# Multiple Perspectives on Linear Regression

Jayanth Koushik
Carnegie Mellon University
jkoushik@cs.cmu.edu

## 1 Introduction

Linear regression is one of the most common regression methods. It is simple to understand, and easy to implement making it a suitable baseline to evaluate. Further, by using basis functions, it can actually be used to model very complex relations. This is a tutorial on simple linear regression and ridge regression. However, we will go beyond simply deriving results. We will consider multiple perspectives to gain a deeper understanding of linear regression, and see how these different perspectives are related. We also consider the relation between ordinary least squares and ridge regression.

## 2 Problem setting

We will consider a very general problem setting independent of any particular method. We have a set of $n$ data points $\mathcal{D} \equiv \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$, and $y_i \in \mathbb{R}$. This set is referred to as the training set. Let $\mathbb{X} \equiv \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix}^\mathrm{T}$, and $\mathbf{y} \equiv \begin{bmatrix}y_1 & \dots & y_n\end{bmatrix}^\mathrm{T}$. The goal is to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ to make predictions for points from $\mathbb{R}^d$ not in $\mathcal{D}$. Given a data point $(\mathbf{x}, y)$, the error in predicting $f(\mathbf{x})$ is the squared difference $(y - f(\mathbf{x}))^2$.

## 3 Best predictor

We can assume that the training set points are iid samples of a random variable pair $(\boldsymbol{X}, Y)$ distributed according to some unknown joint probability distribution. We define the risk of $f$ as the expected error $\mathrm{E}[(Y - f(\boldsymbol{X}))^2]$. With this, it is natural to seek the function with minimum risk. Define the regression function $m(\mathbf{x}) \equiv \mathrm{E}[Y|\boldsymbol{X} = \mathbf{x}]$ and write the risk as

$$\begin{aligned}
\mathrm{E}[(Y - f(\boldsymbol{X}))^2] &= \mathrm{E}[\mathrm{E}[(Y - f(\mathbf{x}))^2|\boldsymbol{X}]] \\
&= \mathrm{E}[\mathrm{E}[(Y - m(\boldsymbol{X}) + m(\boldsymbol{X}) - f(\boldsymbol{X}))^2|\boldsymbol{X}]] \\
&= \mathrm{E}[\mathrm{E}[(Y - m(\boldsymbol{X}))^2|\boldsymbol{X}]] + \mathrm{E}[\mathrm{E}[(m(\boldsymbol{X}) - f(\boldsymbol{X}))^2|\boldsymbol{X}]] + \\
&\quad 2\mathrm{E}[\mathrm{E}[(Y - m(\boldsymbol{X}))(m(\boldsymbol{X}) - f(\boldsymbol{X}))|\boldsymbol{X}]]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}[\mathrm{V}[Y|\boldsymbol{X}]] + \mathrm{E}[(m(\boldsymbol{X}) - f(\boldsymbol{X}))^2] + \\
&\quad 2\mathrm{E}[(m(\boldsymbol{X}) - f(\boldsymbol{X}))(\mathrm{E}[Y|\boldsymbol{X}] - m(\boldsymbol{X}))] \\
&= \mathrm{E}[\mathrm{V}[Y|\boldsymbol{X}]] + \mathrm{E}[(m(\boldsymbol{X}) - f(\boldsymbol{X}))^2] \qquad (1)
\end{aligned}
$$

Note that the first term is independent of $f$, and the second term is minimum (and equal to 0) if $f(\mathbf{x}) = m(\mathbf{x})\ \forall \mathbf{x}$. So the regression function is the ideal predictor in some sense; but since it depends on the unknown data distribution, we cannot actually compute it. This result is only theoretical.

