תרגיל 4

שם: תומר גבאי
ת.ז: 315315523

4) 1)

$$\forall \varepsilon,\delta>0 \quad \exists m(\varepsilon,\delta) \text{ s.t. } \forall m \geq m(\varepsilon,\delta) \quad P_{S\sim D^m}[L_D(A(S))\leq\varepsilon] \geq 1-\delta \qquad (a)$$

$$\lim_{m\to\infty} E_{S\sim D^m}[L_D(A(S))]=0 \qquad (b)$$

כיוון ראשון $(a)\Rightarrow(b)$

יהי: $\varepsilon,\delta>0$ לפי (a) קיים $m(\varepsilon,\delta)$ כך שלכל $m\geq m(\varepsilon,\delta)$ מתקיים כי: $P_{S\sim D^m}[L_D(A(S))\leq\varepsilon]\geq 1-\delta$

ולכן גם לפי המשלים $P_{S\sim D^m}[L_D(A(S))\geq\varepsilon]\leq\delta$ כעת נבחר $\varepsilon'=\frac{\varepsilon}{2}$, $\delta=\frac{\varepsilon}{2}$, ולכל $m\geq m(\frac{\varepsilon}{2},\frac{\varepsilon}{2})$ נקבל:

$$P_{S\sim D^m}\left(L_D(A(S))\geq\frac{\varepsilon}{2}\right)\leq\frac{\varepsilon}{2}$$

כעת לכל m גדול מספיק:

תכונות אינטגרל +
הסתברות נשארת בערך נמוך

$$E_{S\sim D^m}[L_D(A(S))]=\int_0^1 t\,P(L_D(A(S))=t)\,dt \leq \int_0^{\frac{\varepsilon}{2}}\frac{\varepsilon}{2}P(L_D(A(S))=t)\,dt+\int_{\frac{\varepsilon}{2}}^1 P[L_D(A(S))=t]\,dt$$

$m\geq m(\frac{\varepsilon}{2},\frac{\varepsilon}{2})$

$$\leq \int_0^{\frac{\varepsilon}{2}}\frac{\varepsilon}{2}P(L_D(A(S))=t)\,dt+\frac{\varepsilon}{2}=\frac{\varepsilon}{2}P\left(L_D(A(S))\leq\frac{\varepsilon}{2}\right)+\frac{\varepsilon}{2}\leq\frac{\varepsilon}{2}\cdot 1+\frac{\varepsilon}{2}=\varepsilon$$

קיבלנו ש $E_{S\sim D^m}[L_D(A(S))]\leq\varepsilon$ לכל $\varepsilon>0$ ולכן קיבלנו את הגבול

כיוון שני $(b)\Rightarrow(a)$

נניח כי $\lim_{m\to\infty}E_{S\sim D^m}[L_D(A(S))]=0$

יהי $\varepsilon,\delta>0$ מהגבול ידוע שקיים $m\in\mathbb{N}$ עבורו

$$E_{S\sim D^m}[L_D(A(S))]<\varepsilon\delta$$

א"ש מרקוב

$$\Rightarrow \quad P_{S\sim D^m}[L_D(A(S))\geq\varepsilon]\leq\frac{E_{S\sim D^m}[L_D(A(S))]}{\varepsilon}<\frac{\cancel{\varepsilon}\delta}{\cancel{\varepsilon}}=\delta$$

ולכן סה"כ נקבל ש

$$P_{S\sim D^m}[L_D(A(S))\leq\varepsilon]\geq 1-\delta$$

כנדרש ∎

2) נראה כי $h_r(x)=\mathbb{1}_{[\|x\|_2\le r]}$ $\mathcal{H}=\{h_r : r\in\mathbb{R}_+\}$ למידה פא״ק

והקומפלקסיטי מקיים $m_H(\epsilon,\delta)\le\frac{\log(\frac{1}{\delta})}{\epsilon}$

נגדיר אלגו A שבהינתן $S=\{(x_i,y_i)\}_{i=1}^{m}$ יחזיר את h_r הקטן ביותר

$$r=\max_{x\in S}\|x\|_2$$

וגם h^* בעל סיכון $=0$, נסמן את הרדיוס שלה r^*

כעת נשים לב כי $L_D(h_r)=D\left(x : r<\|x\|_2\le r^*\right)$

כעת יהי δ,ϵ ניקח X בעל m דגימות עבורו מתקיים

$$\mathbb{P}_x\left(x : r<\|x\|_2\le r^*\right)\le\epsilon$$

עבור X לא מתקיים שההסתברות שהסיכון יהיה גבוה מ ϵ חסומה מלמעלה ע״י $(1-\epsilon)^m$ שכן כך בחרנו
את m הדגימות שלא אף אחת מהן בתוך הטווח בחירה

$$\mathbb{P}_{X\sim D^m}\left[L_D(h_r)\ge\epsilon\right]\le(1-\epsilon)^m\le e^{-m\epsilon}\le\delta$$

$$\Updownarrow$$

$$m\le\frac{\log(\frac{1}{\delta})}{\epsilon}$$

לכן קיבלנו כי $m_H(\epsilon,\delta)\le\frac{\log(\frac{1}{\delta})}{\epsilon}$ כנדרש $\blacksquare$

3) H בעלת תכונת UC עם $m_H^{UC}:(0,1)^2 \longrightarrow \mathbb{N}$

3) צ"ל H אגנוסטית פא"ק

נראה כי $D^m(\{S\in Z^m \mid S \text{ }\varepsilon\text{-מייצג}\}) \geq 1-\delta$ $\forall \varepsilon,\delta\in(0,1), D, m\geq m_A^{UC}(\varepsilon,\delta)$

