# פתרון תרגיל מספר 5 - מערכות לומדות

**שם:** מיכאל גרינבאום, **ת.ז:** 211747639

13 ביוני 2020

1. **פתרון:**

(א) **צ"ל:** $L_D(h^*) \le \min_{h\in\mathcal{H}_k} L_D(h) + \sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}$

**הוכחה:**

נגדיר $\mathcal{H} = \mathcal{H}_k$ ונגדיר $\varepsilon = \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}}$, ראינו בהרצאה 5 בהוכחה של המשפט היסודי של ה- $IML$ שמתקיים (ראינו בכך ש- $L_S$ הוא $L_D$ בתוחלת, והוא ממוצע של משתנים מקרים חסומים ב- $0-1$ מאותה התפלגות ובלתי תלויים)

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(S \mid \exists h\in\mathcal{H} \text{ s.t. } \left|L_D(h) - L_S(h)\right| > \varepsilon\right) \le 2\cdot|\mathcal{H}|\cdot\exp\left(-2m\cdot\varepsilon^2\right)$$

נעזר בתכונה זאת ונראה כי

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^m}\left(S \mid \exists h\in\mathcal{H} \text{ s.t. } \left|L_D(h) - L_S(h)\right| > \varepsilon\right) \le 2\cdot|\mathcal{H}|\cdot\exp\left(-2m\cdot\varepsilon^2\right)\\
&= 2\left|\mathcal{H}\right|\cdot\exp\left(-2m\cdot\frac{\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}\right) = 2\left|\mathcal{H}\right|\cdot\exp\left(-\ln\left(\frac{2\left|\mathcal{H}\right|}{\delta}\right)\right) = 2\left|\mathcal{H}\right|\cdot\frac{\delta}{2\cdot\left|\mathcal{H}\right|} = \delta
\end{aligned}$$

כלומר נקבל כי

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}\left(S \mid S \text{ is } \varepsilon \text{ representative}\right) &= 1 - \mathbb{P}_{S\sim\mathcal{D}^m}\left(S \mid S \text{ is not } \varepsilon \text{ representative}\right)\\
&= 1 - \mathbb{P}_{S\sim\mathcal{D}^m}\left(S \mid \exists h\in\mathcal{H} \text{ s.t. } \left|L_D(h) - L_S(h)\right| > \varepsilon\right) = 1-\delta
\end{aligned}$$

תהי $S\sim\mathcal{D}^m$ כך ש- $S$ is $\varepsilon$ representative, תהי $h\in\mathcal{H}$, נשים לב כי

$$\begin{aligned}
L_D(h^*) &\overset{\text{S is } \varepsilon \text{ representative}}{\le} L_S(h^*) + \varepsilon\\
&\overset{h^* \text{ minimizes the error on S}}{\le} L_S(h) + \varepsilon\\
&\overset{\text{S is } \varepsilon \text{ representative}}{\le} L_D(h) + \varepsilon + \varepsilon = L_D(h) + 2\cdot\varepsilon\\
&= L_D(h) + 2\cdot\sqrt{\frac{\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2m}} = L_D(h) + \sqrt{\frac{2\cdot\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{m}}
\end{aligned}$$

ולכן מהיות וזה נכון לכל $h\in\mathcal{H}$ נקבל כי $L_D(h^*) \le \min_{h\in\mathcal{H}} L_D(h) + \sqrt{\frac{2\cdot\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{m}}$ ולכן מתקיים

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(h^*) \le \min_{h\in\mathcal{H}_k} L_D(h) + \sqrt{\frac{2\cdot\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}\right) \ge \mathbb{P}_{S\sim\mathcal{D}^m}\left(S \mid S \text{ is } \varepsilon \text{ representative}\right) = 1-\delta$$

מ.ש.ל.א.☺

**(ב) צ"ל:** $L_{\mathcal{D}}\left(h^{*}\right)\leq\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4|\mathcal{H}_{j}|}{\delta}\right)}$

**הוכחה:**

תחילה נשים לב ש- $|V|=\alpha\cdot m$ וגם $|S|=(1-\alpha)\cdot m$,

עתה נסתכל על שלב הולידציה, יש לנו $|\{h_{1},\ldots,h_{k}\}|=k$ היפותזות ואנחנו ממזערים את השגיאה על $|V|$, לכן לפי מה שהוכח בסעיף הקודם מתקיים כי

$$\begin{aligned}
&\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2\cdot\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot m}}\right)\\
=&\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2\cdot\ln\left(\frac{2|\{h_{1},\ldots,h_{k}\}|}{\left(\frac{\delta}{2}\right)}\right)}{|V|}}\right)<\frac{\delta}{2}
\end{aligned}$$

עתה, נסתכל על שלה האימון של $\mathcal{H}_{j}$, נשים לב כי מחלקת ההיפותזות היא $\mathcal{H}_{j}$ ואנחנו ממזערים את השגיאה כל $|S|$ ולכן לפי מה שהוכח בסעיף הקודם נקבל כי

$$\begin{aligned}
&\mathbb{P}\left(L_{\mathcal{D}}\left(h_{j}\right)>\min_{h\in\mathcal{H}_{j}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2\cdot\ln\left(\frac{4|\mathcal{H}_{j}|}{\delta}\right)}{(1-\alpha)\cdot m}}\right)\\
=&\mathbb{P}\left(L_{\mathcal{D}}\left(h_{j}\right)>\min_{h\in\mathcal{H}_{j}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2\cdot\ln\left(\frac{2|\mathcal{H}_{j}|}{\left(\frac{\delta}{2}\right)}\right)}{|S|}}\right)<\frac{\delta}{2}
\end{aligned}$$