# 4 Ordinary least squares

Ordinary least squares (OLS) is a very simple approach to regression. We model $f$ as a linear function of $\mathbf{x}$ i.e. $f(\mathbf{x}) = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$ for some $\boldsymbol{\theta} \in \mathbb{R}^d$. We do not add a separate bias variable, since we can always just add a constant 1 to each $\mathbf{x}$. This linear model reduces the problem from finding $f$ to finding $\boldsymbol{\theta}$. We can consider this from several perspectives, which all lead to the same solution.

## 4.1 Minimization of squared error sum

Under the linear model, the error corresponding to a point $(\mathbf{x}, y)$ is $(y - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x})^2$. So we can seek to find the $\boldsymbol{\theta}$ that minimizes the sum of errors over the training set.

$$
\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i)^2 = \arg\min_{\boldsymbol{\theta}}\ (\mathbf{y} - \mathbb{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbb{X}\boldsymbol{\theta}) \qquad (2)
$$

The derivative of this error with respect to $\boldsymbol{\theta}$ is $-2\mathbb{X}^{\mathrm{T}}\mathbf{y} + 2\mathbb{X}^{\mathrm{T}}\mathbb{X}\boldsymbol{\theta}$, so

$$
2\mathbb{X}^{\mathrm{T}}\mathbb{X}\boldsymbol{\theta}_* = 2\mathbb{X}^{\mathrm{T}}\mathbf{y} \implies \boldsymbol{\theta}_* = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} \qquad (3)
$$

provided $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ is not singular. In case of singularity, we say that the least squares solution does not exist. The second derivative of the error is $2\mathbb{X}^{\mathrm{T}}\mathbb{X}$. As shown in Lemma 1, $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ is positive semi-definite for any matrix A; so the solution given by Equation 3 is indeed a minima. With this expression for $\boldsymbol{\theta}$, the prediction for any $\mathbf{x}$ is given by

$$
f(\mathbf{x}) = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{x}^{\mathrm{T}}(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} = \mathbf{h}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} h_i y_i \qquad (4)
$$

where $\mathbf{h}^{\mathrm{T}} \equiv \mathbf{x}^{\mathrm{T}}(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}$. We see that the prediction is just a linear combination of the training set labels. Such predictors are called linear smoothers.

## 4.2 Generalization of linear fit

Suppose we want to fit a line through the data points. So we have $y_i = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i$ for $i \in \{1, \ldots, n\}$. This system of equations can be expressed concisely as

$$
\mathbb{X}\boldsymbol{\theta} = \mathbf{y} \qquad (5)
$$

From this equation, it is clear that $\mathbb{X}$ must be square for a solution to exist; this corresponds to the number of equations being equal to the number of parameters we are trying to estimate. Furthermore, $\mathbb{X}$ must invertible in which case the solution of Equation (5) is given by

$$\boldsymbol{\theta} = \mathbb{X}^{-1}\mathbf{y} \tag{6}$$

Now we can see the OLS solution given by Equation (3) as a generalization of the above solution with $(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}$ being the Moore-Penrose pseudoinverse of $\mathbb{X}$.

### 4.3 Maximum likelihood estimation

Let us now look at OLS from a probabilistic viewpoint. Specifically, we assume $y = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} + \epsilon$, where $\epsilon$ is independent Gaussian noise with 0 mean and variance $\sigma^2$. So given $\mathbf{x}$, $y$ is a random variable distributed as $\mathcal{N}(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}, \sigma^2)$ where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean $\mu$ and variance $\sigma^2$. This is a parametric model for $y$ characterized by $\boldsymbol{\theta}$, and we can find the likelihood of $\boldsymbol{\theta}$ given the training data as

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbb{X}, \mathbf{y}) \equiv p(\mathbf{y}|\mathbb{X}, \boldsymbol{\theta}) &\propto \prod_{i=1}^{n} \exp\Big(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i)^2\Big) \\ &= \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i)^2\Big) \\ &= \exp\Big(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbb{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbb{X}\boldsymbol{\theta})\Big) \end{aligned} \tag{7}$$

The maximum likelihood solution for $\boldsymbol{\theta}$ is the value that maximizes $L$, or equivalently minimizes $l \equiv -2\sigma^2 \log L$. But this is the same as Equation (2) which shows that the maximum likelihood solution is the same as that obtained by minimizing the sum of squared errors (Equation (3)).

