

: גודל גודל

1.1 Regularization

Based on Lecture 7 and Recitation 9

1. In the following question we will show that although the Ridge estimator is biased it can achieve lower MSE compared to the LS estimator.

Let $\mathbf{X} \in \mathbb{R}^{m \times d}$ be a **constant** design matrix, $\mathbf{y} \in \mathbb{R}^d$ a response vector, and assume that $\mathbf{X}^\top \mathbf{X}$ is invertible. Denote $\hat{\mathbf{w}}$ the LS solution and $\hat{\mathbf{w}}_\lambda$ the ridge solution for the regularization parameter $\lambda \geq 0$ (where $\hat{\mathbf{w}}_0 \equiv \hat{\mathbf{w}}$)

- Assume the linear model is correct, namely $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- Recall that in this case: $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}$.

- (a) Show that $\hat{\mathbf{w}}_\lambda = A_\lambda \hat{\mathbf{w}}$ where $A_\lambda := (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} (\mathbf{X}^\top \mathbf{X})$

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

: מינימיזציה של פונקציית כפיפה

$$\begin{aligned} A_\lambda \hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} (\underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{I}_d}) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{y} = \hat{\mathbf{w}}(\lambda) \end{aligned}$$

: גודל גודל

- (b) From the above, conclude that for any $\lambda > 0$ the ridge estimator is a biased estimator of \mathbf{w} . That is, show that for any $\lambda > 0$ $\mathbb{E}[\hat{\mathbf{w}}_\lambda] \neq \mathbf{w}$.

$$\begin{aligned} E[\hat{\mathbf{w}}_\lambda] &= E[A_\lambda \hat{\mathbf{w}}] = \text{האפקט של } \lambda \text{ על } \mathbf{w} \\ &= A_\lambda E[\hat{\mathbf{w}}] = \text{האפקט של } A_\lambda \text{ על } \mathbf{w} \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} (\mathbf{X}^\top \mathbf{X}) \mathbf{w} \end{aligned}$$

$E[\hat{\mathbf{w}}] \neq \mathbf{w}$ כי $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} (\mathbf{X}^\top \mathbf{X}) \neq \mathbf{I}_d$ $\lambda > 0$ בזק הערך

- (c) Show that: $\text{Var}(\hat{\mathbf{w}}_\lambda) = \sigma^2 A_\lambda (\mathbf{X}^\top \mathbf{X})^{-1} A_\lambda^\top$, for σ^2 the variance of the assumed noise.

Hint: Recall that for a constant matrix B and a random vector \mathbf{z} it holds that $\text{Var}(B\mathbf{z}) = B \cdot \text{Var}(\mathbf{z}) \cdot B^\top$ and that $\text{Var}(\hat{\mathbf{w}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

$$\text{Var}(\hat{w}_\lambda) = \text{var}(A_\lambda \hat{w}) = A_\lambda \text{var}(\hat{w})(A_\lambda)^T = \\ A_\lambda \sigma^2 (X^T X)^{-1} (A_\lambda)^T$$

(d) Derive explicit expressions for the (squared) bias and variance of \hat{w}_λ as a function of λ , i.e. write a bias-variance decomposition for the mean square error of \hat{w}_λ .

Hint: recall that for the multivariate case the MSE defined to be:

$$MSE(\hat{y}) = \mathbb{E}[\|\hat{y} - y\|^2] = \mathbb{E}[(\hat{y} - y)^\top (\hat{y} - y)]$$

where y is the true value, and \hat{y} is our estimation.

Given A_λ is full rank, $E[\hat{w}] = w$

$$\text{a.s.o} \quad \downarrow \quad \begin{matrix} \text{definition} \\ E[XV] = E[X]E[V] \end{matrix} \quad : \text{Prop} \quad \text{pcd} \\ E[\hat{w}_\lambda] = E[A_\lambda \hat{w}] = E[A_\lambda]E[\hat{w}] = A_\lambda w$$

$$\text{Var}(\lambda) = \text{Tr}(\text{Var}(\hat{w}_\lambda)) = \text{Tr}(A_\lambda \sigma^2 (X^T X)^{-1} (A_\lambda)^T) = \sigma^2 \text{Tr}(A_\lambda (X^T X)^{-1} (A_\lambda)^T)$$

$$\text{Bias}^2(\lambda) = \|E[\hat{w}_\lambda] - w\|^2 = \|E[\hat{w}_\lambda] - E[w]\|^2 = \\ = \|A_\lambda w - w\|^2 = \|(A_\lambda - I_d)w\|^2 = \\ = ((A_\lambda - I_d)w)^\top (A_\lambda - I_d)w = w^\top (A_\lambda - I_d)^\top (A_\lambda - I_d)w$$

$$\text{MSE}(\lambda) = E[\|w - \hat{w}_\lambda\|^2] = \text{Var}(\hat{w}_\lambda) + \underbrace{\|w - \hat{w}_\lambda\|^2}_{\text{Bias}^2(\hat{w}_\lambda)} =$$

$$= \sigma^2 \text{Tr} \left(A_\lambda (X^T X)^{-1} (A_\lambda)^T \right) - \omega^T (A_\lambda - I_d)^T (A_\lambda - I_d) \omega$$