עתה נבחין כי מתקיים $\min_{h\in\mathcal{H}_{j}}L_{\mathcal{D}}\left(h\right)=\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)$ וגם מהנתון $h_{j}=h^{*}$ ולכן

$$\begin{aligned}
&\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_{j}\right|}{\delta}\right)}\right)\\
\leq&\mathbb{P}\left(\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}\right)\cup\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_{j}\right|}{\delta}\right)}\right)\right)\\
\leq&\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}\right)+\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)>\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_{j}\right|}{\delta}\right)}\right)\\
<&\frac{\delta}{2}+\mathbb{P}\left(L_{\mathcal{D}}\left(h_{j}\right)>\min_{h\in\mathcal{H}_{j}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_{j}\right|}{\delta}\right)}\right)<\frac{\delta}{2}+\frac{\delta}{2}=\delta
\end{aligned}$$

ולכן נקבל כי

$$\begin{aligned}
&\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)\leq\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_{j}\right|}{\delta}\right)}\right)\\
=&1-\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)\leq\min_{h\in\mathcal{H}_{k}}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_{j}\right|}{\delta}\right)}\right)\\
&\geq1-\delta
\end{aligned}$$

כלומר קיבלנו

$$\mathbb{P}\left(L_{\mathcal{D}}\left(h^{*}\right)\leq\min_{h\in\mathcal{H}_k}L_{\mathcal{D}}\left(h\right)+\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4\left|\mathcal{H}_j\right|}{\delta}\right)}\right)\geq1-\delta$$

כנדרש

מ.ש.ל.ב.☺

(ג) **צ"ל:** אי אפשר להשוות בין 2 החסמים
**הוכחה:**
נסמן את החסמים ב־ $\varepsilon_1=\sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}$, $\varepsilon_2=\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}$,
נסתכל על המנה שלהם ונקבל

$$\frac{\varepsilon_2}{\varepsilon_1}=\frac{\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}+\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}}{\sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}}=\frac{\sqrt{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}}{\sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}}+\frac{\sqrt{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}}{\sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}}$$

$$=\sqrt{\frac{\frac{2}{\alpha\cdot m}\cdot\ln\left(\frac{4k}{\delta}\right)}{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}}+\sqrt{\frac{\frac{2}{(1-\alpha)\cdot m}\cdot\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}}=\sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}}+\sqrt{\frac{\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}{(1-\alpha)\cdot\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}}$$

$\Leftarrow$: מקרה שבו החסם הראשון יותר טוב, נבחר $\mathcal{H}_1=\cdots=\mathcal{H}_k=\mathcal{H}$ (כש $\mathcal{H}$ בגודל קבוע ללא תלות ב־ $k$) ואז נקבל כי

$$\frac{\varepsilon_2}{\varepsilon_1}=\sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}}+\sqrt{\frac{\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}{(1-\alpha)\cdot\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}}$$

$$=\underbrace{\sqrt{\frac{1}{\alpha\cdot\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}}}_{\text{constant in k}}\cdot\underbrace{\sqrt{\ln\left(\frac{4k}{\delta}\right)}}_{\text{goes to infinity when }k\to\infty}+\underbrace{\sqrt{\frac{\ln\left(\frac{4|\mathcal{H}|}{\delta}\right)}{(1-\alpha)\cdot\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}}}_{\text{constant in k}}$$

כלומר נקבל כי $\lim_{k\to\infty}\frac{\varepsilon_2}{\varepsilon_1}=\infty$, כלומר $\varepsilon_1$ חסם יותר טוב מ־ $\varepsilon_2$ במקרה הזה
$\Rightarrow$: מקרה שבו החסם השני יותר טוב, נבחר $|\mathcal{H}_i|=e^{i^2}$ לכל $i\in[k]$ ונקבל כי

$$\frac{\varepsilon_2}{\varepsilon_1}=\sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot\ln\left(\frac{2\cdot e^{k^2}}{\delta}\right)}}+\sqrt{\frac{\ln\left(\frac{4e^{j^2}}{\delta}\right)}{(1-\alpha)\cdot\ln\left(\frac{2e^{k^2}}{\delta}\right)}}$$

$$=\sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot\left[\ln\left(\frac{2}{\delta}\right)+\ln\left(e^{k^2}\right)\right]}}+\sqrt{\frac{\ln\left(\frac{4}{\delta}\right)+\ln\left(e^{j^2}\right)}{(1-\alpha)\cdot\left[\ln\left(\frac{2}{\delta}\right)+\ln\left(e^{k^2}\right)\right]}}$$

$$=\sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot\left[\ln\left(\frac{2}{\delta}\right)+k^2\right]}}+\sqrt{\frac{\ln\left(\frac{4}{\delta}\right)+j^2}{(1-\alpha)\cdot\left[\ln\left(\frac{2}{\delta}\right)+k^2\right]}}$$

$$=\underbrace{\sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\cdot\left[\ln\left(\frac{2}{\delta}\right)+k^2\right]}}}_{\text{goes to 0 when }k\to\infty\text{ because }\frac{\ln(k)}{k^2}\to0}+\underbrace{\sqrt{\frac{\ln\left(\frac{4}{\delta}\right)+j^2}{(1-\alpha)\cdot\left[\ln\left(\frac{2}{\delta}\right)+k^2\right]}}}_{\text{goes to 0 when }k\to\infty\text{ because }\frac{j^2}{k^2}\to0\text{ for big enough k, j will stay fixed}}$$

