ASSESSING GENERALIZABILITY OF MAXIMUM LIKELIHOOD ESTIMATION

DAESEOK LEE AND HIS COLLEAGUES

Abstract. This note describes our ongoing project for building a learning theory for probabilistic models. Our focus at the moment is on maximum likelihood estimation, and the note contains preliminary results about the generalisability of this estimation technique. Specifically, we start by noting that assessing a distribution class's ability to approximate a density via maximum likelihood estimation is mathematically similar to assessing the generalisability of a hypothesis class in PAC learning, but the theory for the latter does not directly transfer to the form due to a challenging technical problem: we cannot apply Azuma's lemma as in the proof of Rademacher bound, since the logarithm can take an unbounded value unlike the 0/1 loss function. In the note, we describe multiple directions for circumventing the problem, which we have been pursuing so far. If one of these does work, then we might obtain a measure like Rademacher's complexity that bounds the generalisability of maximum likelihood estimation. We list concrete open problems that we work on in each direction.

Acknowledgement

I thank professor Hongseok Yang for carefully reading the draft up to section 7 and improving the writing significantly.

1. Background

Let p be an unknown probability distribution on $\mathbb{R}^d$ equipped with its density (with respect to the Lebesgue measure), and $\mathcal{H}$ be a *hypothesis class*, which is either given a priori or chosen implicitly through a general density estimation algorithm. Given independent samples $x_1, \cdots, x_m$ from p, *maximum likelihood estimation* tries to approximate p by finding

$$h^* = \operatorname*{argmax}_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \log h(x_i)$$

with the hope that with a high probability for all $h' \in \mathcal{H}$,

$$\frac{1}{m} \sum_{i=1}^{m} \log h'(x_i) \simeq \mathbb{E}_{X \sim p}[\log h'(X)] \tag{1.1}$$

is true.

One rationale behind maximum likelihood estimation is that the estimation picks a hypothesis h^* close to p in terms of the KL divergence, if such a hypothesis exists in the hypothesis set $\mathcal{H}$. This is because, letting $h^\dagger = \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{KL}[p||h]$,

$$\begin{aligned}
\mathrm{KL}[p||h^*] &= \mathbb{E}_{X \sim p}\left[\log\left(\frac{p(x)}{h^*(x)}\right)\right] \\
&\simeq \mathbb{E}_{X \sim p}\left[\log(p(x))\right] - \frac{1}{m}\sum_{i=1}^{m} \log(h^*(x_i)) \\
&\leq \mathbb{E}_{X \sim p}\left[\log(p(x))\right] - \frac{1}{m}\sum_{i=1}^{m} \log(h^\dagger(x_i)) \\
&\simeq \mathbb{E}_{X \sim p}\left[\log\left(\frac{p(x)}{h^\dagger(x)}\right)\right] = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{KL}[p||h],
\end{aligned}$$

Our discussion so far shows that the problem of justifying the generalisability of MLE for $\mathcal{H}$ boils down to the relationship in (1.1) with a high probability. That is, it amounts to showing that for independent $X_1, \cdots, X_m \sim p$, both

$$\sup_{h \in \mathcal{H}} \left(\mathbb{E}_{X \sim p}[\log h(X)] - \frac{1}{m} \sum_{i=1}^{m} \log h(X_i) \right) \tag{1.2}$$

and

$$\sup_{h \in \mathcal{H}} \left(\frac{1}{m} \sum_{i=1}^{m} \log(h(X_i)) - \mathbb{E}_{X \sim p}[\log h(X)] \right)$$

Date: March 27, 2021.

are small with a high probability. If only the first quantity is small, it is guaranteed that the obtained sample log-likelihood of the MLE solution is as large as the real log-likelihood of the optimal solution. , but the actual expecation might be small and the argument about the KL divergence from above cannot be made. On the other hand, if only the second quantity is small, it is guaranteed that the obtaind log-likelihood is as large as the sample log-likelihood of the optimal solution, but the optimal solution may have a much larger real log-likelihood than its sample estimation. However, methods for bounding these quantities are more or less the same, and so we will focus only on the first one here for simplicity.

If the reader is familiar with the use of Rademacher complexity in the generalisation results in [Mohri et al., 2018], she or he would see the similarity between it and what we do here. In that case, the reader might guess that usng a quantity such as

$$\mathbb{E}_{X_1,\cdots,X_m\sim p}\left[\mathbb{E}_{\sigma\sim\{-1,1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\log h(X_i)\right]\right]$$

would lead to a desired upper bound of (1.2), and this derivation would be just a simple repetation of what is written in [Mohri et al., 2018]. However, it is not the case. One key difference here is that the loss function is defined in terms of logarithm, and may take an unbounded value, so that neither the McDiarmid's bound nor Azuma's theorem for concentration of Martingale sequences is inapplicable in our setting, unlike in the setting of [Mohri et al., 2018]. We explain how we have been trying to overcome this technical challenge in the rest of this note.

2. Setup

Let Θ be a topological space. We write $\mathbb{R}_+$ for the set of positive real numbers.

Definition 2.1 (Pointedwise Continuously-Parameterised Hypothesis class)**.** When a hypothesis class $\mathcal{H}$ can be written as $\mathcal{H}=\{h_\theta\,|\,\theta\in\Theta\}$ for some h and the map $\theta\longmapsto h_\theta(x)$ here is continuous for every $x\in\mathbb{R}^d$, we call $\mathcal{H}$ *pointwise continuously-parametrized (PCP) hypothesis class* and Θ its *parameter space*.

Every hypothesis class $\mathcal{H}$ induces a binary non-negative map on the space $\mathbb{R}^d$ for x, which has some of the defining properties of distance or metric.

Definition 2.2 (Hypothesis Distance)**.** The *hypothesis distance d under $\mathcal{H}$* is a binary function from $\mathcal{H}$ to $\mathbb{R}_+\cup\{\infty\}$ defined by

$$d(x,x')=d_{\mathcal{H}}(x,x')=\sup_{h\in\mathcal{H}}|\log h(x)-\log h(x')|.$$

Remark 2.3. The hypothesis distance is symmetric. Also, it satisfies the triangular inequality as shown below:

$$\begin{aligned}d(x,x')+d(x',x'')&=\sup_{h\in\mathcal{H}}|\log h(x)-\log h(x')|+\sup_{h\in\mathcal{H}}|\log h(x')-\log h(x'')|\\&\geq\sup_{h\in\mathcal{H}}\Big(|\log h(x)-\log h(x')|+|\log h(x')-\log h(x'')|\Big)\\&\geq\sup_{h\in\mathcal{H}}\Big(|\log h(x)-\log h(x'')|\Big)=d(x,x'').\end{aligned}$$

In general, we cannot say that $d_{\mathcal{H}}$ is a semimetric because $d_{\mathcal{H}}(x,x')=\infty$ for some x and x'. However, we have gut feeling that if we can ignore this corner case, our life later will become easier. This brings a natural question of how to achieve this, at least at the support of a distribution p that we care about.

Problem 1. When is it true that $d_{\mathcal{H}}(x,x')<\infty$ for all x,x' in the support of a distribution p?.

Remark 2.4. If we are dealing with a PCP hypothesis class with a compact parameter space, then the condition of Problem 1 is satisfied.

Remark 2.5. One way to get around Problem 1 is to put a blame on p, instead of $\mathcal{H}$. For any $\mathcal{H}$, we consider only those distributions for which the answer to the problem is yes. That is, we consider the following family of probability distributions and study the structure of the family:

$$\left\{p\;\middle|\;p\ll\text{Leb and }\int\mathbf{1}[d_{\mathcal{H}}(x,x')=\infty]p(x)p(x')\,\mathrm{d}x\mathrm{d}x'=0\right\}.$$

3. cutting a bit of head and tail

4. decomposing into bounded random variables

5. Generalisation Bound using an Approximate Version of Azuma's Inequality

We recall Theorem 33 of [Chung and Lu, 2006]:

Lemma 5.1. *Let $\{0,\Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_m$ be a filtration (i.e. sequence of increasing σ-fields), and $Y_0, Y_1, \cdots, Y_m$ be a martingale sequence corresponding to this filtration $\{\mathcal{F}\}_{i=0}^m$, i.e. Y_i is $\mathcal{F}_i$-measurable and $Y_i = \mathbb{E}[Y_{i+1}|\mathcal{F}_i]$ for all $i \geq 0$. If*

$$\sum_{i=1}^{m} \mathbb{P}(|Y_i - Y_{i-1}| \geq c_i) \leq \eta$$

then we have

$$\mathbb{P}\big(Y_m \geq \mathbb{E}[Y_m] + \epsilon\big) \leq \exp\bigg(-\frac{\epsilon^2}{2\sum_{i=1}^m c_i^2}\bigg) + \eta$$

Note that when $\eta = 0$, this inequality reduces to the one of Azuma's theorem.

In the rest of this section, we adopt the convention that $\mathbb{E}$ and $\mathbb{P}$ are taken with respect to the target distribution p, unless specified otherwise.

