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 $\boldsymbol{x}$. For example, we may be interested in predicting blood pressure given age, sex, weight, exercise, and calorie intake. Applications include prediction as well as understanding the relationship between inputs and output, for example, identifying the most important input components.

Linear regression relies on the assumption that $y \simeq \boldsymbol{x}^T\boldsymbol{\theta}$, where $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are elements of $\mathbb{R}^d$. We formulate the problem as follows: Find

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[L(y, \boldsymbol{x}^T\boldsymbol{\theta})],$$

for a given loss function L. We typically do not have the joint distribution for $\boldsymbol{x}, y$. Thus, given a training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, we aim to find

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, \boldsymbol{x}_i^T\boldsymbol{\theta}). \tag{5.1}$$

The linear form, which assumes $y_i \simeq \sum_{j=1}^{d} \theta_j x_{ij}$, may appear restrictive since it apparently excludes dependence on, for example, x_{ij}^2. This, however, 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 $y_i \simeq \sum_{j=1}^{e} \theta_j g_j(\boldsymbol{x}_i)$, where g_j are any function of $\boldsymbol{x}_i$ such as x_{i1}^2 and $x_{i1}x_{i2}x_{i4}$.

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

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

Furthermore, let $\boldsymbol{\epsilon}$ be such that

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

and $\hat{\boldsymbol{y}} = X\boldsymbol{\theta}$. With this notation, our goal is to find $\boldsymbol{\theta}$ such that $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2 = \|\boldsymbol{\epsilon}\|_2$ is minimized, where, for a vector $\boldsymbol{v} \in \mathbb{R}^d$,

$$\|\boldsymbol{v}\|_2^2 = \boldsymbol{v}^T\boldsymbol{v} = \sum_{j=1}^{d} v_j^2.$$

Example 5.1. Suppose

$$\boldsymbol{x}_1 = \begin{pmatrix}0\\1\end{pmatrix}, \quad y_1 = -1, \qquad \boldsymbol{x}_2 = \begin{pmatrix}2\\0\end{pmatrix}, \quad y_2 = 1, \qquad \boldsymbol{x}_3 = \begin{pmatrix}1\\1\end{pmatrix}, \quad y_3 = 0.$$

Then

$$X = \begin{pmatrix}0 & 1\\ 2 & 0\\ 1 & 1\end{pmatrix}, \qquad \hat{\boldsymbol{y}} = X\theta = \theta_1\begin{pmatrix}0\\2\\1\end{pmatrix} + \theta_2\begin{pmatrix}1\\0\\1\end{pmatrix}, \qquad \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{pmatrix}-1-\theta_2\\ 1-2\theta_1\\ -\theta_1-\theta_2\end{pmatrix}. \tag{5.2}$$

△

5.2 Least-squares

A common choice for the loss function is

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

This choice is relatively easy to deal with from a computational perspective and also has the same solution as the MLE for a common probabilistic model, thus providing an additional rationale for the resulting approach.

This choice leads to mean squared loss minimization:

$$\mathcal{L}(\boldsymbol{\theta}) = \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2, \qquad \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}).$$

We define

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\theta}}$$

as the predicted value or estimate based on the model.

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 **??**, $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to each column of X.

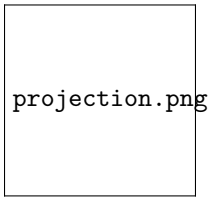


Figure 5.1: Error is minimized by projecting $\boldsymbol{y}$ onto the column space of X, Span(col(X)).

This orthogonality can be written as $X^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$. Then

$$\begin{aligned} X^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0 &\iff X^T(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}) = 0 \\ &\iff X^T\boldsymbol{y} = X^TX\hat{\boldsymbol{\theta}} \\ &\iff \hat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\boldsymbol{y} \end{aligned}$$

Here we have assumed that X^TX is invertible. This holds if the columns of X are linearly independent. To see this, suppose that X^TX is not invertible. Then there exists a nonzero vector $\boldsymbol{\alpha}$ such that $X^TX\boldsymbol{\alpha} = 0$ and hence

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

where the last step follows from the fact that the columns of X are linearly independent. But this contradicts $\boldsymbol{\alpha} \neq \mathbf{0}$. If the columns of X are not linearly independent, then the solution is not unique.