יהי $\varepsilon,\delta\in(0,1)$ יהי $m_H^{UC}(\frac{\varepsilon}{2},\delta)$ כאשר לכל $m\geq m_H^{UC}(\frac{\varepsilon}{2},\delta)$ מתקיים

$$D^m(\{S \mid \forall h\in H \ |L_{S^m}(h)-risk(h)| < \tfrac{\varepsilon}{2}\}) \geq 1-\delta$$

נסמן $h_S=A(S)$ כאשר A הוא האלג'

בנוסף מתקיים לכל $h\in H$

$$L_D(h_S) \leq L_D(h)+\frac{\varepsilon}{2}$$

כעת

$$L_D(h_S) \leq L_S(h_S)+\frac{\varepsilon}{2} \leq L_D(h_S)+\frac{\varepsilon}{2} \leq L_D(h)+\frac{\varepsilon}{2}+\frac{\varepsilon}{2} = L_D(h)+\varepsilon$$

לפי S_m

הוא m מייצג

כי $m>m(\frac{\varepsilon}{2},\delta)$

כלומר קיבלנו

$$L_D(h_S) \leq \min_{h\in H} L_D(h)+\varepsilon$$

לפי הנ"ל לכל נכון לכל $h\in H$.

כלומר קיבלנו

$$P_{D^m}\left(L_D(h_S) \leq \min_{h\in H} L_D(h)+\varepsilon\right) \geq 1-\delta$$

כלומר H למידה אגנוסטית פא"ק, ומתקיים $m_H(\varepsilon,\delta) \leq m^{UC}(\frac{\varepsilon}{2},\delta)$

4) $h_I(x) = \left(\sum_{i\in I} x_i\right) \bmod 2$

נראה שמימד ה VC הוא n.
ראשית נראה קבוצה מנותצת בגודל n. יהי $C=(e_1,\dots,e_n)$ עבור הסטנדרטיים $y = \{0,1\}^n$
כאשר לכל $i\in[n]$ $h_I(e_i)=y_i$
ונשים ש y_i יכול להיות $\{0,1\}$ ויש n איברים ולכן כי $|H_C| = 2^{|C|}$ כלומר C מנותצת ע"י H
ולכן $VC\text{-}Dim(H) \geq n$
ונשים שיש לכל היותר 2^n אפשרויות שונות לפונקציות ב H ולכן ש $|H| = 2^n$ כלומר
כי קבוצה מנותצת צריכה להיות לכל היותר בגודל $\log_2|H|$ כלומר $n \geq VC\text{-}Dim$ ולכן

$$VC\text{-}Dim(H) = n$$

כנדרש $\blacksquare$

ב) נתון $H_1 \subseteq H_2$ צ"ל כי $VC\text{-}Dim(H_1) \le VC\text{-}Dim(H_2)$

תהי $C=(c_1,\dots,c_k)$ קבוצה המנותצת ע"י H_2. כלומר C מקיימת כל תיוג אפשרי ב H_2
נשים לב ש $H_1 \subseteq H_2$ של כפרה קבוצה בגודל k שקיים תיוג שלא ב H_1 ולכן קיים לכל תיוג של C בגודל k שמנותצת ע"י H_2

ולכן

$$VC\text{-}Dim(H_1 \le VC\text{-}Dim(H_2)$$

כנדרש $\blacksquare$

2.3 רגרסיית רידג'

1) א.

$$\hat{w}_\lambda = V\Sigma_\lambda U^T y$$
$$\hat{w} = [X^TX]^{-1}X^Ty$$

$$
\begin{aligned}
A_\lambda \hat{w} &= (X^TX+\lambda I_d)^{-1}(X^TX)\hat{w}\\
&= (V\Sigma^T U^T U\Sigma V^T + V\lambda I_d V^T)^{-1}(X^TX)\hat{w}\\
&= V(\Sigma^T\Sigma+\lambda I)^{-1}V^T \, \cancel{(X^TX)}\cancel{(X^TX)^{-1}}X^Ty\\
&= V(\Sigma^T\Sigma+\lambda I)^{-1}\cancel{V^T}\cancel{V}\Sigma^T U^T y\\
&= V\Sigma_\lambda U^T y = \hat{w}_\lambda
\end{aligned}
$$

ב) יהי $\lambda>0$

$$
\begin{aligned}
\mathbb{E}[\hat{w}_\lambda] &= \mathbb{E}(V\Sigma_\lambda U^T y) = \mathbb{E}(V\Sigma_\lambda U^T(Xw+\varepsilon))\\
&\overset{\text{לינאריות}}{=} \mathbb{E}(V\Sigma_\lambda U^T U\Sigma V^T w + V\Sigma_\lambda U^T\varepsilon)\\
&= \mathbb{E}(V\Sigma_\lambda\Sigma V^T w) + \underbrace{V\Sigma_\lambda U^T\mathbb{E}(\varepsilon)}_{0}\\
&= V\Sigma_\lambda\Sigma V^T\mathbb{E}(w)\\
&= V\Sigma_\lambda\Sigma V^T w \neq w
\end{aligned}
$$

$$[\Sigma_\lambda\Sigma]_{ii} = \frac{\sigma_i^2}{\sigma_i^2+\lambda} \neq 1$$

ג)

$$
\begin{aligned}
Var(\hat{w}_\lambda) &\overset{\text{לינאריות}}{=} Var(\underbrace{A_\lambda}_{B}\underbrace{\hat{w}}_{Z}) = A_\lambda Var(\hat{w})A_\lambda^T = A_\lambda\sigma^2(X^TX)^{-1}A_\lambda^T\\
&= \sigma^2 A_\lambda(X^TX)^{-1}A_\lambda^T
\end{aligned}
$$