(e) Show by differentiation that

$$\frac{\partial}{\partial \lambda} \text{MSE}(\hat{w}_\lambda)|_{\lambda=0} = \frac{\partial}{\partial \lambda} \text{bias}^2(\hat{w}_\lambda)|_{\lambda=0} + \frac{\partial}{\partial \lambda} \text{Var}(\hat{w}_\lambda)|_{\lambda=0} < 0$$

That is, calculate the derivative of the functions above with respect to λ at point $\lambda = 0$.
Update (24/05/22) - Hints:

- For the variance term: The EVD might be useful.
- For the Bias term: try to use the product rule $\partial XY = (\partial X)Y + X(\partial Y)$ (this can also be solved using the EVD)

נוסף ונה פה רצוי מינימום יי'ה. נכון ש $X^T X \in \mathbb{R}^{d \times d}$ ו- $\omega \in \mathbb{R}^d$

$$X^T X = U D U^T \quad \text{ובן-זיהוי } U, D \in \mathbb{R}^{d \times d} \text{ מושג} \leftarrow \text{EVD} \text{ רלוונט}$$

$$\text{Var}(\lambda) = \sigma^2 \text{Tr} \left(A_\lambda (X^T X)^{-1} A_\lambda^T \right) = \text{כפיה}$$

$$= \sigma^2 \text{Tr} \left((X^T X + \lambda I_d)^{-1} \underbrace{(X^T X)}_I (X^T X)^{-1} ((X^T X + \lambda I_d)^{-1})^T \right) =$$

$$= \sigma^2 \text{Tr} \left((X^T X + \lambda I_d)^{-1} (X^T X)^T ((X^T X + \lambda I_d)^{-1})^T \right) =$$

$$= \sigma^2 \text{Tr} \left(((U D U^T + \lambda I_d)^{-1} (U D U^T)^T ((U D U^T + \lambda I_d)^{-1})^T \right) =$$

$$= \sigma^2 \text{Tr} \left(U (D + \lambda I_d)^{-1} U^T U D^T U^T ((D + \lambda I_d)^{-1})^T U^T \right) =$$

$$= \sigma^2 \text{Tr} \left(U (D + \lambda I_d)^{-1} D^T ((D + \lambda I_d)^{-1})^T U^T \right) =$$

. Trace ל- ω יתנו

$$= \sigma^2 \text{Tr} \left(U^T U (D + \lambda I_d)^{-1} D^T ((D + \lambda I_d)^{-1})^T \right) =$$

$$= \sigma^2 \text{Tr} \left((D + \lambda I_d)^{-1} D^T ((D + \lambda I_d)^{-1})^T \right) =$$

$$= \sigma^2 \sum_P \frac{P}{(P + \lambda)^2}$$

$P \in \text{eigenvalues}(X^T X)$

$$\frac{\partial \text{Var}(\lambda)}{\partial \lambda} = \frac{\partial \sigma^2 \sum_P \frac{P}{(P+\lambda)^2}}{\partial \lambda} = : \text{Var} \quad \mu \sigma$$

$$= \sum_P \frac{-2(P+\lambda)P}{(P+\lambda)^4} = - \sum_P \frac{2P}{(P+\lambda)^3}$$

$$\left. \frac{\partial \text{Var}(\lambda)}{\partial \lambda} \right|_{\lambda=0} = - \sum_P \frac{\partial P}{P^3} = - \sum_P \frac{2}{P^2} : \text{Var} \mu \sigma$$

$$A_\lambda = (X^T X + \lambda I_d)^{-1} (X^T X) = : \text{e. j. m. p. } P_0 \geq$$

$$= (U D U^{-1} \lambda I_d)^{-1} (U D U^T) =$$

$$= (U D U^T + \lambda U I_d U^T)^{-1} (U D U^T) = U (D + \lambda I_d)^{-1} D U^T$$

$$\text{Bias}^2(\lambda) = w^T (A_\lambda - I_d)^T (A_\lambda - I_d) w = : \text{Var} \quad \text{BEN}$$

$$= w^T (U (D + \lambda I_d)^{-1} D U^T - I_d)^T (U (D + \lambda I_d)^{-1} D U^T - I_d) w =$$

$$= w^T U ((D + \lambda I_d)^{-1} D - I_d)^T \underbrace{U^T U}_{I} ((D + \lambda I_d)^{-1} D - I_d) U^T w =$$

$$= w^T U ((D + \lambda I_d)^{-1} D - I_d)^T ((D + \lambda I_d)^{-1} D - I_d) U^T w =$$

$$= w^T U \| (D + \lambda I_d)^{-1} D - I_d \|_F^2 U^T w$$

$$\frac{\partial \text{Bias}^2(\lambda)}{\partial \lambda} = \frac{\partial w^T U \| (D + \lambda I_d)^{-1} D - I_d \|_F^2 U^T w}{\partial \lambda} = : \text{JG}$$

$$= w^T U \frac{\partial \| (D + \lambda I_d)^{-1} D - I_d \|_F^2}{\partial \lambda} U^T w$$

$$\frac{\partial \|(D + \lambda I_d)^{-1} D - I_d\|^2}{\partial \lambda} = \frac{\partial \sum_i \left(\sum_j ((D + \lambda I_d)^{-1} D - I_d)_{ij}^2 \right)}{\partial \lambda} =$$