Example 5.2. From Example 5.1, we have

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

and so

$$\begin{aligned} X^TX &= \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}, & (X^TX)^{-1} &= \frac{1}{9}\begin{pmatrix} 2 & -1 \\ -1 & 5 \end{pmatrix} \\ (X^TX)^{-1}X^T &= \frac{1}{9}\begin{pmatrix} -1 & 4 & 1 \\ 5 & -2 & 4 \end{pmatrix} & \hat{\boldsymbol{\theta}} &= \begin{pmatrix} 5/9 \\ -7/9 \end{pmatrix} \\ \hat{\boldsymbol{y}} &= \begin{pmatrix} -7/9 \\ 10/9 \\ -2/9 \end{pmatrix}, & \boldsymbol{y} - \hat{\boldsymbol{y}} &= \begin{pmatrix} -2/9 \\ -1/9 \\ 2/9 \end{pmatrix}. \end{aligned}$$

△

Gradient descent. Alternatively, we can 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

$$\begin{aligned}\mathcal{L} &= (\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta}),\\ \nabla\mathcal{L} &= 2\big[(\boldsymbol{y} - X\boldsymbol{\theta})^T(-X)\big]^T = -2X^T(\boldsymbol{y} - X\boldsymbol{\theta}).\end{aligned}$$

Setting the gradient equal to 0 again gives $\hat{\boldsymbol{\theta}} = (X^TX)^{-1}X^T\boldsymbol{y}$. (Note that the Hessian is X^TX, which is positive-semi-definite.)

Computing $(X^TX)^{-1}$ may be prohibitively expensive computationally. An alternative approach is to start from an arbitrary value $\boldsymbol{\theta}^{(0)}$ and move towards the solution in steps:

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \rho X^T(\boldsymbol{y} - X\boldsymbol{\theta}^{(t)})\\ &= \boldsymbol{\theta}^{(t)} + \rho\sum_{i=1}^{n}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}^{(t)}),\end{aligned}$$

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.

5.3 Probabilistic Models for Regression

So far we haven't made any assumptions regarding the statistics of the data. Let us now assume that $\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I$. (This is the case if $y_i = \boldsymbol{x}^T\boldsymbol{\theta} + \epsilon_i$, where ϵ_i are iid with mean 0 and variance σ^2.) Then

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\theta}}] &= \mathbb{E}[(X^TX)^{-1}X^T\boldsymbol{y}]\\ &= (X^TX)^{-1}X^T\,\mathbb{E}[\boldsymbol{y}]\\ &= (X^TX)^{-1}X^T\,\mathbb{E}[X\boldsymbol{\theta} + \boldsymbol{\epsilon}]\\ &= \boldsymbol{\theta},\end{aligned}$$

using the properties of covariance given in the appendix, so $\hat{\boldsymbol{\theta}}$ is unbiased. The covariance is given by

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\theta}}) &= \mathrm{Cov}((X^TX)^{-1}X^T\boldsymbol{y})\\ &= (X^TX)^{-1}X^T\,\mathrm{Cov}(\boldsymbol{y})X(X^TX)^{-1}\\ &= (X^TX)^{-1}\sigma^2.\end{aligned}$$

The Gauss-Markov theorem. The Gauss-Markov theorem states that under the assumptions that $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 I$, $\hat{\boldsymbol{\theta}}$ is the best *linear* unbiased estimator. More precisely, for any[1] vector $\boldsymbol{u}$, $\boldsymbol{u}^T\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{u}^T\boldsymbol{\theta}$ with the smallest possible variance.

[1] This isn't entirely precise!

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:

$$p(\boldsymbol{y}; \boldsymbol{\theta}, \sigma^2) \sim \mathcal{N}(X\boldsymbol{\theta}, \sigma^2 I).$$

Exercise 5.3. Prove that if $p(\boldsymbol{y}; \boldsymbol{\theta}, \sigma^2) \sim \mathcal{N}(X\boldsymbol{\theta}, \sigma^2 I)$, 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.