ד)

$$
\begin{aligned}
MSE(\hat{y}) &= \mathbb{E}((\hat{y}-y)^T(\hat{y}-y))\\
&= \mathbb{E}((\hat{y}-\mathbb{E}(\hat{y})+\mathbb{E}(\hat{y})-y)^T(\hat{y}-\mathbb{E}(\hat{y})+\mathbb{E}(\hat{y})-y))\\
&= \mathbb{E}((\hat{y}-\mathbb{E}(\hat{y}))^T(\hat{y}-\mathbb{E}(\hat{y})) + (\hat{y}-\mathbb{E}(\hat{y}))^T(\mathbb{E}(\hat{y})-y) + (\mathbb{E}(\hat{y})-y)^T(\hat{y}-\mathbb{E}(\hat{y})) + (\mathbb{E}(\hat{y})-y)^T(\mathbb{E}(\hat{y})-y))\\
&= \mathbb{E}((\hat{y}-\mathbb{E}(\hat{y}))^T(\hat{y}-\mathbb{E}(\hat{y}))) + \mathbb{E}((\hat{y}-\mathbb{E}(\hat{y}))^T(\mathbb{E}(\hat{y})-y)) + \mathbb{E}((\mathbb{E}(\hat{y})-y)^T(\hat{y}-\mathbb{E}(\hat{y}))) + \mathbb{E}((\mathbb{E}(\hat{y})-y)^T(\mathbb{E}(\hat{y})-y))
\end{aligned}
$$

$(\mathbb{E}(\hat{y})-\hat{y})=0$

$$
\begin{aligned}
&= \mathbb{E}[(\hat{y}-\mathbb{E}(\hat{y}))^T(\hat{y}-\mathbb{E}(\hat{y}))] + (\mathbb{E}(\hat{y})-y)^T(\mathbb{E}(\hat{y})-y) = \mathbb{E}\left(\sum_i(\hat{y}_i-\mathbb{E}(\hat{y}_i))^2\right) + Bias(\hat{y})^T Bias(\hat{y})\\
&= \sum_i \mathbb{E}((\hat{y}_i-\mathbb{E}(\hat{y}_i))^2) + \left(\sqrt{Bias(\hat{y})^T Bias(\hat{y})}\right)^2 = \sum_i Var(\hat{y}_i) + (\|Bias(\hat{y})\|^2) =\\
&= tr(Var(\hat{y})) + \|Bias\|^2
\end{aligned}
$$

ה) את הנגזרת שלפי λ של, אז נתרחק קצת מ $\lambda=0$ הפונקציה תרד כלומר MSE יקטן
ולכן אם λ יגדל יותר מידי כבר נצטרך לחשב את הנגזרת בנקודה אחרת ולא לאינטואיציה
למען הדיוק.

חלק 3.1

1. נשים לב שכאשר הערך קטן: הלוס מאוד גבוה וקופץ מאוד כי לא על בסוף יש יותר השפעה מאשר כאשר אנו מגדילים את גודל הלמבדה. אנו מקבלים ירידה משמעותית בלוס והלמבדה מתכנסת לערך קרוב ל 0.

2. כי אנו שוב רואים שככל שגודל הערך גדל. אנו מצליחים לזהות את הקבוצות של תחום הנתונים שלנו בצורה טובה יותר ובכך מצליחים לתפוס יותר תבניות ולקבל אותם נכון

3. התוצאה הטובה ביותר היתה של 238 חברים בערך

4. כתוצאה אנו לוקחים את כל ה[illegible] מדגמים שהופיעו בסט האימון, מסננים בכבד היום
קירלננו [illegible] נכון, ובנוסף מגדילים את הגודל של הסט בהתאם למשקל של הדגימה.
אבל בוסט נותן יותר משקל לדוגמאות שטעו בהם ולכן ניתן לראות שהדגימות שלא נחזו בתחילה
ובכך המודל שלמד קודם משקל גבוה יותר

5. בגרף בראשון אנחנו רואים שאנו מצליחים ללמוד יותר טוב את האימון כלומר הביאס יורד
אך הלוס גדל משמעותית עבור סט המבחן שכן אנו רוצים אחרי האפוק, כלומר ה Variance גדל.

בגרף 4 ניתן לראות שאין לנו צורה מוגדרת שבה אנו מקבלים ובגלל שהרעש מאוד גבוה
אנחנו לא מצליחים ללמוד את הסיגנל שמאחורי הנתונים. המשקל נראה מפוזר מאוד בין כל הדגימות
שכן כל מודל "הורס" את התוצאות של המודל הקודם בערך.

הגרפים בסוף

חלק 3.2

ב) מצאנו שהטווח 0.01 - 0.00001 נמצא יותר רגיש! 0.1-2 יותר למודל
שכן אנו יכולים לראות ירידה בטעות לפני שהיא מגיעה לנקודת הפסקה

ניתן לראות בגרפים שאם שהטעות על האימון נמצאת יותר מהמבחן
ובנוסף הטעות על המבחן יורדת מעט כשמוסיפים את $\lambda > 0$ ואז גדלה מחדש λ