$$\sum_i \frac{\partial \left(\frac{\lambda_i}{\lambda_i + \lambda} - 1 \right)^2}{\partial \lambda} = \sum_i \frac{\partial \left(\frac{-\lambda}{\lambda_i + \lambda} \right)^2}{\partial \lambda} =$$

$$= \sum_i \frac{2\lambda(\lambda_i + \lambda)^2 - 2(\lambda_i + \lambda)\lambda^2}{(\lambda_i + \lambda)^4}$$

$$= \sum_i \frac{2\lambda\lambda_i}{(\lambda_i + \lambda)^3}$$

$$\frac{\partial \text{Bias}^2(\lambda)}{\partial \lambda} \Big|_{\lambda=0} = w^T U \frac{\partial \|(D + \lambda I_d)^{-1} D - I_d\|^2}{\partial \lambda} U^T w \Big|_{\lambda=0} =$$

: $\lambda > 0$

$$= w^T U \sum_i \frac{2\lambda\lambda_i}{(\lambda_i + \lambda)^3} \Big|_{\lambda=0} = w^T U \circ U^T w = 0$$

$$\frac{\partial \text{MSE}}{\partial \lambda} \Big|_{\lambda=0} = 0 + \sum_p \frac{-2}{p^2} = \sum_p \frac{-2}{p^2}$$

: סדר גודל

$(X^T X \neq 0)$

$P^2 > 0 \leftarrow P \neq 0 \text{ ו } 0 \neq P^T P \text{ ו } \sum_p \frac{-2}{p^2} < 0$

$$\sum_p \frac{-2}{p^2} < 0 \quad \text{ור } \frac{-2}{p^2} < 0 \text{ סדר}$$

כזה $\frac{\partial \text{MSE}}{\partial \lambda} \Big|_{\lambda=0} < 0$ יגזר

(f) Conclude that, if the linear model is correct, a little Ridge regularization helps to reduce the MSE.

סדר 1 מילון פס $\lambda=0$ ו. מינימום MSE ב. מינימום יקן

Ridge וריג'ג נס' MSE $\lambda=0$ נס'

$\lambda = 0$ מינימיזציית נקודות כוונתית נס' MSE

אם $\lambda \neq 0$ אז MSE \geq MSE($\lambda=0$)

$\lambda \neq 0$ מינימיזציית MSE

MSE \leq MSE($\lambda=0$) \Rightarrow מינימיזציית MSE($\lambda=0$)

מיון מינימיזציית MSE($\lambda=0$)

MSE($\lambda=0$) \geq MSE($\lambda > 0$)

לפיכך מינימיזציית MSE($\lambda > 0$) מינימיזציית MSE($\lambda=0$)

1.2 PCA

Based on Lecture 8 and Recitation 11

- Let $X : \Omega \rightarrow \mathbb{R}^d$ be a random variable with zero mean and covariance $\Sigma \in \mathbb{R}^{d \times d}$. Show that for any $v \in \mathbb{R}^d$, where $\|v\|_2 = 1$, the variance of $\langle v, X \rangle$ is not larger than variance obtained by the PCA embedding of X into a one-dimensional subspace (assume that the PCA uses the actual Σ).

$$\|v\|_2 = 1 \quad \text{וראgle} \quad v \in \mathbb{R}^d$$

$$\therefore \text{רמ"נ} \quad E[x] = (0, \dots, 0) \in \mathbb{R}^d \quad \text{רמ"נ}$$

$$E[\langle v, x \rangle] = E[\langle (0, \dots, 0), x \rangle] = E[0] = 0$$

כך נוכיח הטענה:

$$\text{Var}(\langle v, x \rangle) = E \left[(\langle v, x \rangle - E[\langle v, x \rangle])^2 \right] =$$

$$= E[\langle v, x \rangle^2] = v^T E[x x^T] v = v^T \Sigma v$$

Σ סהgal הול'ה מוקע $u_1 \sim \mathcal{N}(0)$

1. גאנטס PCA און וורבעת השוואת שערות one בפונקציית

$$u_1^T \Sigma u_1 : \text{הציג}$$

$$\text{ונר...} \quad v^T \Sigma v \leq u_1^T \Sigma u_1, \quad v \in \mathbb{R}^d \text{ סהpn}$$

1.3 Kernels

Based on Lecture 9 and Recitation 12

3. Let $k(\mathbf{x}, \mathbf{x}')$ be a valid PSD kernel. Provide a valid PSD kernel $\tilde{k}(\mathbf{x}, \mathbf{x}')$, constructed from k , which is guaranteed to be normalized. That is, for all \mathbf{x} it holds that $\tilde{k}(\mathbf{x}, \mathbf{x}) = 1$. Prove your answer.