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

$$\begin{aligned} p(\boldsymbol{y}; \boldsymbol{\theta}, \sigma^2) &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta})\right) \\ &\propto \frac{1}{\sigma^n} \exp\left(-\frac{\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2}{2\sigma^2}\right) \\ \ell(\boldsymbol{\theta}, \sigma^2) &\doteq -n \ln(\sigma) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2. \end{aligned}$$

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

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

We can similarly show that

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

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

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

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). To see this,

note that, for Fisher information $I(\boldsymbol{\theta})$, we have

$$\begin{aligned}\ell(\boldsymbol{\theta}, \sigma^2) &\doteq -n\ln(\sigma) - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta}).\\ \nabla_{\boldsymbol{\theta}}\ell &= \left(-\frac{1}{\sigma^2}(\boldsymbol{y} - X\boldsymbol{\theta})^T(-X)\right)^T\\ &= \frac{1}{\sigma^2}X^T(\boldsymbol{y} - X\boldsymbol{\theta})\\ \mathsf{H}_{\boldsymbol{\theta}}\ell &= \frac{d\nabla_{\boldsymbol{\theta}}\ell}{d\boldsymbol{\theta}} = -\frac{1}{\sigma^2}X^TX.\end{aligned}$$

and so $I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}X^TX$. Hence, $I(\boldsymbol{\theta})^{-1} = \sigma^2(X^TX)^{-1}$, which matches the variance of covariance of $\hat{\theta}$.

Bayesian Linear Regression

In Bayesian linear regression, the Gaussian likelihood

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

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

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

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

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

We start with

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

This 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

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^TK^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \doteq (X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}).$$

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

$$\boldsymbol{\theta}^TK^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^TK^{-1}\hat{\boldsymbol{\theta}} \doteq \boldsymbol{\theta}^TX^TX\boldsymbol{\theta} - 2\boldsymbol{\theta}^TX^T\boldsymbol{y},$$

which is satisfied by $K^{-1} = X^TX$ and

$$\begin{aligned}-2\boldsymbol{\theta}^TK^{-1}\hat{\boldsymbol{\theta}} &= -2\boldsymbol{\theta}^TX^T\boldsymbol{y},\\ -2\boldsymbol{\theta}^TX^TX\hat{\boldsymbol{\theta}} &= -2\boldsymbol{\theta}^TX^T\boldsymbol{y},\\ X^TX\hat{\boldsymbol{\theta}} &= X^T\boldsymbol{y},\\ \hat{\boldsymbol{\theta}} &= (X^TX)^{-1}X^T\boldsymbol{y}.\end{aligned}$$

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

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

Now we need to find $p(\sigma^2|\boldsymbol{y})$. It turns out this is a distribution called a scaled inverse-χ^2,

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

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

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

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 $p(y_{n+1}|\boldsymbol{\theta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{x}_{n+1}^T \boldsymbol{\theta}, \sigma^2)$.

Standardization

In some texts, an intercept term is also included in the loss function. For ease of notation, we instead 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 $\boldsymbol{y}^T\mathbf{1} = 0$. This assumption also holds in the following sections.

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

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

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,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2, \tag{5.3}$$

can be shown to be equivalent to

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2,$$
$$\text{subject to} : \|\boldsymbol{\theta}\|_2^2 \leq t,$$

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 (5.3),

$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{\theta}) &= -2X^T(\boldsymbol{y} - X\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta},\\ \nabla\mathcal{L}(\hat{\boldsymbol{\theta}}) = 0 &\iff X^T(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}) = \lambda\hat{\boldsymbol{\theta}}\\ &\iff \hat{\boldsymbol{\theta}} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}.\end{aligned}$$

Exercise 5.4. Prove that for $\lambda > 0$, $X^TX + \lambda I$ 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

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

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

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

Then

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

Based on the previous discussion, it is immediately clear that the mode of the posterior distribution for $\boldsymbol{\theta}$ is $(X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$. 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,

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

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

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

or equivalently

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2,$$
$$\text{subject to : } \|\boldsymbol{\theta}\|_1 \leq t.$$

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

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

Discussion and generalization

In general we could choose the regularization penalty to be of the form[2]

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

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

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

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 Bias-Variance Trade-off

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?

