Project 1

Isak O. Rukan* and Edvard B. Rørnes†
Institute of Physics, University of Oslo,
0371 Oslo, Norway
(Dated: September 24, 2024)

We investigate and compare three regression models[a]: Ordinary Least Squares, Ridge, and Least Absolute Shrinkage and Selection Operator. These were applied and analyzed on a two-dimensional Franke function and later tested on real terrain data from the US Geological Survey Earth Explorer [1]. The models' performances are evaluated using metrics such as the Mean Squared Error and the coefficient of determination R^2. The bias-variance trade-off for OLS is studied in detail where it was found that ... (TBD). We used resampling techniques such as bootstrapping and cross-validation to probe the quality of the evaluations and determine the predictiveness and generalizability of the models. In our analysis the best performing method was ... (TBD).

1. INTRODUCTION

Regression models are essential tools in data analysis and prediction, particularly within the realm of physics. They are used to enable understanding of the relationships between different variables and improve our ability to create predictions. In this report we study and compare three common regression techniques: Ordinary Least Squares (OLS) regression, Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression. Each come with their unique strengths and weaknesses that make them suitable for different data sets.

OLS is the simplest and most foundational method which estimates relationship by minimizing the difference between observed and predicted values. A large downside with OLS is when there are many related variables as this can lead to unstable coefficient estimates [2]. Ridge regression can partially fix this issue by adding a penalty to large coefficients, effectively shrinking them. This in turn creates a more stable model in the event of correlated variables. LASSO regression takes this a step further by once again shrinking coefficients, but also setting some of them to zero. This allows LASSO to effectively choose important variables, which may be helpful when pursuing a simpler model.

To compare these regression model, we use a two-dimensional Franke function which allows us to test the performance of each under controlled conditions. Later we apply each model to real terrain data which is obtained from [1].

The models are evaluated by considering their Mean Squared Error (MSE) and coefficient of determination R^2. Further to make our evaluation more reliable, resampling techniques such as bootstrapping and cross-validation are used. These methods provide a better understanding of how well each model performs on different data.

* Insert Email
† e.b.rornes@fys.uio.no
[a] GitHub Repository: `https://github.com/EdvardRornes/FYS-STK4155/tree/main/Project1`

2. THEORY

All lengthy claims made in this section are derived in Appendix A. The general structure of all our models is that we have some data set $\{x_i, y_i\}$ where $i \in \{0, 1..., n-1\}$ where x_i are independent variables whilst y_i are dependent variables. The data is assumed to be described by

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} \tag{1}$$

where f is some continuous function which takes $\boldsymbol{x}$ as input and $\boldsymbol{\varepsilon}$ is a normal distributed error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$. The function f will then be approximated with a model $\tilde{\boldsymbol{y}}$ in which we will consider a polynomial expansion with coefficients β_i:

$$\tilde{y}_i = \sum_{j=0}^{p-1} \beta_j x_i^j \tag{2}$$

defining the $n \times p$ design matrix $(\boldsymbol{X})_{ij} = (x_i)^j$ we can rewrite this as

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} \tag{3}$$

Further, each model will be defined with a different cost function $C(\boldsymbol{\beta})$ which we minimize to find the coefficients for each respective model.

2.1. OLS

OLS is a primitive method used in linear regression to estimate coefficients of a linear model. The cost function in OLS is simply defined as the residual sum of squares (RSS)

$$C_{\text{OLS}}(\boldsymbol{\beta}) = \text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \tilde{\boldsymbol{y}})^2 = (y_i - X_{ij}\beta_j)^2$$

where we employ the summation notation where repeated indices are summed over. As mentioned prior, the coefficients $\boldsymbol{\beta}$ are found by minimizing the cost function, i.e. taking the derivative w.r.t. $\boldsymbol{\beta}$. This results in

$$\boldsymbol{\beta}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

which yields the model

$$\tilde{\boldsymbol{y}}_{\text{OLS}} = \boldsymbol{X}\boldsymbol{\beta}_{\text{OLS}} \tag{4}$$