Theorem 5.2. *Let c and η be positive reals, and m a natural number. Assume that*

$$\mathbb{P}_X\Big(\mathbb{E}_{X'}[d_{\mathcal{H}}(X,X')] \geq mc\Big) \leq \eta.$$

Let

$$\mathcal{R}_{\mathcal{H}}(m) \stackrel{\text{def}}{=} \mathbb{E}_{X_1,\cdots,X_m}\bigg[\mathbb{E}_{\sigma\sim\{-1,1\}^m}\bigg[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \log h(X_i)\bigg]\bigg].$$

Then for all $\epsilon > 0$, except for at most the probability

$$\exp\bigg(-\frac{\epsilon^2}{2mc^2}\bigg) + m\eta,$$

we have

$$\sup_{h\in\mathcal{H}}\bigg(\mathbb{E}_X[\log h(X)] - \frac{1}{m}\sum_{i=1}^m \log h(X_i)\bigg) \leq \epsilon + 2\mathcal{R}_{\mathcal{H}}(m).$$

Proof. Let

$$f(x_1,\cdots,x_m) \stackrel{\text{def}}{=} \sup_{h\in\mathcal{H}}\bigg(\mathbb{E}_X[\log h(X)] - \frac{1}{m}\sum_{i=1}^m \log h(x_i)\bigg)$$

$$F(x_1,\cdots,x_m) \stackrel{\text{def}}{=} f(x_1,\cdots,x_m) - \mathbb{E}_{X_1,\cdots,X_m}[f(X_1,\cdots,X_m)].$$

Consider i.i.d samples $X_1,\cdots,X_m$ from p. For each $0 \leq i \leq m$, let

$$\mathcal{F}_i \stackrel{\text{def}}{=} \sigma(X_1,\cdots,X_i),$$

$$Y_i \stackrel{\text{def}}{=} \mathbb{E}[F(X_1,\cdots,X_m)\,|\,\mathcal{F}_i] = \mathbb{E}_{X'_{i+1},\cdots,X'_m}[F(X_1,\cdots,X_i,X'_{i+1},\cdots,X'_m)].$$

We can apply Lemma 5.1 to $(F_i)_i$, $(Y_i)_i$, c, and $\sum_{i=1}^m \eta(i)$. To see why, note that for all $1 \leq i \leq m$,

$$\begin{aligned}
|Y_i - Y_{i-1}| &= \Big|\mathbb{E}_{X_{i+1},\cdots,X_m}[F(X_1,\cdots,X_m)] - \mathbb{E}_{X_i,\cdots,X_m}[F(X_1,\cdots,X_m)]\Big| \\
&\leq \bigg|\mathbb{E}_{X'}\bigg[\sup_{h\in\mathcal{H}}\bigg(\frac{1}{m}(\log h(X') - \log h(X))\bigg)\bigg]\bigg| \\
&\leq \frac{1}{m}\mathbb{E}_{X'}\bigg[\sup_{h\in\mathcal{H}}|\log h(X') - \log h(X)|\bigg] \\
&= \frac{1}{m}\mathbb{E}_{X'_i}[d_{\mathcal{H}}(X,X')].
\end{aligned}$$

Thus,

$$\mathbb{P}\Big(|Y_i - Y_{i-1}| \geq c\Big) \leq \mathbb{P}\Big(\mathbb{E}_{X'}[d_{\mathcal{H}}(X,X')] \geq mc\Big) \leq \eta.$$

The second inequality follows from the assumption of the theorem. Applying Lemma 5.1 to the data as mentioned gives us the following bound on probability:

$$\begin{aligned}
\mathbb{P}\Big(f(X_1,\cdots,X_m) \leq \mathbb{E}_{X'_1,\cdots,X'_m}[f(X'_1,\cdots X'_m)] + \epsilon\Big) &\leq \exp\bigg(-\frac{\epsilon^2}{2\sum_{i=1}^m c^2}\bigg) + m\eta \\
&\leq \exp\bigg(-\frac{\epsilon^2}{2mc^2}\bigg) + m\eta.
\end{aligned}$$

The last step of the proof of the theorem is to bound $\mathbb{E}[f(X'_1,\cdots X'_m)]$ by $2\mathcal{R}_{\mathcal{H}}(m)$. This can be done as in the proof of Theorem 3.1 of the first edition of [Mohri et al., 2018]. □

Remark 5.3. We note two points. First, $\mathcal{R}_{\mathcal{H}}$ is nothing but the Rademacher complexity of $\{\log h \mid h \in \mathcal{H}\}$, the hypothesis class consisting of log-pdfs. Next, in the following development, we assume that c and η are given for each m. Let us make this dependency on m explicit and write $c(m)$ and $\eta(m)$. To get a meaningful bound in this setting, it is likely that we should have $c(m) \ll m^{1/2}$, as in the case that $c(m) = m^a$ for $0 < a < 1/2$.

Problem 2. For simple distributions such as Gaussian, when the parameters are appropriately restricted, what is a possible choice of the pair $(c(m), \eta(m))$?

Problem 3. We don't know yet how to calculate the quantity similar to the Rademacher complexity. Since the range of logarithm is an infinite set, we cannot rely on Massart's lemma as in [Mohri et al., 2018] to bound in terms of growth function and VC dimension. Are there any analogous definitions that work for this case?

6. Generalisation Bound using a Maurey-Pisier Type Inequality

An alternative to the assumption of boundedness with high probability seen in the previous section is a condition that the hypothesis distance $d_{\mathcal{H}}(x, x')$ between x and x' is small (e.g. $d_{\mathcal{H}}(x, x') \le C \|x - x'\|$) for all x and x'. In this section, we explore this alternative.

We recall Maurey-Pisier Theorem and its proof from [Naor, 2008] (Theorem 8).

We combine this theorem with a common strategy of using Martingale difference for deriving a bound, and obtain the next theorem:

Theorem 6.1. *Let*

$$p \overset{\text{def}}{=} \mathcal{N}(0, I_d), \quad \delta_{\mathcal{H}}(x) \overset{\text{def}}{=} \sup_{h \in \mathcal{H}} \|\nabla \log h(x)\|, \quad f(x_1, \cdots, x_m) \overset{\text{def}}{=} \sup_{h \in \mathcal{H}} \left[\mathbb{E}_X[\log(h(X))] - \frac{1}{m} \sum_{i=1}^{m} \log(h(x_i)) \right].$$

If $\delta_{\mathcal{H}}(x) < \infty$ for all $x \in \mathbb{R}^d$ and $\mathcal{H}$ is sufficiently regular in the sense clarified in the proof, then

$$\mathbb{E}_{X_1, \cdots, X_m} \left[\exp\left(t\big(f(X_1, \cdots, X_m) - \mathbb{E}_{X_1, \cdots, X_m}[f(X_1, \cdots, X_m)]\big) \right) \right] \le \left(\mathbb{E}_G \left[\exp\left(\frac{\pi^2 t^2}{8m^2} \delta_{\mathcal{H}}(G)^2 \right) \right] \right)^m.$$

Proof. Let

$$\begin{aligned}
\Delta_1(x_1) &\overset{\text{def}}{=} \mathbb{E}_{X_1, \cdots, X_m}[f(x_1, X_2, \cdots, X_m) - f(X_1, \cdots, X_m)], \\
\Delta_2(x_1, x_2) &\overset{\text{def}}{=} \mathbb{E}_{X_2, \cdots, X_m}[f(x_1, x_2, X_3, \cdots, X_m) - f(x_1, X_2, \cdots, X_m)], \\
&\vdots \\
\Delta_m(x_1, \cdots, x_m) &\overset{\text{def}}{=} \mathbb{E}_{X_m}[f(x_1, \cdots, x_m) - f(x_1, \cdots, x_{m-1}, X_m)].
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E}_{X_1, \cdots, X_m} \left[\exp\left(t\big(f(X_1, \cdots, X_n) - \mathbb{E}_{X_1, \cdots, X_m}[f(X_1, \cdots, X_n)]\big) \right) \right] \\
&= \mathbb{E}_{X_1, \cdots, X_m} \left[\exp\left(t \sum_{i=1}^{m} \Delta_i(X_1, \cdots, X_i) \right) \right] \\
&= \mathbb{E}_{X_1} \left[e^{t\Delta_1(X_1)} \, \mathbb{E}_{X_2} \left[e^{t\Delta_2(X_1, X_2)} \, \mathbb{E}_{X_3} \left[\cdots \mathbb{E}_{X_m} \left[e^{t\Delta_m(X_1, \cdots, X_m)} \;\middle|\; X_1, \cdots, X_{m-1} \right] \cdots \;\middle|\; X_2, X_1 \right] \;\middle|\; X_1 \right] \right].
\end{aligned}$$

We upper-bound the quantity step by step from the inner-most expression to the out-most one by the same constant $C \overset{\text{def}}{=} \mathbb{E}_G[\exp(\frac{\pi^2 t^2}{8m^2} \delta_{\mathcal{H}}(G)^2)]$. To do so, we have to show that for each $1 \le i \le m$,

$$\mathbb{E}_{X_i}[\exp(t\Delta_i(X_1, \cdots, X_i)) \mid X_1, \cdots, X_{i-1}] \le C.$$