Our goal is 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 estimator. For linear regression this is of the form $f(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\theta}}$, so finding the estimator is the same as finding $\hat{\boldsymbol{\theta}}$. For a *specific* estimator f (e.g., a specific value for $\hat{\boldsymbol{\theta}}$), assuming quadratic loss, we have

$$\mathcal{L}(f) = \mathbb{E}[(y - f(\boldsymbol{x}))^2]. \tag{5.4}$$

Recall from **??** that

$$\mathcal{L}(f) = \mathbb{E}\Big[(y - \bar{y}(\boldsymbol{x}))^2\Big] + \mathbb{E}\Big[(\bar{y}(\boldsymbol{x}) - f(x))^2\Big],$$

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

where $\bar{y}(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$.

An important observation is that the second term in the expected loss for f is a function of f while the first term is not. The first term is an intrinsic noise term which we cannot reduce by choosing a better f or by collecting more data. This term can be viewed as the accumulated effect of all factors that are not included in $\boldsymbol{x}$. Given that this terms is not a function of f, define

$$\bar{\mathcal{L}}(f) = \mathbb{E}\big[(f(\boldsymbol{x}) - \bar{y}(\boldsymbol{x}))^2\big]. \tag{5.5}$$

This compares our estimator to the best possible. So we should choose f to minimize the above quantity.

Let us consider how f is chosen:

1. Determine a set $\mathcal{F}$ from which f can be chosen, e.g., all linear functions.
2. Define an empirical loss function that is related to the expected loss (5.4), but not necessarily identical, e.g., ridge loss.
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, by extension, 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). In this context, we write the loss as

$$\bar{\mathcal{L}}(\mathcal{M}) = \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{y}(\boldsymbol{x}))^2\big] = \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}_{n+1}) - \bar{y}(\boldsymbol{x}_{n+1}))^2\big],$$

where we have added the subscript $n+1$ to emphasize that the loss can be viewed as the prediction loss for the next sample. With a similar trick as above, we have

$$\begin{aligned}
\bar{\mathcal{L}}(\mathcal{M}) &= \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{y}(\boldsymbol{x}))^2\big] \\
&= \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})] + \mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})] - \bar{y}(\boldsymbol{x}))^2\big] \\
&= \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})])^2\big] + \mathbb{E}\big[(\mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})] - \bar{y}(\boldsymbol{x}))^2\big] + \\
&\quad\; 2\,\mathbb{E}[(f_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})])(\mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})] - \bar{y}(\boldsymbol{x}))]. \\
&= \mathbb{E}\big[(\mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})] - \bar{y}(\boldsymbol{x}))^2\big] + \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})])^2\big],
\end{aligned}$$

where the last equality follows from conditioning on $\boldsymbol{x}$.

We can understand this loss better by assuming a given value of $\boldsymbol{x}$ (or more precisely, a given value of $\boldsymbol{x}_{n+1}$). The loss then becomes[3]

$$(\mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})] - \bar{y}(\boldsymbol{x}))^2 + \mathbb{E}\big[(f_{\mathcal{D}}(\boldsymbol{x}) - \mathbb{E}[f_{\mathcal{D}}(\boldsymbol{x})])^2\big] = (\text{bias})^2 + \text{variance}$$

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.

[3] Technically the expectation terms need to be conditioned on $\boldsymbol{x}$. But I have dropped those for simplicity.

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.

Example 5.5 (Regularization)**.** 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. In particular, with $\mathcal{D} = (X, \boldsymbol{y})$, if $X^T X = I$, then

$$\begin{aligned}
\mathbb{E}[\hat{y}_{n+1}] &= \mathbb{E}[\boldsymbol{x}_{n+1}^T \hat{\boldsymbol{\theta}}] \\
&= \boldsymbol{x}_{n+1}^T (X^T X + \lambda I)^{-1} E[X^T \boldsymbol{y}] \\
&= \boldsymbol{x}_{n+1}^T (X^T X + \lambda I)^{-1} E[X^T X \boldsymbol{\theta}] \\
&= \frac{\boldsymbol{x}_{n+1}^T \boldsymbol{\theta}}{1+\lambda} \\
&\neq \boldsymbol{x}_{n+1}^T \boldsymbol{\theta} \\
&= \mathbb{E}[y_{n+1}].
\end{aligned}$$

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

Example 5.6 (Overfitting and Underfitting)**.** Suppose the true relationship between two scalar variables x and y is

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

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

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

We consider three predictors of the forms

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

and find $\theta, \theta_1, \theta_2$ to minimize the square loss for our data,

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

We then find the error for 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.