כלומר נקבל כי $\lim_{k\to\infty}\frac{\varepsilon_2}{\varepsilon_1}=0$, כלומר $\varepsilon_2$ חסם יותר טוב מ־ $\varepsilon_1$ במקרה הזה

מ.ש.ל.ג.☺

2. **פתרון:**

(א) **צ"ל:** $\hat{w}_\lambda^{ridge} = \frac{\hat{w}^{LS}}{1+\lambda}$

**הוכחה:**

נזכר שבמקרה שלנו מתקיים $\hat{w}^{LS} = \left(X \cdot X^T\right)^{-1} \cdot X \cdot y = X \cdot y$, נזכר כי

$$f_{l_2}(w) = \left\|y - X^T \cdot w\right\|^2 + \lambda \cdot \|w\|^2$$

לכן

$$\begin{aligned}
J_w\left(f_{l_2}\right) &= J_w\left(\left\|y - X^T \cdot w\right\|^2 + \lambda \cdot \|w\|^2\right) = J_w\left(\left\|y - X^T \cdot w\right\|^2\right) + J_w\left(\lambda \cdot \|w\|^2\right) \\
&= J_{y - X^T \cdot w}\left(\left\|y - X^T \cdot w\right\|^2\right) \cdot J_w\left(y - X^T \cdot w\right) + 2 \cdot \lambda \cdot w^T \\
&= 2 \cdot \left(y - X^T \cdot w\right)^T \cdot \left[-X^T\right] + 2 \cdot \lambda \cdot w^T \\
&= 2 \cdot \left[w^T X X^T - y^T X^T + \lambda w^T\right] = 2 \cdot \left[X \cdot X^T \cdot w - X \cdot y + \lambda w\right]^T \\
&= 2 \cdot \left[I_n \cdot w - \hat{w}_{LS} + \lambda w\right]^T = 2 \cdot \left[(1+\lambda) \cdot w - \hat{w}^{LS}\right]^T
\end{aligned}$$

לכן אנחנו רוצים ש־ $\nabla\left(f_{l_2}\right) = 0$, לכן

$$\begin{aligned}
0 = \nabla\left(f_{l_2}\right) = J_w\left(f_{l_2}\right)^T &= \left[2 \cdot \left[(1+\lambda) \cdot w - \hat{w}^{LS}\right]^T\right]^T = 2 \cdot \left[(1+\lambda) \cdot w - \hat{w}^{LS}\right] \\
\Rightarrow 2 \cdot (1+\lambda) \cdot w &= 2 \cdot \hat{w}^{LS} \Rightarrow \boxed{w = \frac{1}{1+\lambda} \cdot \hat{w}^{LS}}
\end{aligned}$$

כלומר כל נקודת קיצון $w$ של $f_{l_2}$ חייבת לקיים $w = \frac{1}{1+\lambda} \cdot \hat{w}^{LS}$,

ולכן $\boxed{\hat{w}_\lambda^{ridge} = \frac{1}{1+\lambda} \cdot \hat{w}^{LS}}$ כי הוא נקודת מינימום של $f_{l_2}$ ובפרט נקודת קיצון, כנדרש

מ.ש.ל.א.☺

(ב) **צ"ל:** $\hat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{hard}\left(\hat{w}^{LS}\right)$

**הוכחה:**

נזכר שבמקרה שלנו מתקיים $\hat{w}^{LS} = \left(X \cdot X^T\right)^{-1} \cdot X \cdot y = X \cdot y$, נזכר כי

$$\begin{aligned}
f_{l_0}(w) &= \left\|y - X^T \cdot w\right\|^2 + \lambda \cdot \|w\|_0 = \left\|X^{-1} \cdot \left[X \cdot y - X \cdot X^T \cdot w\right]\right\|^2 + \lambda \cdot \|w\|_0 \\
&= \left(X^{-1} \cdot \left[X \cdot y - X \cdot X^T \cdot w\right]\right)^T \cdot \left(X^{-1} \cdot \left[X \cdot y - X \cdot X^T \cdot w\right]\right) + \lambda \cdot \|w\|_0 \\
&= \left[X \cdot y - X \cdot X^T \cdot w\right]^T \cdot \left[X^{-1}\right]^T X^{-1} \cdot \left[X \cdot y - X \cdot X^T \cdot w\right] + \lambda \cdot \|w\|_0 \\
&= \left[X \cdot y - X \cdot X^T \cdot w\right]^T \cdot \left(X \cdot X^T\right)^{-1} \cdot \left[X \cdot y - X \cdot X^T \cdot w\right] + \lambda \cdot \|w\|_0 \\
&= \left[X \cdot y - X \cdot X^T \cdot w\right]^T \cdot I_n \cdot \left[X \cdot y - X \cdot X^T \cdot w\right] + \lambda \cdot \|w\|_0 \\
&= \left\|X \cdot y - X \cdot X^T \cdot w\right\|^2 + \lambda \cdot \|w\|_0 = \left\|\hat{w}^{LS} - X \cdot X^T \cdot w\right\|^2 + \lambda \cdot \|w\|_0 \\
&= \sum_{i=1}^{|w|} \left(\hat{w_i}^{LS} - w_i\right)^2 + \lambda \cdot 1_{w_i \neq 0}
\end{aligned}$$