For a given $x_1, \cdots, x_{i-1} \in \mathbb{R}^d$, define $g_i(x) \overset{\text{def}}{=} \Delta_i(x_1, \cdots, x_{i-1}, x)$. Then, we have to show that

$$\mathbb{E}_X[\exp(t g_i(X))] \le C.$$

For $x, x' \in \mathbb{R}^d$, we have

$$\begin{aligned}
|g_i(x) - g_i(x')| &= |\Delta_i(x_1, \cdots, x_{i-1}, x) - \Delta_i(x_1, \cdots, x_{i-1}, x')| \\
&= |\mathbb{E}_{X_{i+1}, \cdots, X_m}[f(x_1, \cdots, x_{i-1}, x, X_{i+1}, \cdots, X_m) - f(x_1, \cdots, x_{i-1}, x', X_{i+1}, \cdots, X_m)]| \\
&\leq \mathbb{E}_{X_{i+1}, \cdots, X_m}\left[\sup_{h \in \mathcal{H}} \left| \frac{\cdots + \log h(x_{i-1}) + \log h(x) + \log h(X_{i+1}) + \cdots}{m} \right.\right. \\
&\qquad \left.\left. - \frac{\cdots + \log h(x_{i-1}) + \log h(x') + \log h(X_{i+1}) + \cdots}{m} \right|\right] \\
&= \sup_{h \in \mathcal{H}} \left| \frac{\log h(x) - \log h(x')}{m} \right| \\
&= \frac{1}{m} d_{\mathcal{H}}(x, x') \\
&= \frac{1}{m} \left| \sup_{h \in \mathcal{H}} \int_x^{x'} \nabla_x l_h(x) \cdot dx \right| \\
&\leq \frac{1}{m} \sup_{h \in \mathcal{H}} \int_x^{x'} \|\nabla_x l_h(x)\| \, \|\mathrm{d}x\| \\
&\leq \frac{1}{m} \int_x^{x'} \delta_h(x) \, \|\mathrm{d}x\| \, .
\end{aligned}$$

Now, assume that there exists a continuously differentiable function g that approximates g_i quite well. More precisely, we assume for arbitarily small $M, M' > 0$, we can find a continuously differentiable g such that $|g(x) - g_i(x)| \leq M$ and $|g(x) - g_i(x) - g(x') + g_i(x')| \leq M' \|x - x'\|$ for each x, x'. Fix M, M', and g. For all $x, x' \in \mathbb{R}^d$, we have

$$|g(x) - g(x')| \leq |g_i(x) - g_i(x')| + M' \|x - x'\| \leq \frac{1}{m} \int_x^{x'} \delta_{\mathcal{H}}(x) \, \|dx\| + M' \|x - x'\|$$

where the integration is taken with respect to the direct line segment from x to x'.

Setting x' in the direction of $\nabla g(x)$, using the mean value theorem, and taking limit using the continuous differentiability, we get $\|\nabla g(x)\| \leq \frac{1}{m} \delta_{\mathcal{H}}(x) + M'$. We now apply the previous lemma, and obtain

$$\begin{aligned}
\mathbb{E}[\exp(t g_i(X))] &\leq e^{Mt} \mathbb{E}[\exp(t g(X))] \\
&\leq e^{Mt} \mathbb{E}_G\left[\exp\left(\frac{1}{8}\pi^2 t^2 \|\nabla g(G)\|^2\right)\right] \\
&\leq e^{Mt} \mathbb{E}_G\left[\exp\left(\frac{1}{8}\pi^2 t^2 \left(\frac{1}{m}\delta_{\mathcal{H}}(x) + M'\right)^2\right)\right].
\end{aligned}$$

Taking $M, M' \to 0$, we're done. □

Remark 6.2. If the expectation in the upper bound amounts to $\leq \exp(Ct^2/m^2)$, then we get a concentration inequality of usual order.

Problem 4. Example calculation?

Problem 5. Maurey-Pisier type inequalities might hold not only for Gaussian random variables. At least, it might hold also for bounded random variables, because of the McDiarmid's inequality. Is there any broader class of such distributions? How about a mixture of Gaussians?

7. Analogue of VC Dimension and Bound on Rademacher Complexity

The goal of this section is to solve Problem 3 partially using an analogue of VC dimension. Let $l_h(x) = \log h(x)$. In the section, we give a bound on the following Rademacher complexity

$$\mathbb{E}_{\sigma \sim \{-1,1\}^m}\left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i l_h(x_i)\right]$$

when $x_1, \cdots, x_m$ are in a given compact set $K \subset \mathbb{R}^d$.

To achieve our goal, we recall the standard strategy in [Mohri et al., 2018] for bounding the Rademacher complexity for a class of 0/1 loss functions. In this case, a hypothesis class $\mathcal{F}$ is a family of functions from $\mathbb{R}^d$ to $\{0, 1\}$, and we want to bound the Rademacher complexity of $\mathcal{F}$, that is,

$$\mathbb{E}_{\sigma \sim \{-1,1\}^m}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(x_i)\right].$$

The strategy is to find a desired bound in two steps:

(1) Show that $|\{(f(x_1),\cdots,f(x_m)) \mid f\in\mathcal{F}\}|$ is much smaller than 2^m.
(2) Show that the Rademacher complexity is small whenever $|\{(f(x_1),\cdots,f(x_m)) \mid f\in\mathcal{F}\}|$ is small.

The fist step is typically done by Sauer's lemma. Recall the definition of the VC dimension:

$$\mathrm{VC}(\mathcal{F}) \stackrel{\text{def}}{=} \max\left\{d \;\middle|\; \exists x_1,\cdots,x_d\in K.\forall S\subseteq[d].\exists f\in\mathcal{F}.S=\{i \mid f(x_i)=1\}\right\}.$$

Lemma 7.1 (Sauer). *Let $d=\mathrm{VC}(\mathcal{F})$. If $d<\infty$, then for all $m\geq d$,*

$$\left|\{(f(x_1),\cdots,f(x_m)) \mid f\in\mathcal{F}\}\right| \leq \sum_{i=0}^{d}\binom{m}{i} \leq \left(\frac{em}{d}\right)^d.$$

The second step is done by Massart's lemma or the maximal inequality:

Lemma 7.2 (Massart). *For all $S\subseteq\{0,1\}^m$,*

$$\mathbb{E}_{\sigma\sim\{-1,1\}^m}\left[\max_{(y_1,\cdots,y_m)\in S}\frac{1}{m}\sum_{i=1}^{m}\sigma_i y_i\right] \leq \sqrt{\frac{2\log|S|}{m}}.$$

The main obstacle that stops us from using the same strategy is the fact that the set

$$\{(l_h(x_1),\cdots,l_h(x_m)) \mid h\in\mathcal{H}\}$$

is usually infinite. Our tool for overcoming the obstacle is to use the following notion of ϵ-cover to analyse $\mathcal{F}=\{l_h \mid h\in\mathcal{H}\}$.

Definition 7.3 (ϵ-cover). For given $x_1,\cdots,x_m\in K$, a subset $\mathcal{C}\subset\mathbb{R}^m$ is an *ϵ-cover* or *$(x_1,\cdots,x_m)$-ϵ-cover of* $\mathcal{F}$ if for every $f\in\mathcal{F}$, there exists $y\in\mathcal{C}$ such that

$$\max_{i\in[m]}|f(x_i)-y_i|<\epsilon.$$

Note that when $\mathcal{C}$ is an ϵ-cover of $\{l_h \mid h\in\mathcal{H}\}$,

$$\mathbb{E}_{\sigma\sim\{-1,1\}^m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i l_h(x_i)\right] \leq \epsilon+\mathbb{E}_{\sigma\sim\{-1,1\}^m}\left[\sup_{y\in\mathcal{C}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i y_i\right]$$

Thus, in this case, we can achieve our goal by finding an ϵ-cover of a small finite size. In the following subsections, we describe two approaches for finding such a cover. Currently, we think that the second approach is more promising.

7.1. a continuous Sauer's lemma.

Warning: This subsection is about a method turned out to be apparently useless because when the method is used alone, the resulting bound is too crude. However, you may get inspired by the proof of the main result.

First, let us define a seemingly wierd quantity.

Definition 7.4 (sequential strict shatter dimension). The *sequential strict shatter dimension* $s_K^\epsilon(\mathcal{F})$ of a family $\mathcal{F}$ of functions $K\to\mathbb{R}$ is the maximum integer $m\geq 0$ such that there exist points $x_1,\cdots,x_s\in K$ and values $c:\bigcup_{k=0}^{s-1}\{-1,1\}^k\to\mathbb{R}$ such that for any choice of $\sigma_1,\cdots,\sigma_s\in\{-1,1\}$, there exists $f\in\mathcal{F}$ such that for all $1\leq k\leq s$, the following holds:

$$\sigma_k(f(x_k)-c(\sigma_1,\cdots,\sigma_{k-1}))\geq\epsilon/2$$

This is an analogue of the VC dimension, while the following and its proof are analogues of those of Lemma 7.1

Theorem 7.5 (A continuous Sauer's lemma). *Let K be a given set. Let $\mathcal{F}$ be a given function family with $\{f(x)|f\in\mathcal{F},x\in K\}\subseteq[a,b]$. Let $n_{m,s}^{\epsilon,\epsilon'}$ be the supremum over $((x_1,\cdots,x_m),\mathcal{F}')$ of the minimum size of $(x_1,\cdots,x_m)$-ϵ-cover of $\mathcal{F}'$ when $x_1,\cdots,x_m\in K$ and $\mathcal{F}'\subseteq\mathcal{F}$ is a function family such that $s^{\epsilon'}_{\{x_1,\cdots,x_m\}}(\mathcal{F}')\leq s$. Since $n_{m,s}^{\epsilon,\epsilon'}\leq(\frac{b-a}{\epsilon})^m$, it is finite. Then we have*

$$n_{m,s}^{\epsilon,\epsilon'}\leq\frac{\epsilon'}{2\epsilon}n_{m-1,s}^{\epsilon,\epsilon'}+\frac{b-a}{2\epsilon}n_{m-1,s-1}^{\epsilon,4\epsilon+\epsilon'}$$