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \frac{k(\mathbf{x}, \mathbf{x}')}{\sqrt{k(\mathbf{x}, \mathbf{x}) \cdot k(\mathbf{x}', \mathbf{x}')}} : \text{אץ'}$$

(יכי' כי \tilde{k} מוגדר נכון)

$$\begin{aligned} \tilde{k}(\mathbf{x}, \mathbf{x}') &= \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\sqrt{\langle \psi(\mathbf{x}), \psi(\mathbf{x}) \rangle \langle \psi(\mathbf{x}'), \psi(\mathbf{x}') \rangle}} = \\ &= \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\sqrt{\|\psi(\mathbf{x})\|^2 \|\psi(\mathbf{x}')\|^2}} = \frac{\langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle}{\|\psi(\mathbf{x})\| \|\psi(\mathbf{x}')\|} = \\ &= \left\langle \frac{\psi(\mathbf{x})}{\|\psi(\mathbf{x})\|}, \frac{\psi(\mathbf{x}')}{\|\psi(\mathbf{x}')\|} \right\rangle : 1 \text{ מיל' גודלה} \end{aligned}$$

$$\tilde{k}(\mathbf{x}, \mathbf{x}) = \left\langle \frac{\psi(\mathbf{x})}{\|\psi(\mathbf{x})\|}, \frac{\psi(\mathbf{x})}{\|\psi(\mathbf{x})\|} \right\rangle = \frac{\|\psi(\mathbf{x})\|^2}{\|\psi(\mathbf{x})\|^2} = 1$$

4. Consider a data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$, and a feature map $\psi: \mathbb{R}^d \rightarrow \mathcal{F}$ where \mathcal{F} is some feature space. Give an example of a data set S and a feature map ψ such that S is not linearly separable in \mathbb{R}^d (for $d \geq 2$) but that the transformed data set $S_\psi = \{(\psi(\mathbf{x}_i), y_i)\}_{i=1}^m$ is linearly separable in \mathcal{F} .

$$\text{: בפניהם } S \quad \text{הו } y = \{-1, 1\} \quad x \in \mathbb{R}^2 \quad \text{הו}$$

$$S = \left\{ \left(\begin{pmatrix} -2 \\ 0 \end{pmatrix}, 1 \right), \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, -1 \right), \left(\begin{pmatrix} 2 \\ 0 \end{pmatrix}, 1 \right) \right\}$$

\mathbb{R}^3 -ה יתקיים פה שולח $\mathbb{R}^2 \rightarrow \mathbb{R}$ מוגדר ב- \mathbb{R}^2 ו- \mathbb{R}^3 קיימת פונקציית

$$\Psi \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) = \begin{pmatrix} x_1 \\ x_2 \\ (x_1 + x_2)^2 \end{pmatrix} \quad \Psi: \mathbb{R}^2 \rightarrow \mathbb{R}^3 \quad \text{: פונקציית}$$

$$S_\psi = \left\{ \left(\begin{pmatrix} -2 \\ 0 \\ 4 \end{pmatrix}, 1 \right), \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, -1 \right), \left(\begin{pmatrix} 2 \\ 0 \\ 4 \end{pmatrix}, 1 \right) \right\} \quad \text{הו}$$

. בזיהוי $b=-1$, $\begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}$ "הו" מוגדר ב- \mathbb{R}^3 ו- \mathbb{R}^2 נס庭ה

5. For each of the following functions, prove it is a valid PSD kernel or show a counter example:

(a) $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2)$

$$\text{לכל } \mathbf{x}, \mathbf{y} \in \mathbb{R}^2 \quad k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2)$$

train set = $\begin{bmatrix} \ln(1) \\ \ln(2) \end{bmatrix}$ ב- \mathbb{R}^2

$$G_k = \begin{bmatrix} e^{-\|\ln(1) - \ln(1)\|} & e^{-\|\ln(1) - \ln(2)\|} \\ e^{-\|\ln(1) - \ln(2)\|} & e^{-\|\ln(2) - \ln(2)\|} \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad \text{: (ב- \mathbb{R}^2)}$$

. G הוא מושג חיובי רציף

$$\det(G - \lambda I) = (1 - \lambda)^2 - 4 = \lambda^2 - 2\lambda - 3 = (\lambda+3)(\lambda-5)$$

$$\lambda_1 = -3, \lambda_2 = 5 \quad \text{רַא שְׁגָן שְׁגָן}$$

. PSD מatrix G כולל (-3) ו-5 שוגר ו-0

PSD-kernel הוא K-ה גראן שוגר שוגר Mercer סע אוניברסיטה

(b) $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) - k_2(\mathbf{x}, \mathbf{y})$ for any two valid kernels $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$

הוכחה - קונtradiction:

K=2 גורגי K=1 גורגי . קונtradiction שוגר שוגר שוגר שוגר

$$k_1 = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^1 = \langle \mathbf{x}, \mathbf{y} \rangle + 1$$

$$k_2 = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^2 = \langle \mathbf{x}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + 1 =$$

$$= (\mathbf{y}^T \mathbf{x})(\mathbf{x}^T \mathbf{y}) + 2\langle \mathbf{x}, \mathbf{y} \rangle + 1 =$$

$$= \mathbf{y}^T \mathbf{x} \mathbf{x}^T \mathbf{y} + 2\langle \mathbf{x}, \mathbf{y} \rangle + 1 =$$

$$= \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + 1$$

: שוגר שוגר

$$k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) - k_2(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle + 1 - (\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + 1) =$$

$$= -\langle \mathbf{x}, \mathbf{y} \rangle - \|\mathbf{x}\|^2 \|\mathbf{y}\|$$

הנחות על פונקציית הגרדיאנט

. PSD של y ו- x

$$: \text{רפלקסיה כפולה} \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, y = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad \text{ולכן } K \text{ חיובי}$$

$$K = \begin{bmatrix} K(x, x) & K(x, y) \\ K(y, x) & K(y, y) \end{bmatrix} = \begin{bmatrix} -6 & -20 \\ -20 & -72 \end{bmatrix}$$

$$: V = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{ORTHOGONAL}$$

$$V^T K V = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} -6 & -20 \\ -20 & -72 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -6 & -20 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -6 < 0$$