נרצה למצוא מזעור ל־ $f_{l_0}$, נשים לב שנוכל למזער כל אינדקס בנפרד.

יהי $i \in \left[\left|\hat{w_i}^{LS}\right|\right]$, נחלק ל־ 2 מקרים:

i. אם $\sqrt{\lambda} > \left|\hat{w}_i{}^{LS}\right|$: אם $\hat{w}_i \neq 0$ אז

$$\left[f_{l_0}\left(\hat{w}\right)\right]_i = \left(\hat{w}_i{}^{LS} - \hat{w}_i\right)^2 + \lambda \cdot 1_{w_i \neq 0} \geq \lambda \cdot 1_{w_i \neq 0} = \lambda$$

ואם נבחר $\hat{w}_i = 0$ נקבל

$$\left[f_{l_0}\left(\hat{w}\right)\right]_i = \left(\hat{w}_i{}^{LS} - \hat{w}_i\right)^2 + \lambda \cdot 1_{w_i \neq 0} = \left(\hat{w}_i{}^{LS} - 0\right)^2 + \lambda \cdot 0 = \left|\hat{w}_i{}^{LS}\right|^2 < \sqrt{\lambda}^2 = \lambda$$

כלומר המינימום מתקבל כאשר $\hat{w}_i = 0$ ונשים לב כי $\hat{w}_i = 0 = \eta^{hard}_{\sqrt{\lambda}}\left(\hat{w}^{LS}\right)$, כנדרש

ii. אחרת $\sqrt{\lambda} \leq \left|\hat{w}_i{}^{LS}\right|$: אם $\hat{w}_i = 0$ אז

$$\left[f_{l_0}\left(\hat{w}\right)\right]_i = \left(\hat{w}_i{}^{LS} - \hat{w}_i\right)^2 + \lambda \cdot 1_{w_i \neq 0} \geq \left(\hat{w}_i{}^{LS} - 0\right)^2 = \left|\hat{w}_i{}^{LS}\right|^2 = \sqrt{\lambda}^2 = \lambda$$

עתה נשים לב שעבור $\hat{w}_i \neq 0$ נקבל

$$\left[f_{l_0}\left(\hat{w}\right)\right]_i = \left(\hat{w}_i{}^{LS} - \hat{w}_i\right)^2 + \lambda \cdot 1_{w_i \neq 0} = \left(\hat{w}_i{}^{LS} - \hat{w}_i\right)^2 + \lambda$$

המינימום יתקבל כאשר $\hat{w}_i = \hat{w}_i{}^{LS}$ ובמקרה זה נקבל

$$\left[f_{l_0}\left(\hat{w}\right)\right]_i = \left(\hat{w}_i{}^{LS} - \hat{w}_i\right)^2 + \lambda \cdot 1_{w_i \neq 0} = \left(\hat{w}_i{}^{LS} - \hat{w}_i{}^{LS}\right)^2 + \lambda = \lambda$$

כלומר המינימום מתקבל כאשר $\hat{w}_i = \hat{w}_i{}^{LS}$ ונשים לב כי $\hat{w}_i = \hat{w}_i{}^{LS} = \eta^{hard}_{\sqrt{\lambda}}\left(\hat{w}^{LS}\right)$, כנדרש

מ.ש.ל.ב.☺

3. **פתרון:**

(א) **צ"ל:** $A_\lambda = \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right)$

**הוכחה:**

ראינו בתרגול שמתקיים $\hat{w}\left(\lambda\right) = \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot X \cdot y$ ולכן

$$\begin{aligned}
\hat{w}\left(\lambda\right) &= \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot X \cdot y \\
&= \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left[\left(X \cdot X^T\right) \cdot \left(X \cdot X^T\right)^{-1}\right] \cdot X \cdot y \\
&= \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right) \cdot \left[\left(X \cdot X^T\right)^{-1} \cdot X \cdot y\right] \\
&= \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right) \cdot \hat{w} = A_\lambda \cdot \hat{w}
\end{aligned}$$

מ.ש.ל.א.☺

(ב) **צ"ל:** $\mathbb{E}\left[\hat{w}\left(\lambda\right)\right]$

**הוכחה:**

ראינו בתרגול שמתקיים

$$\mathbb{E}\left[\hat{w}\left(\lambda\right)\right] = \mathbb{E}\left[A_\lambda \cdot \hat{w}\right] = A_\lambda \cdot \mathbb{E}\left[\hat{w}\right] = A_\lambda \cdot w = \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right) \cdot w$$

