Chapter 5

Linear Regression

5.1 Introduction

The goal of *regression* is to predict a real value y as a function of the input variable $\boldsymbol{x}$. (The vector $\boldsymbol{x}$ is referred to as the feature vector, while y is called the target variable.) For example, in a marketing campaign, we may be interested in predicting total sales, given ad budgets in various platforms based on prior experience. Our data is a set of pairs $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, collected from n prior ad campaigns for n previous products,

Product		TV ads	Print ads	Web ads	Sale
1	$\boldsymbol{x}_1^T =$	\$20k	\$10k	\$10k	$y_1 = \$500\text{k}$
$\vdots$			$\vdots$		$\vdots$
i	$\boldsymbol{x}_i^T =$	x_{i1}	x_{i2}	x_{i3}	y_i
$\vdots$			$\vdots$		$\vdots$
n	$\boldsymbol{x}_n^T =$	x_{n1}	x_{n2}	x_{n3}	y_n

If we can predict y for any given value of $\boldsymbol{x}$, we can predict the outcome of a marketing campaign or optimize the marketing budget.We can also study what types of ads are more helpful, etc.

In *linear regression* our prediction for y is $\hat{y} = \boldsymbol{x}^T\boldsymbol{\theta}$, where $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are elements of $\mathbb{R}^d$. In our marketing example, our goal becomes to find $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ such that $\hat{y} = \boldsymbol{x}^T\boldsymbol{\theta} = \boldsymbol{\theta}^T\boldsymbol{x} = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ is a good predictor for y.

From a probabilistic standpoint, we may consider each $(\boldsymbol{x}_i, y_i)$ to be an independent realization of the random pair $(\boldsymbol{X}, Y)$ with some joint distribution $p_{\boldsymbol{X},Y}$. We then formulate the linear regression problem as follows: Find

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[L(Y, \boldsymbol{X}^T\boldsymbol{\theta})], \tag{5.1}$$

for a given loss function L. As we typically do not have the joint distribution for $\boldsymbol{X}, Y$, we aim to find

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, \boldsymbol{x}_i^T\boldsymbol{\theta}). \tag{5.2}$$

The linear form, $\hat{y} = \boldsymbol{x}^T\boldsymbol{\theta} = \sum_{j=1}^{d} \theta_j x_j$, may appear restrictive since it apparently excludes dependence on, for example, x_j^2. Imagine, in our marketing example, that buyers are likely to purchase a product if they see both TV ads and Web ads. In other words, y is large when $x_1 x_3$ is large. It seems that this case is not covered well by linear regression. However, this is not the case since we can transform the input variable using a set of functions $g_1, \ldots, g_e$ and reformulate our assumption as $\hat{y} = \sum_{j=1}^{e} \theta_j g_j(\boldsymbol{x})$, where g_j are any

functions of $\boldsymbol{x}$, such as x_1^2 and $x_1 x_3$. (But finding approrpiate fatures is a challenging problem.) Note that the expression $\hat{y} = \sum_{j=1}^{e} \theta_j g_j(\boldsymbol{x})$ is still linear in $\boldsymbol{\theta}$, which is what matters, since we need to optimize $\boldsymbol{\theta}$.

Notation. Define $\mathsf{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y}$ as

$$\mathsf{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{5.3}$$

For a given value of $\boldsymbol{\theta}$, we let $\hat{\boldsymbol{y}} = \mathsf{X}\boldsymbol{\theta}$ to be the predicted value. Furthermore, let $\boldsymbol{\epsilon}$ be the error vector such that

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{\epsilon} = \mathsf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{5.4}$$

Example 5.1. Suppose

$$\boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{x}_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \qquad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{5.5}$$

$$y_1 = -1, \qquad y_2 = 1, \qquad y_3 = 0. \tag{5.6}$$

Then

$$\mathsf{X} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \hat{\boldsymbol{y}} = \mathsf{X}\boldsymbol{\theta} = \theta_1 \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} + \theta_2 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{pmatrix} -1 - \theta_2 \\ 1 - 2\theta_1 \\ -\theta_1 - \theta_2 \end{pmatrix}. \tag{5.7}$$

△

5.2 Least-squares

A common choice for the loss function is

$$L(y_i, \boldsymbol{x}_i^T \boldsymbol{\theta}) = (y_i - \boldsymbol{x}_i^T \boldsymbol{\theta})^2. \tag{5.8}$$

The empirical risk can be written as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\theta})^2 = \|\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}\|_2^2, \tag{5.9}$$

where we have dropped the $1/n$ factor present in (5.2) as it does not affect our choice of $\boldsymbol{\theta}$. Denote

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \tag{5.10}$$

and define $\hat{\boldsymbol{y}} = \mathsf{X}\hat{\boldsymbol{\theta}}$ as the predicted value or estimate based on the model.

Least-squares is relatively easy to deal with from a computational perspective. It also has the same solution as the MLE for a common probabilistic model as we will see, thus providing an additional rationale for the resulting approach.

Projection onto the column space of X. Our first observation is that $\hat{\boldsymbol{y}}$ is in the column space of X, i.e., it is a linear combination of the columns of X. We can thus restate our goal as finding $\hat{\boldsymbol{y}}$ in the column space of X such that $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|$ is minimized. Hence, $\hat{\boldsymbol{y}}$ is the projection of $\boldsymbol{y}$ onto the column space of X as shown in Figure 5.1. Then, from the Projection Lemma in the Appendix, $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to each column of X.

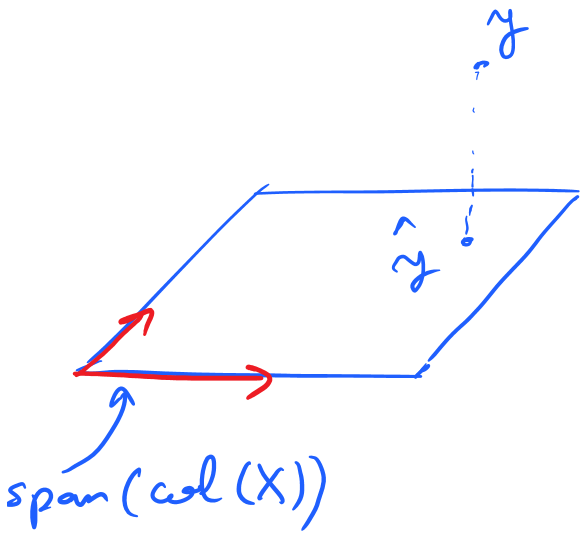


Figure 5.1: Error is minimized by projecting $\boldsymbol{y}$ onto the column space of X, Span(col(X)).

