Deadline: Th. November 28th , 13:00 Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz. Alternatively upload a `.pdf` file via LearnWeb. (e.g. exported Jupyter notebook)

1. L^1 regularization (10 points)

A. [7p] Fit a linear regression model (including bias) with L^1 regularization to the dataset from Table 1 by performing 2 iterations of coordinate descent (update each parameter twice). Use $\beta^{(0)} = 0$ and $\lambda = 1$.

x_1	x_2	y
1	1	1.4
1	-1	1.6
-1	0	0.5
-1	-1	0.6

Table 1

B. [3p] The **elastic-net** model is a linear model with a mix of L^1 and L^2 regularization.

$$L^{\text{enet}}(\beta) = \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\Big(\alpha\|\beta\|_1 + (1-\alpha)\tfrac{1}{2}\|\beta\|_2^2\Big)$$

Note that if $\alpha = 1$, elastic net is the same as LASSO and for $\alpha = 0$ it is the same as RIDGE regression. For $\alpha \in (0, 1)$ it is something in between. We trained an Elastic Net model 4 times on a regression task, each time choosing a different trade-off $\alpha \in \{0, 0.25, 0.5, 1\}$. The resulting regularization paths, as well as the number of non-zero coefficients at different total regularization strength λ is shown in Figure 1. Explain which figure corresponds to which choice of α.

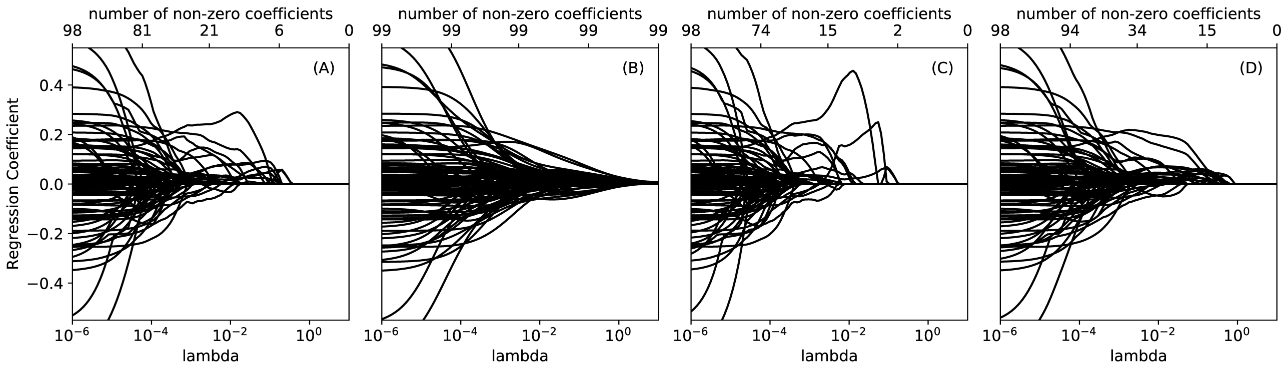

Figure 1: Regularization paths of the 4 models

2. Variable Selection – Programming (10 points)

Use the following code to load the famous "Boston Housing" dataset (alternatively data-files will be uploaded to the LearnWeb as well)

```
import numpy as np
from sklearn.datasets import load_boston
np.random.seed(2019)
dataset = load_boston()
Xdata = dataset['data']
Ydata = dataset['target']
N, M = Xdata.shape
ridx = np.random.permutation(N)
split = int(0.8*N)
Xtrain = Xdata[ridx[:split]]
Ytrain = Ydata[ridx[:split]]
Xvalid = Xdata[ridx[split:]]
Yvalid = Ydata[ridx[split:]]
```

A. [7p] Implement both forward search and backward search and apply them to the provided data using a linear regression model. At the start of each outer loop, report the currently selected variables V as well as the loss on the training and validation set.

B. [3p] Repeat the experiment 100 times using random train/valid splits (you'll need to remove `np.random.seed(2019)`). Whats the average improvement of forward/backward search compared to fitting with the whole dataset? How often does each of the 13 variables end up in the final selection?

3* Parameter Variance – OLS vs Ridge Regression (5 points)

For the following problem, we assume that the ground truth is is a linear function $y(x) = x^\intercal\hat{\beta}+\epsilon$ with $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and we are given a **finite** data sample (X, Y). From the lecture we know that the ordinary least squares (OLS) estimator $\hat{\beta}^{\text{OLS}} = (X^\intercal X)^{-1}X^\intercal Y$ satisfies:

- $\mathbb{E}[\hat{\beta}^{\text{OLS}}] = \hat{\beta}$
- $\mathbb{V}[\hat{\beta}^{\text{OLS}}] = (X^\intercal X)^{-1}\sigma^2$

In particular, we note that the OLS estimator is unbiased!

A. [2p] Show that the RIDGE estimator $\hat{\beta}^{\text{RIDGE}} = (X^\intercal X + \lambda\,\mathbb{I})^{-1}X^\intercal y$ satisfies

- $\mathbb{E}[\hat{\beta}^{\text{RIDGE}}] = (X^\intercal X + \lambda\,\mathbb{I})^{-1}X^\intercal X\hat{\beta}$
- $\mathbb{V}[\hat{\beta}^{\text{RIDGE}}] = (X^\intercal X + \lambda\,\mathbb{I})^{-1}X^\intercal X(X^\intercal X + \lambda\,\mathbb{I})^{-1}\sigma^2$

In particular, we note that the RIDGE estimator is biased!

B. [3p] Given two covariance matrices Σ_A and Σ_B, we say that Σ_A is strictly greater than Σ_B (in symbols $\Sigma_A > \Sigma_B$) iff $\Sigma_A - \Sigma_B$ is positive definite. (This is the so called Löwner order). Show that $\hat{\beta}^{\text{OLS}}$ has stricly greater variance than $\hat{\beta}^{\text{RIDGE}}$

Hint: Note that $(X^\intercal X)^{-1}$ and $X^\intercal X + \lambda\,\mathbb{I}$ commute. More generally, if p and q are polynomial functions, then $p(A)q(A) = q(A)p(A)$ and likewise $q(A)^{-1}p(a) = p(A)q(A)^{-1}$ for any square matrix A.