Assuming our data takes the form of (1) then the expectation value $\boldsymbol{y}$ is

$$\mathbb{E}(y_i) = \mathbb{E}(f(x_i)) = \boldsymbol{X}_{i,*}\boldsymbol{\beta}$$

since $\mathbb{E}(\varepsilon_i) = 0$ follows from its definition. The variance of $\boldsymbol{y}$ is given by

$$\text{Var}(y_i) = \text{Var}(\varepsilon_i^2) = \sigma^2$$

which yields $y_i \sim \mathcal{N}(\boldsymbol{X}_{i,*}\boldsymbol{\beta}, \sigma^2)$. The expectation value of the optimal parameters $\hat{\boldsymbol{\beta}}$ can be found to be

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \boldsymbol{\beta}.$$

with the variance

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

2.2. Ridge

Ridge regression is an extension of OLS where define the cost function as a modified version of the OLS cost function with an added penalty term which is proportional to the coefficients β_i^2:

$$C_{\text{Ridge}}(\boldsymbol{\beta}) = C_{\text{OLS}}(\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^2 \tag{5}$$

Here $\lambda \geq 0$ is a regularization parameter which controls the strength of this additional penalty. This regulator essentially drives the magnitude of these coefficients allowing for more tweaking in the parameter space. This parametrization of course includes the constraint that $\boldsymbol{\beta}^2 \leq t$ for some $t < \infty$ such that we can choose our arbitrary parameter $\lambda \geq 0$ to be sufficiently small s.t. the cost function (5) does not diverge. The optimal parameters for Ridge regressions can again be found by the same process as for OLS:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Here we can see that the effect of adding this penalty term is essentially taking $(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ when compared to the OLS case. In the past this was generally the starting point for Ridge regression in the cases where the matrix $\boldsymbol{X}^T\boldsymbol{X}$ was not invertible. A direct way of seeing the effect of the regulator is by considering the singular value decomposition of $\boldsymbol{X}$. Doing so one can show that

$$\tilde{\boldsymbol{y}}_{\text{Ridge}} = \sum_{j=0}^{p-1} \boldsymbol{u}_j\boldsymbol{u}_j^T \frac{\sigma_j^2}{\sigma_j^2 + \lambda}\boldsymbol{y}$$

Since $\lambda \geq 0$ then this added factor compared to OSL is ≤ 1. We can then see that Ridge regression effectively suppresses all the coefficients, thus λ is often called the "shrinkage" factor.

2.3. LASSO

Similarly to Ridge, LASSO also includes a penalty factor. The cost function in this case is instead defined to be

$$C_{\text{LASSO}}(\boldsymbol{\beta}) = C_{\text{OLS}}(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1 \tag{6}$$

where

$$||\boldsymbol{\beta}||_k \equiv \sum_{i=0}^{n-1} |\beta_i|^k$$

is the L^k norm of $\boldsymbol{\beta}$. Taking the derivative of (6) w.r.t. $\boldsymbol{\beta}$ and requiring that this becomes zero we have

$$0 = \frac{\partial C_{\text{LASSO}}}{\partial\boldsymbol{\beta}} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\,\text{sgn}(\boldsymbol{\beta}) \tag{7}$$

This has the added benefit of being able to set certain parameters to be 0 instead of suppressing them, at the cost of losing analytical expressions for $\hat{\boldsymbol{\beta}}$ in non-trivial cases.

2.4. Resampling

2.5. Bias-Variance

The so-called Bias-Variance Trade-Off can be summarized in a single equation:

$$\mathbb{E}[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2] = \text{Bias}[\tilde{\boldsymbol{y}}] + \text{Var}[\tilde{\boldsymbol{y}}] + \sigma^2 \tag{8}$$

The LHS of (8) is the expected value of the MSE which tells us how well the model's predictions match the true data on average. The equation shows that we can decompose this expected MSE into 3 different components.

- Bias: This quantity measures how much the model's average prediction differs from its true value. A high bias implies that the model is underfitting the data of is simply too simplistic.

- Var: The variance measures how much the model's predictions vary when trained on different datasets. It captures the sensitivity of the model to small changes in the training data. A high variance suggests overfitting, meaning it performs well on the training data but may be capturing noise or false patterns.