נשים לב שעבור $\lambda > 0$ מתקיים

$$\begin{aligned}
& w - \mathbb{E}\left[\hat{w}\left(\lambda\right)\right] = w - \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right) \cdot w \\
& = w \cdot \left(I_n - \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right)\right) \\
& = w \cdot \left(\left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T + \lambda \cdot I_n\right) - \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right)\right) \\
& = w \cdot \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(\left(X \cdot X^T + \lambda \cdot I_n\right) - \left(X \cdot X^T\right)\right) \\
& = w \cdot \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(\lambda \cdot I_n\right) \neq 0 \\
& \Rightarrow \boxed{w \neq \mathbb{E}\left[\hat{w}\left(\lambda\right)\right]}
\end{aligned}$$

כלומר קיבלנו כי $\boxed{\mathbb{E}\left[\hat{w}\left(\lambda\right)\right] = \left(X \cdot X^T + \lambda \cdot I_n\right)^{-1} \cdot \left(X \cdot X^T\right) w}$ ולכל $\lambda > 0$ מתקיים $\mathbb{E}\left[\hat{w}\left(\lambda\right)\right] \neq w$, כנדרש

מ.ש.ל.ב.☺

(ג) **צ"ל:** $Var\left[\hat{w}\left(\lambda\right)\right]$

**הוכחה:**

$$\begin{aligned}
Var\left[\hat{w}\left(\lambda\right)\right] = Var\left[A_\lambda \cdot \hat{w}\right] & \overset{A_\lambda \text{ constant non random matrix}}{=} A_\lambda \cdot Var\left[\hat{w}\right] \cdot A_\lambda^T \\
& = A_\lambda \cdot \sigma^2 \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T = \sigma^2 \cdot A_\lambda \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T
\end{aligned}$$

כלומר קיבלנו כי

$$\boxed{Var\left[\hat{w}\left(\lambda\right)\right] = \sigma^2 \cdot A_\lambda \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T}$$

מ.ש.ל.ג.☺

(ד) **צ"ל:** $MSE\left(\lambda\right)$

**הוכחה:**

בהרצאה ראינו ש־

$$MSE\left(\hat{y}\right) = \left\|y - \hat{y}\right\|^2 = \left\|y - \mathbb{E}\left[\hat{y}\right]\right\|^2 + \sum_{i=1}^{|y|} Var\left(\hat{y}_i\right)$$

נשים לב ש־ $Var\left(\hat{w}\left(\lambda\right)\right)$ זאת מטריצת וריאס ולכן $Var\left(\hat{y}_i\right)$ במקרה שלנו הוא באלכסון במקום $i,i$ כלומר ב־ $\left[Var\left(\hat{w}\left(\lambda\right)\right)\right]_{i,i}$, נציב ונקבל

$$\begin{aligned}
MSE\left(\hat{w}\left(\lambda\right)\right) & = \left\|w - \mathbb{E}\left[\hat{w}\left(\lambda\right)\right]\right\|^2 + \sum_{i=1}^{|w|} \left[Var\left(\hat{w}\left(\lambda\right)\right)\right]_{i,i} \\
& = \left\|w - A_\lambda \cdot w\right\|^2 + Tr\left(Var\left(\hat{w}\left(\lambda\right)\right)\right) \\
& = \left\|\left(I_n - A_\lambda\right) \cdot w\right\|^2 + Tr\left(\sigma^2 \cdot A_\lambda \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T\right) \\
& = \left[\left(I_n - A_\lambda\right) \cdot w\right]^T \cdot \left[\left(I_n - A_\lambda\right) \cdot w\right] + Tr\left(\sigma^2 \cdot A_\lambda \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T\right) \\
& = w^T \cdot \left(I_n - A_\lambda\right)^T \cdot \left(I_n - A_\lambda\right) \cdot w + Tr\left(\sigma^2 \cdot A_\lambda \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T\right) \\
& = w^T \cdot \left[I_n - A_\lambda^T - A_\lambda + A_\lambda^T \cdot A_\lambda\right] \cdot w + Tr\left(\sigma^2 \cdot A_\lambda \cdot \left(X \cdot X^T\right)^{-1} \cdot A_\lambda^T\right)
\end{aligned}$$

עתה נגזור

$$\frac{\partial A_\lambda}{\partial\lambda}=\frac{\partial\left(X\cdot X^T+\lambda\cdot I_n\right)^{-1}\cdot\left(X\cdot X^T\right)}{\partial\lambda}=\frac{\partial\left(X\cdot X^T+\lambda\cdot I_n\right)^{-1}}{\partial\lambda}\cdot\left(X\cdot X^T\right)$$
$$=-1\cdot\left(X\cdot X^T+\lambda\cdot I_n\right)^{-2}\cdot\frac{\partial\left(X\cdot X^T+\lambda\cdot I_n\right)}{\partial\lambda}\cdot\left(X\cdot X^T\right)$$
$$=-1\cdot\left(X\cdot X^T+\lambda\cdot I_n\right)^{-2}\cdot I_n\cdot\left(X\cdot X^T\right)$$
$$=-\left(X\cdot X^T+\lambda\cdot I_n\right)^{-2}\cdot\left(X\cdot X^T\right)$$
$$=-\left(X\cdot X^T+\lambda\cdot I_n\right)^{-1}\cdot\left(X\cdot X^T+\lambda\cdot I_n\right)^{-1}\cdot\left(X\cdot X^T\right)$$
$$=-\left(X\cdot X^T+\lambda\cdot I_n\right)^{-1}\cdot A_\lambda$$