### 4.4 Consistent estimation of best linear predictor

In the probabilistic view of Section 3, we made no assumptions about the relation between $\boldsymbol{X}$ and $Y$. Particularly, we did not assume it to be linear. However, we can still talk about the *best* linear predictor i.e. the linear predictor with the least risk.

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \mathrm{E}[(Y - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X})^2] = \arg\min_{\boldsymbol{\theta}} \mathrm{E}[(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})^2] \tag{8}$$

The derivative of the risk with respect to $\boldsymbol{\theta}$ is $-2\mathrm{E}[\boldsymbol{X}(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})]$. So $\boldsymbol{\theta}_*$ is given by

$$\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}]\boldsymbol{\theta}_* = \mathrm{E}[\boldsymbol{X}Y] \implies \boldsymbol{\theta}_* = (\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}])^{-1}\mathrm{E}[\boldsymbol{X}Y] \tag{9}$$

provided the inverse exists. The second derivative is $2\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}]$ which is positive semi-definite since, from Lemma 1, $\mathbf{b}\mathbf{b}^{\mathrm{T}}$ is positive semi-definite for any vector $\mathbf{b}$. But like before, this value depends on the unknown data distribution, and cannot be computed.
However, we can try to estimate the expression in Equation (9). Let $\mathrm{S} \equiv {}^{\mathbb{X}^{\mathrm{T}}\mathbb{X}}/_{n}$. Let $\mathrm{A}_i^j$ denote the element in row $i$, column $j$ of matrix A. For any $i, j \in \{1, \ldots, d\}$, we have by the law of large numbers:

$$\mathrm{S}_i^j = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k^i \mathbf{x}_k^j \xrightarrow{P} \mathrm{E}[\boldsymbol{X}^i \boldsymbol{X}^j] \implies \mathrm{S} \xrightarrow{P} \mathrm{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}] \tag{10}$$

So by the continuous mapping theorem,

$$\mathrm{S}^{-1} = n(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1} \xrightarrow{P} (\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}])^{-1} \tag{11}$$

Similarly define $\mathbf{q} \equiv {}^{\mathbb{X}^{\mathrm{T}}\mathbf{y}}/_{n}$, and note that for any $i \in \{1, \ldots, d\}$, we have

$$\mathbf{q}^i = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k^i \mathbf{y}_k \xrightarrow{P} \mathrm{E}[\boldsymbol{X}^i Y] \implies \mathbf{q} \xrightarrow{P} \mathrm{E}[\boldsymbol{X}Y] \tag{12}$$

Finally we can combine the results of Equations (11) and (12) with the continuous mapping theorem to get

$$\mathrm{S}^{-1}\mathbf{q} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} \xrightarrow{P} (\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}])^{-1}\mathrm{E}[\boldsymbol{X}Y] = \boldsymbol{\theta}_* \tag{13}$$

$\mathrm{S}^{-1}\mathbf{q}$ is the same value given by Equation (3) which shows that the OLS solution is a consistent estimator of the best linear predictor.

# 5 Ridge regression

One problem with the OLS solution (Equation (3)) is that it depends on $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ being invertible. The original motivation for ridge regression was to directly fix this problem. As shown in Lemma 2, for any positive semi-definite matrix A, and any positive number $\lambda$, the matrix $\mathrm{A} + \lambda\mathrm{I}$ is invertible. So the ridge regression solution modifies the OLS solution as

$$\boldsymbol{\theta}_* = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \lambda\mathrm{I})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} \tag{14}$$

for some positive $\lambda$. This solution always exists, independent of $\mathbb{X}$ and $\mathbf{y}$. The constant $\lambda$ is a hyperparameter that is usually picked by cross-validation. However, in the following subsections, we will look at ridge regression from two different perspectives to gain some intuition about this hyperparameter, and the overall solution in general. Note that setting $\lambda$ to 0 reduces this solution to the OLS solution. Given a new point $\mathbf{x}$, the prediction is given by

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_* = \mathbf{x}^{\mathrm{T}}(\mathbb{X}^{\mathrm{T}}\mathbb{X} + \lambda\mathrm{I})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} \tag{15}$$

Note that this too is a linear smoother.

