2.1.2 Regularized Linear Models

Now we will briefly discuss so-called regularized linear models. In the case of a very large number of input variables, one may only be interested in a few variables that have the strongest influence. Thus, the model becomes

$$\min_{\beta \in \mathbb{R}^{n \times m}} \frac{1}{2} ||X\beta - Y||^2 \text{ such that most entries of } \beta \text{ are small or zero}$$

To enforce that the entries of β are small, we add a penalty term and obtain

2.1.4

$$\min_{\beta \in \mathbb{R}^{n \times m}} \frac{1}{2} ||X\beta - Y||^2 + \frac{\lambda}{2} ||\beta||^2$$

where $\lambda \geq 0$ is a freely selectable parameter. This model is also called Ridge Regression. Here, $|| \cdot ||$ denotes the Frobenius norms, but we omit the corresponding designation.
What is remarkable about this model is that for all $\lambda > 0$, it has a unique solution, regardless of the invertibility of $X^T X$

Proposition 2.1.4

For all $\lambda > 0$, the problem 2.1.4 has a unique solution given by

2.1.5

$$\hat{\beta}_\lambda := (X^T X + \lambda \mathbb{I})^{-1} X^T Y$$

Also, $\hat{\beta}_\lambda$ represents the **Lasso estimate** of the regression coefficient β for a given regularization parameter λ. Note that β is the true coefficient in a regression model, $\hat{\beta}_\lambda$ is the estimated value of β obtained by solving the Lasso optimization problem.
(This is derived by expanding the squared norms from 2.1.4, then compute the gradient w.r.t beta and rearranging terms)

Expanding

The squared Frobenius norm $\|X\beta - Y\|^2$ expands as follows:

$$\|X\beta - Y\|^2 = \text{Trace}\left((X\beta - Y)^T(X\beta - Y)\right).$$

Expanding the product:

$$(X\beta - Y)^T(X\beta - Y) = \beta^T X^T X\beta - \beta^T X^T Y - Y^T X\beta + Y^T Y.$$

Taking the trace (linear operator):

$$\|X\beta - Y\|^2 = \underbrace{\text{Trace}(\beta^T X^T X\beta)}_{\text{Quadratic Term}} - \underbrace{2\text{Trace}(Y^T X\beta)}_{\text{Linear Term}} + \underbrace{\text{Trace}(Y^T Y)}_{\text{Constant}}.$$

Key Simplifications:

1. **Quadratic Term**:
   $\text{Trace}(\beta^T X^T X\beta) = \beta^T X^T X\beta$ (if β is a vector).
2. **Linear Term**:
   $\text{Trace}(Y^T X\beta) = Y^T X\beta$ (if β is a vector).
3. **Constant Term**:
   $\text{Trace}(Y^T Y) = \|Y\|^2$.

For simplicity, assume β is a vector ($m = 1$). Then:

$$\|X\beta - Y\|^2 = \beta^T X^T X\beta - 2Y^T X\beta + Y^T Y.$$

2. Expanding the Regularization Term

The regularization term $\|\beta\|^2$ is the Frobenius norm squared of β:

$$\|\beta\|^2 = \beta^T \beta \quad (\text{for } \beta \text{ as a vector}).$$

---

3. Full Loss Function

Substitute the expanded terms into the loss function:

$$\mathcal{L}(\beta) = \frac{1}{2}\left(\beta^T X^T X\beta - 2Y^T X\beta + Y^T Y\right) + \frac{\lambda}{2}\beta^T \beta.$$

---

4. Computing the Gradient

To find the optimal β, take the derivative of $\mathcal{L}(\beta)$ with respect to β and set it to zero:

1. **Quadratic Term**:
$$\frac{\partial}{\partial \beta} \left(\frac{1}{2} \beta^T X^T X \beta \right) = X^T X \beta.$$
2. **Linear Term**:
$$\frac{\partial}{\partial \beta} \left(-Y^T X \beta \right) = -X^T Y.$$
3. **Regularization Term**:
$$\frac{\partial}{\partial \beta} \left(\frac{\lambda}{2} \beta^T \beta \right) = \lambda \beta.$$
4. **Constant Term**:
$\frac{\partial}{\partial \beta} \left(\frac{1}{2} Y^T Y \right) = 0.$

Combine these results:

$$\nabla \mathcal{L}(\beta) = X^T X \beta - X^T Y + \lambda \beta = 0.$$

---

5. Solving for β

Rearrange the equation:

$$(X^T X + \lambda \mathbb{I}) \beta = X^T Y.$$

Since $X^T X + \lambda \mathbb{I}$ is always invertible for $\lambda > 0$, the solution is:

$$\boxed{\hat{\beta}_\lambda = (X^T X + \lambda \mathbb{I})^{-1} X^T Y}.$$

Back to the script

Proof

The objective function above 2.1.4 is convex, thus the optimality condition is

2.1.6

$$0 = X^T (X\hat{\beta} - Y) + \lambda \hat{\beta} = (X^T X + \lambda \mathbb{I}) \hat{\beta} - X^T Y$$