נשים לב כי $A_0=\left(X\cdot X^T+0\cdot I_n\right)^{-1}\cdot\left(X\cdot X^T\right)=\left(X\cdot X^T\right)^{-1}\cdot\left(X\cdot X^T\right)=I_n$, נציב $\lambda=0$ ונקבל כי

$$\frac{\partial A_\lambda}{\partial\lambda}\mid_{\lambda=0}=-\left(X\cdot X^T+0\cdot I_n\right)^{-1}\cdot A_0=-\left(X\cdot X^T\right)^{-1}\cdot I_n=-\left(X\cdot X^T\right)^{-1}$$

עתה נשתמש בנגזרת הזאת כדי לגזור את הביטויים שקיבלנו

$$\frac{\partial Tr\left(\sigma^2\cdot A_\lambda\cdot\left(X\cdot X^T\right)^{-1}\cdot A_\lambda^T\right)}{\partial\lambda}\mid_{\lambda=0}=Tr\left(\frac{\partial\sigma^2\cdot A_\lambda\cdot\left(X\cdot X^T\right)^{-1}\cdot A_\lambda^T}{\partial\lambda}\right)\mid_{\lambda=0}$$
$$=Tr\left[\sigma^2\cdot\frac{\partial A_\lambda\cdot\left(X\cdot X^T\right)^{-1}\cdot A_\lambda^T}{\partial\lambda}\right]\mid_{\lambda=0}=\sigma^2\cdot Tr\left[\frac{\partial A_\lambda\cdot\left(X\cdot X^T\right)^{-1}\cdot A_\lambda^T}{\partial\lambda}\right]\mid_{\lambda=0}$$
$$=\sigma^2\cdot Tr\left[\frac{\partial A_\lambda}{\partial\lambda}\cdot\left(X\cdot X^T\right)^{-1}\cdot A_\lambda^T+A_\lambda\cdot\left(X\cdot X^T\right)^{-1}\cdot\frac{\partial A_\lambda^T}{\partial\lambda}\right]\mid_{\lambda=0}$$
$$=\sigma^2\cdot Tr\left[-\left(X\cdot X^T\right)^{-1}\cdot\left(X\cdot X^T\right)^{-1}\cdot A_0^T+A_0\cdot\left(X\cdot X^T\right)^{-1}\cdot-\left(X\cdot X^T\right)^{-1}\right]$$
$$=\sigma^2\cdot Tr\left[-\left(X\cdot X^T\right)^{-2}-\left(X\cdot X^T\right)^{-2}\right]=-2\cdot\sigma^2\cdot Tr\left[\left(X\cdot X^T\right)^{-2}\right]$$

נשים לב כי $\left(X\cdot X^T\right)^{-2}$ היא מטריצה סימטרית ולכן הטרייס שלה אי שלילי והייותה הפיכה הוא לא 0, כלומר נקבל כי

$$\frac{\partial Tr\left(\sigma^2\cdot A_\lambda\cdot\left(X\cdot X^T\right)^{-1}\cdot A_\lambda^T\right)}{\partial\lambda}\mid_{\lambda=0}<0$$

עתה נגזור את החלק השני של הביטוי

$$\frac{\partial w^T\cdot\left[I_n-A_\lambda^T-A_\lambda+A_\lambda^T\cdot A_\lambda\right]\cdot w}{\partial\lambda}\mid_{\lambda=0}=w^T\cdot\frac{\partial\left[I_n-A_\lambda^T-A_\lambda+A_\lambda^T\cdot A_\lambda\right]}{\partial\lambda}\cdot w\mid_{\lambda=0}$$
$$=w^T\cdot\left[\frac{\partial I_n}{\partial\lambda}-\frac{\partial A_\lambda}{\partial\lambda}-\frac{\partial A_\lambda^T}{\partial\lambda}+\frac{\partial A_\lambda^T\cdot A_\lambda}{\partial\lambda}\right]\cdot w\mid_{\lambda=0}$$
$$=w^T\cdot\left[\frac{\partial I_n}{\partial\lambda}-\frac{\partial A_\lambda}{\partial\lambda}-\frac{\partial A_\lambda^T}{\partial\lambda}+\frac{\partial A_\lambda^T}{\partial\lambda}\cdot A_\lambda+A_\lambda^T\cdot\frac{\partial A_\lambda}{\partial\lambda}\right]\cdot w\mid_{\lambda=0}$$
$$=w^T\cdot\left[0-\left(-\left(X\cdot X^T\right)^{-1}\right)-\left(-\left(X\cdot X^T\right)^{-1}\right)^T+\left(-\left(X\cdot X^T\right)^{-1}\right)^T\cdot A_0+A_0^T\cdot\left(-\left(X\cdot X^T\right)^{-1}\right)\right]\cdot w$$
$$=w^T\cdot\left[0+\left(X\cdot X^T\right)^{-1}+\left[\left(X\cdot X^T\right)^{-1}\right]^T-\left[\left(X\cdot X^T\right)^{-1}\right]^T\cdot I_n-I_n\cdot\left(X\cdot X^T\right)^{-1}\right]\cdot w$$
$$=w^T\cdot[0]\cdot w=0$$