## 5.1 Tikhonov regularization

When the OLS solution does not exist, the linear regression problem is said to be ill-posed. This can happen if $n < d$, or if the columns of $\mathbb{X}$ are highly correlated. A common strategy for solving ill-posed problems is Tikhonov regularization. It seeks to make the problem well posed by adding further constraints on the model. In the case of linear regression, this is achieved by adding another term to the OLS objective (Equation (2)) which penalizes the norm of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbb{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2 = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbb{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbb{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta} \quad (16)$$

Here $\lambda$ is a hyperparameter that controls the amount regularization. The derivative of the above expression with respect to $\boldsymbol{\theta}$ is $-2\mathbb{X}^{\mathrm{T}}\mathbf{y} + 2\mathbb{X}^{\mathrm{T}}\mathbb{X}\boldsymbol{\theta} + 2\lambda\boldsymbol{\theta}$. So,

$$(\mathbb{X}^{\mathrm{T}}\mathbb{X} + \lambda\mathrm{I})\boldsymbol{\theta}_* = \mathbb{X}^{\mathrm{T}}\mathbf{y} \implies \boldsymbol{\theta}_* = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \lambda\mathrm{I})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} \quad (17)$$

While this view motivates ridge regression as a way to make the problem well posed, the role of $\lambda$ is not very clear. This will be made more explicit in the next section.

## 5.2 Maximum a posteriori estimation

In Section 4.3, we assumed Gaussian noise, and found $\boldsymbol{\theta}$ by maximizing the likelihood. Let us now take this further with a Bayesian approach. We will put a prior distribution on $\boldsymbol{\theta}$; a normal distribution with mean $\mathbf{0}$ and variance $\Sigma$. With the Bayesian approach, we are interested in the posterior distribution of $\boldsymbol{\theta}$ i.e. the distribution given the data. We will assume that the covariates $\mathbf{x}_i$ are fixed, so by Bayes' theorem

$$p(\boldsymbol{\theta}|\mathbb{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbb{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbb{X})} \propto L(\boldsymbol{\theta}|\mathbb{X}, \mathbf{y})p(\boldsymbol{\theta}) \quad (18)$$

Here $L$ is the same likelihood calculated in Equation (7). The denominator $p(\mathbf{y}|\mathbb{X})$ can be ignored since it does not depend on $\boldsymbol{\theta}$ – it is just a normalizing constant to ensure that $p(\boldsymbol{\theta}|\mathbb{X}, \mathbf{y})$ is a valid distribution. Multiplying the likelihood and the prior, we see that the posterior distribution is proportional to

$$\begin{aligned} & \exp\Big(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbb{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbb{X}\boldsymbol{\theta})\Big) \exp\Big(-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta}\Big) \\ = \ & \exp\Big(-\frac{1}{2}\Big(\frac{1}{\sigma^2}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{2}{\sigma^2}\mathbf{y}^{\mathrm{T}}\mathbb{X}\boldsymbol{\theta} + \frac{1}{\sigma^2}\boldsymbol{\theta}^{\mathrm{T}}\mathbb{X}^{\mathrm{T}}\mathbb{X}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta}\Big)\Big) \\ \propto \ & \exp\Big(-\frac{1}{2}\Big(\boldsymbol{\theta}^{\mathrm{T}}(\Sigma^{-1} + \frac{1}{\sigma^2}\mathbb{X}^{\mathrm{T}}\mathbb{X})\boldsymbol{\theta} - \frac{2}{\sigma^2}\boldsymbol{\theta}^{\mathrm{T}}\mathbb{X}^{\mathrm{T}}\mathbf{y}\Big)\Big) \end{aligned} \quad (19)$$