(Note that the first part $0 = X^T(X\hat{\beta} - Y) + \lambda\hat{\beta}$ is the gradient which must be 0)which is necessary and sufficient for minimality. We assert that the square matrix $X^TX + \lambda\mathbb{I}$ is invertible. To show this, it suffices to check injectivity. Thus, assume that $(X^TX + \lambda\mathbb{I})\beta = 0$ or equivalently $X^TX\beta = -\lambda\beta$. Then we obtain

$$-\lambda||\beta||^2 = \langle X^TX\beta, \beta\rangle = \langle X\beta, X\beta\rangle = ||X\beta||^2 \geq 0$$

Since $\lambda > 0$, this is only possible if $\beta = 0$, which means injectivity. Thus, the optimality condition 2.1.6 has the unique solution $\hat{\beta}_\lambda = (X^TX + \lambda\mathbb{I})^{-1}X^TY$We can also express the solution 2.1.5 using the pseudoinverse. Let $X = U\Sigma V^T$ be a singular value decomposition of X. Then we have

$$\begin{aligned} X^TX + \lambda\mathbb{I} &= (U\Sigma V^T)^T(U\Sigma V^T) + \lambda VV^T \\ &= V\Sigma^TU^TU\Sigma V^T + \lambda VV^T \\ &= V\Sigma^T\Sigma V^T + \lambda VV^T \\ &= V(\Sigma^T\Sigma + \lambda)V^T \end{aligned}$$

The matrix in parantheses has the following form:

$$\Sigma^T\Sigma + \lambda = \text{diag}(\sigma_1^2 + \lambda, \ldots, \sigma_k^2 + \lambda, \lambda \ldots, \lambda)$$

In particular, we obtain

$$\begin{aligned} (X^TX + \lambda\mathbb{I})^{-1}X^T &= V\text{diag}(\frac{1}{\sigma_1^2 + \lambda}, \ldots, \frac{1}{\sigma_k^2 + \lambda}, \frac{1}{\lambda}, \ldots \frac{1}{\lambda})V^T \cdot (U\Sigma V^T)^T \\ &= V\text{diag}(\frac{1}{\sigma_1^2 + \lambda}, \ldots, \frac{1}{\sigma_k^2 + \lambda}, \frac{1}{\lambda}, \ldots \frac{1}{\lambda})V^TV\text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)U^T \\ &= \text{now } \mathrm{V}^TV \text{ cancels out and we then multiply diag with the other diag:} \\ &= V\text{diag}(\frac{\sigma_1}{\sigma_1 + \lambda}, \cdots, \frac{\sigma_k}{\sigma_k + \lambda}, 0, \ldots, 0)U^T \end{aligned}$$

Corollary 2.1.1

As $\lambda \to 0$, the solution $\hat{\beta}_\lambda$ of 2.1.4 converges to the minimum-norm solution $\hat{\beta} := (X^TX)^\dagger X^TY$

Proof

According to the above calculation, we have

$$= V\text{diag}(\frac{\sigma_1}{\sigma_1 + \lambda}, \cdots, \frac{\sigma_k}{\sigma_k + \lambda}, 0, \ldots, 0)U^T$$

As $\lambda \to 0$ we have

$$\lim_{\lambda \to 0} \frac{\sigma_i}{\sigma_i^2 + \lambda} = \frac{\sigma_i}{\sigma_i^2} = \frac{1}{\sigma_i}$$

Thus:

$$\begin{aligned}\lim_{\lambda \to 0}(X^TX + \lambda \mathbb{I})^{-1}X^T &= V\text{diag}(\sigma_1^{-1}, \dots, \sigma_k^{-1}, 0, \dots, 0)U^T \\ &= (X^TX)^{\dagger}X^T\end{aligned}$$

Thus, the claim follows from Proposition 2.1.3
It turns out that the Frobenius/Euclidian norms in 2.1.4 are not suitable for reducing the entries of $\hat{\beta}_\lambda$ to zero.

Example 2.1.1

The solution of

$$\min_{\beta \in \mathbb{R}} \frac{1}{2}(x\beta - y)^2 + \frac{\lambda}{2}\beta^2$$

is given by the derivative (chain rule)

$$x(x\beta - y) + \lambda\beta = 0$$

Now solving for β

$$\begin{aligned}x^2\beta - xy + \lambda\beta &= 0 \\ \Rightarrow \beta(x^2 + \lambda) &= xy \quad | \div (x^2 + \lambda) \\ \hat{\beta}_\lambda &= \frac{xy}{x^2 + \lambda}\end{aligned}$$