כלומר קיבלנו כי

$$\frac{\partial MSE\left(\hat{w}\left(\lambda\right)\right)}{d\lambda}\mid_{\lambda=0}=\frac{\partial\left[w^{T}\cdot\left[I_{n}-A_{\lambda}^{T}-A_{\lambda}+A_{\lambda}^{T}\cdot A_{\lambda}\right]\cdot w+Tr\left(\sigma^{2}\cdot A_{\lambda}\cdot\left(X\cdot X^{T}\right)^{-1}\cdot A_{\lambda}^{T}\right)\right]}{\partial\lambda}\mid_{\lambda=0}$$

$$=\frac{\partial\left[w^{T}\cdot\left[I_{n}-A_{\lambda}^{T}-A_{\lambda}+A_{\lambda}^{T}\cdot A_{\lambda}\right]\cdot w\right]}{\partial\lambda}\mid_{\lambda=0}+\frac{\partial\left[Tr\left(\sigma^{2}\cdot A_{\lambda}\cdot\left(X\cdot X^{T}\right)^{-1}\cdot A_{\lambda}^{T}\right)\right]}{\partial\lambda}\mid_{\lambda=0}$$

$$=0+\frac{\partial\left[Tr\left(\sigma^{2}\cdot A_{\lambda}\cdot\left(X\cdot X^{T}\right)^{-1}\cdot A_{\lambda}^{T}\right)\right]}{\partial\lambda}\mid_{\lambda=0}=\frac{\partial\left[Tr\left(\sigma^{2}\cdot A_{\lambda}\cdot\left(X\cdot X^{T}\right)^{-1}\cdot A_{\lambda}^{T}\right)\right]}{\partial\lambda}\mid_{\lambda=0}<0$$

מ.ש.ל.ד.☺

(ה) **צ"ל:** לאמבדה ישפר טיפה אם המודל צודק

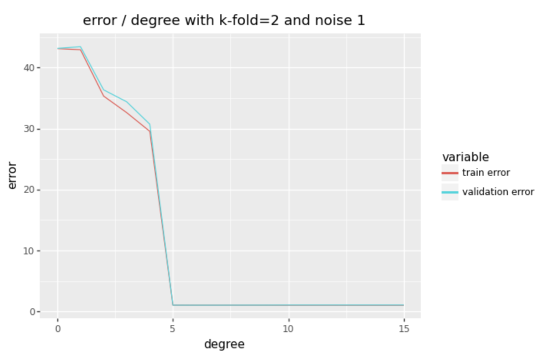
**הוכחה:**

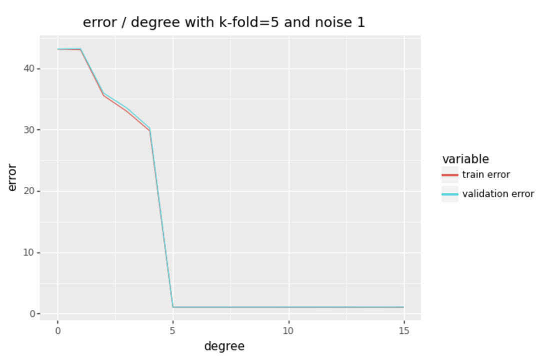
נשים לב שהנגזרת של ה־ $MSE$ שלילי עבור $\lambda=0$, כלומר קיימת $[0,a]$ כאשר $a>0$ שבה פונקצית השגיאה יורדת, ולכל $\lambda\in(0,a)$ יתקיים $MSE\left(\hat{w}\left(\lambda\right)\right)<MSE\left(\hat{w}\left(0\right)\right)$, כלומר טיפה רגולריזציה רק תשפר את המודל.