- σ^2: This is the irreducible error or noise in the data itself which cannot be explained by the model.

The idea is to minimize the LHS of (8), so clearly we want to minimize both the bias and the variance at the same time. However these are correlated to one another, so lowering the e.g. the bias will in general increase the variance. So Bias-Variance Tradeoff is essentially trying to optimize the complexity of the model such that we neither overfit nor underfit the model such that it can be generalized to other cases. These quantities can then be used as means to fine tune a model.

3. IMPLEMENTATION

For this project the surface we will consider is given by the Franke function

$$
\begin{aligned}
f(x,y) &= \frac{3}{4}\exp\left(-\frac{(9x-2)^2}{4}-\frac{(9y-2)^2}{4}\right)\\
&+\frac{3}{4}\exp\left(-\frac{(9x+1)^2}{49}-\frac{(9y+1)^2}{10}\right)\\
&+\frac{1}{2}\exp\left(-\frac{(9x-7)^2}{4}-\frac{(9y-3)^2}{4}\right)\\
&-\frac{1}{5}\exp\left(-(9x-4)^2-(9y-7)^2\right) \qquad (9)
\end{aligned}
$$

This function maps a surface defined on the interval $x, y \in [0,1]$. To perform an analysis on this function we consider a polynomial fit up to degree n where

$$
\begin{aligned}
\tilde{\boldsymbol{z}} &= \frac{1}{n+1}\sum_{i=0}^{n}\Big(\boldsymbol{\beta}_{00} \qquad (10)\\
&+\boldsymbol{\beta}_{10}x_i+\boldsymbol{\beta}_{11}y_i\\
&+\boldsymbol{\beta}_{20}x_i^2+\boldsymbol{\beta}_{21}x_iy_i+\boldsymbol{\beta}_{22}y_i^2\\
&+\ldots\\
&+\boldsymbol{\beta}_{n0}x_i^n+\boldsymbol{\beta}_{n1}x_i^{n-1}y+\ldots+\boldsymbol{\beta}_{n(n-1)}x_iy_i^{n-1}+\boldsymbol{\beta}_{nn}y_i^n\Big)\\
&=\frac{1}{n+1}\sum_{i,j=0}^{n}\sum_{k=0}^{i}\boldsymbol{\beta}_{jk}x_i^{j-k}y_i^k \equiv \boldsymbol{X}\boldsymbol{\beta}
\end{aligned}
$$

where the components x_i and y_i are entries in the input vectors $\boldsymbol{x}^T = \begin{bmatrix} x_0 & \ldots & x_n\end{bmatrix}$ and $\boldsymbol{y}^T = \begin{bmatrix} y_0 & \ldots & y_n\end{bmatrix}$ respectively which are our independent variables. Each $\boldsymbol{\beta}_{ij}$ is a $\frac{(n+1)(n+2)}{2}$ component vector with a single non-zero entry with magnitude β_{ij} and the design matrix $\boldsymbol{X}$ is then an $(n+1)\times\frac{(n+1)(n+2)}{2}$ matrix of the form:

$$
\boldsymbol{X} = \frac{1}{n+1}\begin{bmatrix}
1 & x_0 & y_0 & x_0^2 & x_0y_0 & y_0^2 & \ldots & x_0^n & \ldots & y_0^n\\
1 & x_1 & y_1 & x_1^2 & x_1y_1 & y_1^2 & \ldots & x_1^n & \ldots & y_1^n\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\
1 & x_n & y_n & x_n^2 & x_ny_n & y_n^2 & \ldots & x_n^n & \ldots & y_n^n
\end{bmatrix}
$$

and the $\boldsymbol{\beta}$ vector contains the $\frac{(n+1)(n+2)}{2}$ components

$$
\boldsymbol{\beta}^T = \begin{bmatrix}\beta_{00} & \beta_{10} & \beta_{11} & \beta_{20} & \beta_{21} & \beta_{22} & \ldots & \beta_{n(n-1)} & \beta_{nn}\end{bmatrix}
$$

It should now be clear which unit vectors correspond to each term in (10).