This orthogonality of the error to columns of X can be written as $\mathsf{X}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$. We have

$$\mathsf{X}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0 \iff \mathsf{X}^T(\boldsymbol{y} - \mathsf{X}\hat{\boldsymbol{\theta}}) = 0 \tag{5.11}$$

$$\iff \mathsf{X}^T\boldsymbol{y} = \mathsf{X}^T\mathsf{X}\hat{\boldsymbol{\theta}} \tag{5.12}$$

$$\iff \hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y} \tag{5.13}$$

Here we have assumed that $\mathsf{X}^T\mathsf{X}$ is invertible. This holds if the columns of X are linearly independent. To see this, we will show $\mathsf{X}^T\mathsf{X}\boldsymbol{\alpha} = \mathbf{0}$ implies that $\boldsymbol{\alpha} = \mathbf{0}$ if the columns of X are linearly independent:

$$\mathsf{X}^T\mathsf{X}\boldsymbol{\alpha} = \mathbf{0} \Rightarrow \boldsymbol{\alpha}^T\mathsf{X}^T\mathsf{X}\boldsymbol{\alpha} = 0 \Rightarrow (\mathsf{X}\boldsymbol{\alpha})^T(\mathsf{X}\boldsymbol{\alpha}) = 0 \Rightarrow \mathsf{X}\boldsymbol{\alpha} = \mathbf{0} \Rightarrow \boldsymbol{\alpha} = \mathbf{0}, \tag{5.14}$$

where the last step follows from the fact that the columns of X are linearly independent.

Example 5.2. From Example 5.1, we have

$$\mathsf{X} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \tag{5.15}$$

and so

$$\mathsf{X}^T\mathsf{X} = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}, \qquad (\mathsf{X}^T\mathsf{X})^{-1} = \frac{1}{9}\begin{pmatrix} 2 & -1 \\ -1 & 5 \end{pmatrix} \tag{5.16}$$

$$(\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T = \frac{1}{9}\begin{pmatrix} -1 & 4 & 1 \\ 5 & -2 & 4 \end{pmatrix} \qquad \hat{\boldsymbol{\theta}} = \begin{pmatrix} 5/9 \\ -7/9 \end{pmatrix} \tag{5.17}$$

$$\hat{\boldsymbol{y}} = \begin{pmatrix} -7/9 \\ 10/9 \\ -2/9 \end{pmatrix}, \qquad \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{pmatrix} -2/9 \\ -1/9 \\ 2/9 \end{pmatrix}. \tag{5.18}$$

△

Gradient descent. The closed-form solution provided in (5.13) for finding $\hat{\boldsymbol{\theta}}$ requires taking matrix inverses of possibly very large matrices, which could be computationally expensive. A less expensive solution is gradient descent, where we take the derivative of the loss to minimize it. Let $\nabla\mathcal{L}(\boldsymbol{\theta}) = \left(\frac{d\mathcal{L}}{d\boldsymbol{\theta}}\right)^T$ be the

gradient of $\mathcal{L}$. Recall that the direction of the gradient indicates the direction of maximum increase and its magnitude represents the slope of the increase. We have

$$\mathcal{L} = (\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}), \tag{5.19}$$

$$\nabla\mathcal{L} = 2\big[(\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta})^T(-\mathsf{X})\big]^T = -2\mathsf{X}^T(\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}). \tag{5.20}$$

(Setting the gradient equal to 0 again gives $\hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}$. Note that the Hessian is $\mathsf{X}^T\mathsf{X}$, which is positive-semi-definite.) In gradient descent, we start from an arbitrary value $\boldsymbol{\theta}^{(0)}$ and move towards the solution in steps:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \rho\mathsf{X}^T(\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}^{(t)}) \tag{5.21}$$

$$= \boldsymbol{\theta}^{(t)} + \rho\sum_{i=1}^{n}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}^{(t)}), \tag{5.22}$$

where ρ is the learning rate. This approach gets to the lowest point by moving in the direction of the *steepest descent* as shown in figure below for Example 5.1.

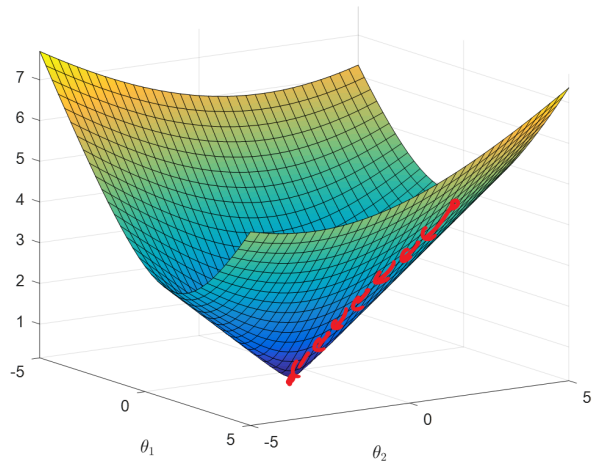


Figure 5.2: Gradient descent for linear regression

Standardization

We sometimes assume that X is standardized, meaning that each column $\boldsymbol{v}$ is shifted and scaled such that $\boldsymbol{v}^T\mathbf{1} = 0$ and $\boldsymbol{v}^T\boldsymbol{v} = 1$ and that $\boldsymbol{y}$ is centered so that $\boldsymbol{y}^T\mathbf{1} = 0$. Standardization of the inputs puts different features under the same scale and can help to reduce the correlation between features when having polynomial/interaction terms. Standardizing inputs can also be shown to be equivalent to minimizing the squared loss with an intercept term.

Example 5.3 (†)**.** We show that standardizing inputs and then finding the solution with no intercept term is equivalent to minimizing the squared loss with an intercept term. By including an intercept term, the loss becomes

$$\mathcal{L}(\boldsymbol{\theta}, \theta_0) = \sum_{i=1}^{n}\big(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta} - \theta_0\big)^2. \tag{5.23}$$

Such formulation has the advantage of allowing y to be nonzero when $\boldsymbol{x}$ is zero. Let the solution to this new loss formulation be $\hat{\boldsymbol{\theta}}, \hat{\theta}_0$:

$$\hat{\theta}_0, \hat{\boldsymbol{\theta}} = \arg\min_{\theta_0,\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \theta_0) \tag{5.24}$$

and let $\hat{\boldsymbol{y}} = \mathsf{X}\hat{\boldsymbol{\theta}} + \hat{\theta}_0\mathbf{1}$ be the predictions.