לפניהם מתקיים הטענה PSD של K

• $\exists k_1, k_2$ כ- L^2 ו-PSD kernel של G

(c) $k(\mathbf{x}, \mathbf{y}) = k_a(\mathbf{x}_a, \mathbf{y}_a) + k_b(\mathbf{x}_b, \mathbf{y}_b)$ for any two valid kernels $k_a(\cdot, \cdot)$ and $k_b(\cdot, \cdot)$, where

$$\mathbf{x} := \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}, \quad \mathbf{y} := \begin{bmatrix} \mathbf{y}_a \\ \mathbf{y}_b \end{bmatrix}$$

וכך

לעתה נסמן $k_a(\mathbf{x}_a, \mathbf{y}_a), k_b(\mathbf{x}_b, \mathbf{y}_b)$ ו- G_a

k_a הוא פונקציית גראדיאנט של G_a ב- \mathcal{H}

, K_b הוא פונקציית קורלציה G_b - א

: וריאנט K שווה ל $(G_a) + (G_b)$

$$(G_K)_{ij} = K(x_i, y_j) = K_a(x_{ai}, y_{aj}) + K_b(x_{bi}, y_{bj})$$

$$= (G_a)_{ij} + (G_b)_{ij}$$

$$\left(\cdot \in \mathbb{R}^{d \times d} \ni \right) G_K = G_a + G_b \quad \text{ו } \parallel G \parallel$$

K_a, K_b הם Mercer בונוס נגיעה מוגדר

PSD מושגון G_a, G_b - וריאנטים מוגדרים

$$V^T G_a V \geq 0, V^T G_b V \geq 0 \quad V \in \mathbb{R}^d \text{ ו } V \neq 0$$

$$V^T G V = V^T (G_a + G_b) V =$$

: $V \neq 0$

$$= V^T G_a V + V^T G_b V \geq 0$$

ונמצא PSD מושגון ל- G ו- G מוגדרים כך

. וריאנט K הוא PSD

חלק פרקי:

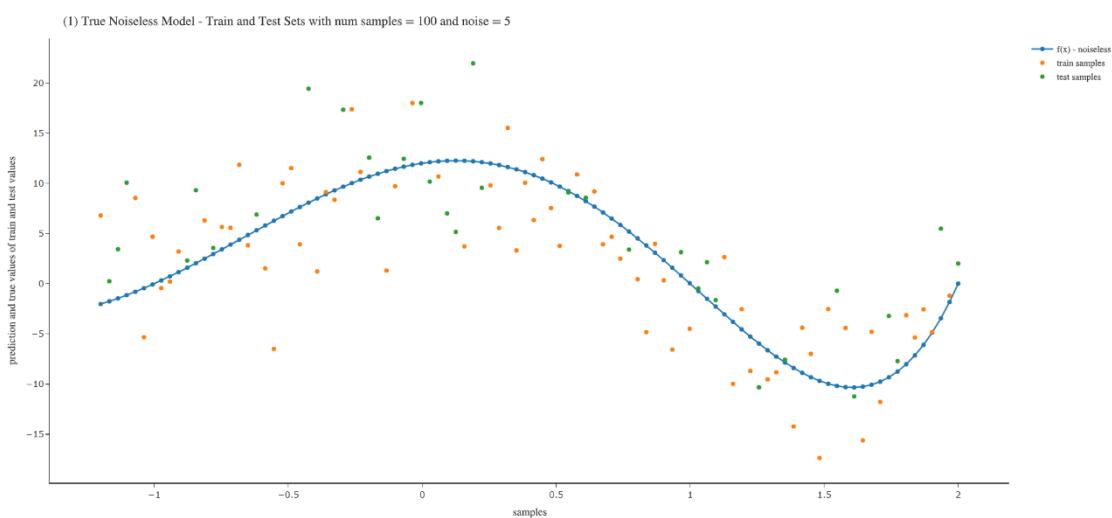
שאלה 1:

1. Generate a dataset of $m = 100$ samples according to the following model: $y = f(x) + \epsilon$ where

$$f(x) := (x+3)(x+2)(x+1)(x-1)(x-2), \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

and x are selected uniformly in the range $x \in [-1.2, 2]$. Use a noise level (i.e. σ^2) of 5. Then, split the dataset into train- and test-sets with a `train_portion` of $\frac{2}{3}$ using your implementation of the `split_train_test` function.