Proof. Let $x_1,\cdots,x_m\in K$ and $\mathcal{F}'$ be a function family with $s^{\epsilon'}_{\{x_1,\cdots,x_m\}}(\mathcal{F}')\leq s$. Define $\rho(f,f')=\max_{i=2}^n|f(x_i)-f'(x_i)|$ and split $\mathcal{F}$ into $\mathcal{F}'=\mathcal{F}_1\sqcup\mathcal{F}_2$ where

$$\mathcal{F}_1=\{f\in\mathcal{F}'|\forall f'\in\mathcal{F}'.\rho(f,f')<2\epsilon\to|f(x_1)-f'(x_1)|<\epsilon'\}$$

and

$$\mathcal{F}_2=\{f\in\mathcal{F}'|\exists f'\in\mathcal{F}'.\rho(f,f')<2\epsilon\wedge|f(x_1)-f'(x_1)|\geq\epsilon'\}$$

First, since $(x_2, \cdots, x_n)$ and $\mathcal{F}_1$ satisfies $s^{\epsilon'}_{\{x_2,\cdots,x_m\}}(\mathcal{F}_1) \leq s^{\epsilon'}_{\{x_1,\cdots,x_m\}}(\mathcal{F}') \leq s$, there exists ϵ-$(x_2, \cdots, x_m)$-cover $\mathcal{C} \subset \mathbb{R}^{m-1}$ of $\mathcal{F}_1$ with $|\mathcal{C}| \leq n^{\epsilon,\epsilon'}_{m-1,s}$. For $y \in \mathcal{C}$, let $\mathcal{F}_y = \{f \in \mathcal{F}_1 | \max_{i=2}^m |f(x_i) - y_i| < \epsilon\}$, so that we have $\mathcal{F}_1 = \bigcup_{y\in\mathcal{C}} \mathcal{F}_y$. Then for arbitrary $f, f' \in \mathcal{F}_y$, we have $\rho(f, f') \leq \max_{i=2}^m |f(x_i) - y_i| + \max_{i=2}^m |f'(x_i) - y_i| < 2\epsilon$, so by the definition of $\mathcal{F}_1$, $|f(x_1) - f'(x_1)| < \epsilon'$. Then $\{f(x_1)|f \in \mathcal{F}_y\}$ is contained in an interval of length ϵ'. Dividing the interval into subintervals of length 2ϵ and extending y by setting the first coordinate as the midpoints of those intervals, and collecting all those extensions of $y \in \mathcal{C}$, we get a $(x_1, \cdots, x_n)$-ϵ-cover $\mathcal{C}_1$ of $\mathcal{F}_1$ with $|\mathcal{C}_1| \leq \frac{\epsilon'}{2\epsilon}|\mathcal{C}| \leq \frac{\epsilon'}{2\epsilon} n^{\epsilon,\epsilon'}_{m-1,s}$.

On the other hand, I claim that $s^{\epsilon,4\epsilon+\epsilon'}_{x_2,\cdots,x_m}(\mathcal{F}_2) \leq s - 1$. Suppose otherwise. Then there exists $\{z_1, \cdots, z_s\} \subset \{x_2, \cdots, x_m\}$ and $c : \bigcup_{k=0}^{s-1} \{-1, 1\}^k \to \mathbb{R}$ such taht for any choice of $\sigma_1, \cdots, \sigma_s \in \{-1, 1\}$ there exists $f \in \mathcal{F}_2$ such that

$$\sigma_k(f(z_k) - c(\sigma_1, \cdots, \sigma_{k-1})) \geq (4\epsilon + \epsilon')/2, (1 \leq k \leq s) \tag{7.1}$$

Let $y_{s+1} = x_1$. For $\sigma_1, \cdots, \sigma_s \in \{-1, 1\}$, let $f \in \mathcal{F}_2$ satisfies (7.1). By the definition of $\mathcal{F}_2$, there exists $f' \in \mathcal{F}'$ such that $|f(y_i) - f'(y_i)| < 2\epsilon$ for all $1 \leq i \leq s$ while $|f(y_{s+1}) - f'(y_{s+1}))| \geq \epsilon'$. Extends c by letting $c(\sigma_1, \cdots, \sigma_s) = \frac{f(y_{s+1})+f'(y_{s+1})}{2}$. Then it is easy to check that for any additional choice of $\sigma_{s+1} \in \{-1, 1\}$, either $g = f$ or $g = f'$ satisfies

$$\sigma_k(g(y_k) - c(\sigma_1, \cdots, \sigma_{k-1})) \geq \epsilon'/2, (1 \leq k \leq s + 1)$$

This contradicts to that $s^{\epsilon'}_{\{x_1,\cdots,x_m\}}(\mathcal{F}') \leq s$, so proves the claim. Therefore, we can find $(x_2, \cdots, x_m)$-ϵ-cover $\mathcal{C} \subset \mathbb{R}^{m-1}$ of $\mathcal{F}_2$ with size $|\mathcal{C}| \leq n^{\epsilon,4\epsilon+\epsilon'}_{m-1,s-1}$. By subdividing the interval $\{f(x)|f \in \mathcal{F}, x \in K\}$ of length $\leq b - a$ into subintervals of length 2ϵ, taking midpoints $x_1, \cdots, x_n$ ($n \leq \frac{b-a}{2\epsilon}$) and setting $\mathcal{C}_2 = \{(x_i, y_2, \cdots, y_m)|1 \leq i \leq n, (y_2, \cdots, y_m) \in \mathcal{C}\}$, $\mathcal{C}_2$ becomes a $(x_1, \cdots, x_m)$-ϵ-cover of $\mathcal{F}_2$.

Summing up, $\mathcal{C}_1 \cup \mathcal{C}_2$ becomes a $(x_1, \cdots, x_m)$-ϵ-cover of $\mathcal{F}'$.

□

Problem 6. Example calculation of the sequential strict shatter dimension?

Problem 7. How does this lead to explicit bounds such as that of Lemma 7.1?

Lemma 7.6 (naive monotone method for calculating sequential strict shatter dimension)**.** *If K can be divided into $K = \bigcup_{i=1}^k K_i$ and for each $1 \leq i \leq k$ there exists a function $F_i : \mathcal{F} \to \mathbb{R}$ of range $[a_i, b_i]$ and a value $\epsilon_i > 0$ such that*

$$\forall x \in K. \forall f, f' \in \mathcal{F}. f(x) \geq f'(x) + \epsilon \implies F_i(f) \geq F_i(f') + \epsilon_i$$

then the strict shatter dimension can be bounded as follows

$$s^\epsilon_K(\mathcal{F}) \leq k \log_2(\max_{i=1}^k \frac{b_i - a_i}{\epsilon_i}) + 1$$

Corollary 7.7. *a*

7.2. **searching for the cover hierarchically using VC dimension of differences.**

Theorem 7.8 (bounding min cover using VC dimension of differences)**.** *Let $\mathcal{F}$ be a family of functions whose domain is K. Let d be the VC dimension of the function function family $s(\mathcal{F} - \mathcal{F}) = \{s \circ (f - f')|f, f' \in \mathcal{F}\}$, where*

$$s(y) = \begin{cases} 1 & , y \geq 0 \\ 0 & , y < 0 \end{cases}$$

Then when $\{f(x)|f \in \mathcal{F}, x \in K\} \subseteq [a, b]$ and $x_1, \cdots, x_n \in K$, there exists a $(x_1, \cdots, x_n)$-ϵ-cover of $\mathcal{F}$ of size at most

$$(\frac{em}{d})^{d\lceil \log_2(\frac{b-a}{2\epsilon})\rceil}$$

Proof. Let $n = \lceil \log_2(\frac{b-a}{2\epsilon})\rceil$. Let $f(L)$ be the smallest integer such that for any choice of subintervals $[a_1, b_1], \cdots, [a_m, b_m] \subset [a, b]$ with $b_i - a_i = L$, there exists $\mathcal{C} \subset \mathbb{R}^m$ with the property that

$$\forall f \in \mathcal{F}. (\forall 1 \leq i \leq m. a_i \leq f(x_i) \leq b_i) \implies \exists y \in \mathcal{C}. \max_{i=1}^m |f(x_i) - y_i| \leq \epsilon \tag{7.2}$$

Then obviously, $f((b - a)/2^n) \leq f(2\epsilon) \leq 1$. We're going to show that

$$f(L) \leq (\frac{em}{d})^d f(L/2)$$

which is enough to finish the proof. Let $[a_1, b_1], \cdots, [a_m, b_m] \subset [a, b]$ be subintervals with $b_i - a_i = L$. Let $c_i = \frac{b_i - a_i}{2} (1 \leq i \leq m)$. Let

$$\mathcal{F}' = \{f \in \mathcal{F} | \forall 1 \leq i \leq m. a_i \leq f(x_i) \leq b_i\}$$

For $f \in \mathcal{F}'$, define

$$\tilde{f}(i) = \begin{cases} 1 & , f(x_i) \geq c_i \\ 0 & , f(x_i) < c_i \end{cases}, 1 \leq i \leq m$$

and $\tilde{\mathcal{F}} = \left\{ \tilde{f} | f \in \mathcal{F}' \right\}$. Then VC dimension of $\tilde{F}$ is at most d because $A \subseteq \{1, \cdots, m\}$ being shattered by $\tilde{\mathcal{F}}$ implies $\{x_i | i \in A\}$ being shattered by $s(\mathcal{F} - \mathcal{F})$. Therefore, by Lemma 7.1, we have

$$| \left\{ (\tilde{f}(1), \ldots, \tilde{f}(m)) | f \in \mathcal{F}' \right\} | \leq (\frac{em}{d})^d$$