Let the columns of X be $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$. We provide an analysis that the optimal prediction $\hat{\boldsymbol{y}}$ found by considering the intercept term is the same as minimizing the normal squared loss (5.9) over the standardized inputs $\tilde{\mathsf{X}}$ and $\tilde{\boldsymbol{y}}$:

$$\check{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\theta}\right)^2, \tag{5.25}$$

$$\check{\boldsymbol{y}} = \tilde{\mathsf{X}}\check{\boldsymbol{\theta}} + \bar{y}\mathbf{1}, \tag{5.26}$$

where

$$\tilde{\mathsf{X}} = \begin{pmatrix} \tilde{\boldsymbol{x}}_1^T \\ \vdots \\ \tilde{\boldsymbol{x}}_n^T \end{pmatrix} = (\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_m), \quad \tilde{\boldsymbol{v}}_j = (\boldsymbol{v}_j - \beta_j\mathbf{1})/\alpha_j, \quad \beta_j = \frac{1}{n}\sum_{i=1}^{n} v_{ji}, \quad \alpha_j = \|\boldsymbol{v}_j - \beta_j\mathbf{1}\|_2, \tag{5.27}$$

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}, \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j. \tag{5.28}$$

Let's first minimize the new loss (5.23). We can fix $\boldsymbol{\theta}$ for now. Then, the loss function is quadratic in θ_0 and the optimal choice for θ_0 can be found as

$$\hat{\theta}_0(\boldsymbol{\theta}) = \bar{y} - \sum_{j=1}^{m} \beta_j\theta_j. \tag{5.29}$$

Substituting θ_0 by $\hat{\theta}_0(\boldsymbol{\theta})$ in the loss (5.23) gives

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\theta} - \bar{y} + \sum_{j=1}^{m} \beta_j\theta_j)^2. \tag{5.30}$$

After rewriting (5.25) as

$$\check{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\theta})^2 = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \bar{y} - \sum_{j=1}^{m} \frac{(x_{ij} - \beta_j)}{\alpha_j}\theta_j)^2 \tag{5.31}$$

$$= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} x_{ij}\frac{\theta_j}{\alpha_j} - \bar{y} + \sum_{j=1}^{m} \beta_j\frac{\theta_j}{\alpha_j})^2, \tag{5.32}$$

we can observe an analogy between (5.30) and (5.32). It follows that $\check{\theta}_j = \alpha_j\hat{\theta}_j$ for all $1 \le j \le m$.

Hence, for all $1 \le i \le n$,

$$\check{y}_i = \tilde{\boldsymbol{x}}_i^T\check{\boldsymbol{\theta}} + \bar{y} = \sum_{j=1}^{m} \tilde{x}_{ij}\check{\theta}_j + \bar{y} = \sum_{j=1}^{m} (x_{ij} - \beta_j)\hat{\theta}_j + \bar{y} = \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}} + \hat{\theta}_0 = \hat{y}_i, \tag{5.33}$$

i.e., the predictions we obtained from (5.24) and (5.25) are equal. Note that the second last equality follows from $\hat{\theta}_0(\hat{\boldsymbol{\theta}}) = \hat{\theta}_0$. △

5.3 Probabilistic Models for Regression

So far we haven't made any assumptions regarding the statistics of the data. In this section, we consider two models: i) a model that only characterizes the mean and covariance of the error vector and ii) a Gaussian model.

5.3.1 General model

Let us now assume that

$$Y = \boldsymbol{x}^T\boldsymbol{\theta}^* + \epsilon, \qquad \mathbb{E}[\epsilon] = 0,\ \mathrm{Var}(\epsilon) = \sigma^2.$$

We have n samples $(\boldsymbol{x}_i, y_i)$. For simplicity, we will assume that $\boldsymbol{x}_i$ are deterministic. We will further assume for any $i, j, i \neq j$, ϵ_i and ϵ_j are uncorrelated. In vector form, we have

$$\boldsymbol{Y} = \mathsf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}, \qquad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I.$$

At this point, we are not making any other assumptions related to the distribution of $\boldsymbol{\epsilon}$.[1]

Frequentist evaluation: Consider the estimator $\hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}$. Let us investigate the behavior of $\hat{\boldsymbol{\theta}}$ by viewing it as a random variable $\hat{\boldsymbol{\Theta}}$ under this model. We have

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\Theta}}] &= \mathbb{E}[(\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{Y}] && (5.34)\\
&= (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\,\mathbb{E}[\boldsymbol{Y}] && (5.35)\\
&= (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\,\mathbb{E}[\mathsf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}] && (5.36)\\
&= \boldsymbol{\theta}^*, && (5.37)
\end{aligned}$$

indicating that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}^*$. In particular, each dimension is estimated without bias, i.e., $\mathbb{E}[\hat{\Theta}_i] = \theta_i^*$.

We also find

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\Theta}}) &= \mathrm{Cov}((\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{Y}) && (5.38)\\
&= (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\,\mathrm{Cov}(\boldsymbol{Y})\mathsf{X}(\mathsf{X}^T\mathsf{X})^{-1} && (5.39)\\
&= (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\,\mathrm{Cov}(\boldsymbol{\epsilon})\mathsf{X}(\mathsf{X}^T\mathsf{X})^{-1} && (5.40)\\
&= (\mathsf{X}^T\mathsf{X})^{-1}\sigma^2. && (5.41)
\end{aligned}$$

In particular, element i of the diagonal of $\mathrm{Cov}(\hat{\boldsymbol{\Theta}})$ is the variance of $\hat{\Theta}_i$ and also its MSE, as the estimator is unbiased.

The Gauss-Markov theorem. The Gauss-Markov theorem states that under the assumptions that $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 I$, $\hat{\boldsymbol{\theta}}$ is the best *linear* unbiased estimator. Here, linear means that $\hat{\boldsymbol{\theta}}$ is linear in $\boldsymbol{y}$, i.e., $\hat{\boldsymbol{\theta}} = a_1 y_1 + a_2 y_2 + \cdots + a_m y_m$ for some scalars a_1^m. The Gauss-Markov theorem implies that for any[2] vector $\boldsymbol{u}$, $\boldsymbol{u}^T\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{u}^T\boldsymbol{\theta}^*$ with the smallest possible variance.

5.3.2 Gaussian model

Let us further assume that ϵ_i are iid, with distribution $\mathcal{N}(0, \sigma^2)$, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. In other words, we have:

$$\begin{aligned}
\boldsymbol{Y} &= \mathsf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) && (5.42)\\
p_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}^*, \sigma^2) &\sim \mathcal{N}(\mathsf{X}\boldsymbol{\theta}^*, \sigma^2 I). && (5.43)
\end{aligned}$$

Exercise 5.4. Prove that if $p(\boldsymbol{y}; \boldsymbol{\theta}, \sigma^2) \sim \mathcal{N}(\mathsf{X}\boldsymbol{\theta}, \sigma^2 I)$, then for all i, $p(y_i; \boldsymbol{\theta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\theta}, \sigma^2)$ and the y_i are independent. △

Now we have a probabilistic model with unknown parameters $\boldsymbol{\theta}$ and σ^2.

