Dr. Stefan Richter
Statistische Analyse von maschinellen Lernalgorithmen
Wintersemester 2020/2021

12. Übungsblatt

Aufgabe 1 (Maximalungleichungen aus der Bernstein-Ungleichung, Lemma 7.10(ii)):
Lösung:

(a) Es gilt für beliebiges $t_0 > 0$:

$$\begin{aligned}\mathbb{E}[W^2] &= \int_0^\infty \mathbb{P}(W^2 \geq u)du = \int_0^\infty \mathbb{P}(W \geq \sqrt{u})du \overset{t=\sqrt{u}}{=} 2\int_0^\infty t\mathbb{P}(W \geq t)dt \\ &= \underbrace{2\int_0^{t_0} t\,\underbrace{\mathbb{P}(W \geq t)}_{\leq 1}\,dt}_{\leq t_0^2} + 2\int_{t_0}^\infty t\mathbb{P}(W \geq t)dt.\end{aligned}$$

(b) Es gilt für jedes $g \in \mathcal{G}$:

$$\|\tilde{g}\|_\infty \leq \frac{\|g\|_\infty}{M\sqrt{\frac{H}{n}}} \leq \sqrt{\frac{n}{H}},$$

und

$$\mathbb{E}[g(Z_1)^2] = \frac{\mathbb{E}[g(Z_1)^2]}{\mathbb{E}[g(Z_1)^2]} \leq 1.$$

Mit Bernstein-Ungleichung gilt

$$\begin{aligned}\mathbb{P}\Big(\Big|\sum_{i=1}^n \{\tilde{g}(Z_i) - \mathbb{E}\tilde{g}(Z_i)\}\Big| \geq t\Big) &\leq 2\exp\Big(-\frac{t^2}{2n + \frac{2}{3}\sqrt{\frac{n}{H}}\cdot t}\Big) \\ &\overset{t\geq 3\sqrt{nH}}{\leq} 2\exp\Big(-\frac{t^2}{\frac{4}{3}\sqrt{\frac{n}{H}}\cdot t}\Big) = 2\exp\Big(-\frac{3t}{4\sqrt{\frac{n}{H}}}\Big).\end{aligned}$$

Damit folgt

$$\mathbb{P}(W \geq t) \leq |\mathcal{G}|\mathbb{P}\big(\big|\sum_{i=1}^n \{\tilde{g}(Z_i) - \mathbb{E}\tilde{g}(Z_i)\}\big| \geq t\big) \leq 2|\mathcal{G}|\exp\Big(-\frac{3t}{4\sqrt{\frac{n}{H}}}\Big).$$

(c) Wähle $t_0 = 3\sqrt{nH}$ und $a = \frac{3}{4}\sqrt{\frac{H}{n}}$.
Partielle Integration $\Rightarrow \int te^{-at}dt = -(a^{-2} + a^{-1}t)e^{-at} \Rightarrow \int_{t_0}^\infty te^{-at}dt = (a^{-2} + a^{-1}t_0)e^{-at_0}$.
Es folgt

$$\begin{aligned}2\cdot\int_{t_0}^\infty t\mathbb{P}(W \geq t)dt &\leq 4(a^{-2} + a^{-1}t_0)|\mathcal{G}|\underbrace{e^{-at_0}}_{=e^{-\frac{9}{4}H}\leq|\mathcal{G}|^{-1}} \\ &\leq 4\big\{(\frac{4}{3})^2\frac{n}{H} + 4n\big\} \\ &\overset{|\mathcal{G}|\geq 2}{\leq} 24n.\end{aligned}$$

Zusammen mit $t_0^2 = 9nH$ folgt aus (a) die Behauptung.

Aufgabe 2 (Maximalungleichung für Normalverteilungen, Lemma 7.11):
Lösung:

(a) $Z_1, ..., Z_n \sim N(0, v^2)$ i.i.d. $\Rightarrow$

$$W_j = \frac{1}{v\sqrt{n}} \frac{\sum_{i=1}^n a_{ij} Z_i}{\left(\frac{1}{n}\sum_{i=1}^n a_{ij}^2\right)^{1/2}} \sim N(0,1).$$

(b) Für $W_j \sim N(0,1)$ gilt $\mathbb{E}\exp(\frac{W_j^2}{4}) = \sqrt{2}$. Mit der konvexen, monoton wachsenden Funktion $\varphi_2(x) = \exp(x^2) - 1$ folgt

$$\varphi_2(\mathbb{E}\max_{j=1,...,N} \frac{|W_j|}{2}) \leq \mathbb{E}\max_{j=1,...,N} \varphi_2(\frac{|W_j|}{2}) \leq \sum_{j=1}^N \mathbb{E}\varphi_2(\frac{|W_j|}{2}) \leq N\cdot(\sqrt{2}-1) \leq N.$$

$\Rightarrow$

$$\mathbb{E}\max_{j=1,...,N} |W_j| \leq 2\varphi_2^{-1}(N) = 2\sqrt{\log(N+1)}.$$