Let $\mathrm{A} \equiv \Sigma^{-1} + {}^{\mathbb{X}^{\mathrm{T}}\mathbb{X}}/_{\sigma^2}$ and $\mathbf{b} \equiv {}^{\mathbb{X}^{\mathrm{T}}\mathbf{y}}/_{\sigma^2}$. We can use a very handy completion of squares trick to simplify the above expression as

$$\exp\Big(-\frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{A}\boldsymbol{\theta} - 2\boldsymbol{\theta}^{\mathrm{T}}\mathbf{b})\Big)$$

$$
\begin{aligned}
&= \exp\Big(-\frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{A}\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{T}}\mathrm{A}\mathrm{A}^{-1}\mathbf{b} - \mathbf{b}^{\mathrm{T}}\mathrm{A}^{-1}\mathrm{A}\boldsymbol{\theta})\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}\mathrm{A}\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{T}}\mathrm{A}\mathrm{A}^{-1}\mathbf{b} - \mathbf{b}^{\mathrm{T}}\mathrm{A}^{-1}\mathrm{A}\boldsymbol{\theta} + \mathbf{b}^{\mathrm{T}}\mathrm{A}^{-1}\mathrm{A}\mathrm{A}^{-1}\mathbf{b})\Big) \\
&= \exp\Big(-\frac{1}{2}(\boldsymbol{\theta} - \mathrm{A}^{-1}\mathbf{b})^{\mathrm{T}}\mathrm{A}(\boldsymbol{\theta} - \mathrm{A}^{-1}\mathbf{b})\Big) \qquad (20)
\end{aligned}
$$

Note that this expression has the form of a normal distribution with mean $\mathrm{A}^{-1}\mathbf{b}$ and variance $\mathrm{A}^{-1}$. This is not a coincidence. The prior distribution of $\boldsymbol{\theta}$ was chosen specifically for this property, and such distributions are called conjugate priors. The Bayesian approach does not give us a single estimate of $\boldsymbol{\theta}$; rather, it gives us a whole distribution. This can be very useful, and in fact, Bayesian regression directly uses this distribution to make predictions. However, for now we are just interested in a point estimate. This is typically obtained by taking the mean or mode of the posterior distribution. In the case of a Gaussian the two are equivalent, so our estimate is given by

$$
\boldsymbol{\theta}_* = \mathrm{A}^{-1}\mathbf{b} = \Big(\Sigma^{-1} + \frac{1}{\sigma^2}\mathbb{X}^{\mathrm{T}}\mathbb{X}\Big)^{-1}\frac{1}{\sigma^2}\mathbb{X}^{\mathrm{T}}\mathbf{y} = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \sigma^2\Sigma^{-1})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y} \qquad (21)
$$

This expression bears a lot of resemblance to the ridge regression solution (Equation (14)); in fact if we assume a diagonal variance for $\boldsymbol{\theta}$ i.e. $\Sigma = \nu^2\mathrm{I}$, then the expression becomes equivalent to Equation (14) with $\lambda = {}^{\sigma^2}/_{\nu^2}$. This captures the role of the regularization hyperparameter as the amount of noise relative to the uncertainty in $\boldsymbol{\theta}$. $\nu$ controls the breadth of the prior distribution, and signifies our uncertainty in $\boldsymbol{\theta}$. A large value indicates that we do not have strong prior beliefs, and in the limit $\nu \to \infty$ produces what is known as an "uninformative" prior. It causes $\lambda$ to shrink, and the solution goes closer to the OLS solution. Conversely, a small value of $\nu$ indicates a strong prior belief that $\boldsymbol{\theta}$ is close to $\mathbf{0}$ – it increases $\lambda$ and makes $\boldsymbol{\theta}$ remain close to $\mathbf{0}$.