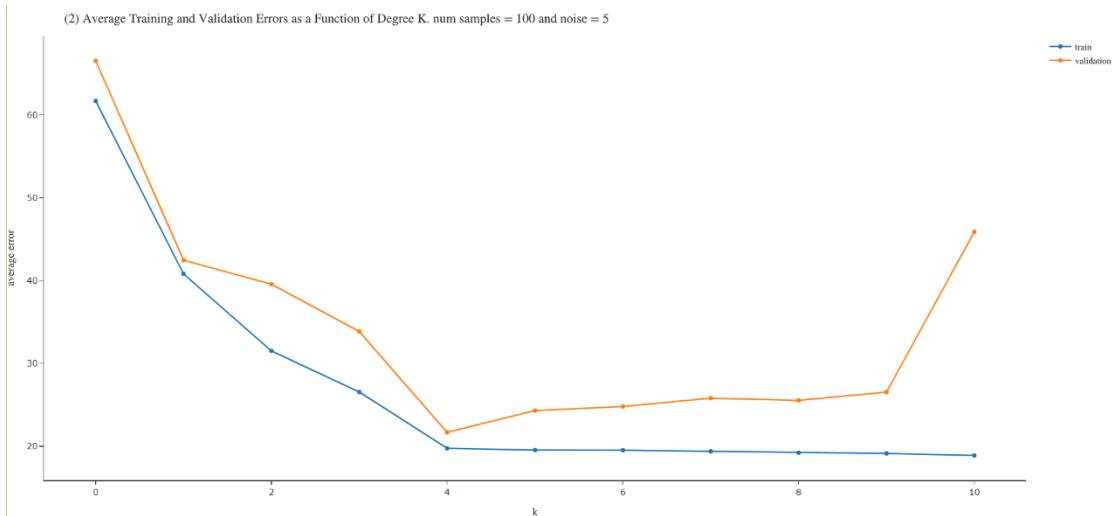
Scatter plot the true (noiseless) model and the two sets using different colors for train and test samples.



שאלה 2:

2. Using only the training portion of the samples, perform 5-fold cross-validation for each of the polynomial degrees $k = 0, 1, \dots, 10$. Plot for each value of k the average training- and validation errors. Explain your results and address the differences between the train and validation errors

ניתן לראות שהשגיאה על ה Train הולכת וקטנה מהר כשמתק Robbins לדרגה הנכונה (היא נשארת נמוכה). לעומת זאת, ב Validation ניתן לראות שהשגיאה אמנים הולכת וקטנה כשמתק Robbins לדרגה הנכונה אבל כשמתרחקים ממנה היא עולה, ב Gegוד ל Train שהיא נשארת נמוכה.



שאלה 3 :

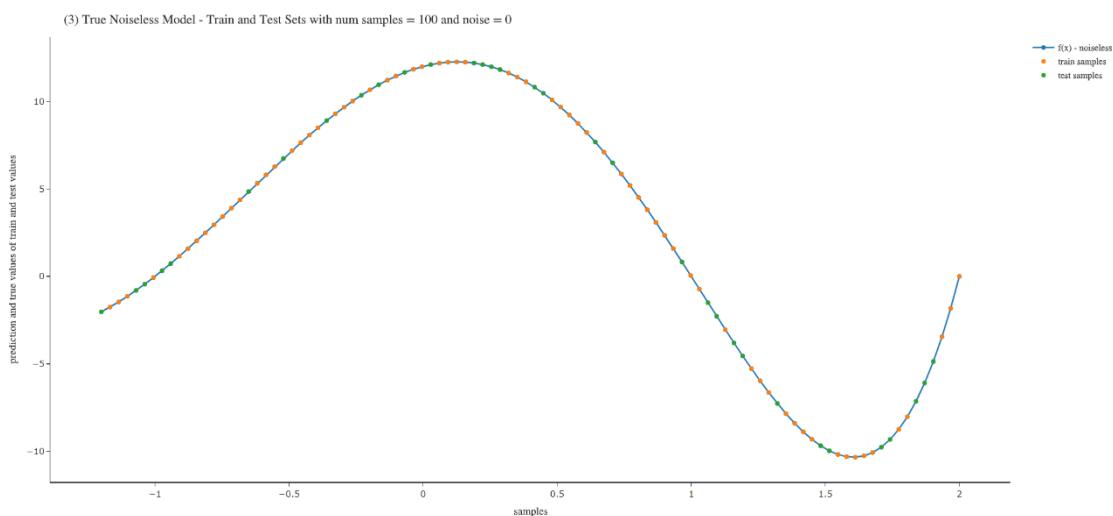
3. Using k^* , the polynomial degree for which the lowest validation error was achieved, fit a polynomial model over the entire training set (that is, not only on specific folds). Report what was the value of k^* and the *test* error of the fitted model. Is the test error similar to the validation error previously achieved? Round the test error to two decimal places.

ערך השגיאה על הטעט אכן שונה מערך השגיאה על ה K שנבחר הוא 4 ולא 5. Validation נימן לראות שבגלל הרעש ה K שנבחר הוא 4 ולא 5.