[1] For ϵ, we will not follow the convention that random variables are shown as capital letters since capital ϵ can be confused with Latin E.

[2] This isn't entirely precise!

Maximum Likelihood

Given that the covariance matrix is $\sigma^2 I$ and assuming that $\boldsymbol{y}$ is n-dimensional, the density and the likelihood are

$$p(\boldsymbol{y};\boldsymbol{\theta},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta})^T(\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta})\right) \tag{5.44}$$

$$\propto \frac{1}{\sigma^n} \exp\left(-\frac{\|\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}\right) \tag{5.45}$$

$$\ell(\boldsymbol{\theta},\sigma^2) \doteq -n\ln(\sigma) - \frac{1}{2\sigma^2}\|\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta}\|_2^2. \tag{5.46}$$

So maximizing for $\boldsymbol{\theta}$ leads to minimizing $\|\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta}\|_2^2 = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\boldsymbol{\theta})^2$ which we already know the solution to:

$$\hat{\boldsymbol{\theta}}_{ML} = \hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}. \tag{5.47}$$

We can similarly show that

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}})^2. \tag{5.48}$$

The mean and covariance of $\hat{\boldsymbol{\theta}}$ are the same as in §5.3.1. But now we also know that $\hat{\boldsymbol{\theta}}$ *is Gaussian*. This is because the linear combination of Gaussian variables is Gaussian. Hence,

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathsf{X}^T\mathsf{X})^{-1}). \tag{5.49}$$

Cramer-Rao Lower Bound. With the additional Gaussian assumption in this section, using Cramer-Rao lower bound, a stronger result compared to the Gauss-Markov theorem can be obtained. Namely, $\hat{\boldsymbol{\theta}}$ is the best unbiased estimator (not just the best linear unbiased estimator).

Example 5.5 (†)**.** For an unbiased vector estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, the CRLB has the form

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \succcurlyeq I^{-1}(\boldsymbol{\theta}), \tag{5.50}$$

where $A \succcurlyeq B$ denotes that $A - B$ positive semidefinite. Let us find the CRLB for $\hat{\boldsymbol{\theta}}$ of (5.47). We have

$$\ell(\boldsymbol{\theta},\sigma^2) \doteq -n\ln(\sigma) - \frac{1}{2\sigma^2}(\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta})^T(\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta}). \tag{5.51}$$

$$\nabla_{\boldsymbol{\theta}}\ell = \left(-\frac{1}{\sigma^2}(\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta})^T(-\mathsf{X})\right)^T \tag{5.52}$$

$$= \frac{1}{\sigma^2}\mathsf{X}^T(\boldsymbol{y}-\mathsf{X}\boldsymbol{\theta}) \tag{5.53}$$

$$\mathsf{H}_{\boldsymbol{\theta}}\ell = \frac{d\nabla_{\boldsymbol{\theta}}\ell}{d\boldsymbol{\theta}} = -\frac{1}{\sigma^2}\mathsf{X}^T\mathsf{X}. \tag{5.54}$$

and so $I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\mathsf{X}^T\mathsf{X}$. Hence, $I(\boldsymbol{\theta})^{-1} = \sigma^2(\mathsf{X}^T\mathsf{X})^{-1}$, which matches the covariance of $\hat{\boldsymbol{\theta}}$ found in (5.41). △

Bayesian Linear Regression

In Bayesian linear regression, the Gaussian likelihood

$$\boldsymbol{y}|\boldsymbol{\theta},\sigma^2 \sim \mathcal{N}(\mathsf{X}\boldsymbol{\theta}, \sigma^2 I) \tag{5.55}$$

is a common choice. But we also need to choose priors for $\boldsymbol{\theta}$ and σ^2. A possible non-informative choice is

$$p(\boldsymbol{\theta},\sigma^2) \propto 1/\sigma^2, \tag{5.56}$$

or equivalently, $p(\sigma^2) \propto \frac{1}{\sigma^2}$, $p(\boldsymbol{\theta}) \propto 1$ and σ^2, and $\boldsymbol{\theta}$ are independent.

We are interested in finding

$$p(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = p(\boldsymbol{\theta}|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y}) \tag{5.57}$$

We start with

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \sigma^2) = \frac{p(\boldsymbol{\theta}, \boldsymbol{y}|\sigma^2)}{p(\boldsymbol{y}|\sigma^2)} \propto p(\boldsymbol{\theta}, \boldsymbol{y}|\sigma^2) = p(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)p(\boldsymbol{\theta}|\sigma^2) \propto \exp\left(-\frac{(\mathsf{X}\boldsymbol{\theta} - \boldsymbol{y})^T(\mathsf{X}\boldsymbol{\theta} - \boldsymbol{y})}{2\sigma^2}\right). \tag{5.58}$$

Note that in the above expression, we are viewing $\boldsymbol{y}$ and σ^2 as given. So the expression $p(\boldsymbol{y}|\sigma^2)$ is treated as a constant and discarded. Furthermore, $p(\boldsymbol{\theta}|\sigma^2) = p(\boldsymbol{\theta}) \propto 1$.

The right-hand expression in (5.58) is quadratic in $\boldsymbol{\theta}$. So we'll try to see if we can write it in terms of a Gaussian distribution. With foresight, let the mean and the covariance of this distribution be denoted $\hat{\boldsymbol{\theta}}$ and $K\sigma^2$. We need

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T K^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \doteq (\mathsf{X}\boldsymbol{\theta} - \boldsymbol{y})^T(\mathsf{X}\boldsymbol{\theta} - \boldsymbol{y}). \tag{5.59}$$

Ignoring terms that are constant in $\boldsymbol{\theta}$, we require

$$\boldsymbol{\theta}^T K^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T K^{-1}\hat{\boldsymbol{\theta}} \doteq \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - 2\boldsymbol{\theta}^T X^T \boldsymbol{y}, \tag{5.60}$$

which is satisfied by $K^{-1} = \mathsf{X}^T\mathsf{X}$ and

$$-2\boldsymbol{\theta}^T K^{-1}\hat{\boldsymbol{\theta}} = -2\boldsymbol{\theta}^T X^T \boldsymbol{y}, \tag{5.61}$$

$$-2\boldsymbol{\theta}^T X^T X\hat{\boldsymbol{\theta}} = -2\boldsymbol{\theta}^T X^T \boldsymbol{y}, \tag{5.62}$$

$$\mathsf{X}^T\mathsf{X}\hat{\boldsymbol{\theta}} = \mathsf{X}^T\boldsymbol{y}, \tag{5.63}$$

$$\hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}. \tag{5.64}$$

So it suffices to set $\hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\boldsymbol{y}$ and $K = (\mathsf{X}^T\mathsf{X})^{-1}$,

