Proofs, Exercises and Literature - Regularization in Regression

1 Proofs

1.1 Deriving the Regression Solution if $X^\top X$ is not Invertible

The global minimizers $\boldsymbol{\beta}$ to the linear regression problem with design matrix X are given by

$$\{\boldsymbol{\beta} \in \mathbb{R}^p \mid X^\top X \boldsymbol{\beta} = X^\top \mathbf{y}\}. \tag{1}$$

If the matrix $X^\top X$ is invertible, then we can solve the system of linear equations directly and get

$$\boldsymbol{\beta} = (X^\top X)^{-1} X^\top \mathbf{y}.$$

However, if the matrix $X^\top X$ is not invertible, then there are infinitely many solutions of $\boldsymbol{\beta}$. We discuss now how we can derive the solution vectors $\boldsymbol{\beta}$ in this case.

The $(p \times p)$ matrix $X^\top X$ is not invertible if this matrix has $r < p$ nonzero singular values. The singular values of $X^\top X$ are specified by the SVD of $X = U\Sigma V^\top$, we have

$$X^\top X = V\Sigma^\top \underbrace{U^\top U}_{=I} \Sigma V^\top = V\Sigma^\top \Sigma V^\top.$$

The singular value decomposition is uniquely defined and the decomposition $V\Sigma^\top\Sigma V^\top$ satisfies the requirements for the singular value decomposition of $X^\top X$. Hence, the singular values of $X^\top X$ are given by the diagonal elements of the matrix $\Sigma^\top\Sigma$, which can be decomposed into an invertible part Σ_r^2 and a non-invertible part:

$$\Sigma^\top\Sigma = \left(\begin{array}{ccc:c} \sigma_1^2 & \dots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \dots & \sigma_r^2 & \\ \hdashline & \mathbf{0} & & \mathbf{0} \end{array}\right) = \left(\begin{array}{c:c} \Sigma_r^2 & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} \end{array}\right). \tag{2}$$

Given the singular value decomposition of X and $X^\top X$, we can try to solve Eq. (1) for $\boldsymbol{\beta}$:

$$X^\top X\boldsymbol{\beta} = X^\top \mathbf{y} \quad \Leftrightarrow \quad V\Sigma^\top\Sigma V^\top \boldsymbol{\beta} = V\Sigma^\top U^\top \mathbf{y} \quad \Leftrightarrow \quad \Sigma^\top\Sigma V^\top\boldsymbol{\beta} = \Sigma^\top U^\top \mathbf{y}, \tag{3}$$

where the last equality follows from multiplying with $V^\top$ from the left. If we want to solve for $\boldsymbol{\beta}$, then the next step would be to multiply with the inverse of the matrix $\Sigma^\top\Sigma$. However, this is not possible since only $r < p$ singular values are nonzero. We set

$$\boldsymbol{\beta} = VA\Sigma^\top U^\top \mathbf{y}, \tag{4}$$

for a $p \times p$ matrix A, which we will specify later. Substituting this definition of $\boldsymbol{\beta}$ in Eq. (3) yields

$$\begin{aligned} &\Sigma^\top\Sigma V^\top(VA\Sigma U^\top \mathbf{y}) = \Sigma^\top U^\top \mathbf{y} \\ \Leftrightarrow\ &(\Sigma^\top\Sigma)A\Sigma^\top U^\top \mathbf{y} = \Sigma^\top U^\top \mathbf{y}. \end{aligned}$$

That is, if the matrix A satisfies $(\Sigma^\top\Sigma)A\Sigma^\top = \Sigma^\top$, then the vector $\boldsymbol{\beta}$ in Eq. (4) is a global minimizer of the regression objective. We specify now for which matrices A this equality is satisfied. Therefore, we write A as a collection of four sub-matrices: $A_1 \in \mathbb{R}^{r\times r}, A_2 \in \mathbb{R}^{r\times(p-r)}, A_3 \in \mathbb{R}^{(p-r)\times r}$ and $A_4 \in \mathbb{R}^{(p-r)\times(p-r)}$:

$$\begin{aligned} (\Sigma^\top\Sigma)A\Sigma^\top &= p\Bigg\{\underbrace{\left(\begin{array}{c|c} \Sigma_r^2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)}_{p}\underbrace{\left(\begin{array}{c|c} A_1 & A_2 \\ \hline A_3 & A_4 \end{array}\right)}_{p}\underbrace{\left(\begin{array}{c|c} \Sigma_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)}_{n} \\ &= \left(\begin{array}{c|c} \Sigma_r^2 A_1 & \Sigma_r^2 A_2 \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)\left(\begin{array}{c|c} \Sigma_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right) \\ &= \left(\begin{array}{c|c} \Sigma_r^2 A_1 \Sigma_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right). \end{aligned}$$

We can see that the matrix above equals $\Sigma^\top$ if $A_1 = \Sigma_r^{-2}$:

$$(\Sigma^\top\Sigma)A\Sigma^\top = \Sigma^\top \Leftrightarrow A_1 = \Sigma_r^{-2}.$$

