rangle}P_\theta(y) = \mathbb{E}_\theta[Y] - \Phi'(\langle x,\theta\rangle).
\end{aligned}
$$

Thus substitute the result into the calculation of the KL divergence, we have

$$
D(P_\theta||P_{\theta'}) = \frac{1}{s(\sigma)}\sum_{i=1}^n \Phi'(\langle x,\theta\rangle)\left(y_i(\langle x_i,\theta-\theta'\rangle - \Phi(x_i,\theta)+\Phi(x_i,\theta'))\right)
$$

(b) By Taylor's expansion, we have

$$
\Phi(\langle x,\theta'\rangle) = \Phi(\langle x,\theta\rangle) + \Phi'(\langle x,\theta\rangle)\left(\langle x,\theta'\rangle - \langle x,\theta\rangle\right) + \frac{\Phi''(\xi)}{2}\left(\langle x,\theta'\rangle - \langle x,\theta\rangle\right)^2.
$$

Thus, the KL-divergence can be bounded by

$$
\begin{aligned}
D(P_\theta||P_{\theta'}) &= \frac{1}{s(\sigma)}\sum_{i=1}^n \Phi'(\langle x,\theta\rangle)\left(y_i(\langle x_i,\theta-\theta'\rangle - \Phi(x_i,\theta)+\Phi(x_i,\theta'))\right)\\
&\le \frac{1}{s(\sigma)}\frac{L}{2}\sigma_{\max}(X)\|\theta'-\theta\|_2^2.
\end{aligned}
$$

(c) By Fano's method, we have

$$
\inf_{\widehat{\theta}}\sup_{\theta\in\mathbb{B}_2^d(1)}\mathbb{E}\left[\|\widehat{\theta}-\theta\|_2^2\right] \ge \delta^2\left(1-\frac{\max_{\theta,\theta'\in\mathscr{M}} D(P_\theta||P_{\theta'})+\log 2}{\log M}\right),
$$

where $\mathscr{M}$ is a 2δ packing. Notice $\Theta = \{\theta\in\mathbb{R}^d : \|\theta\|_2\le 1\}$, thus choose the net with covering number

$$
\begin{aligned}
|\mathscr{M}| &= \left(1+\frac{1}{2\delta}\right)^d\\
\Longrightarrow \log|\mathscr{M}| &\ge d\log\left(1+\frac{1}{2\delta}\right)
\end{aligned}
$$

By part (b), we have

$$
D(P_\theta||P_{\theta'}) \le \frac{1}{s(\sigma)}\frac{L}{2}\sigma_{\max}^2(X)\|\theta-\theta'\|_2^2 = \underbrace{\frac{1}{s(\sigma)}\frac{L}{2}\eta_{\max}^2}_{c} n\|\theta-\theta'\|_2^2 \le cn\cdot 16\delta^2
$$

Combine them together and we get

$$
\inf_{\widehat{\theta}}\sup_{\theta\in\mathbb{B}_2^d(1)}\mathbb{E}\left[\|\widehat{\theta}-\theta\|_2^2\right] \ge \delta^2\left(1-\frac{cn\cdot 16\delta^2+\log 2}{d\log(1+\frac{1}{2\delta})}\right).
$$

Case 1 *If $cn \leq 100d$, then choose $\delta^2 = c_0$, where c_0 is some constant number. Thus we have*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \geq 1$$

Case 2 *If $cn > 100d$, then choose $\delta^2 = \frac{d}{cn}$. Thus we have*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \geq c \cdot \frac{s(\sigma)}{L\eta_{\max}^2} \cdot \frac{d}{n}.$$

Thus, we have

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \geq \min\left\{1, c \cdot \frac{s(\sigma)}{L\eta_{\max}^2} \cdot \frac{d}{n}\right\}.$$

(d) Choose $y_i = \langle x_i, \theta \rangle + \sigma w_i$ where $w_i \sim \mathscr{N}(0,1)$. We have

$$P_\theta(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y_i^2}{2\sigma^2}} e^{\frac{1}{\sigma}\left(y_i \langle x_i, \theta\rangle - \frac{\langle x_i, \theta\rangle^2}{2}\right) \Phi(\langle x_i, \theta \rangle)}.$$

Choose $\Phi(t) = \frac{t^2}{2}$, whose second derivative is bounded, and we get a special case of part (c).

Remark 6.2. *The upper bound and minimax lower bound of regression are both the order $\frac{d}{n}$.*

References

[1] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[2] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.