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \sigma^2) \sim \mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, K\sigma^2). \tag{5.65}$$

Now we need to find $p(\sigma^2|\boldsymbol{y})$. Using the fact that $p(\sigma^2|\boldsymbol{y}) = p(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y})/p(\boldsymbol{\theta}|\sigma^2, \boldsymbol{y})$, it can be shown that $p(\sigma^2|\boldsymbol{y})$ has a scaled inverse-χ^2 distribution,

$$p(\sigma^2|\boldsymbol{y}) \sim \text{Inv-}\chi^2(n - m, s^2), \tag{5.66}$$

where m is the dimension of $\boldsymbol{x}_i$ and

$$s^2 = \frac{1}{n - m}(\boldsymbol{y} - \mathsf{X}\hat{\boldsymbol{\theta}})^T(\boldsymbol{y} - \mathsf{X}\hat{\boldsymbol{\theta}}). \tag{5.67}$$

While we can continue analytically and find $p(\boldsymbol{\theta}|\boldsymbol{y})$, in practice, we proceed computationally by generating samples from $p(\sigma^2|\boldsymbol{y})$ and then $p(\boldsymbol{\theta}|\boldsymbol{y}, \sigma^2)$. With this sampling approach we can also perform prediction for a given input vector $\boldsymbol{x}_{n+1}$ of by producing samples from $p(y_{n+1}|\boldsymbol{\theta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{x}_{n+1}^T\boldsymbol{\theta}, \sigma^2)$.

5.4 Regularized Linear Regression

Sometimes we are interested in reducing the flexibility of the model to avoid over-fitting, especially when the size of the data set is small. Alternatively, we may be interested in putting restrictions (e.g., forcing small coefficients to become 0) so that only the most important aspects of the data appear in the learned model, thus increasing its interpretability. These can be done by altering the loss function by adding a regularization term.

Ridge Regression

Ridge regression adds a penalty for the magnitude of the coefficients. Specifically, the loss function is

$$\mathcal{L}(\boldsymbol{\theta}) = \|\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2, \tag{5.68}$$

where λ is a parameter determining the relative importance of the square error versus the regularization loss term $\|\boldsymbol{\theta}\|_2^2$. The problem of minimizing this loss,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2, \tag{5.69}$$

can be shown to be equivalent to

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}\|_2^2, \tag{5.70}$$

$$\text{subject to : } \|\boldsymbol{\theta}\|_2^2 \leq t, \tag{5.71}$$

for some t. There is a one-to-one correspondence between λ and t. The second form is perhaps easier to understand because of the explicit constraints on $\|\boldsymbol{\theta}\|_2^2$.

From (5.69),

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = -2X^T(\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta}, \tag{5.72}$$

$$\nabla\mathcal{L}(\hat{\boldsymbol{\theta}}) = 0 \iff \mathsf{X}^T(\boldsymbol{y} - \mathsf{X}\hat{\boldsymbol{\theta}}) = \lambda\hat{\boldsymbol{\theta}} \tag{5.73}$$

$$\iff \hat{\boldsymbol{\theta}} = (\mathsf{X}^T\mathsf{X} + \lambda I)^{-1}\mathsf{X}^T\boldsymbol{y}. \tag{5.74}$$

Exercise 5.6. Prove that for $\lambda > 0$, $\mathsf{X}^T\mathsf{X} + \lambda I$ is invertible, even if the columns of X are not linearly independent. △

Bayesian Interpretation

We will now view the regularization penalty from a Bayesian point of view. As before, assume the Gaussian likelihood

$$\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\mathsf{X}\boldsymbol{\theta}, \sigma^2 I). \tag{5.75}$$

For simplicity, we focus on estimating only $\boldsymbol{\theta}$ and not σ^2. For the prior on $\boldsymbol{\theta}$, let

$$p(\boldsymbol{\theta}|\sigma^2) \sim \mathcal{N}(0, (\sigma^2/\lambda)I) \propto e^{-\frac{\lambda\boldsymbol{\theta}^T\boldsymbol{\theta}}{2\sigma^2}}. \tag{5.76}$$

Then

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \sigma^2) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2)p(\boldsymbol{\theta}|\sigma^2) \propto \exp\left(-\frac{(\mathsf{X}\boldsymbol{\theta} - \boldsymbol{y})^T(\mathsf{X}\boldsymbol{\theta} - \boldsymbol{y}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}}{2\sigma^2}\right). \tag{5.77}$$

Based on the previous discussion, it is immediately clear that the mode of the posterior distribution for $\boldsymbol{\theta}$ is $(\mathsf{X}^T\mathsf{X} + \lambda I)^{-1}\mathsf{X}^T\boldsymbol{y}$. Furthermore, since the distribution is quadratic, and hence Gaussian, this is also the mean of the posterior. Hence the formulation for ridge regression is equivalent to assuming a zero-mean Gaussian distribution for $\boldsymbol{\theta}$, which assigns high prior probabilities to smaller length of $\boldsymbol{\theta}$.

Lasso

In lasso, the regularization penalty has the form of the ℓ_1 norm,

$$\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{m} |\theta_i|, \tag{5.78}$$

where m is the length of $\boldsymbol{\theta}$. The problem is to find

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1, \tag{5.79}$$

or equivalently

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \mathsf{X}\boldsymbol{\theta}\|_2^2, \tag{5.80}$$

$$\text{subject to : } \|\boldsymbol{\theta}\|_1 \leq t. \tag{5.81}$$

Lasso does not have a closed form solution but efficient computational methods exist.

From a Bayesian point of view, lasso is equivalent to finding the *mode* of the posterior for $\boldsymbol{\theta}$ assuming the same model as above but with the double exponential (Laplace) prior

$$p(\boldsymbol{\theta}|\sigma^2) \propto e^{-\frac{\lambda\|\boldsymbol{\theta}\|_1}{2\sigma^2}}. \tag{5.82}$$

Discussion and generalization

In general we could choose the regularization penalty to be of the form[3]

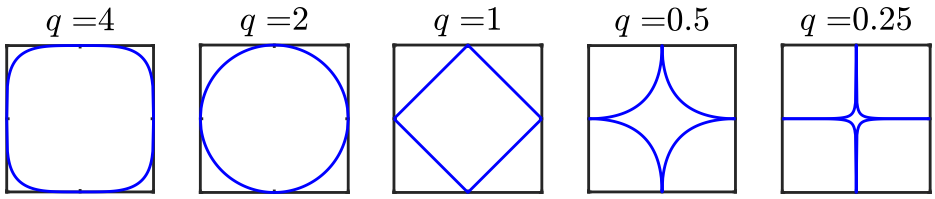
$$\|\boldsymbol{\theta}\|_q^q = \sum_{i=1}^{m} |\theta_i|^q, \tag{5.83}$$

where m is the length of $\boldsymbol{\theta}$. For $q = 1$ and $q = 2$, we get lasso and ridge regression, respectively.