Therefore, among the 2^m half sub-hypercubes of $[a_1, b_1] \times \cdots \times [a_m, b_m]$, only $\frac{em}{d})^d$ are occupied by $(f(x_1), \cdots, f(x_m))$ for some $f \in \mathcal{F}$. For i-th such hypercube, we find $\mathcal{C}_i$ of size at most $f(L/2)$ with the property (7.2). By taking $\mathcal{C} = \bigcup_i \mathcal{C}_i$, we're done. □

I don't know whether it is useful, but the following lemma gives a broad class of examples. According to [Anthony and Bartlett, 2009] theorem 8.3,

Lemma 7.9. *A feed-forward sigmoid neural network(with sigmoid activation functions for non-output nodes and threshold activation function for the output node) with k non-input nodes and W parameters(linear coefficients+thresholds) has VC dimension $O((kW)^2)$*

Example 7.10. *Let $\mathcal{H} = \left\{ x \mapsto h_\theta(x) = e^{f_\theta(x)}/Z(\theta) | \theta \in \mathbb{R}^W \right\}$, where $Z(\theta)$ is the partition function and f_θ is a feed-forward sigmoid neural network(with all activation functions including that of the output node is sigmoid) with k non-input nodes and W parameters, than* $\mathrm{VC}(s(\log(\mathcal{H}) - \log(\mathcal{H}))) = O((kW)^2)$ *where*

$$s(y) = \begin{cases} 1 & , y \geq 0 \\ 0 & , y < 0 \end{cases}$$

Proof. $s(\log(h_\theta(x)) - \log(h_{\theta'}(x))) = s(f_\theta(x) - f_{\theta'}(x) + \log \frac{Z(\theta')}{Z(\theta)})$ can be realized as a computation of neural network made out of that of $\{f_\theta\}$ by duplicating all non-input nodes and connections and creating a new output node connected from the two previous output nodes, where the weights are fixed to 1 and -1 respectively, and fixing the final threshold to $-\log \frac{Z(\theta')}{Z(\theta)}$. Family of such functions is a subfamily of a sigmoid neural network in the sense of Lemma 7.9 with $2W + 3$ parameters $2k + 1$ non-input nodes. Then the VC dimension of our functions is not greater than that of the VC dimension of the same neural networks with all parameters free, which is $O(((2k + 1)(2W + 3))^2) = O((kW)^2)$ □

7.3. **parameter space covering for PCP hypotheses.** In this subsection, I will denote $l_\theta(x) = \log(h_\theta(x))$. Actually for a PCP hypthesis class(Definition 2.1) $\{h_\theta | \theta \in \Theta\}$ with compact parameter space Θ, finding a smaller $(x_1, \cdots, x_m)$-ϵ-cover (where $x_1, \cdots, x_m \in K$) can be much easier. This is done by finding a near-dense finite subset $\Theta_0 \subset \Theta$, using an assumption that $\max_{x \in K} |l_\theta(x) - l_{\theta'}(x)|$ is small whenever θ and θ' are close to each other in Θ.

For example, assume that $\theta \mapsto l_\theta(x)$ is differentiable for each x and

$$L = \sup_{x \in K, \theta \in \Theta} \|\nabla_\theta l_\theta(x)\| < \infty$$

and Θ is a compact convex set contained in a ball in $\mathbb{R}^D$ of radius $0 < R < \infty$. We can find a (ϵ/L)-l^2 cover $\Theta_0 \subset \Theta$ of size $C_D(CR/\epsilon)^D$ (where $C_D > 0$ is a constant that only depends on D). Then

$$\mathcal{C} = \{(l_\theta(x_1), \cdots, l_\theta(x_m)) | \theta \in \Theta_0\}$$

becomes a $(x_1, \cdots, x_m)$-ϵ-cover. To show this, let $\theta \in \Theta$. Since Θ_0 is a l^2 cover, there exists $\theta_0 \in \Theta_0$ with $\|\theta - \theta_0\| \leq \epsilon/L$. Then for each $1 \leq i \leq m$,

$$|l_\theta(x_i) - l_{\theta_0}(x_i)| \leq \int_{\theta_0}^{\theta} \|\nabla_\theta(x_i)\| \, |d\theta| \leq L \|\theta - \theta_0\| \leq \epsilon$$

where the integral is taken with respect to the straight line segment, which completes the proof. Therefore, we found a desired cover of size

$$C_D(CR/\epsilon)^D$$

which depends on m only through ϵ.

8. FULL EXAMPLE 1 - LIGHT TAIL CONCEPT + GAUSSIAN MIXTURE HYPOTHESIS

The following case will be studied in this section.

- For a given function $t : [r_0, \infty) \to \mathbb{R}_{>0}$, p, the density on $\mathbb{R}^d$ to estimate, has tail bound
$$r \geq r_0 \implies \mathbb{P}_{X\sim p}(\|X\| > r) \leq t(r)$$
- The hypothesis class $\mathcal{H}$ is composed of densities of the form

$$h_\theta(x) = \sum_{i=1}^{k} a_i (\frac{\det(K_i)}{2\pi})^{d/2} e^{-\frac{1}{2}(x-\mu_i)^T K_i (x-\mu_i)} \tag{8.1}$$

where

$$\theta \in \Theta = \left\{ \{(a_i, \mu_i, K_i)\}_{i=1}^k \,|a_i \geq 0, \sum_{i=1}^k a_i = 1, \|\mu_i\| \leq R, K_i\text{:positive definite}, \lambda_{min} \leq \Lambda(K_i) \leq \lambda_{max} \right\} \subset \mathbb{R}^{(1+d+d^2)k}$$

8.1. notations.

- $\Lambda(A)$: the multiset of eigenvalues of A.
- $|||A||| \equiv \max_{\lambda\in\Lambda(A)} |\lambda|$: the operator norm
- $\|A\|_{\text{nuc}} \equiv \sum_{\lambda\in\Lambda(A)} |\lambda|$: the nuclear norm, equal to $tr(A)$ when A is semipositive definite
- $A \odot B \equiv \sum_{i,j} A_{i,j} B_{i,j}$: matrix dot product
- $\tilde{A}$: the adjoint matrix. $\tilde{A}_{i,j} = (-1)^{i+j} \det(A^{(j),(i)})$, where $A^{(j),(i)}$ is the matrix obtaiend from A by omitting j-th row and i-th column. $A\tilde{A} = \det(A)I$.
- $h_{\mu,K}(x) \equiv (\frac{\det(K)}{2\pi})^{d/2} e^{-\frac{1}{2}(x-\mu)^T K(x-\mu)}$
- $l_h(x) \equiv \log(h(x))$
- $l_\theta(x) \equiv l_{h_\theta}(x)$
- $d_{\mathcal{H}}(x, x') \equiv \sup_{h\in\mathcal{H}} |l_h(x) - l_h(x')|$
- For convenience, $h_\theta(x)$ will refer to the function as in the expression (8.1) even for $\theta \in \mathbb{R}^{(1+d+d^2)k} \setminus \Theta$.

8.2. part 1.

Lemma 8.1. *For any $x, y \in \mathbb{R}^d$ and symmetric $A \in \mathbb{R}^{d\times d}$, we have*

$$\left|x^T Ax - y^T Ay\right| \leq |||A||| \cdot \|x - y\| \cdot \sqrt{2(\|x\|^2 + \|y\|^2)}$$

Proof. Let $v_1, \cdot, v_d$ be a orthogonal unit eigenvectors and $\lambda_1, \cdots, \lambda_d$ be corresponding eigienvalues, and $x = \sum_{i=1}^d x_i v_i$ and $y = \sum_{i=1}^d y_i v_i$ where $x_i, y_i \in \mathbb{R}$. Then

$$\begin{aligned}
\left|x^T Ax - y^T Ay\right| &= \left|\sum_{i=1}^d \lambda_i (x_i^2 - y_i^2)\right| \\
&\leq \sqrt{\sum_{i=1}^d \lambda_i^2 (x_i - y_i)^2} \sqrt{\sum_{i=1}^d (x_i + y_i)^2} \\
&\leq \max_{i=1}^d |\lambda_i| \sqrt{\sum_{i=1}^d (x_i - y_i)^2} \sqrt{\sum_{i=1}^d 2(x_i^2 + y_i^2)} \\
&= |||A||| \, \|x - y\| \sqrt{2(\|x\|^2 + \|y\|^2)}
\end{aligned}$$

□

Lemma 8.2.

$$d_{\mathcal{H}}(x, x') \leq \lambda_{max} \|x - x'\| (R + \max(\|x\|, \|x'\|))$$

Proof. For $\theta = \{(a_i, \mu_i, K_i)\}_{i=1}^k \in \Theta$, using Lemma 8.1,

$$\begin{aligned}
|l_\theta(x) - l_\theta(x')| &\leq \max_{i=1}^k |\log h_{\mu_i,K_i}(x) - \log h_{\mu_i,K_i}(x')| = |\log h_{\mu,K}(x) - \log h_{\mu,K}(x')| \\
&= \frac{1}{2} |(x-\mu)^T K(x-\mu) - (x'-\mu)^T K(x'-\mu)| \\
&\leq \frac{1}{2} |||K||| \, \|x - x'\| \sqrt{2(\|x-\mu\|^2 + \|x'-\mu\|^2)} \\
&\leq \lambda_{max} \|x - x'\| (R + \max(\|x\|, \|x'\|))
\end{aligned}$$