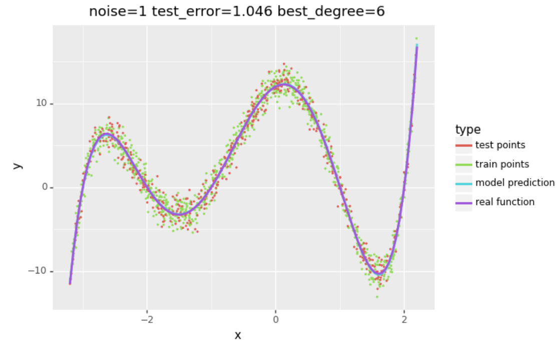
מ.ש.ל.ה.☺

4. **צ"ל:** גרפים ומסקנות

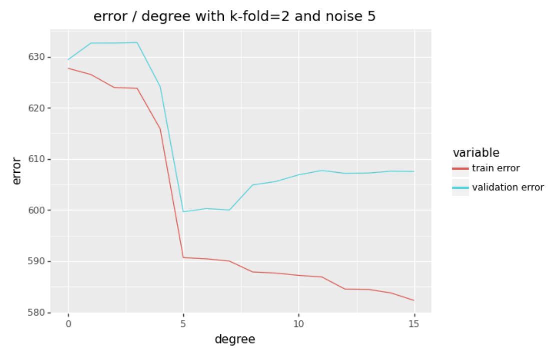
**הוכחה:**

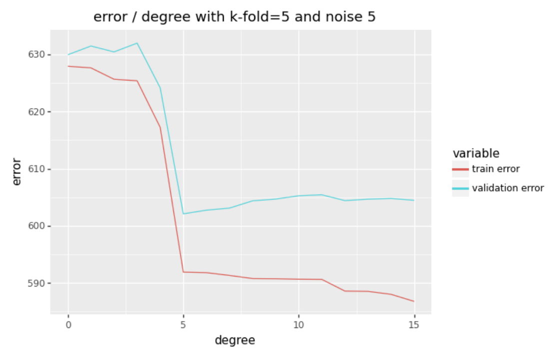


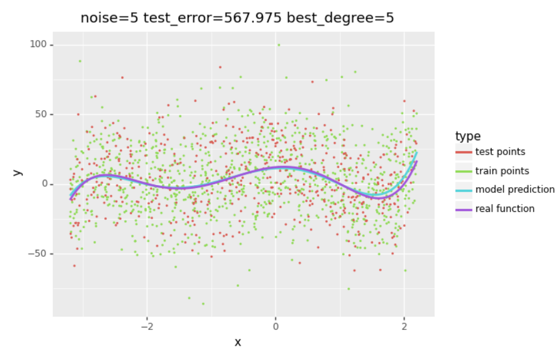






אפשר לשים לב שעם טיפה רעש, הוא מצליח ללמוד את הפולינום הנכון גם לולידציה וגם לאימון ואף מהדרגה הכמעט נכונה עם שגיאה אפסית, אפשר לראות כמה הפולינומים דומים בגרף השלישי שצורף.



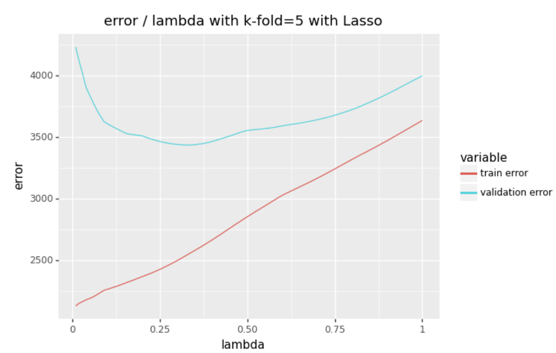



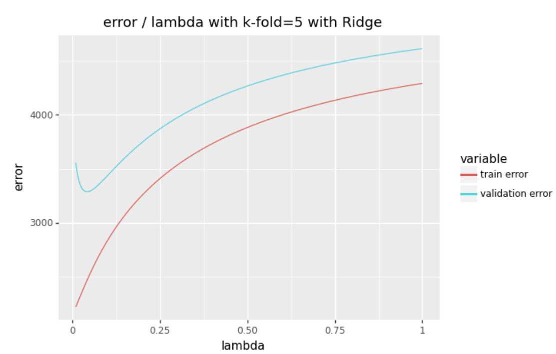



לאחר שהוספנו הרבה רעש, אפשר לראות שעבור דרגות פולינום גבוהות, הוא מתחיל ללמוד את הרעש והשגיאה על ה־ $valdation$ עולה (כי הוא מאבד את ה־ $generalization$) בזמן שהאימון יורד כי יש לו מחלקת היפותזות יותר מורכבת לתאר את הדגימות.
באותו הזמן אפשר לראות את הגרף השלישי המצורף ולראות שעל אף שהיה רעש מטורף כמו שאפשר לראות, הוא עדיין הצליח ללמוד פולינום מאוד קרוב לפולינום המקורי עם הדרגה האמיתית של הפולינום.
יחסית לטיפה רעש, השגיאה גדולה מאוד, אבל ביחס למידע הניתן, זה שגיאה מדהימה.

מ.ש.ל.☺

5. **צ"ל:** גרפים ומסקנות
**הוכחה:**




| | best lasso error | best ridge error | best no regularization error |
|---|---|---|---|
| 0 | 3320.250889 | 3265.039781 | 3490.007801 |

תחילה אפשר לראות שגם הרידג' וגם הלאסו שיפרו את השגיאה על פני רגרסיה ללא רגולריזציה והאחד הכי טוב הוא ה־ $ridge$, רוב הסיכויים בגלל שכל הפרמטרים חשובים ולא נרצה את התכונה של $lasso$ שמאפסת פרמטרים.
אני בהתחלה הסתכלתי על פרמטרים $\lambda$ בקטע $[1, 500]$ בקפיצות של 10, ושמתי לב שהשגיאה הכי קטנה על פני ה־ $validation$ היא דווקא בהתחלה בהתחלה, ולאחר חיפוש מספר פעמים, הבחנתי שהשגיאה הכי טובה על פני ה־ $validation$ היא בקטע $[0, 1]$, אז החלטתי לבדוק בקטע $[0.01, 1]$ עם $10000$ דגימות בדרך ומצאתי את המינימום שחיפשתי מלכתחילה.
אפשר לראות שלשניהם המינימום הוא עם $\lambda$ נמוך, שזה הגיוני כי לפי מה שהוכחנו בשאלה $3.4$, כשהמודל יכול לתאר את המחלקה עד כדי רעש, אז טיפה רגולרזציה יכולה רק לעזור.

מ.ש.ל.☺