We then generated data with the Franke function and used an x-$(x-1)$ train-test split. This was chosen because... Next we

4. RESULTS

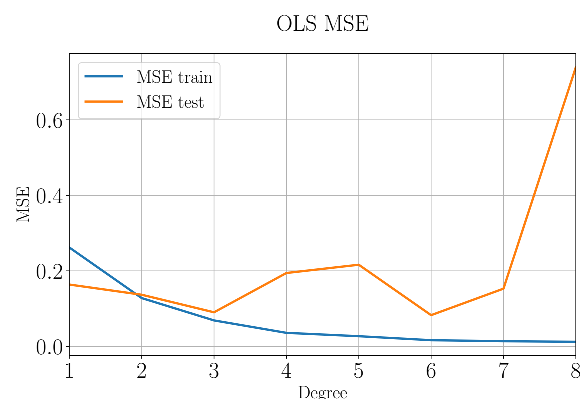

Fig. 1. Caption

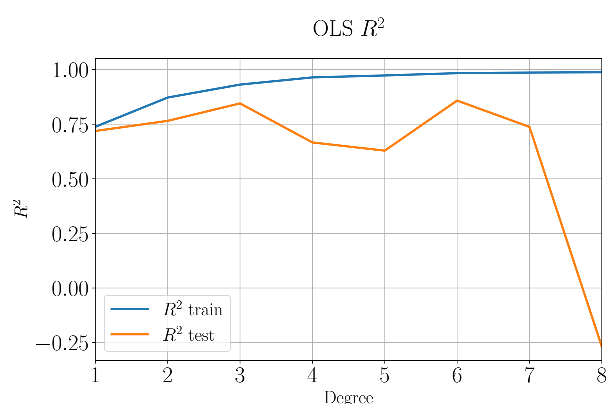

Fig. 2. Caption

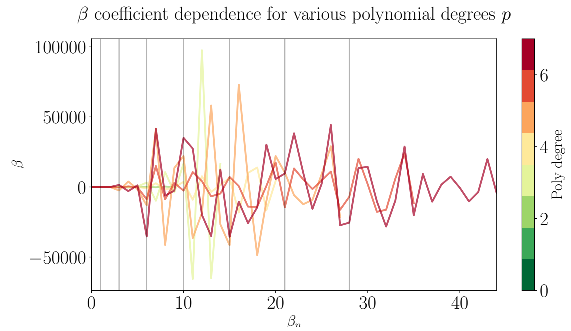

Fig. 3. Caption

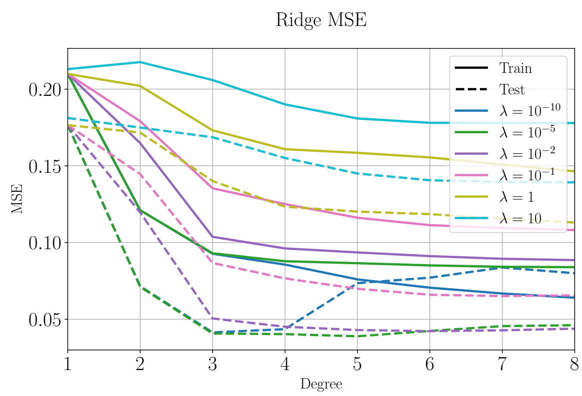

Fig. 4. Caption

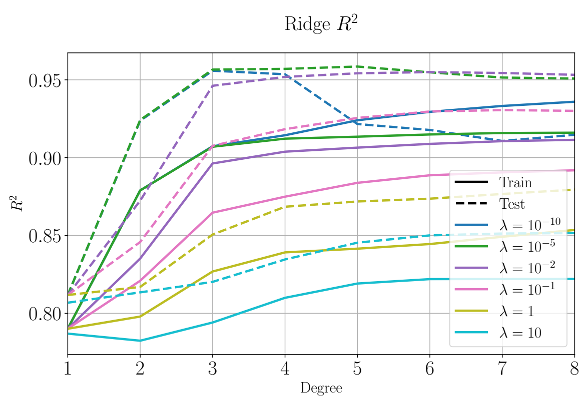

Fig. 5. Caption

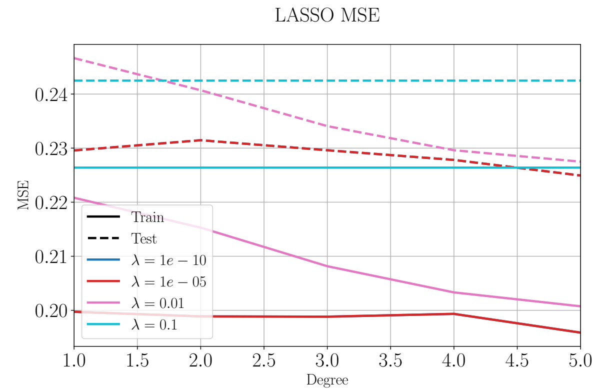

Fig. 6. Caption

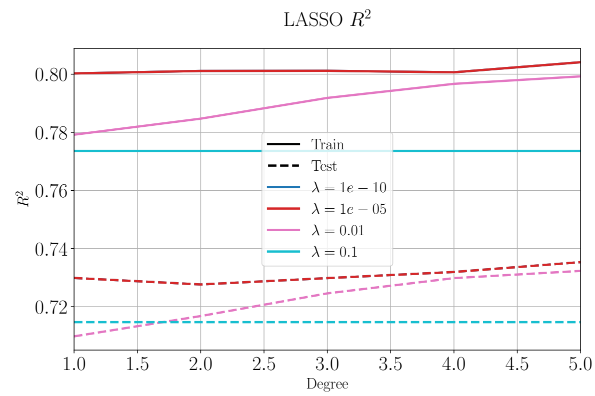

Fig. 7. Caption

4.1. OLS

4.2. Ridge

4.3. LASSO

5. RESULTS

6. DISCUSSION

7. CONCLUSION

---

[1] U.S. Geological Survey, "EarthExplorer: Terrain Elevation Data." `https://earthexplorer.usgs.gov`, 2024. Accessed: 20/09/2024. Extracted terrain elevation data.
[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 1. Springer, 2006.

Appendix A: Derivations

1. Model

The variance of $\boldsymbol{y}$ calculated as follows:

$$\begin{aligned}
\mathrm{Var}(y_i) &= \mathbb{E}\{[y_i - \mathbb{E}(y_i)]^2\} \\
&= \mathbb{E}\{(\boldsymbol{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i)^2\} - (\boldsymbol{X}_{i,*}\boldsymbol{\beta})^2 \\
&= (\boldsymbol{X}_{i,*}\boldsymbol{\beta})^2 + \mathbb{E}(\varepsilon_i^2) + 2\mathbb{E}(\varepsilon_i)\boldsymbol{X}_{i,*}\boldsymbol{\beta} - (\boldsymbol{X}_{i,*}\boldsymbol{\beta})^2 \\
&= \mathrm{Var}(\varepsilon_i^2) = \sigma^2
\end{aligned}$$

A direct way of seeing the effect of the regulator is by considering the singular value decomposition of $\boldsymbol{X}$. Writing $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$ where $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal and $\boldsymbol{\Sigma}$ only contains elements on the diagonal the proof goes as follows:

$$\begin{aligned}
\tilde{\boldsymbol{y}}_{\text{Ridge}} &= \boldsymbol{X}\boldsymbol{\beta}_{\text{Ridge}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T + \lambda\boldsymbol{I})^{-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\boldsymbol{I})\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \\
&= \sum_{j=0}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\boldsymbol{y}
\end{aligned}$$

where the last step is valid due to the orthogonality of $\boldsymbol{U}$ and σ_j are the elements on the diagonal of $\boldsymbol{\Sigma}$.