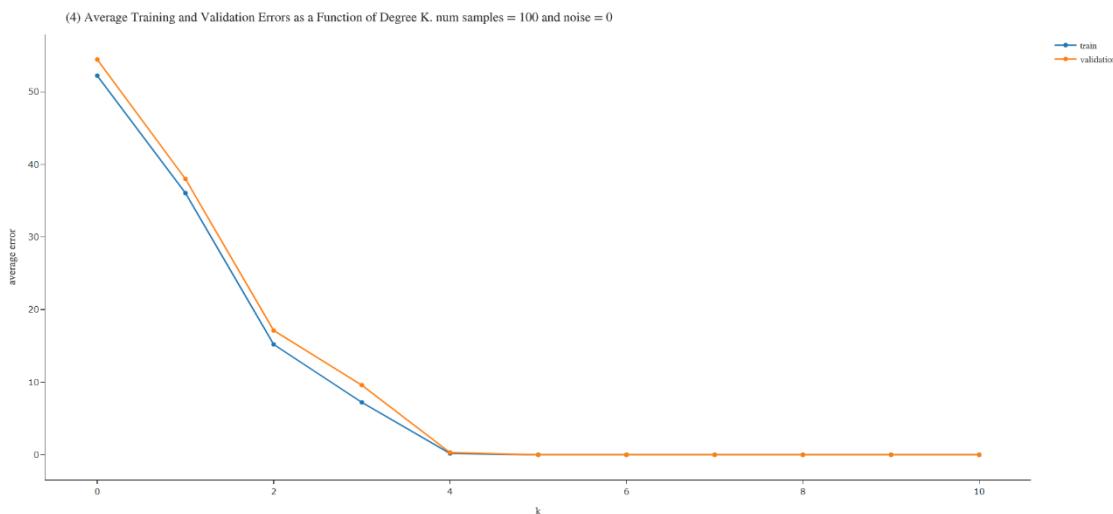
```
num samples: 100 , noise: 5
k: 4
test error: 25.85
validation error: 21.69
```

שאלה 4 :

4. Repeat the questions above but using a noise level of 0. Show the figures obtained for such scenario. Compare your results to the noisy case and explain. Address the differences between the train-, validation- and test errors for the different values of k .



נימן לראות שכשאין רעש השגיאה גם על ה *train* וגם על ה *validation* הולכת וקטנה עד שmagimim לדרגת הפולינום הנכונה ולאחר מכן השגיאה נשארת נמוכה.

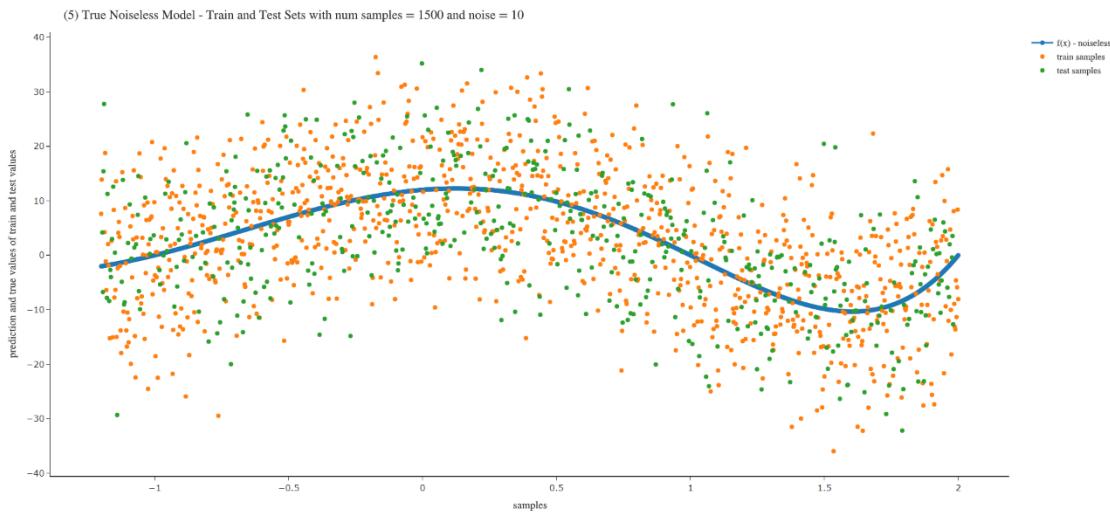


ניתן לראות שבגלל שאין רוש ה K שנבחר הוא הנכון והשגיאה 0 גם על הทดสอบ וגם על הvalidation.

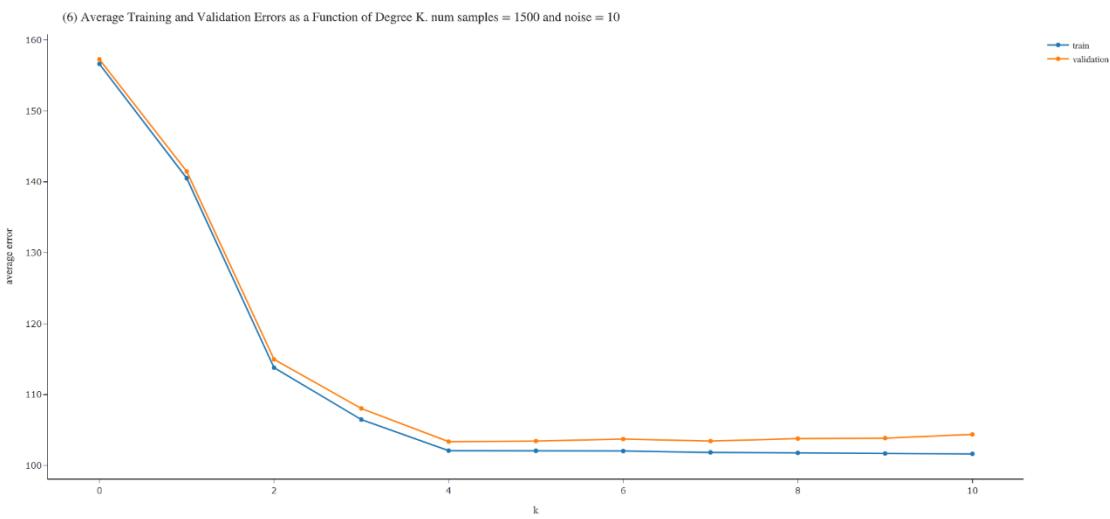
```
num samples: 100 , noise: 0
k: 5
test error: 0.0
validation error: 0.0
```