□

Lemma 8.3. *For any continuously differentiable $f : [r_0, \infty) \to \mathbb{R}$, if $r \geq r_0$ then*

$$\mathbb{E}_{X\sim p}\mathbf{1}_{\|X\|\geq r} f(\|X\|) \leq f(r)t(r) + \int_r^\infty f'(s)t(s)ds$$

Proof. Let μ be the push forward of the measure associated with p with respect to the map $x \mapsto \|x\|$.

$$\begin{aligned}
\mathbb{E}_{X\sim p}\mathbf{1}_{\|X\|\geq r} f(\|X\|) &= \int_r^\infty f(R)\mu(dR) \\
&= \int_r^\infty (f(r) + \int_r^\infty \mathbf{1}_{s\leq R} f'(s)ds)\mu(dR) \\
&\leq f(r)t(r) + \int_r^\infty f'(s)(\int_r^\infty \mathbf{1}_{s\leq R}\mu(dR))ds \\
&\leq f(r)t(r) + \int_r^\infty f'(s)t(s)ds
\end{aligned}$$

□

Theorem 8.4. *Let $\alpha_0 > 0$. If $\alpha(m) \geq r_0$ and $\int_{\alpha(m)}^\infty (R+2s)t(s)ds \leq 2\alpha_0$, then*

$$\mathbb{P}_{X\sim p}[\mathbb{E}_{X'\sim p} d_{\mathcal{H}}(X, X') \geq 4\lambda_{max}(\alpha_0 + \alpha(m)(R + \alpha(m)))] \leq t(\alpha(m))$$

Proof. Using Lemma 8.2, it is enough to show that $\|X\| \leq \alpha(m)$ implies

$$\mathbb{E}_{X'\sim p} \|X - X'\| (R + \max(\|X\|, \|X'\|)) < 4(\alpha_0 + \alpha(m)(R + \alpha(m)))$$

Let $f(s) = s(R + s)$. Then using Lemma 8.3,

$$\begin{aligned}
&\mathbb{E}_{X'\sim p} \|X - X'\| (R + \max(\|X\|, \|X'\|)) \\
&= \mathbb{E}_{X'\sim p}\mathbf{1}_{\|X'\|\leq\alpha(m)} \|X - X'\| (R + \max(\|X\|, \|X'\|)) + \mathbb{E}_{X'\sim p}\mathbf{1}_{\|X'\|>\alpha(m)} \|X - X'\| (R + \max(\|X\|, \|X'\|)) \\
&\leq 2\alpha(m)(R + \alpha(m)) + 2\mathbb{E}_{X'\in p}\mathbf{1}_{\|X'\|>\alpha(m)} f(\|X'\|) \\
&\leq 2\alpha(m)(R + \alpha(m)) + 2f(\alpha(m))t(\alpha(m)) + 2\int_{\alpha(m)}^\infty f'(s)t(s)ds \\
&< 4\alpha(m)(R + \alpha(m)) + 2\int_{\alpha(m)}^\infty (R + 2s)t(s)ds \\
&\leq 4(\alpha_0 + \alpha(m)(R + \alpha(m)))
\end{aligned}$$

□

8.3. **part 2.** Assume $\|x_i\| \leq R'(1 \leq i \leq m)$. We want to find a $(x_1, \cdots, x_m)$-ϵ_2-cover of $\{x \mapsto l_h(x) | h \in \mathcal{H}\}$. As mentioned before, a bound on the size of the gradients $\nabla_\theta l_\theta(x_i)$ and compactness of Θ can be utilized to find one such cover. Therefore, we first calculate as follows. Let $\theta = \{(a_i, \mu_i, K_i)\}_{i=1}^k \in \Theta$. When $1 \leq i \leq k$ and $1 \leq j \leq m$,

- (1)

$$\begin{aligned}
\left|\frac{\partial l_\theta(x_j)}{\partial a_i}\right| &= \frac{h_{\mu_i,K_i}(x_j)}{h_\theta(x_j)} \leq \sup_{\mu,K} \frac{h_{\mu_i,K_i}(x_j)}{h_{\mu,K}(x_j)} \\
&= \sup_{\mu,K} (\frac{\det(K_i)}{\det(K)})^{d/2} e^{\frac{1}{2}(x_j-\mu)^T K(x_j-\mu) - \frac{1}{2}(x_j-\mu_i)^T K_i(x_j-\mu_i)} \\
&= (\frac{\lambda_{max}}{\lambda_{min}})^{d^2/2} \sup_K e^{\frac{1}{2}(\lambda_{max}-\lambda_{min})\cdot\|x_j-\mu\|^2} \\
&\leq (\frac{\lambda_{max}}{\lambda_{min}})^{d^2/2} e^{\frac{1}{2}(\lambda_{max}-\lambda_{min})(R+R')^2} = M_a
\end{aligned}$$

  (2) If $a_i \geq \delta$, for some not too small $\delta > 0$, there is another bound:

$$\left|\frac{\partial l_\theta(x_j)}{\partial a_i}\right| = \frac{h_{\mu_i,K_i}(x_j)}{h_\theta(x_j)} \leq \frac{1}{a_i} \leq \frac{1}{\delta}$$

- $$\begin{aligned}\sum_{i=1}^{k}\left\|\frac{\partial l_\theta(x_j)}{\partial \mu_i}\right\| &= \sum_{i=1}^{k}\left\|\frac{a_i h_{\mu_i,K_i}(x_j)}{h_\theta(x_j)}\cdot\frac{1}{2}(K_i(x_j-\mu_i)+K_i^T(x_j-\mu_i))\right\| \\ &\le (\sum_{i=1}^{k}\frac{a_i h_{\mu_i,K_i}(x_j)}{h_\theta(x_j)})\max_{i=1}^{k}|||K_i|||\cdot\|x_j-\mu_i\| \\ &\le \lambda_{max}(R+R') = M_\mu\end{aligned}$$

- $$\begin{aligned}\sum_{i=1}^{k}\left|\left|\left|\frac{\partial l_\theta(x_j)}{\partial K_i}\right|\right|\right| &= \sum_{i=1}^{k}\left|\left|\left|\frac{a_i h_{\mu_i,K_i}(x_j)}{h_\theta(x_j)}\cdot(\frac{d}{2}(\frac{\det K_i}{2\pi})^{-1}\frac{\tilde{K_i}^T}{2\pi}-\frac{1}{2}(x_j-\mu_i)(x_j-\mu_i)^T)\right|\right|\right| \\ &\le (\sum_{i=1}^{k}\frac{a_i h_{\mu_i,K_i}(x_j)}{h_\theta(x_j)})\max_{i=1}^{k}\left|\left|\left|\frac{d}{2}K_i^{-1}-\frac{1}{2}(x_j-\mu_i)(x_j-\mu_i)^T\right|\right|\right| \\ &\le \frac{1}{2}(\frac{d}{\lambda_{min}}+(R+R')^2) = M_K\end{aligned}$$

  where $\frac{\partial}{\partial K_i}$ are interpreted as $d\times d$ matrices whose entries are partial differential with respect to corresponding entries of K_i.

Our cover-searching strategy is to find an appropriate cover for each kind of parameters, and combine them by taking the product.

Lemma 8.5. *There exists a ϵ-l^2-cover of $B=\{x\in\mathbb{R}:\|x\|<R\}$ of size*

$$(1+\frac{2R}{\epsilon})^d$$

Proof. Let $S\subset B$ be a maximal set with the property

$$x,y\in S\wedge x\neq y \implies \|x-y\|>\epsilon$$

Then for any $x\in B\setminus S$, there exists $y\in S$ such that $\|x-y\|\le\epsilon$, since otherwise the set $S\cup\{x\}$ has the above property, contradicting to the maximality of S, which shows that S is a ϵ-cover. Then since the balls $B_x(\epsilon/2)$ $(x\in S)$ are all disjoint, we have

$$|S|\,(\frac{\epsilon}{2})^d\operatorname{vol}(B)=\operatorname{vol}(\bigcup_{x\in S}B_x(\epsilon/2))\le\operatorname{vol}(B_0(R+\frac{\epsilon}{2}))=(R+\frac{\epsilon}{2})^d\operatorname{vol}(B)$$

□

Lemma 8.6. *There exists a ϵ-l^2-cover of $S^d=\left\{x\in\mathbb{R}^d|\,\|x\|=R\right\}$ of size*

$$2d(1+\frac{2R}{\epsilon})^{d-1}$$

Proof. We take $S\subset S^d$ just as in the proof of Lemma 8.5, and it becomes a cover. Now we have

$$|S|(\frac{\epsilon}{2})^d\operatorname{vol}(B)=\operatorname{vol}(\bigcup_{x\in S}B_x(\epsilon/2))\le\operatorname{vol}(B_0(R+\frac{\epsilon}{2})\backslash B_0(R-\frac{\epsilon}{2}))=((R+\frac{\epsilon}{2})^d-(R-\frac{\epsilon}{2})^d)\operatorname{vol}(B)\le\epsilon d(R+\frac{\epsilon}{2})^{d-1}\operatorname{vol}(B)$$