2. Coefficients

The expectation value of the optimal parameters $\hat{\boldsymbol{\beta}}$ can be found to be

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{y}] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.
\end{aligned}$$

with the variance

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) &= \mathbb{E}\{[\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta})][\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta})]^T\} \\
&= \mathbb{E}\{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\beta}] \\
&\quad \times [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\beta}]^T\} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}\{\boldsymbol{y}\boldsymbol{y}^T\}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&\quad - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T + \sigma^2]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&\quad - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= \boldsymbol{\beta}\boldsymbol{\beta}^T + \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

The optimal parameters for Ridge regressions can again be found by the same process as for OLS:

$$\begin{aligned}
0 &= \frac{\partial C_{\text{Ridge}}}{\partial \boldsymbol{\beta}} = -\frac{2}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T \\
&= \frac{2}{n}(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{y}^T\boldsymbol{X}) + 2\lambda\boldsymbol{\beta}^T \\
0 &= \boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \tilde{\lambda}\boldsymbol{I}) - \boldsymbol{y}^T\boldsymbol{X} \\
\boldsymbol{\beta}^T &= \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \tilde{\lambda}\boldsymbol{I})^{-1} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

where we defined $\tilde{\lambda} \equiv n\lambda$, renamed $\tilde{\lambda} \to \lambda$ and used that the matrix in the parenthesis is a symmetric matrix and thus its inverse must also be symmetric.

3. Bias-Variance

For ease of notation we write $f(\boldsymbol{x}) = f$ and simply ignore vector notation since everything is a scalar in the end. Then we have

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2] &= \mathbb{E}[(f + \boldsymbol{\varepsilon} - \tilde{\boldsymbol{y}})^2] \\
&= \mathbb{E}[(f - \tilde{\boldsymbol{y}})^2] + 2\underbrace{\mathbb{E}[(f - \tilde{\boldsymbol{y}})\boldsymbol{\varepsilon}]}_{=0} + \underbrace{\mathbb{E}[\boldsymbol{\varepsilon}^2]}_{=\sigma^2} \\
&= \mathbb{E}[((f - \mathbb{E}[\tilde{\boldsymbol{y}}]) - (\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}]))^2] + \sigma^2 \\
&= \mathbb{E}[(f - \mathbb{E}[\tilde{\boldsymbol{y}}]))^2] + \mathbb{E}[(\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}])^2] \\
&\quad - 2\,\mathbb{E}[((f - \mathbb{E}[\tilde{\boldsymbol{y}}])(\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}]))] + \sigma^2 \\
&= \text{Bias}[\tilde{\boldsymbol{y}}] + \text{Var}[\tilde{\boldsymbol{y}}] + \sigma^2 \\
&\quad - 2\,\mathbb{E}[(f - \mathbb{E}[\tilde{\boldsymbol{y}}])(\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}])]
\end{aligned}$$

where $\mathbb{E}[(f - \tilde{\boldsymbol{y}})\boldsymbol{\varepsilon}] = 0$ is justified by $\boldsymbol{\varepsilon}$ being independent and we note that the wrong definition of the Bias is given in the problem text (with that definition σ^2 gets put into the 'Bias'). All that remains is to show that the last term is 0. Since $\mathbb{E}[f] = f$ and $\mathbb{E}[f\,\mathbb{E}[\tilde{\boldsymbol{y}}]] = f\,\mathbb{E}[\mathbb{E}[\tilde{\boldsymbol{y}}]] = f\,\mathbb{E}[\tilde{\boldsymbol{y}}]$ then

$$\begin{aligned}
\mathbb{E}[(f - \mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{y}])] &= \mathbb{E}[f\tilde{\boldsymbol{y}} - f\,\mathbb{E}[\tilde{\boldsymbol{y}}] - \tilde{\boldsymbol{y}}\,\mathbb{E}[\tilde{y}] + \mathbb{E}^2[\tilde{\boldsymbol{y}}]] \\
&= f\,\mathbb{E}[\tilde{\boldsymbol{y}}] - f\,\mathbb{E}[\tilde{\boldsymbol{y}}] - \mathbb{E}^2[\tilde{\boldsymbol{y}}] + \mathbb{E}^2[\tilde{\boldsymbol{y}}] \\
&= 0
\end{aligned}$$

which proves the claim.

Appendix B