This can easily be seen by differentiation. Thus, for all $\lambda \in [0, \infty)$, it holds that $\hat{\beta}_\lambda \neq 0$.In contrast, let us consider the problem

$$\min_{\beta \in \mathbb{R}} \frac{1}{2}(x\beta - y)^2 + \lambda|\beta|$$

This problem can be equivalently reformulated as

$$\min_{\beta \in \mathbb{R}} f(\beta)$$

where

$$\begin{aligned} f(\beta) &= \frac{1}{2}(x^2\beta^2 - 2yx\beta + y^2) + \lambda|\beta| \\ &= \frac{1}{2}x^2\beta^2 - xy\beta + \frac{1}{2}y^2 + \lambda|\beta| \qquad\qquad |\div x^2 \\ &= \frac{1}{2}\beta^2 - \frac{y}{x}\beta + \frac{y^2}{2x^2} + \frac{\lambda}{x^2}|\beta| \\ &\text{since constants do not affect the minimizer, we drop} \frac{y^2}{2x^2} \\ f(\beta) &:= \frac{1}{2}\beta^2 - \frac{y}{x}\beta + \frac{\lambda}{x^2}|\beta| \end{aligned}$$

Suppose that $\frac{y}{x} > 0$. Then the solution $\hat{\beta}_\lambda > 0$ because otherwise, putting $\beta = -\beta_\lambda$ would yield $f(\beta) < f(\hat{\beta}_\lambda)$. Then for $\beta \geq 0$, we can rewrite

$$f(\beta) := \frac{1}{2}\beta^2 - \frac{y}{x}\beta + \frac{\lambda}{x^2}\beta$$

Differentiation leads to

$$f'(\beta) = \beta - \left(\frac{y}{x} - \frac{\lambda}{x^2}\right)$$

If $(\frac{y}{x} - \frac{\lambda}{x^2}) < 0$ then $f'(\beta) > 0$ for all $\beta > 0$ hence f is increasing on $[0, +\infty]$ and the minimum is attained for $\hat{\beta}_\lambda = 0$. On the other hand, if $(\frac{y}{x} - \frac{\lambda}{x^2}) > 0$, then $f'(\hat{\beta}_\lambda) = 0 \implies \hat{\beta}_\lambda = \frac{y}{x} - \frac{\lambda}{x^2}$. We can then pout

$$\hat{\beta}_\lambda = \max\left\{\frac{y}{x} - \frac{\lambda}{x^2}, 0\right\}$$

Now supposing that $\frac{y}{x} \leq 0$ and using the same reasoning, we can show that

$$\hat{\beta}_\lambda = -\max\left\{-\frac{y}{x} - \frac{\lambda}{x^2}, 0\right\}$$

Using this formula, we see that for $\lambda \geq |xy|$, the solution $\hat{\beta}_\lambda$ is equal to zero.
Motivated by this example, instead of 2.1.4, we now consider the following model

2.1.7 (LASSO regression)

$$\min_{\beta \in \mathbb{R}^{n \times m}} \frac{1}{2}||X\beta - Y||^2 + \lambda||\beta||_1$$

where $||\beta||_1 := \sum_{i,j} |\beta_{i,j}|$. This model is also known as LASSO regression, where LASSO stands for *Least Absolute Shrinkage and Selection Operator*Since the objective function above is not differentiable, we cannot derive optimality conditions using our methods. Moreover, (except for the uninteresting case where X is an orthogonal matrix) it it is not possible to derive an explicit solution. Solutions must be approximated using an iterative procedure. Here, we

consider the **ISTA** (**I**terative **S**oft **T**hresholding **A**lgorithm), which successively reduces the first and second terms in 2.1.7. For a step size $\tau > 0$, one computes:

2.1.8 (ISTA Algorithm Steps)

$$\begin{cases} \text{Initialize } \beta^0 \\ \beta^{k-1/2} := \beta^{k-1} - \tau X^T(X\beta^{k-1} - Y) \\ \beta^k := \mathcal{S}_{\tau\lambda}\left(\beta^{k-1/2}\right) \end{cases}$$

As an initialization one might choose for example, $\beta^0 = 0$ or $\beta^0 = (X^T X)^\dagger X^T Y$.The operator $\mathcal{S}_\lambda$ is the so-called *Soft Thresholding* operator, which is applied entry-wise and is defined as follows:

2.1.9

$$\mathcal{S}_\mu(x) = \operatorname{sgn}(x)\max\{|x| - \mu, 0\}$$

The optimal step size τ, which guarantees the convergence of 2.1.8 to the solution of 2.1.7 can be precomputed and is given by

2.1.10 (step size computation)

$$\tau := \frac{1}{\sigma_1(X)^2}$$

where $\sigma_1(X)^2$ is the largest singular value of X. An interesting special case is $\lambda = 0$ and thus $\mathcal{S}_{\tau\lambda} = \mathrm{id}$. In this case, 2.1.8 reduces to

$$\begin{cases} \text{Initialize } \beta^0 \\ \beta^k := \beta^{k-1} - \tau X^T(X\beta^{k-1} - Y) \quad k \in \mathbb{N} \end{cases}$$

which is a very slow algorithm for iteratively solving the linear system $X\beta = Y$. The use of the pseudoinverse or better iterative methods is preferred for $\lambda = 0$