□

Lemma 8.7. *Let $A=\sum_{i=1}^{d}\lambda_i u_i u_i^T$ and $B=\sum_{i=1}^{d}\mu_i v_i v_i^T$ where $\lambda_i,\mu_i\ge 0$ and u_i,v_i are unit vectors, not necessarily eigenvectors.Then we have the following bound on the nuclear norm of the difference.*

$$\|A-B\|_{nuc}\le\sum_{i=1}^{d}|\lambda_i-\mu_i|+2\min(\lambda_i,\mu_i)\,\|u_i-v_i\|$$

Without lose of generality, we may assume $\lambda_1 \geq \mu_1, \cdots, \lambda_a \geq \mu_a, \lambda_{a+1} < \mu_{a+1}, \cdots, \lambda_d < \mu_d$.

$$\begin{aligned}
\|A-B\|_{nuc} &= \left\| \sum_{i=1}^{d} \lambda_i u_i u_i^T - \sum_{i=1}^{d} \mu_i v_i v_i^T \right\|_{nuc} \\
&= \left\| \sum_{i=1}^{d} \min(\lambda_i, \mu_i)(u_i u_i^T - v_i v_i^T) + \sum_{i=1}^{a} (\lambda_i - \mu_i) u_i u_i^T - \sum_{i=a+1}^{d} (\mu_i - \lambda_i) v_i v_i^T \right\|_{nuc} \\
&\leq \sum_{i=1}^{d} \min(\lambda_i, \mu_i) \left\| u_i u_i^T - v_i v_i^T \right\|_{nuc} + \sum_{i=1}^{a} (\lambda_i - \mu_i) \left\| u_i u_i^T \right\|_{nuc} + \sum_{i=a+1}^{d} (\mu_i - \lambda_i) \left\| v_i v_i^T \right\|_{nuc} \\
&\leq \sum_{i=1}^{d} \min(\lambda_i, \mu_i) \|u_i - v_i\| \sqrt{2(\|u_i\|^2 + \|v_i\|^2)} + \sum_{i=1}^{a} (\lambda_i - \mu_i) + \sum_{i=a+1}^{d} (\mu_i - \lambda_i) \\
&= \sum_{i=1}^{d} |\lambda_i - \mu_i| + 2\min(\lambda_i, \mu_i) \|u_i - v_i\|
\end{aligned}$$

where in the last inequality we used (1)Lemma 8.1 together with the fact that $|||\cdot|||$ *and* $\|\cdot\|_{nuc}$ *are dual to each other.*

Lemma 8.8. *Let* X *be a metric space and* $Y \subset X$*. If* $C \subset X$ *is a* ϵ*-cover of* Y*, then there exists an* 3ϵ*-cover* C' *of* Y *such that* $C' \subset Y$ *and* $|C'| \leq |C|$

Proof. Let $f : X \to Y$ be any function such that for all $x \in X$, $f(x)$ satisfies

$$d(x, f(x)) \leq \epsilon + \inf_{y \in Y} d(x, y)$$

Define

$$|C'| = \{f(x) | x \in C\}$$

We claim that C' is a 3ϵ-cover of Y. For $y \in Y$, let $x \in C$ be such that $d(x, y) \leq \epsilon$. Then

$$\begin{aligned}
d(y, f(x)) &\leq d(y, x) + d(x, f(x)) \\
&\leq d(y, x) + \epsilon + \inf_{y' in Y} d(x, y') \\
&\leq d(y, x) + \epsilon + d(x, y) \\
&\leq 3\epsilon
\end{aligned}$$

□

Lemma 8.9. *Let* $X = \left\{A \in \mathbb{R}^{d \times d} | A \text{ is symmetric positive definite and } \lambda_{min} \leq \Lambda(A) \leq \lambda_{max}\right\}$*, then there exists a* ϵ*-cover with respect to the nuclear norm of* X *within* X *of size*

$$(1 + \frac{24\lambda_{max} d}{\epsilon})^{d(d-1)} (\frac{6d(\lambda_{max} - \lambda_{min})}{\epsilon})^d$$

Proof. a □

Lemma 8.10. *Let* $0 < \epsilon \leq \sqrt{120}\lambda_{max}$*,* $X = \left\{A \in \mathbb{R}^{d \times d} | A \text{ is symmetric positive definite and } \lambda_{min} \leq \Lambda(A) \leq \lambda_{max}\right\}$*, then there exists a* ϵ*-cover with respect to the nuclear norm of* X *of size*

$$2^d d! (1 + \frac{\lambda_{max} 2^{d+3}}{\epsilon})^{\binom{d}{2}} (\frac{2d(\lambda_{max} - \lambda_{min})}{\epsilon})^d$$

Proof. The proof is composed of (1)finding a cover for a set of orthogonal unit eigenvectors, (2)finding a cover for a set of eigenvalues, and (3)combining them via Lemma 8.7

(1) Let $\epsilon_0 = \lambda_{max}^{-1} 2^{-d-2} \epsilon$. Define

$$\Phi : \left\{A \in \mathcal{P}(\mathbb{R}^d) : |A| < d\right\} \to \mathcal{P}(\left\{x \in \mathbb{R}^d : \|x\| = 1\right\})$$

in such a way that for each A, $\Phi(A)$ is a minimum size ϵ_0-l^2-cover of $\left\{x \in A^\perp : \|x\| = 1\right\}$. By Lemma 8.6, we have

$$|\Phi(A)| \leq 2(d - |A|)(1 + \frac{2}{\epsilon_0})^{d-|A|-1}$$

Then define

$$C_{vec} = \left\{(x_1, \cdots, x_d) \in (\mathbb{R}^d)^d : \forall 1 \leq i \leq d. x_i \in \Phi(\{x_1, \cdots, x_{i-1}\})\right\}$$

which has size

$$|C_{vec}| \leq \prod_{i=0}^{d-1} 2(d-i)(1 + \frac{2}{\epsilon_0})^{d-i-1} = 2^d d! (1 + \frac{\lambda_{max} 2^{d+3}}{\epsilon})^{\binom{d}{2}}$$

We claim that for any set of orthogonal unit vectors $(y_1, \cdots, y_d) \in (\mathbb{R}^d)^d$, we can find $(x_1, \cdots, x_d) \in C_{vec}$ in such a way that

$$\sum_{i=1}^{d} \|x_i - y_i\| \leq 2^d \epsilon_0 = \frac{\epsilon}{4\lambda_{max}} \tag{8.2}$$

First, note the following simple geometric property, which can be shown easily.

$$\forall x, y, y' \in \mathbb{R}^d. \|x\| = \|x'\| = \|y\| = 1 \wedge y \perp y' \wedge \|x - y\| \leq \epsilon' \implies |x \cdot y'| \leq \sqrt{\epsilon'^2 - \epsilon'^4/4} \tag{8.3}$$

For $1 \leq i \leq d$, let us assume that we have constructed $x_1, \cdots, x_{i-1}$ such that
(a) $x_1 \in \Phi(\phi), x_2 \in \Phi(\{x_1\}), \cdots, x_{i-1} \in \Phi(\{x_1, \cdots, x_{i-2}\})$
(b) $\|x_1 - y_1\| \leq \epsilon_0 = \epsilon_1, \|x_2 - y_2\| \leq 2\epsilon_0 = \epsilon_2, \cdots, \|x_{i-1} - y_{i-1}\| \leq 2^{i-2}\epsilon_0 = \epsilon_{i-1}$
Using (8.3), we get

$$|x_k \cdot y_i| \leq \sqrt{\epsilon_k^2 - \epsilon_k^4/4} \quad (1 \leq k \leq i-1) \tag{8.4}$$

Let $\pi : \mathbb{R}^d \to \{x_1, \cdots, x_{i-1}\}^\perp$ be the projection onto $\{x_1, \cdots, x_{i-1}\}^\perp$,

$$\alpha = \|\pi(y_i)\|$$

, and

$$\beta = \sqrt{\sum_{k=1}^{i-1} (x_k \cdot y_i)^2}$$

. Decomposing y_i into components, we have

$$y_i = \pi(y_i) + \sum_{k=1}^{i-1} (x_k \cdot y_i) x_k$$

which shows

$$\alpha^2 + \beta^2 = 1$$

Since

$$\beta^2 \leq \sum_{k=1}^{i-1} \epsilon_k^2 \leq 2^{2d-3}\epsilon_0^2 = \lambda_{max}^{-2} 2^{-7}\epsilon \leq 15/16 \tag{8.5}$$

, we have $\alpha \neq 0$ so we may define

$$z = \alpha^{-1}\pi(y_i)$$

Let $A = 1 - \sum_{k=1}^{i-1} \epsilon_k^2/2$, $B = \sum_{j \neq k} \epsilon_j^2 \epsilon_k^2/2$. Then

$$\begin{aligned}
\|y_i - z\|^2 &= (1 - \alpha^{-1})^2 \alpha^2 + \beta^2 \\
&= 2 - 2\sqrt{1 - \beta^2} \\
&= 2 - 2\sqrt{1 - \sum_{k=1}^{i-1} \epsilon_k^2 + \sum_{k=1}^{i-1} \epsilon_k^4/4} \\
&= 2 - 2\sqrt{A^2 - B} \\
&\leq 2 - 2(A - \frac{B}{2\sqrt{A^2 - B}}) \\
&= \sum_{k=1}^{i-1} \epsilon_k^2 + \frac{B}{\sqrt{A^2 - B}} \\
&= \sum_{k=1}^{i-1} \epsilon_k^2 + \frac{\sum_{j \neq k} \epsilon_j^2 \epsilon_k^2/2}{\sqrt{1 - \sum_{k=1}^{i-1} \epsilon_k^2 + \sum_{k=1}^{i-1} \epsilon_k^4/4}} \\
&\leq \sum_{k=1}^{i-1} \epsilon_k^2 + 2\sum_{j \neq k} \epsilon_j^2 \epsilon_k^2 \quad \text{(by (8.5))} \\
&\leq (\sum_{k=1}^{i-1} \epsilon_k)^2
\end{aligned}$$