(c) Mit $\varphi_1(x) = e^x - 1$:

$$\varphi_1(\mathbb{E}\max_{j=1,...,N} \frac{|W_j|^2}{4}) \leq \mathbb{E}\max_{j=1,...,N} \varphi_1(\frac{|W_j|^2}{4}) \leq \sum_{j=1}^N \mathbb{E}\varphi_1(\frac{|W_j|^2}{4}) \leq N\cdot(\sqrt{2}-1) \leq N.$$

$\Rightarrow$

$$\mathbb{E}\max_{j=1,...,N} |W_j|^2 \leq 4\varphi_1^{-1}(N) = 4\log(N+1).$$

Aufgabe 3 (Covering Numbers von neuronalen Netzwerken, Lemma 7.12):
Lösung:

(a) Schreibe $\theta = (v^1, ..., v^{(L)}, W^{(0)}, ..., W^{(L)})$ (Matrizen zeilenweise) als Vektor. Dann gilt $\theta \in [-1,1]^T$ mit

$$T = \sum_{l=1}^L \text{Anz. Einträge } v^{(l)} + \sum_{l=0}^L \text{Anz. Einträge } W^{(l)} = \sum_{l=1}^L p_l + \sum_{l=0}^L p_l p_{l+1}.$$

(b) Für jedes $f_\theta \in \mathcal{F}(L, p, s, \infty)$ sind jeweils höchstens s Einträge von θ Nicht Null, d.h. $\mathcal{F}(L,p,s,\infty) = \{f_\theta : \theta \in \Theta\}$ mit

$$\begin{aligned} \Theta &\subset \{\theta \in [-1,1]^T : \leq s \text{ Einträge von } \theta \text{ nicht Null}\} \\ &= \{\theta \in [-1,1]^T : \exists S \subset \{1,...,T\}, |S| \leq s : \forall j \in S^c : \theta_j = 0\} \\ &= \bigcup_{S\subset\{1,...,T\}:|S|\leq s} \underbrace{\{\theta \in [-1,1]^T : \forall j \in S^c : \theta_j = 0\}}_{=:\Theta_S}. \end{aligned}$$

(c) Wir wählen eine geeignete Gitterapproximation von Θ_S. OBdA. sei $S = \{1, ..., s\}$ (damit $\Theta_S \subset [-1,1]^s \times \{0\}^{T-s}$). Dann setze

$$\tilde{\Theta}_S = \Big\{-1 + j\cdot a : j = 1, ..., m\Big\}^s \times \{0\}^{T-s}.$$

Falls $-1 + m\cdot a \geq 1 - a$, erfüllt $\tilde{\Theta}_S$ die Approximationsaussage. Dies ist erfüllt mit $(m+1)\cdot a \geq 2$, d.h. $m \geq \frac{2}{a} - 1$ bzw. falls $m \geq \lfloor \frac{2}{a} \rfloor$.

Es folgt

$$|\tilde{\Theta}_S| \leq m^s = \left\lfloor \frac{2}{a} \right\rfloor^s.$$

(d) Es gilt

$$\begin{aligned}|\tilde{\Theta}| &\leq \sum_{S\subset\{1,...,T\}:|S|\leq s} |\tilde{\Theta}_S| \leq \sum_{k=0}^{s} \sum_{S\subset\{1,...,T\}:|S|=k} |\tilde{\Theta}_S| \\ &\overset{(c)}{=} \sum_{k=0}^{s} \#\{S\subset\{1,...,T\}:|S|=k\}\cdot \left\lfloor\frac{2}{a}\right\rfloor^{s}.\end{aligned}$$

Es gilt

$$\#\{S\subset\{1,...,T\}:|S|=k\} = \binom{T}{k} \leq T^k \leq V^k,$$

denn

$$T = \sum_{l=0}^{L} p_l p_{l+1} + \sum_{l=1}^{L} p_l \leq \sum_{l=0}^{L}(p_l+1)p_{l+1} \leq \prod_{l=0}^{L+1}(p_l+1) = V.$$

Einsetzen:

$$|\tilde{\Theta}| \leq \sum_{k=0}^{s} V^k \cdot \big(\frac{2}{a}\big)^k \leq \big(\frac{2V}{a}\big)^{s+1}.$$

(e) Für $\gamma > 0$ wähle $a = \frac{\gamma}{(L+1)V}$. Dann gibt es für jedes $f \in \mathcal{F}(L,p,s,F)$ ein $f_{\tilde{\theta}}$ mit $\tilde{\theta}\in\tilde{\Theta}$, so dass

$$\|f_\theta - f_{\tilde{\theta}}\|_\infty \leq a\cdot(L+1)\cdot V \leq \gamma.$$

$\Rightarrow$

$$N(\gamma,\mathcal{F}(L,p,s,F),\|\cdot\|_\infty) \leq \big(\frac{2V}{a}\big)^{s+1} = \big(2\gamma^{-1}V^2(L+1)\big)^{s+1}.$$

