ELE 535
Machine Learning and Pattern Recognition [1]
Handout #3: Convexity and Least Squares Regression

Peter J. Ramadge

Fall 2016, version 2.1

Chapter 7

Learning a Function

7.1 Introduction

So far we have examined learning simple structure in the data such a lying near a low dimensional subspace. This form of data analyses is called unsupervised learning. Unsupervised learning uses the geometry of the data to construct a more useful organization or representation of the data. Data clustering is another form of unsupervised learning. In a sense, these methods are exploratory in nature.

We now return the situation of labelled data and examine the problem of learning an unknown function from a finite set of (noisy) examples of its action. Specifically, we want to predict the value of a scalar variable y using measurements of a related set of variables $x \in \mathbb{R}^n$. To do so we have available a set of previously collected instances $\{(x_j, y_j)\}_{j=1}^m$ of the relationship between x and y. This is what makes the problems supervised: a "teacher" has supplied an output label for each example input.

Our objective is to use the labeled examples learn a function $\hat{g}$ that takes x as input and outputs a "good" estimate $\hat{y} = \hat{g}(x)$ of y. Of course, this presupposes that there is an underlying relationship between some (or all) of the variables in x and the variable y, and that this relationship is adequately captured in the labeled data $\{(x_j, y_j)\}_{j=1}^m$.

There are two common situations of interest. In the first, y takes values in a finite set and these values are interpreted as indicating category or class membership. The estimation of y from x is then termed *classification* and the function $\hat{g}$ that specifies the estimate is called a *classifier*. For example, x_j may be an embedding of an email message in $\mathbb{R}^n$ and its corresponding label y_j indicates whether or not the email is spam. Given a sufficiently large collection of labeled examples from both classes we want to learn a classifier that will accurately predict if a new email x is or is not spam. In the second situation y takes values in a real (or possibly complex) Euclidian space. In this case, the estimation of y from x is often termed *prediction* or *estimation* and the function $\hat{g}$ specifying the estimate is called a *predictor* or *estimator*. For example, $x_j \in \mathbb{R}^n$ could be the values of a set of image pixels surrounding a center group of pixels and its corresponding label y_j is the value of the center group of pixels. Given a sufficiently rich collection of labeled examples $\{(x_j, y_j)\}_{j=1}^m$ we want to learn a predictor that accurately predicts the value of the center pixels y given the values x of the surrounding pixels. At a surface level, these two problems are similar in nature. However, in the second case, the label space is a inner product space with a concepts of distance, geometry and similarity.

We first consider the situation of predicting a real value $y \in \mathbb{R}$. A common model in this case is that x and y are related by

$$y = f(x) + v, \tag{7.1}$$

where $f \colon \mathbb{R}^n \to \mathbb{R}$ is an unknown function and $v \in \mathbb{R}$ is a noise term. Typically one assumes the noise term is random with zero mean and finite variance. For this model, it is natural to call the vector variable x

the *input variable*, and the scalar variable y the *output variable*. The components of x are called *features* or *predictor variables* and the variable y is called the *response* or *dependent variable*.

Partial information about the relationship between x and y is provided by a set of example input-output pairs $D = \{(x_j, y_j)\}_{j=1}^m$, with $y_j = f(x_j) + v_j$, $j = 1, \dots, m$. The set of labelled examples D is called the *training data*, and each instance $(x_j, y_j) \in D$, is a *training example*. Typically one assumes that these examples are drawn independently from an unknown joint probability distribution $p(x, y)$.

The objective is to use the training data to learn a function $\hat{g} \colon \mathbb{R}^n \to \mathbb{R}$ from some class of functions $\mathcal{G}$, such that $\hat{g}(x)$ provides a good approximation to y. For example, $\mathcal{G}$ might be the set of linear functions $g(x) = w^T x$ for $w \in \mathbb{R}^n$. In this case, $\hat{g} \in \mathcal{G}$ can be selected by learning its parameter vector $\hat{w} \in \mathbb{R}^n$.

One way to gauge how well $\hat{g}$ is predicting the output value is through the average squared prediction error on the training data:

$$\frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{f}(x_j))^2.$$

Indeed, this particular metric is often used in the training process since it is a simple convex function of the vector of prediction errors. However, what is ultimately important is how well $\hat{g}$ *generalizes* to new data. By this we mean that on average $(y - \hat{g}(x))^2$ should be small for previously unseen examples drawn from $p(x, y)$. In practice, this metric could be approximated by averaging $(y - \hat{g}(x))^2$ over a set of *test data* $T = \{(x_i, y_i)\}_{i=1}^k$. Each *test example* $(x_i, y_i) \in T$ is assumed to be drawn independently from the same distribution as the training data but has not used in the training process.

Several aspects determine how well the selected function $\hat{g}$ generalizes to new data. One aspect is how well one can fit the function f in (7.1) using a function $g \in \mathcal{G}$. Typically $f \notin \mathcal{G}$, so approximating f using $\hat{g} \in \mathcal{G}$ incurs a structural error $(f(x) - \hat{g}(x))^2$. In principle, this error could be reduced by enlarging $\mathcal{G}$. A second aspect is the noise term v. Even if $f \in \mathcal{G}$ and we could learn $\hat{g} = f$ exactly, we would still incur the squared error v^2. This imposes a fundamental limit on the performance of any predictor. A third aspect is the size and richness of the training set. The more complex the class $\mathcal{G}$, the more difficult it will be select $\hat{g} \in \mathcal{G}$ to best match f using a finite set of training data. This suggests that there will be a trade-off between the complexity of the class $\mathcal{G}$ and the generalization performance of $\hat{g}$. Making $\mathcal{G}$ simpler (more constrained) encourages selecting $\hat{g}$ to fit the overall pattern of the training data, rather than idiosyncratic aspects (noise, errors). The latter phenomenon is known as *overfitting*. But it also potentially results in a greater structural error.