# 6 Kernel ridge regression

Linear regression tries to fit a very simple model to the data, and so will not perform very well if the data deviates strongly from a line. However, it is possible to implicitly fit more complex models by making use of basis functions. Instead of finding a line as a function of $\mathbf{x}$, we can first transform $\mathbf{x}$ with a function $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$, and find a line as a function of $\boldsymbol{\phi}(\mathbf{x})$. Note that $m$ can be much larger than $d$ or even infinite. This simple change allows for much more complex functions to be fit through the data. As a simple example, consider unidimensional covariates $x$, and let $\phi(x) \equiv x^2$. Then a line in $\phi$-space is a quadratic curve in $x$-space. The regression solution remains the same. Let $\boldsymbol{\Phi} \equiv \begin{bmatrix}\boldsymbol{\phi}(\mathbf{x}_1) & \dots & \boldsymbol{\phi}(\mathbf{x}_n)\end{bmatrix}^{\mathrm{T}}$. Then the ridge regression solution in $\boldsymbol{\phi}$-space is simply

$$
\boldsymbol{\theta}_* = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda\mathrm{I})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} \qquad (22)
$$

However there is a problem with this solution. It requires explicitly transforming points with $\boldsymbol{\phi}$ and working with $\boldsymbol{\Phi}$. This is computationally expensive if $m$ is

very large, and impossible if it is infinite. So we will make use of the *kernel trick* which will let us make *implicit* transformations. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be any symmetric positive semi-definite function. Then by Mercer's theorem,

$$k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle \tag{23}$$

for some $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$. Mercer's theorem allows us to create an implicit mapping by using a kernel function. So, in our regression we will assume that $\boldsymbol{\phi}$ is generated by some Mercer kernel $k$. A common choice is the Gaussian kernel $k(\mathbf{x}, \mathbf{x}') \equiv \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2)$ for some constant $\sigma$. Let $\mathbb{K}$ be a matrix such that $\mathbb{K}_i^j \equiv \boldsymbol{\phi}(\mathbf{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$. Observe that $\mathbb{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$. Next we will make use of the following identity from Lemma 3. If P and R are positive definite matrices, then

$$(\mathrm{P}^{-1} + \mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B})^{-1}\mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1} = \mathrm{P}\mathrm{B}^{\mathrm{T}}(\mathrm{B}\mathrm{P}\mathrm{B}^{\mathrm{T}} + \mathrm{R})^{-1} \tag{24}$$

where B is such that the inverses are well defined. Taking $\mathrm{P}^{-1}$ equal to $\lambda\mathrm{I}$, R equal to I, and B equal to $\boldsymbol{\Phi}$, we get

$$(\lambda\mathrm{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}} = \frac{1}{\lambda}\boldsymbol{\Phi}^{\mathrm{T}}(\frac{1}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \mathrm{I})^{-1} = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda\mathrm{I})^{-1} \tag{25}$$

For any new point $\mathbf{x}$, let $\mathbf{k}_{\mathbf{x}} \equiv \begin{bmatrix} k(\mathbf{x}, \mathbf{x}_1) & \dots & k(\mathbf{x}, \mathbf{x}_n) \end{bmatrix}^{\mathrm{T}} = \boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x})$. From Equations (22) and (25), we have that the prediction at $\mathbf{x}$ is given by

$$\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}_* = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda\mathrm{I})^{-1}\mathbf{y} = \mathbf{k}_{\mathbf{x}}{}^{\mathrm{T}}(\mathbb{K} + \lambda\mathrm{I})^{-1}\mathbf{y} \tag{26}$$

$\mathbb{K}$ and $\mathbf{k}_{\mathbf{x}}$ both only make use of the kernel function $k$; so the above expression allows us to use kernel ridge regression without explicitly mapping points using $\boldsymbol{\phi}$. Observe that kernel ridge regression is also a form of linear smoothing.