The effect of the regularization can be viewed from a Bayesian framework, by setting the prior

$$\exp\left(-\frac{\lambda}{2\sigma^2}\|\boldsymbol{\theta}\|_q^q\right). \tag{5.84}$$

The contours for the priors for different values of q are given below.



In all cases, as we get further from the origin, the prior probability drops. But when q is small, the probability falls slower along the axes, encouraging solutions in which some of the coordinates are small or zero.

5.5 Error analysis and model selection

If our goal is to minimize the square of the prediction error, why would we use a different loss function for empirical risk minimization, as we did for ridge regression and lasso? How does this choice affect the error? Given that we have different choices for the form of the model and its parameters, how do we choose?

[3] $\|\boldsymbol{\theta}\|_q = \left(\sum_{i=1}^{m} |\theta_i|^q\right)^{1/q}$ is called the ℓ_q-*norm* of $\boldsymbol{\theta}$.

5.5.1 Bias-variance trade-off for quadratic error

Let us consider a general regression problem where we want to predict a value Y given an input vector $\boldsymbol{x}$. Let the prediction/estimate $\hat{y}$ for Y given $\boldsymbol{x}$ be denoted by $\hat{y} = f(\boldsymbol{x})$, where f is the function predicting Y given $\boldsymbol{x}$. For linear regression this is of the form $f(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\theta}}$, so finding the predictor is the same as finding $\hat{\boldsymbol{\theta}}$.

For a *specific* estimator f (e.g., one found based on a given data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$), the expected value of the quadratic loss for the next data point is

$$\mathcal{L}(f) = \mathbb{E}\big[(Y - f(\boldsymbol{X}))^2 | \boldsymbol{X} = \boldsymbol{x}\big] \tag{5.85}$$
$$= \mathbb{E}\big[(Y - f(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}\big], \tag{5.86}$$

which is called the **test error** for $\boldsymbol{x}$. We can view this as the loss for the $n+1$st data point, where we are given $\boldsymbol{x} = \boldsymbol{x}_{n+1}$ and are interested in the loss of predicting Y_{n+1}. So, we are interested in evaluating f for a given input. For instance, in our marketing example from the beginning of the chapter, $\boldsymbol{x}_{n+1}$ would indicate a specific budget, e.g., $\boldsymbol{x}_{n+1} = (\$10k, \$20k, \$5k)$.

Let $\bar{y}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$. Then, using the fact that $\mathbb{E}[(Z-c)^2] = \mathrm{Var}(Z) + (\mathbb{E}[Z] - c)^2$, we have

$$\mathcal{L}(f) = \mathbb{E}\big[(Y - f(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x}\big] \tag{5.87}$$
$$= \mathrm{Var}(Y|\boldsymbol{X} = \boldsymbol{x}) + \big(\bar{y}(\boldsymbol{x}) - f(\boldsymbol{x})\big)^2. \tag{5.88}$$

Note that the error has two parts: an irreducible part, referred to as intrinsic error, which is not under our control, and a part that depends on the choice of the predictor. The intrinsic error results from the noise in "nature," i.e., the fact that $\boldsymbol{X}$ does not have enough information to fully determine Y. In other words, this term can be viewed as the accumulated effect of all factors that are not included in $\boldsymbol{X}$. Having a larger dataset or choosing a better f does not affect this term. The reducible part compares the performance of our predictor with the best possible. This error is minimized by setting $f(\boldsymbol{x}) = \bar{y}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$. However, doing so exactly is only possible if we have the distribution or an infinite amount of data.

To summarize, we can write **the test error for given f and $\boldsymbol{x}$** as

$$\mathcal{L}(f) = \text{irred.} + \big(f(\boldsymbol{x}) - \bar{y}(\boldsymbol{x})\big)^2. \tag{5.89}$$

We should choose f to minimize the above quantity. Let us consider how f is chosen through empirical risk minimization.

1. Determine a set $\mathcal{F}$ from which f can be chosen, e.g., all linear functions.
2. Define an empirical loss. Typically, this reflects the loss function in the expected loss (5.86), but may include a regularization term, i.e., $\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \text{Reg}$.
3. Collect data, $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, and find $f \in \mathcal{F}$ that minimizes the empirical loss.

Consider a thought experiment in which this process is repeated many times. In each trial, the set $\mathcal{F}$ and the definition of the empirical loss stay the same, while $\mathcal{D}$ and consequently, f are random. Since f is a function of $\mathcal{D}$, let us denote it as $f_{\mathcal{D}}$. Let $\mathcal{M}$ denote the fixed components of this process, i.e., the set $\mathcal{F}$ and the definition of the empirical loss. We are interested to find the loss as a function of $\mathcal{M}$, which is under our control, averaged over all possible datasets (which is outside our control). This is called the **expected test error**

$$\mathcal{L}(\mathcal{M}) = \mathbb{E}[\mathcal{L}(f_{\mathcal{D}})] = \text{irred.} + \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{y}(\boldsymbol{x}))^2\big], \tag{5.90}$$

where the expectation is taken over all possible datasets. Note that the irreducible part, $\mathrm{Var}(y|\boldsymbol{x})$ is a constant.With a similar trick as above, we have

$$\mathcal{L}(\mathcal{M}) - \text{irred.} = \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{y}(\boldsymbol{x}))^2\big] \tag{5.91}$$
$$= \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}\, f_{\mathcal{D}}(\boldsymbol{x}))^2\big] + (\mathbb{E}\, f_{\mathcal{D}}(\boldsymbol{x}) - \bar{y}(\boldsymbol{x}))^2, \tag{5.92}$$

where the last equality follows from the fact that $\mathbb{E}\, f_{\mathcal{D}}(\boldsymbol{x})$ and $\bar{y}(\boldsymbol{x})$ are constants for a given $\boldsymbol{x}$. The first term on the last line is the square of the bias and the second term is the variance of $f_{\mathcal{D}}(\boldsymbol{x})$. So

$$\mathcal{L}(\mathcal{M}) = \text{irred.} + (\text{bias})^2 + \text{variance} \tag{5.93}$$

Now, the loss is written as the sum of squared bias term, which compares the average prediction across all possible datasets with the best possible predictor, and a variance term, which quantifies how different the estimate for each dataset is from the average, across all datasets.