3) התוצאות הן

$\lambda_{ridge} = 0.0006$

$\lambda_{lasso} = 0.5962$

ולטעויות של כל אחד מהמודלים

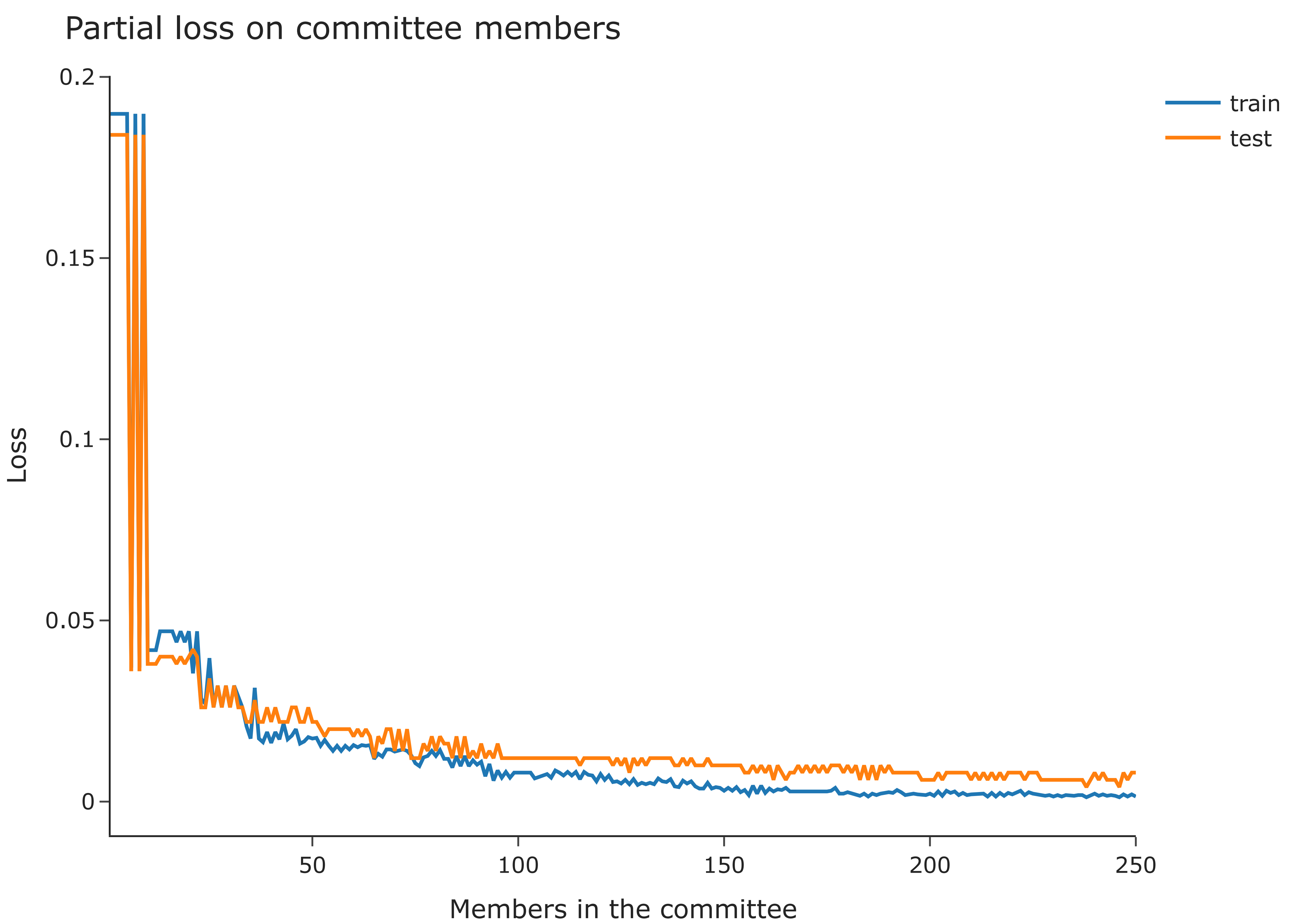
$L.S = 3614.2496$