As a result, we get for any $p \times p$ matrix of the form

$$A = \left(\begin{array}{c|c} \Sigma_r^{-2} & A_2 \\ \hline A_3 & A_4 \end{array}\right)$$

a global minimizer $\boldsymbol{\beta}$ in the form of Eq. (4). In fact, only the sub-matrix A_3 has an impact on the definition of $\boldsymbol{\beta}$, since

$$\begin{aligned} \boldsymbol{\beta} &= VA\Sigma^\top U^\top \mathbf{y} \\ &= V\left(\begin{array}{c|c} \Sigma_r^{-2} & A_2 \\ \hline A_3 & A_4 \end{array}\right)\left(\begin{array}{c|c} \Sigma_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)U^\top \mathbf{y} \\ &= V\left(\begin{array}{c|c} \Sigma_r^{-1} & \mathbf{0} \\ \hline A_3\Sigma & \mathbf{0} \end{array}\right)U^\top \mathbf{y}. \end{aligned}$$

Hence, the set of all global minimizers is defined by the solutions $\boldsymbol{\beta}$ given in Eq. (4) for a matrix A of the form

$$A = \left(\begin{array}{c|c} \Sigma_r^{-2} & \mathbf{0} \\ \hline A_3 & \mathbf{0} \end{array}\right),$$

where $A_3 \in \mathbb{R}^{(p-r)\times r}$, and the last $p - r$ columns of A are equal to zero.

1.2 The Minimizer of Ridge Regression is Unique

The uniqueness of the regression vector $\boldsymbol{\beta}$ for ridge regression follows from the fact that the matrix $X^\top X + \lambda I$ is invertible for any $\lambda > 0$. To show this, we use the SVD of $X = U\Sigma V^\top$. We get then

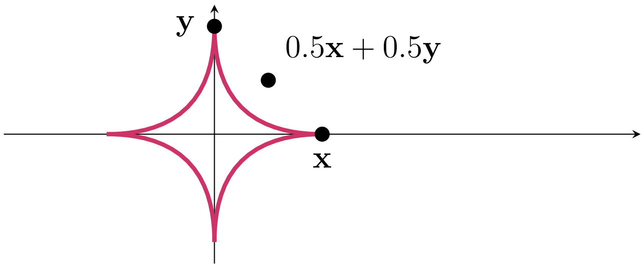
$$\begin{aligned} X^\top X + \lambda I &= V\Sigma^\top \underbrace{U^\top U}_{=I} \Sigma V^\top + \lambda \underbrace{I}_{=VV^\top} \\ &= V\Sigma^\top\Sigma V^\top + \lambda V V^\top \\ &= V(\Sigma^\top\Sigma + \lambda I)V^\top. \end{aligned}$$

The last equation denotes the SVD of the matrix $X^\top X+\lambda I$. The singular values of $X^\top X+\lambda I$ are denoted by the elements on the diagonal of the diagonal matrix $\Sigma^\top\Sigma + \lambda I$. Since these singular values are all nonzero (they are all larger than equal to $\lambda > 0$), the matrix $X^\top X+\lambda I$ is invertible.

2 Exercises

1. Show that the L_p-'norm' is not a real norm for $p = 1/2$.

Solution:



$p = 1/2$

We choose the points $\mathbf{x} = (0, 1)$ and $\mathbf{y} = (1, 0)$. Both points are on the unit circle of the $L_{1/2}$-norm, that is

$$\|\mathbf{x}\|_{\frac{1}{2}} = (|x_1|^{\frac{1}{2}} + |x_2|^{\frac{1}{2}})^2 = \|\mathbf{y}\|_{\frac{1}{2}} = 1.$$

However, the convex combination $0.5\mathbf{x} + 0.5\mathbf{y} = (0.5, 0.5)$ is not on the unit circle:

$$\|0.5\mathbf{x} + 0.5\mathbf{y}\|_{\frac{1}{2}} = (|0.5|^{\frac{1}{2}} + |0.5|^{\frac{1}{2}})^2 = 2 > 1. \tag{5}$$

The inequality above shows that the $p = 1/2$-norm is not a real norm. If the $p = 1/2$-norm would be a real norm, then we would have

$$\begin{aligned} \|0.5\mathbf{x} + 0.5\mathbf{y}\|_{\frac{1}{2}} &\leq \|0.5\mathbf{x}\|_{\frac{1}{2}} + \|0.5\mathbf{y}\|_{\frac{1}{2}} && \text{(triangle inequality)} \\ &= 0.5\|\mathbf{x}\|_{\frac{1}{2}} + 0.5\|\mathbf{y}\|_{\frac{1}{2}} && \text{(homogenity)} \\ &= 1. \end{aligned}$$

However, Eq. (5) shows that this is not the case, hence the $p = 1/2$-norm is not a real norm.

3 Recommended Literature

As always, the best exercise is to go through the lecture and see if you can follow the steps (maybe with pen and paper). If you feel like reading, I can recommend the following two chapters from the recommended books:

Bishop. Pattern recognition and machine learning. 2006.

3.1.4 Regularized Least Squares

Friedman, Hastie, and Tibshirani. The elements of statistical learning. 2001.

3.4 Shrinkage Methods

- **3.4.1** Ridge Regression
- **3.4.2** The Lasso
- **3.4.3** Discussion: Subset Selection, Ridge Regression and the Lasso

Linear Algebra and Optimization for Machine Learning by Charu C. Aggarwal

5.8.1 The Subgradient Method

- **5.8.1.1** Application: L1-Regularization
- **5.8.1.2** Combining Subgradients with Coordinate Descent