Typically, as model complexity/flexibility[4] increases, bias decreases, while variance increases, since it has more freedom to vary based on the dataset. Simple/rigid models on the other hand typically have high bias and low variance. The bottom line is that neither unbiased models nor low variance predictors are necessary the best in terms of minimizing prediction error.

Another factor that affects the error is the size of the data. In many situations, the size of the data does affect the bias term but it may not significant since this term is averaged over data sets. The variance terms is affected because with more data, we get more information, approaching the situation in which $p(y|x)$ is known. So for very large datasets, the prediction reflects the distribution, which is constant.

5.5.2 Model Selection

As discussed above, the complexity of the model can affect the error. How can we choose the best model? In the past, we have used minimizing the empirical risk (training error) to optimize the parameters of a model. Can we again use the same strategy?

The **training error** for a given predictor f, i.e.,

$$\frac{1}{n}\sum L(y_i, f(\boldsymbol{x}_i)). \tag{5.94}$$

The training error itself is difficult to study. Averaged over all datasets, the **expected training error** is

$$\frac{1}{n}\sum \mathbb{E}[L(Y_i, f_D(\boldsymbol{x}_i))], \tag{5.95}$$

where for simplicity, we have assumed that that the $\boldsymbol{x}_i$ are fixed.

A typically behavior is in the table below, (this is not universally true). Training error is usually smaller for more complex models but this is not necessarily true for the test error. This makes choosing the best model based on training error difficult.

	Expected Train Err	Expected Test Err			
		Irred.	Bias^2	Var.	Total
More complex model	↓	–	↓	↑	?
More data	↑	–	↓	↓	↓

5.5.2.1 Overfitting and Underfitting

Suppose the true relationship between two scalar variables x and Y is

$$y = ax + w, \quad w \sim \mathcal{N}(0, \sigma^2). \tag{5.96}$$

We assume that $\sigma < ax$ for typical values of x since otherwise, we cannot predict Y accurately even if a is known (the irreducible error is large relative to the best predicted value).

The data available to us consists of two points

$$\mathcal{D} = \{(x_1 = 1, y_1), (x_2 = 2, y_2)\}. \tag{5.97}$$

We consider predictors of the forms

[4] By flexibility, I mean its responsiveness to changes in the data, i.e., the extent to which the results change when data changes.

- $\hat{y}(x) = 0$,
- $\hat{y}(x) = \theta x$,
- $\hat{y}(x) = \theta_1 x + \theta_2 x^2$.

For each predictor, we will find the parameter values $(\theta, \theta_1, \theta_2)$ that minimize the square loss for our data,

$$\frac{1}{2}\big[(y_1 - \hat{y}(x_1))^2 + (y_2 - \hat{y}(x_2))^2\big]. \tag{5.98}$$

For the third predictor, we will also consider the regularized version with loss

$$\frac{1}{2}\big[(y_1 - \hat{y}(x_1))^2 + (y_2 - \hat{y}(x_2))^2\big] + b\theta_2^2, \tag{5.99}$$

where b is a constant. Here, I chose the form $b\theta_2^2$ instead of the ℓ_2 norm, $b(\theta_1^2 + \theta_2^2)$ to simplify the derivation. We will still be able to see the effect of regularization.

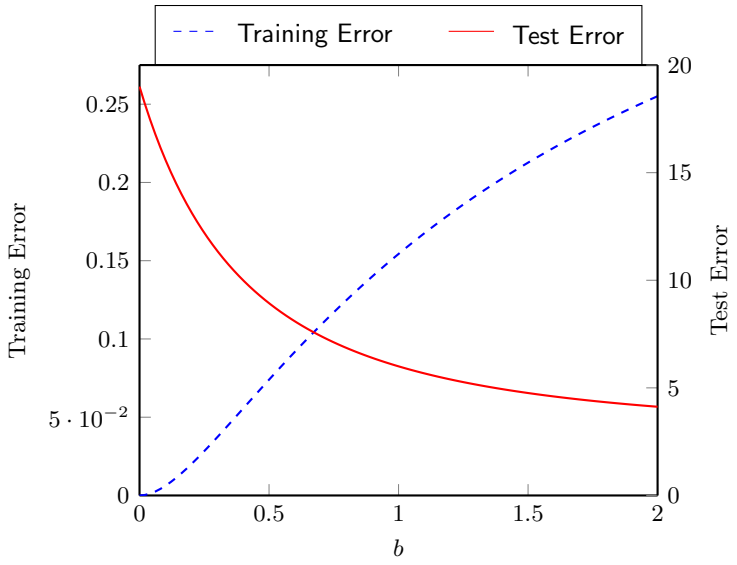
We then find the expected error for the training data and for a test data point (x_3, y_3), where we assume $x_3 = 3$. The expectation is taken over the randomness in y_1, y_2, y_3. The results are given in the table below.

Prediction	Expected Train Err	Expected Test Error for $x_3 = 3$			
		Irr.	Bias^2	Var.	Total
$\hat{y}(x) = 0$	$\frac{5a^2}{2} + \sigma^2$	σ^2	$9a^2$	0	$9a^2 + \sigma^2$
$\hat{y}(x) = \frac{y_1+2y_2}{5}x$	$\frac{\sigma^2}{2}$	σ^2	0	$\frac{9}{5}\sigma^2$	$\frac{14}{5}\sigma^2$
$\hat{y}(x) = \frac{4y_1-y_2}{2}x - \frac{2y_1-y_2}{2}x^2$	0	σ^2	0	$18\sigma^2$	$19\sigma^2$
$\hat{y}(x) = \frac{b(y_1+2y_2)+8y_1-2y_2}{5b+4}x - \frac{4y_1-2y_2}{5b+4}x^2$	$\frac{\sigma^2}{2}\left(1 - \frac{1}{5b/4+1}\right)^2$	σ^2	0	$\frac{9\sigma^2}{5}\left(1 + \frac{9}{(5b/4+1)^2}\right)$	$\frac{\sigma^2}{5}\left(14 + \frac{81}{(5b/4+1)^2}\right)$

As we go down the table, the model complexity increases. This allows the model to fit the training data better, leading to smaller expected training (square) error. The irreducible component of the test error stays the same, regardless of the model. The prediction bias for the test data point decreases, while its variance increases.

Given the assumption that σ is small relative to a, the smallest total error is obtained by the middle predictor. The zero predictor is not complex enough to be able to fit even the training data well. This situation is referred to as **underfitting**. The quadratic predictor is so complex that it can fit the training data, including the noise in the data, perfectly. But it does not generalize well due to its susceptibility to noise and high variance. This is called **overfitting**. In other words, the model memorizes this specific dataset rather than looking for patterns in it.