$Ridge = 3214.3397$

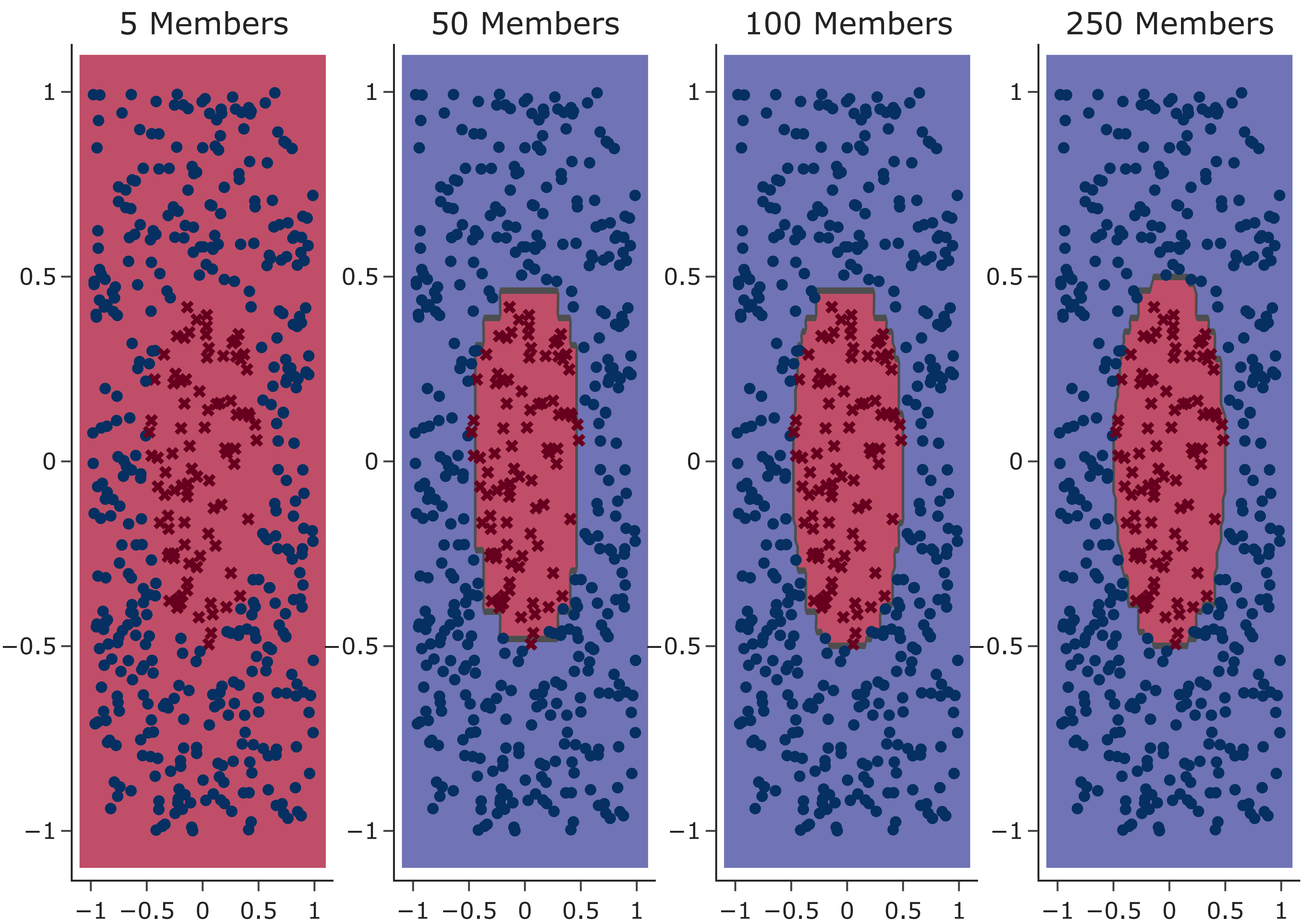
$Lasso = 3640.4921$

קיבלנו שRidge הביא למודל את הגבול בצורה הטובה ביותר והשיג את הטעות הנמוכה ביותר.