Since $z \in \{x_1, \cdots, x_{i-1}\}^\perp$ and due to the cover property of $\Phi(\{x_1, \cdots, x_{i-1}\})$, we can find $x_i \in\in \Phi(\{x_1, \cdots, x_{i-1}\})$ such that $\|x_i - z\| \leq \epsilon_0$. These show

$$\|x_i - y_i\| \leq \|x_i - z\| + \|z - y_i\| \leq \epsilon_0 + \sum_{k=1}^{i-1} \epsilon_k = \epsilon_0 + \sum_{k=1}^{i-1} 2^{k-1}\epsilon_0 = 2^{i-1}\epsilon_0$$

which shows x_i has the desired property. Therefore, we have constructed $x_1, \cdots, x_d \in \mathbb{R}^d$ such that
(a) $x_1 \in \Phi(\phi), x_2 \in \Phi(\{x_1\}), \cdots, x_d \in \Phi(\{x_1, \cdots, x_{d-1}\})$
(b) $\|x_1 - y_1\| \leq \epsilon_0, \|x_2 - y_2\| \leq 2\epsilon_0, \cdots, \|x_d - y_d\| \leq 2^{d-1}\epsilon_0$
These imply both $(x_1, \cdots, x_d) \in C_{vec}$ and (8.2), showing the claim.

(2) For eigenvalues, we simply take a $\frac{\epsilon}{2d}$-l^∞-cover C_{val} of $[\lambda_{min}, \lambda_{max}]^d$ of size

$$(\frac{2d(\lambda_{max} - \lambda_{min})}{\epsilon})^d$$

(3) Let

$$C = \left\{ \sum_{i=1}^{d} \lambda_i u_i u_i^T : (\lambda_1, \cdots, \lambda_d) \in C_{val}, (u_1, \cdots, u_d) \in C_{vec} \right\}$$

We claim that C is a nuclear norm ϵ-cover of X. For $B = \sum_{i=1}^d \mu_i v_i v_i^T$ where $\{v_i\}_{i=1}^d$ is a set of orthogonal unit eigenvectors and μ_i are eigenvalues, find $A = \sum_{i=1}^d \lambda_i u_i u_i^T$ such that
(a) $(\lambda_1, \cdots, \lambda_d) \in C_{val}$ and $(u_1, \cdots, u_d) \in C_{vec}$
(b) $\max_{i=1}^d |\lambda_i - \mu_i| \leq \frac{\epsilon}{2d}$
(c) $\sum_{i=1}^d \|u_i - v_i\| \leq \frac{\epsilon}{4\lambda_{max}}$
Then by Lemma 8.7,

$$\begin{aligned} \|A - B\|_{\text{nuc}} &\leq \sum_{i=1}^{d} |\lambda_i - \mu_i| + 2\min(\lambda_i, \mu_i) \|u_i - v_i\| \\ &\leq d \cdot \frac{\epsilon}{2d} + 2\lambda_{max} \cdot \frac{\epsilon}{4\lambda_{max}} = \epsilon \end{aligned}$$

□

Theorem 8.11. *Suppose $x_1, \cdots, x_m \in \mathbb{R}^d$ satisfies $\|x_i\| \leq R'$, and $\epsilon \leq 60\lambda_{max}^2(\frac{d}{\lambda_{min}} + (R + R')^2)$. As in the beginning of this section, let*

$$M_a = (\frac{\lambda_{max}}{\lambda_{min}})^{d^2/2} e^{\frac{1}{2}(\lambda_{max} - \lambda_{min})(R+R')^2}$$

,

$$M_\mu = \lambda_{max}(R + R')$$

and

$$M_K = \frac{1}{2}(\frac{d}{\lambda_{min}} + (R + R')^2)$$

then there exists a $(x_1, \cdots, x_m)$-ϵ-cover of $\mathcal{H}$, in the sense of the previous section, of size

$$(\frac{6k\sqrt{M_a}}{\epsilon})^k (1 + \frac{6M_\mu R}{\epsilon})^{dk} 2^{\binom{d}{2}k} (d!)^k (1 + \frac{\lambda_{max} 2^{d+3} 3M_K}{\epsilon})^{\binom{d}{2}k} (\frac{6d(\lambda_{max} - \lambda_{min})M_K}{\epsilon})^{dk}$$

Proof. In this proof, as noted earlier, l_θ will be interpreted in the extended sense.

(1) Since $M_a \geq 1$, let $0 < \delta = \frac{1}{\sqrt{M_a}} < 1$. Let $C_a \subset [0, 1]$ be the set where $\frac{3kM_a}{\epsilon}\delta$ points are uniformly distributed in $[0, \delta]$ and $\frac{3k\delta^{-1}}{\epsilon}(1 - \delta)$ points are uniformly distributed in $(\delta, 1]$, which has size

$$|C_a| \leq \frac{3kM_a}{\epsilon}\delta + \frac{3k\delta^{-1}}{\epsilon}(1 - \delta) < \frac{6k\sqrt{M_a}}{\epsilon}$$

(2) Let $C_\mu \subset \mathbb{R}^d$ be a l^2-$\frac{\epsilon}{3M_\mu}$-cover of $B(R)$ of size

$$|C_\mu| \leq (1 + \frac{6M_\mu R}{\epsilon})^d$$

found by Lemma 8.5.

(3) Let C_K be a $\|\cdot\|_{\text{nuc}}$-$\frac{\epsilon}{3M_K}$-cover of $\left\{A \in \mathbb{R}^{d\times d} | A \text{ is a symmetric matix with } \lambda_{min} \leq \Lambda(A) \leq \lambda_{max}\right\}$ of size

$$|C_K| \leq 2^{\binom{d}{2}} d! (1 + \frac{\lambda_{max} 2^{d+3} 3M_K}{\epsilon})^{\binom{d}{2}} (\frac{6d(\lambda_{max} - \lambda_{min})M_K}{\epsilon})^d$$

found by Lemma 8.9.

(4) Let

$$C = \left\{ (l_\theta(x_1), \cdots, l_\theta(x_m)) : \theta = \{(a_i, \mu_i, K_i)\}_{i=1}^k, a_i \in C_a, \mu_i \in C_\mu, K_i \in C_K \right\}$$

We claim that C is a $(x_1, \cdots, x_m)$-ϵ-cover of $\mathcal{H}$.

Let $\theta = \{(a_i, \mu_i, K_i)\}_{i=1}^k \in \Theta$. We find $\theta' = \{(a_i', \mu_i', K_i')\}_{i=1}^k \in C_a^k \times C_\mu^k \times C_K^k$ such that

(a)
- If $a_i \in [0, \delta]$ then $a_i' \in [0, \delta]$ and $|a_i - a_i'| \leq \frac{\epsilon}{3kM_a}$.
- If $a_i \in (\delta, 1]$ then $a_i' \in (\delta, 1]$ and $|a_i - a_i'| \leq \frac{\epsilon}{3k\delta^{-1}}$

(b) $\|\mu_i - \mu_i'\| \leq \frac{\epsilon}{3M_\mu}$

(c) $\|K_i - K_i'\|_{\text{nuc}} \leq \frac{\epsilon}{3M_K}$

Let $1 \leq j \leq m$. By the mean value theorem, there exists $0 < p_j < 1$ such that $\theta_j = p_j\theta + (1 - p_j)\theta'$ satisfies

$$l_{\theta'}(x_j) - l_\theta(x_j) = (\theta' - \theta) \odot \nabla_\theta l_\theta(x_j)|_{\theta=\theta_j}$$

Then

$$\begin{aligned}
|l_{\theta'}(x_j) - l_\theta(x_j)| &\leq \sum_{i=1}^k |a_i' - a_i| \left| \frac{\partial l_\theta(x_j)}{\partial a_i}|_{\theta=\theta_j} \right| + \sum_{i=1}^k \|\mu_i' - \mu_i\| \left\| \frac{\partial l_\theta(x_j)}{\partial \mu_i}|_{\theta=\theta_j} \right\| + \sum_{i=1}^k \|K_i' - K_i\|_{\text{nuc}} \left\| \left\| \frac{\partial l_\theta(x_j)}{\partial K_i}|_{\theta=\theta_j} \right\| \right\| \\
&\leq k \cdot \frac{\epsilon}{3k} + \frac{\epsilon}{3M_\mu} \sum_{i=1}^k \left\| \frac{\partial l_\theta(x_j)}{\partial \mu_i}|_{\theta=\theta_j} \right\| + \frac{\epsilon}{3M_K} \sum_{i=1}^k \left\| \left\| \frac{\partial l_\theta(x_j)}{\partial K_i}|_{\theta=\theta_j} \right\| \right\| \\
&\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon
\end{aligned}$$

□

8.4. **resulting bound.**

Theorem 8.12. *a*

References

[Anthony and Bartlett, 2009] Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

[Chung and Lu, 2006] Chung, F. and Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127.

[Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

[Naor, 2008] Naor, A. (2008). Concentration of measure.