שאלה 5 :

5. Repeat the questions above while generating $m = 1500$ samples using a noise level of 10. What can we learn about the use of Cross-Validation with respect to the sample size? How is this supported by the Law of Large Numbers and the consistency property of estimators?



ניתן לראות שערך השגיאה על ה validation גדול וקטן עד שמגיעים לדרגת הפולינום הנכונה ולאחר מכן מכך השגיאה על ה validation נמשכת יותר מאשר על ה train.



ניתן לראות שאטנו לא נבחר K המדויק אך השגיאה על הבדיקה קטנה מהשגיאה על הvalidation, ניתן להניח שהזיה מכיוון שיש לנו כמו גודלה יחסית של דוגמאות (בניגוד למקרה הראשון).

```
num samples: 1500 , noise: 10
k: 4
test error: 98.75
validation error: 103.38
```

- Load the diabetes dataset and split it to a training and testing portion, using 50 samples for training.

سؤالה 7 :

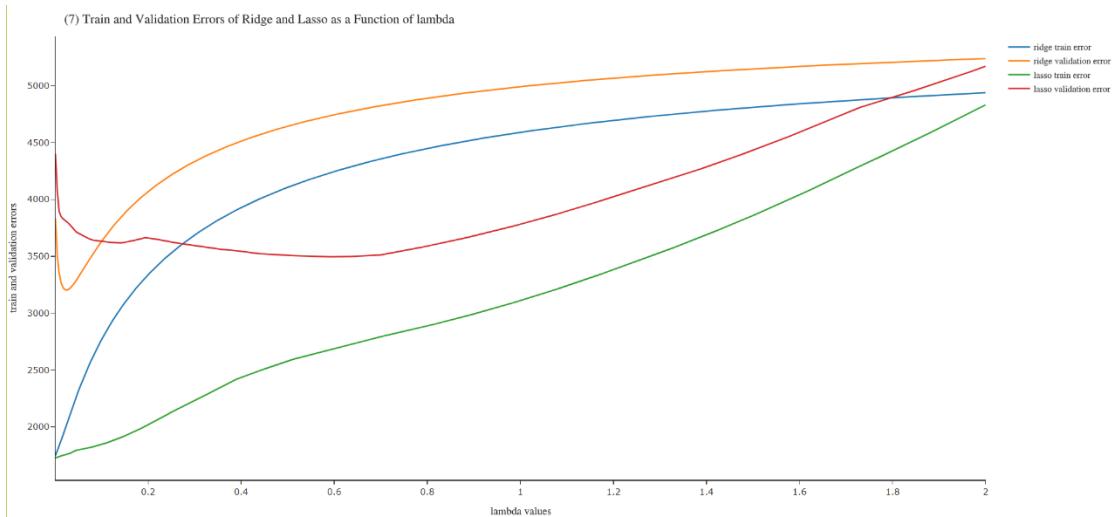
- For both Ridge and Lasso regularizations, run 5-Fold Cross-Validation using different values of the regularization parameter λ . Explore possible ranges of values of λ (say, take `n_evaluations=500` equally spaced values in a range of your choice). Which range of values is meaningful for the given data for each of the algorithms?

Then, over these selected ranges, for each of the two algorithms, plot the train- and validation errors as a function of the tested regularization parameter value. Explain your results. Address

both the differences between the train- and validation errors as well as the differences in the plots for the two algorithms.

בחרתי לבצע Cross-Validation על ערכי `lambda` שבין 0.001 ל 2 שהוא. עבור 0 זה לא מעניין כי אז אין משמעות ל-Cross-Validation. עבור ערכים גדולים יותר השגיאה די קבועה וגדולה ללא תלות ב`lambda`.

הערך `lambda` שמביא למינימום את השגיאה על ה `Validation` עבור Ridge קטן מה `lambda` עבור Lasso ובשני המקרים השגיאה על `Train` קטנה מושגיאת על ה `Validation`.



שאלה 8 :

8. Report which regularization parameter values achieved the best validation errors for the Ridge and Lasso regularizations. Then, using these values, fit Ridge, Lasso and Least Squares regressions over the entire training set. Report the test errors of each of the fitted models.

- Use your *own* previously implemented Least Squares algorithm.

```
--Ridge Regression-----
best lambda:  0.02503607214428858
test error:  3241.5644999313267

-----Lasso Regression-----
best lambda:  0.5978957915831664
test error:  3641.6573474624356

-----Mean Square Error-----
test error for Mean Square Error:  3612.2496883249
```