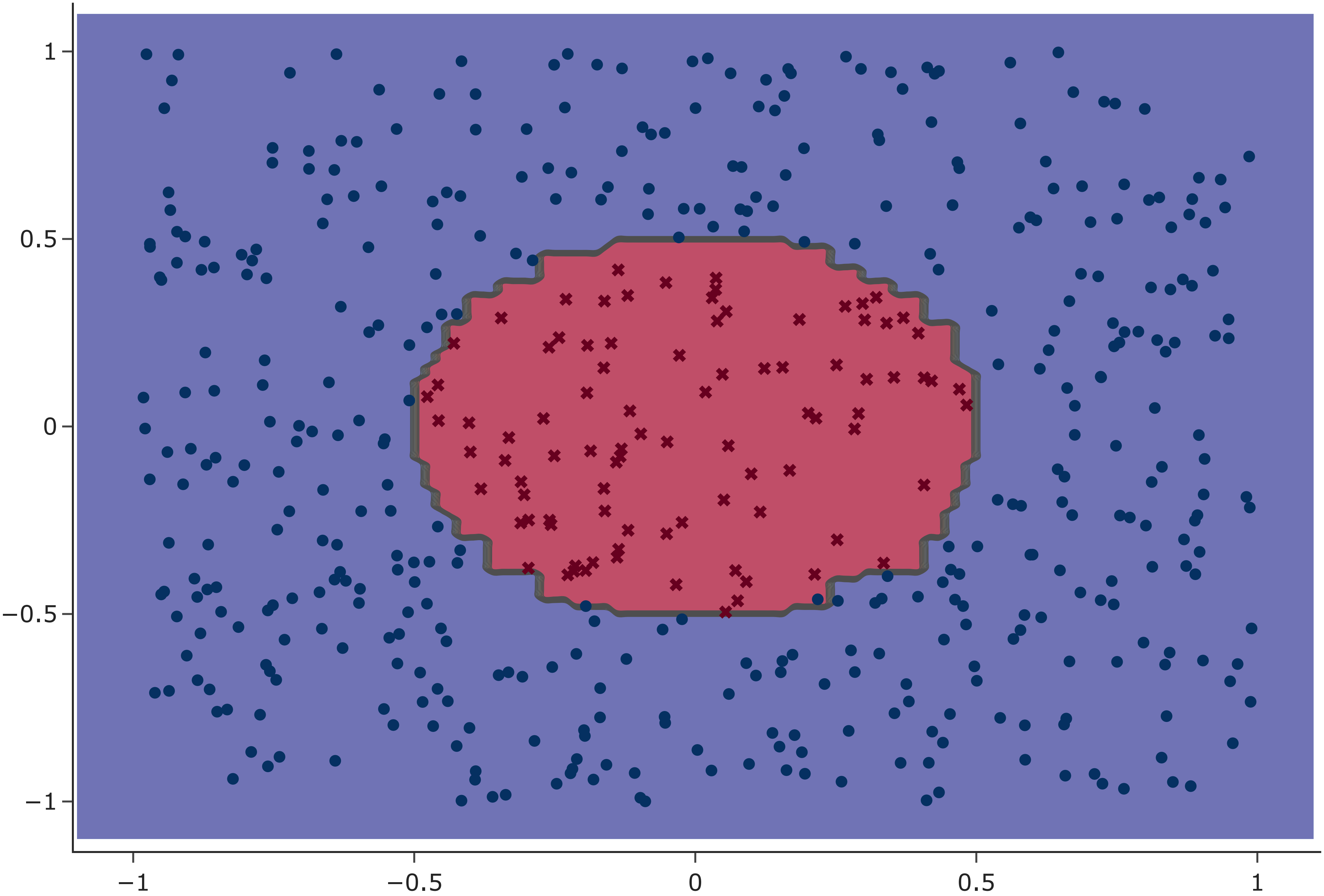
הגרפים בסוף



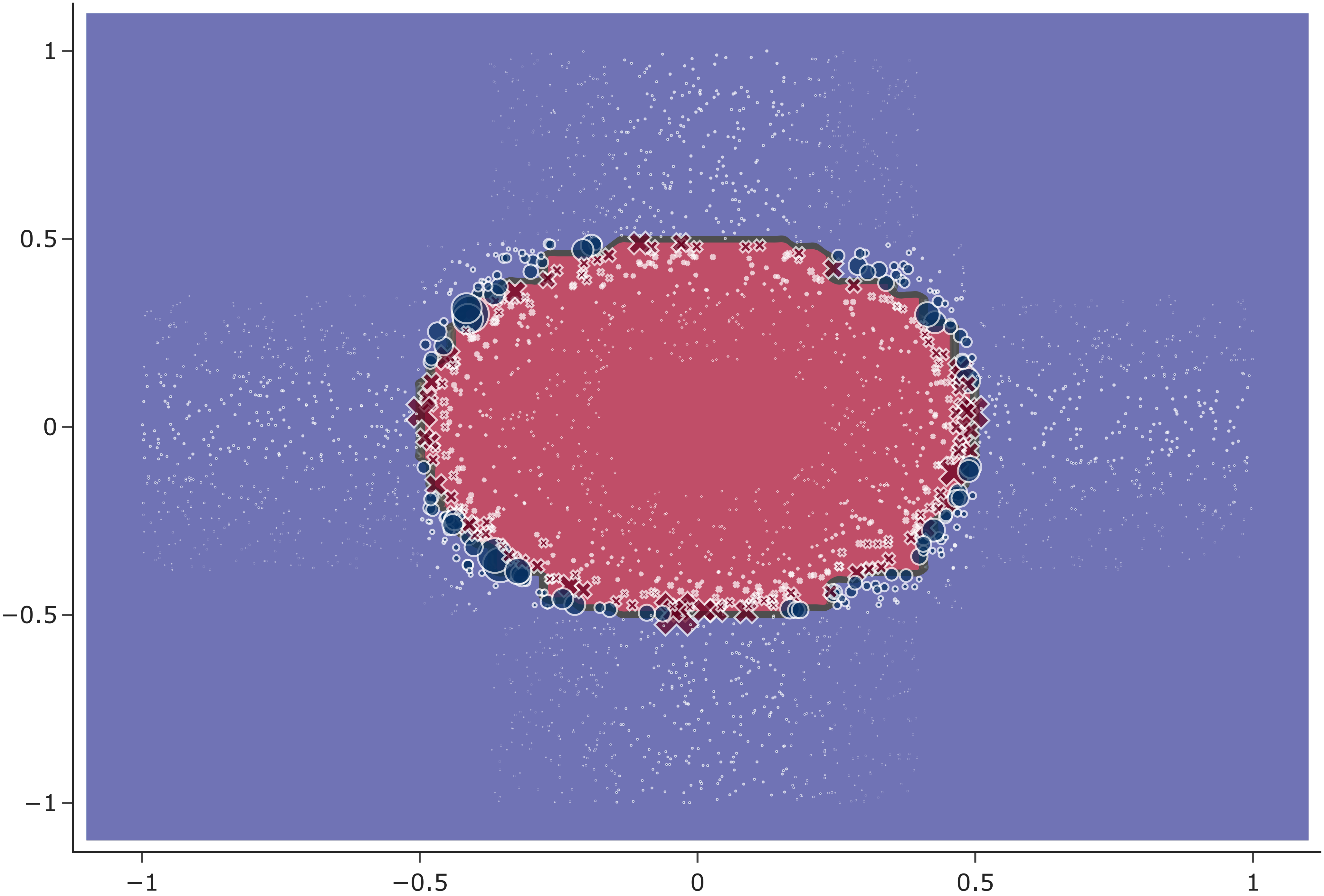
Partial loss on committee members
train
test
Loss
0.2
0.15
0.1
0.05
0
50
100
150
200
250
Members in the committee