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

Prediction	Expected Train Err	Expected Test Error for $x_3 = 3$			
		Irred.	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	$18\sigma^2$	$19\sigma^2$

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.

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. △

5.6 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

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

where a_t satisfies

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

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

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

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)

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

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

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

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

Let

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

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

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

which is the stochastic gradient descent algorithm for linear regression.

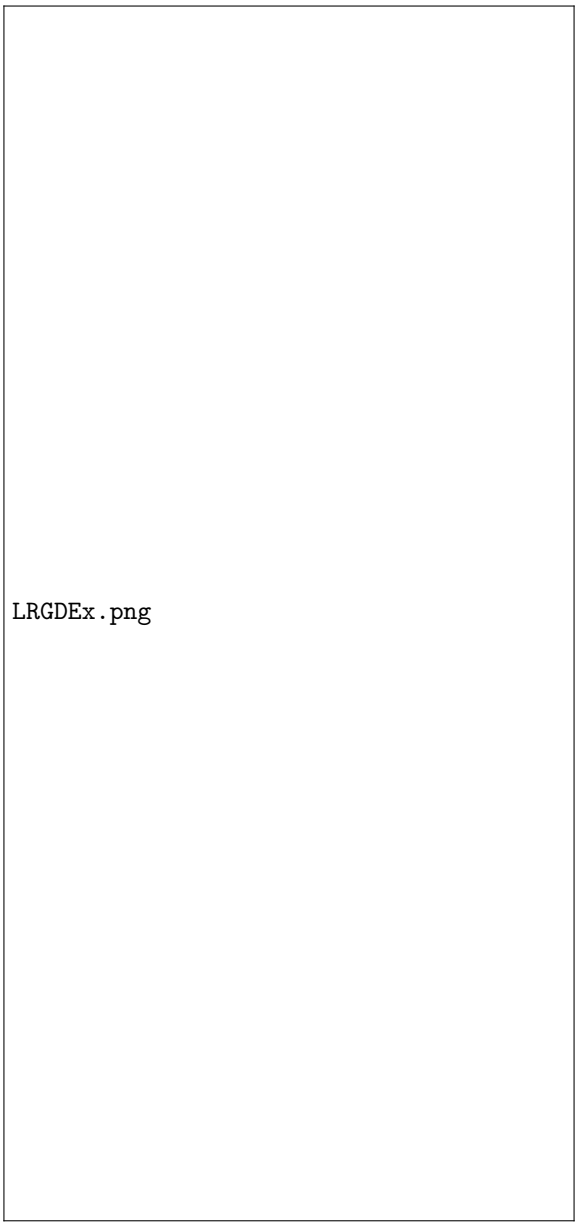
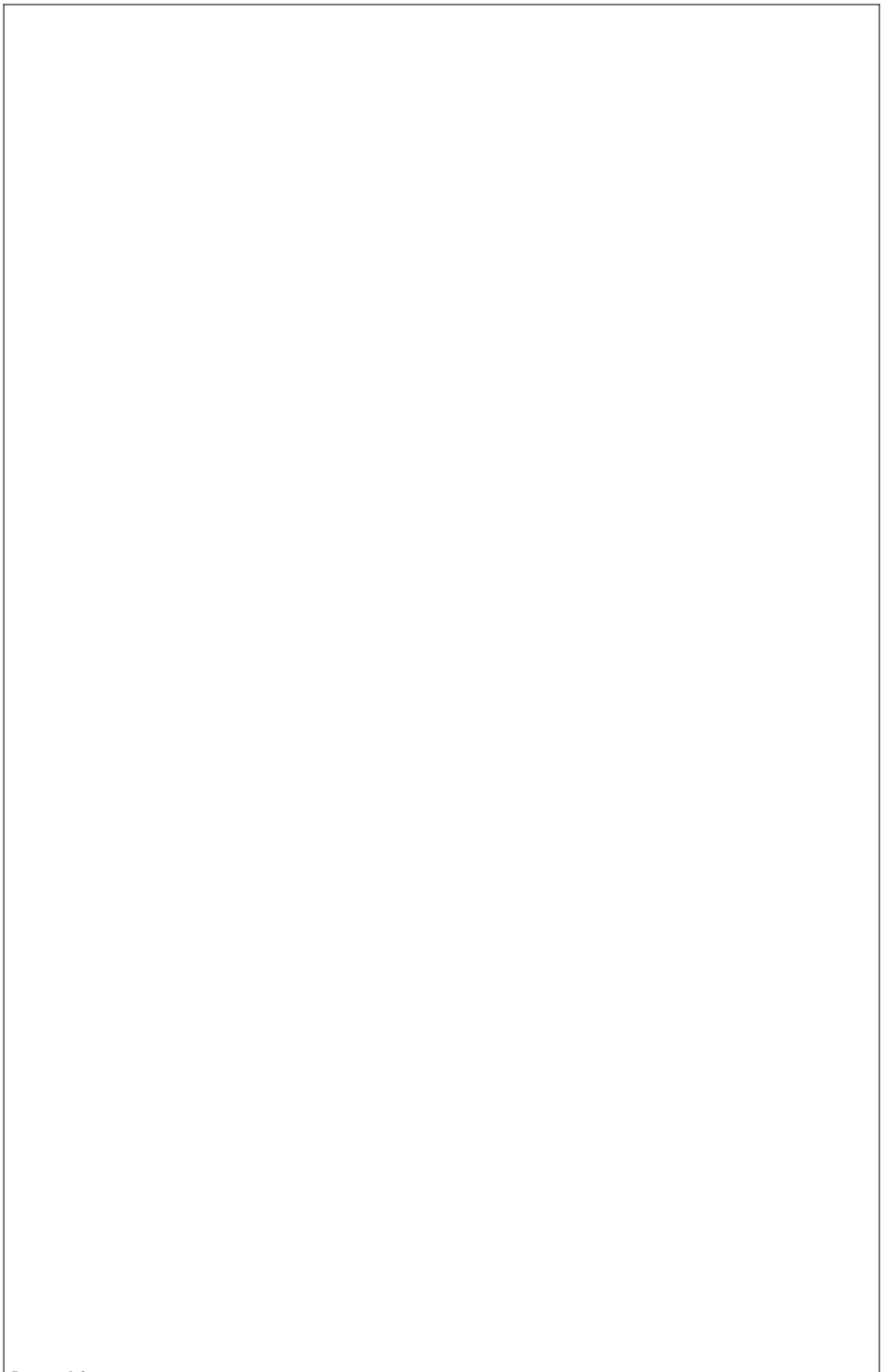
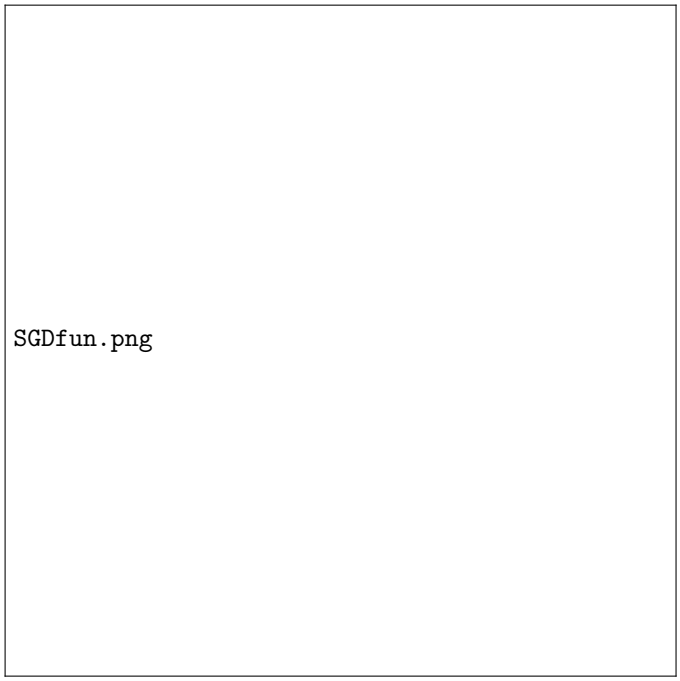


Figure 5.2: Gradient descent for linear regression



Lq.pdf


SGDfun.png