(f) Es gilt

$$\begin{aligned}|A_\theta^{k-}(x) - A_\theta^{k-}(x')| &\leq \|W^{(L)}\|_Z\cdot ... \cdot\|W^{(k-1)}\|_Z\|x-x'\|_\infty \\ &\underset{\|W^{(l)}\|_\infty\leq 1, W^{(l)}\in\mathbb{R}^{p_l\times p_{l+1}}}{\leq} \Big(\prod_{l=k-1}^{L} p_l\Big)\cdot\|x-x'\|_\infty.\end{aligned}$$

(g) Es gilt

$$\begin{aligned}|A_\theta^{k+}(x)|_\infty &\leq \|W^{(k)}\|_Z\cdot\big\{...\|W^{(1)}\|_Z\big\{\|W^{(0)}\|_Z\|x\|_\infty + \|v^{(1)}\|_\infty) + \|v^{(2)}\|_\infty\big\}...\big\} + \|v^{(k)}\|_\infty \\ &\leq \prod_{l=0}^{k-1}(p_l+1).\end{aligned}$$

(h) Es gilt mit $\sigma_{v^{(L+1)}}(x) := x$:

$$\begin{aligned}
|f(x) - f_{\tilde\theta}(x)| &\le \sum_{k=1}^{L+1} \Big| A_\theta^{(k+1)-} \circ \sigma_{v^{(k)}} W^{(k-1)} A_{\tilde\theta}^{(k-1)-}(x) - A_\theta^{(k+1)-} \circ \sigma_{\tilde v^{(k)}} \tilde W^{(k-1)} A_{\tilde\theta}^{(k-1)-}(x) \Big| \\
&\overset{(f)}{\le} \sum_{k=1}^{L+1} \Big(\prod_{l=k}^{L} p_l\Big) \Big\| \sigma_{v^{(k)}} W^{(k-1)} A_{\tilde\theta}^{(k-1)-}(x) - \sigma_{\tilde v^{(k)}} \tilde W^{(k-1)} A_{\tilde\theta}^{(k-1)-}(x) \Big\|_\infty \\
&\le \sum_{k=1}^{L+1} \Big(\prod_{l=k}^{L} p_l\Big) \cdot \big[\|v^{(k)} - \tilde v^{(k)}\|_\infty + \|W^{(k-1)} - \tilde W^{(k-1)}\|_Z \cdot \|A_{\tilde\theta}^{(k-1)}(x)\|_\infty \big] \\
&\overset{(g)}{\le} \sum_{k=1}^{L+1} \Big(\prod_{l=k}^{L} p_l\Big) \cdot \big[a + p_{k-1} \cdot a \cdot \prod_{l=0}^{k-2} (p_l + 1) \big] \\
&\le a \cdot (L+1) \cdot \prod_{l=0}^{L} (p_l + 1) = a \cdot (L+1) \cdot V.
\end{aligned}$$

(i) Ist $f \in \mathcal{F}(L, p, s, F)$, so gilt für jeden Layer $l \in \{1, ..., L\}$: Höchstens s Spalten von $W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}}$ sind keine kompletten Nullspalten. Ist aber die j-te Spalte von $W^{(l)}$ eine komplette Nullspalte, kann diese zusammen mit der j-ten Zeile von $W^{(l-1)}$ und dem j-ten Eintrag aus $v^{(l)}$ aus dem Modell entfernt werden, und es wird immer noch dieselbe Funktion dargestellt. Daher gilt in diesem Fall

$$f \in \mathcal{F}(L, (p_0, p_1, ..., p_{l-1}, p_l - 1, p_{l+1}, ..., p_L, p_{L+1}), s, F).$$

Iteration dieses Vorgehens liefert

$$f \in \mathcal{F}(L, (p_0, p_1 \wedge s, ..., p_L \wedge s, p_{L+1}), s, F).$$

(j) Mit $\tilde V = (p_0 + 1) \cdot (p_{L+1} + 1) \cdot \prod_{l=1}^{L} ((p_l \wedge s) + 1) \le 2^{L+2} p_0 p_{L+1} s^L$ folgt:

$$\begin{aligned}
H(\gamma) &= \log N(\gamma, \mathcal{F}(L, p, s, F), \|\cdot\|_\infty) \\
&\overset{(i)}{=} \log N(\gamma, \mathcal{F}(L, (p_0, p_1 \wedge s, ..., p_L \wedge s, p_{L+1}), s, F), \|\cdot\|_\infty) \\
&\overset{(e)}{\le} (s+1) \log(2\gamma^{-1} \tilde V^2 (L+1)) \\
&= (s+1) \log(2^{2L+5} \gamma^{-1} (L+1) p_0^2 p_{L+1}^2 s^{2L}) \\
&\le 2s \cdot \big\{ (2L+5)\log(2) + \log(\gamma^{-1}) + \log(2L) + 2\log(p_0 p_{L+1}) + 2L \log(s) \big\} \\
&\le c \cdot s \cdot \big\{ L \log(s) + \log(\gamma^{-1}) + \log(p_0 p_{L+1}) \big\}
\end{aligned}$$

mit $c \ge 1$ eine universelle Konstante, groß genug und $s \ge 2, L \ge 1$.

Homepage der Vorlesung:
`https://ssp.math.uni-heidelberg.de/sam-ws2020/`