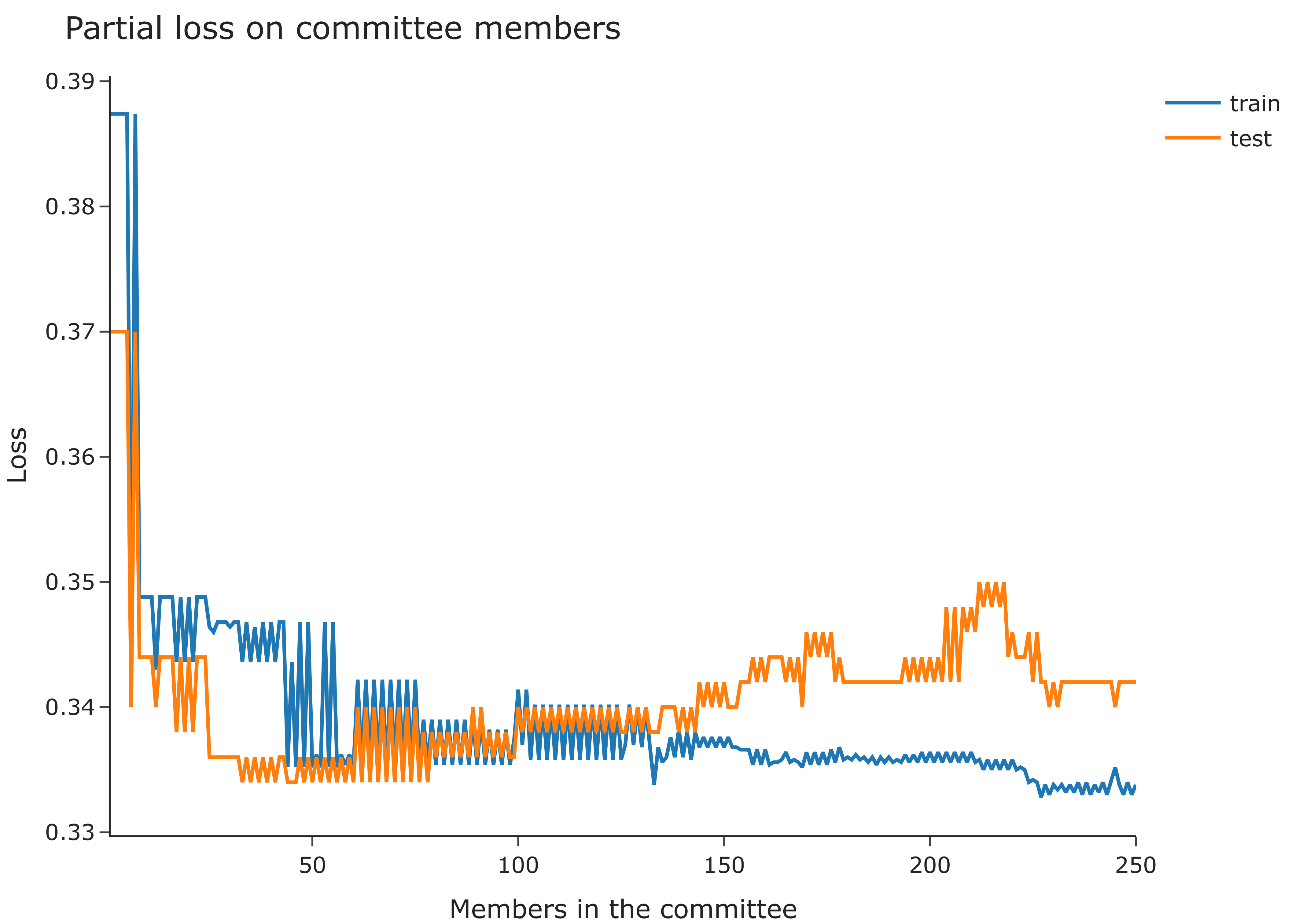

5 Members
50 Members
100 Members
250 Members
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

Best committee, 238 members, Accuracy: 1.0

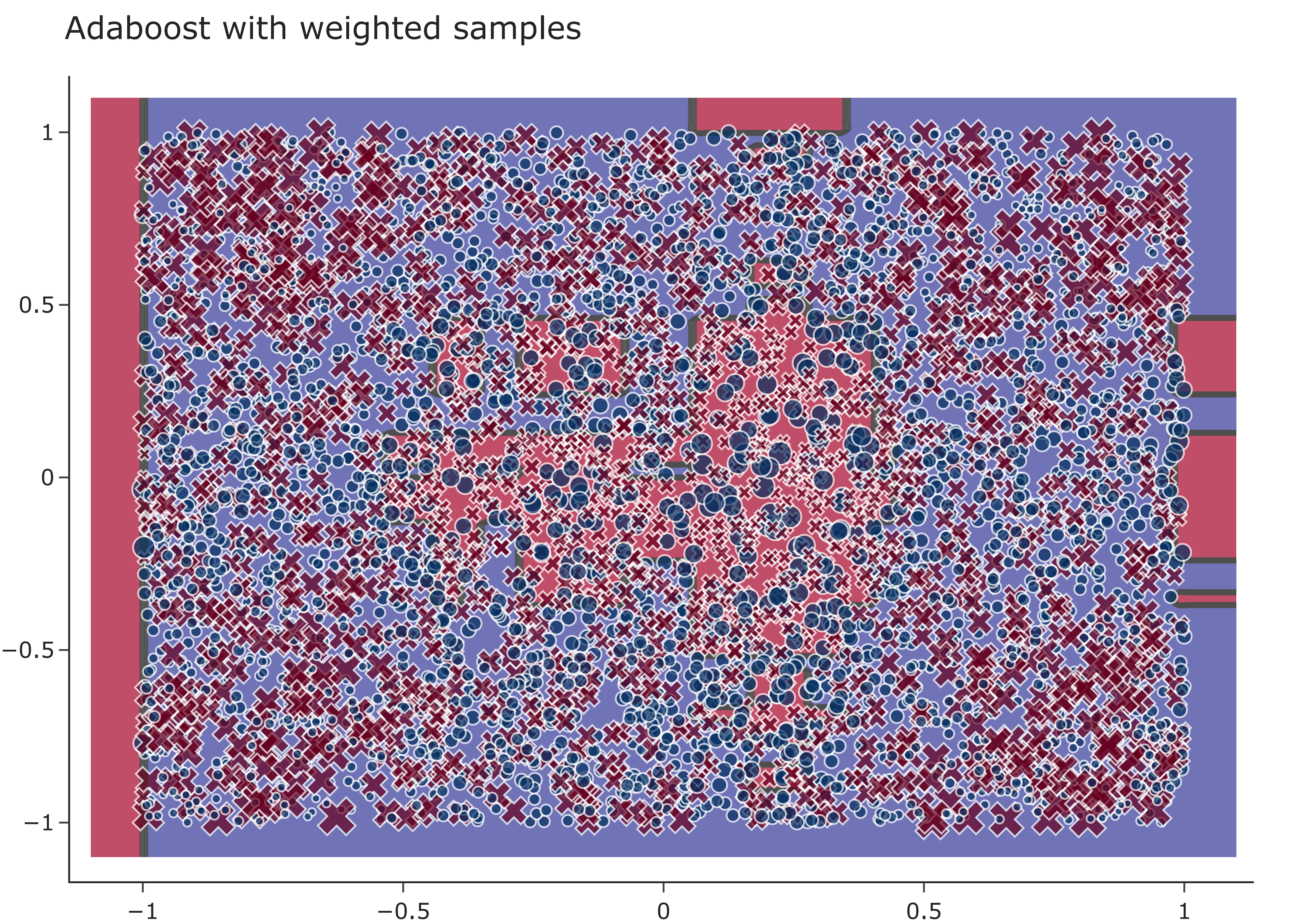


Adaboost with weighted samples

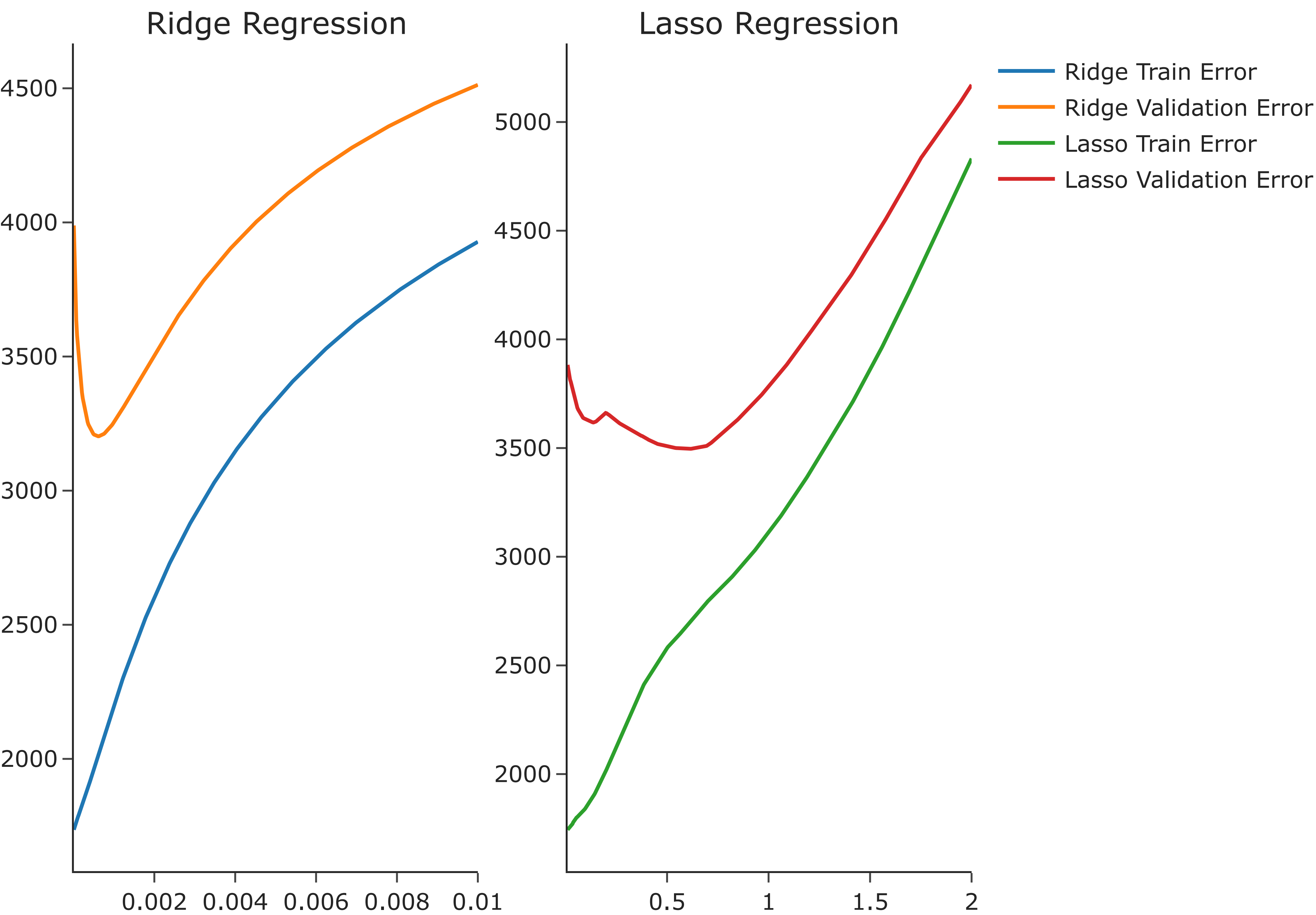




Partial loss on committee members
Loss
Members in the committee
train
test
0.39
0.38
0.37
0.36
0.35
0.34
0.33
50
100
150
200
250



Adaboost with weighted samples
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1



Ridge Regression
Lasso Regression
4500
4000
3500
3000
2500
2000
0.002
0.004
0.006
0.008
0.01
5000
4500
4000
3500
3000
2500
2000
0.5
1
1.5
2
Ridge Train Error
Ridge Validation Error
Lasso Train Error
Lasso Validation Error