For the predictor with regularization, the graph below shows how the training and test errors change as a function of b. In this case, the predictor without regularization is overfitting the data. As b increases, overfitting decreases and we obtain a better test error. Note however that here the specific form of regularization prevents underfitting for large b, something that may occur in practice.



It is important to note models could perform poorly for reasons other than over- and under-fitting. For example, if the true distribution of the data is $y = a \sin x + w$, no polynomial predictor will perform well for a wide range of inputs due to the poor match between the true distribution and the learning model.

5.5.2.2 Training, validation, and test sets

In the previous subsection, we could identify the best model because we knew the true model for $(\boldsymbol{x}, y)$, which nature uses to produce them. In practice, however, the true model is not known and we cannot compute the expected test error. We have also seen that the training error is not necessarily a good estimate for the test error.

If we have sufficient data, a good solution is to divide it into three parts, a **training set**, a **validation set**, and a **test set**. For each model, the training set is used to optimize its parameters. Then all optimized models are evaluated on the validation set. Since the validation set is not used in training, this reduces the risk of over-fitting and, so, the errors for the validation set are better estimates for the test error. We choose the best model based on the validation set. We perform a final assessment using the test set, which should provide a good estimate of the error of the selected model for future practical use. Note that the test set cannot be used for any other purpose. If it is used in training or validation, it will not provide a reliable estimate of the error in the wild.

Example 5.7 (Regularization bias-variance trade-off)**.** Regularization allows us to control the flexibility of the model. In ridge regression as λ increases, the model becomes more constrained. For $\lambda > 0$ it can be shown to be biased. With $\mathcal{D} = (\mathsf{X}, \boldsymbol{y})$,

$$\begin{aligned} \mathbb{E}[\hat{y}_{n+1}] &= \mathbb{E}[\boldsymbol{x}_{n+1}^T \hat{\boldsymbol{\theta}}] && (5.100)\\ &= \boldsymbol{x}_{n+1}^T (\mathsf{X}^T\mathsf{X} + \lambda I)^{-1} \mathsf{X}^T \, \mathbb{E}[\boldsymbol{y}] && (5.101)\\ &= \boldsymbol{x}_{n+1}^T (\mathsf{X}^T\mathsf{X} + \lambda I)^{-1} \mathsf{X}^T \mathsf{X} \boldsymbol{\theta}. && (5.102) \end{aligned}$$

Noting that $\mathbb{E}[y_{n+1}] = \boldsymbol{x}_{n+1}^T \boldsymbol{\theta}$, we see that the estimate of $\hat{y}_{n+1}$ is biased. In particular, if $\mathsf{X}^T\mathsf{X} = I$, then

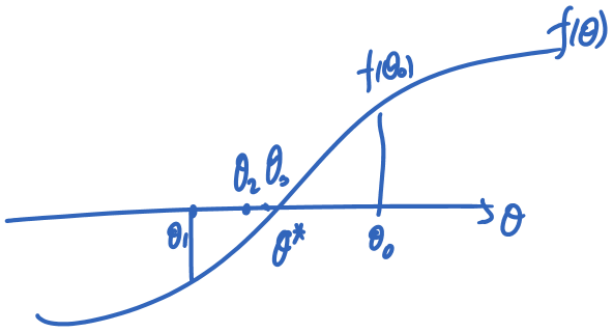
$$\mathbb{E}[\hat{y}_{n+1}] = \frac{\boldsymbol{x}_{n+1}^T \boldsymbol{\theta}}{1+\lambda} < \boldsymbol{x}_{n+1}^T \boldsymbol{\theta} = \mathbb{E}[y_{n+1}]. \tag{5.103}$$

But it can be shown to have lower variance. If the choice of λ is appropriate, it will have a smaller total loss. △

5.6 Stochastic Gradient Descent

Even though that gradient descent is sometimes less computationally expensive than directly finding the solution, its cost may still be high. In such cases, using *stochastic gradient descent* (SGD) may be helpful. SGD tries to improve the estimate by considering one data point (or a small batch of data points) at a time.

First, let's consider finding the root of a function $f(\theta)$ with a simple method. We assume that $f(\theta)$ is bounded and there is a unique root θ^* such that f is increasing at θ^*.



Suppose that we start from a point $\theta^{(0)}$ that is appropriately close to θ^*. We proceed iteratively as

$$\theta^{(t+1)} = \theta^{(t)} - a_t f(\theta^{(t)}), \tag{5.104}$$

where a_t satisfies

$$\sum_{t=1}^{\infty} a_t = \infty, \qquad \sum_{t=1}^{\infty} a_t^2 < \infty. \tag{5.105}$$

For example, $a_t = 1/t$ is a good choice while $a_t = 1/t^2$ isn't. It can then be shown that $\theta^{(t)}$ converges to θ^*.

But what if we cannot compute $f(\theta)$ but instead we have access to a noisy version $F(\theta)$ that satisfies $f(\theta) = \mathbb{E}[F(\theta)]$, where $F(\theta)$ is bounded. It turns out that if we let

$$\theta^{(t+1)} = \theta^{(t)} - a_t F(\theta^{(t)}), \tag{5.106}$$

where in each iteration we sample $F(\theta)$, then $\theta^{(t)}$ again converges to θ^*.

Now let us consider the loss function for linear regression (note that we are using the expected loss as opposed to empirical loss)

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}[(y - \boldsymbol{x}^T\boldsymbol{\theta})^2], \tag{5.107}$$

where we are also assuming that $\boldsymbol{x}$ is random with some distribution. To minimize this loss, we compute the gradient:

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}[-2(y - \boldsymbol{x}^T\boldsymbol{\theta})\boldsymbol{x}] \tag{5.108}$$

We would like to find $\boldsymbol{\theta}$ such that the gradient above is zero.

Let

$$f(\boldsymbol{\theta}) = \mathbb{E}[-2(y - \boldsymbol{x}^T\boldsymbol{\theta})\boldsymbol{x}] \tag{5.109}$$

$$F(\boldsymbol{\theta}) = -2(y - \boldsymbol{x}^T\boldsymbol{\theta})\boldsymbol{x}, \tag{5.110}$$

so that $f(\boldsymbol{\theta}) = \mathbb{E}[F(\boldsymbol{\theta})]$. Now the elements of the data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ can be used to produce samples for $F(\boldsymbol{\theta})$. So we let

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + a_t(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}^{(t)})\boldsymbol{x}_i, \tag{5.111}$$

which is the stochastic gradient descent algorithm for linear regression.