# 7 Conclusion

We considered different approaches to simple linear regression and ridge regression. These different approaches, while leading to the same solution, provide different perspectives on these techniques, and show how they are related.

# A Appendix

**Lemma 1.** *Let* $\mathrm{A}$ *be any matrix, and* $\mathbf{b}$ *be any vector.* $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ *and* $\mathbf{b}\mathbf{b}^{\mathrm{T}}$ *are positive semi-definite.*

*Proof.* For any vector $\mathbf{x}$, we have

$$\mathbf{x}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\mathrm{A}\mathbf{x} = (\mathrm{A}\mathbf{x})^{\mathrm{T}}\mathrm{A}\mathbf{x} = \|\mathrm{A}\mathbf{x}\|^2 \geq 0$$

and

$$\mathbf{x}^{\mathrm{T}}\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{x} = (\mathbf{b}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}\mathbf{b}^{\mathrm{T}}\mathbf{x} = \|\mathbf{b}^{\mathrm{T}}\mathbf{x}\|^2 \geq 0$$

So $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ and $\mathbf{b}\mathbf{b}^{\mathrm{T}}$ are positive semi-definite. □

**Lemma 2.** *Let* $\mathrm{A}$ *be any positive semi-definite matrix, and* $\epsilon$ *be any positive scalar.* $\mathrm{A} + \epsilon\mathrm{I}$ *is invertible.*

*Proof.* Let $\mathbf{v}$ be an eigenvector of A with corresponding eigenvalue $\lambda$. Since A is positive semi-definite, $\lambda \geq 0$. We have

$$(\mathrm{A} + \epsilon\mathrm{I})\mathbf{v} = \mathrm{A}\mathbf{v} + \epsilon\mathbf{v} = (\lambda + \epsilon)\mathbf{v}$$

So $\mathbf{v}$ is also an eigenvector of $\mathrm{A} + \epsilon\mathrm{I}$ with eigenvalue $\lambda + \epsilon$. But $\lambda + \epsilon > 0$, and since the choice of $\mathbf{v}$ was arbitrary, this implies that all eigenvalues of $\mathrm{A} + \epsilon\mathrm{I}$ are positive. Hence it is invertible. □

**Lemma 3.** *Let* $\mathrm{P}$ *and* $\mathrm{R}$ *be positive definite matrices. Then*

$$(\mathrm{P}^{-1} + \mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B})^{-1}\mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1} = \mathrm{P}\mathrm{B}^{\mathrm{T}}(\mathrm{B}\mathrm{P}\mathrm{B}^{\mathrm{T}} + \mathrm{R})^{-1}$$

*for any* $\mathrm{B}$ *such that the inverses are well defined.*

*Proof.* We have

$$\begin{aligned}
&(\mathrm{P}^{-1} + \mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B})^{-1}\mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{B}\mathrm{P}\mathrm{B}^{\mathrm{T}} + \mathrm{R}) \\
&= (\mathrm{P}^{-1} + \mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B})^{-1}(\mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B}\mathrm{P}\mathrm{B}^{\mathrm{T}} + \mathrm{B}^{\mathrm{T}}) \\
&= (\mathrm{P}^{-1} + \mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B})^{-1}(\mathrm{B}^{\mathrm{T}}\mathrm{R}^{-1}\mathrm{B} + \mathrm{P}^{-1})\mathrm{P}\mathrm{B}^{\mathrm{T}} \\
&= \mathrm{P}\mathrm{B}^{\mathrm{T}}
\end{aligned}$$

Now multiplying both sides by $(\mathrm{B}\mathrm{P}\mathrm{B}^{\mathrm{T}} + \mathrm{R})^{-1}$ gives the desired result. □