In summary, we want to examine how to learn a function $\hat{g}$ in some restricted class of functions $\mathcal{G}$ such that $\hat{g}$ does a good job of predicting the labels on the training set. This is typically quantified using a convex loss function L of the vector of errors $[g(x_j) - y_j]$. However, what is ultimately important is how well this predictor generalizes to new test data.

We begin by considering the class $\mathcal{G}$ of linear functions. In addition, since the loss function in both training and testing is often a convex function, we first review the relevant properties of convex functions.

Chapter 8

Convexity

8.1 Preliminaries: Bounded, Closed and Compact Sets

A subset $S \subset \mathbb{R}^n$ is *bounded* if there exists $B > 0$, such that for each $x \in S$, $\|x\|_2 \leq B$. So a bounded set is, as the name implies, bounded in extent. The set S is *closed* if, roughly speaking, it contains its boundary. For example, the interval $[0, 1]$ in $\mathbb{R}$ is closed but the interval $(0, 1)$ is not since it does not contain its boundary points 0 and 1. A more precise definition is that if $\{x_k\}$ is a sequence of points in S that converges as $k \to \infty$ to a point $x \in \mathbb{R}^n$, then $x \in S$. So a closed set contains the limits of its convergent sequences. The subset S is said to be *compact* if it is both closed and bounded. The implications of this are very important. For a start, every real valued continuous function defined on a compact subset $S \subset \mathbb{R}^n$ achieves a minimum and maximum value over S.

8.2 Convex Sets

For $x_1, x_2 \in \mathbb{R}^n$ and each $\alpha \in [0, 1]$, let $x_\alpha = (1 - \alpha)x_1 + \alpha x_2$. As α ranges from 0 to 1, the point x_α traces out the line segment from x_1 to x_2. We can also write $x_\alpha = x_1 + \alpha(x_2 - x_1)$. So x_α starts at x_1 when $\alpha = 0$ and proceeds in a straight line in the direction $x_2 - x_1$ reaching x_2 when $\alpha = 1$.

A subset $S \subset \mathbb{R}^n$ is *convex* if for each $x_1, x_2 \in S$ the line segment joining x_1 and x_2 is contained in S. So S is convex if for each $x_1, x_2 \in S$ and each $\alpha \in [0, 1]$,

$$x_\alpha = (1 - \alpha)x_1 + \alpha x_2 \in S.$$

Example 8.2.1. Here are some simple examples of convex sets:

a) The empty set and $\mathbb{R}^n$ are trivial convex subsets of $\mathbb{R}^n$.

b) Any subspace of $\mathcal{U} \subset \mathbb{R}^n$ is convex set.

c) An *affine manifold* is a set of the form $S = s + \mathcal{U} = \{x = s + u, u \in \mathcal{U}\}$ where $s \in \mathbb{R}^n$ and $\mathcal{U}$ is a subspace of $\mathbb{R}^n$. An affine manifold is a convex set.

d) A *closed half space* $H \subset \mathbb{R}^n$ is a set of the form $\{x \colon a^T x \leq b\}$ where $a \in \mathbb{R}^n$ with $a \neq 0$ and $b \in \mathbb{R}$. Thus a closed half space is one side of an $n - 1$ dimensional plane in $\mathbb{R}^n$ including the plane itself. A closed half space is a (closed) convex set.

8.2.1 Properties of Convex Sets

Theorem 8.2.1. Properties of convex sets:

a) Closure under intersection: If for each $a \in A$, $S_a \subset \mathbb{R}^n$ is convex, then $\cap_{a \in A} S_a$ is convex.

b) Image under a linear map: If $S \subset \mathbb{R}^n$ is convex and F is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, then $F(S) = \{z \colon z = Fs, s \in S\}$ is convex.

Proof. Exercise. □

Corollary 8.2.1. Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then $\{x \colon Ax \leq b\}$ is a convex set.

Proof. Let $S = \{x \colon Ax \leq b\}$. Then S is the intersection of the half spaces $\{x \colon a_j^T x \leq b_j\}$ where a_j^T is the j-th row of A and b_j is the j-th entry of b. Since half spaces are convex, the result then follows by Theorem 8.2.1. □

8.3 Convex Functions

A function f mapping $\mathbb{R}^n$ into $\mathbb{R}$ is a *convex function* if for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0, 1]$:

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y). \tag{8.1}$$

For $\alpha \in [0, 1]$, the point $z_\alpha = (1-\alpha)x + \alpha y$ lies on the line joining x $(\alpha = 0)$ to y $(\alpha = 1)$. On the other hand, the scalar value $c_\alpha = (1-\alpha)f(x) + \alpha f(y)$ is the corresponding linear interpolation of the values $f(x)$ $(\alpha = 0)$ and $f(y)$ $(\alpha = 1)$. Convexity requires that the value of the function f along the line segment from x to y is no greater that the corresponding linear interpolation $(1-\alpha)f(x) + \alpha f(y)$. This is illustrated in Fig. 8.1.

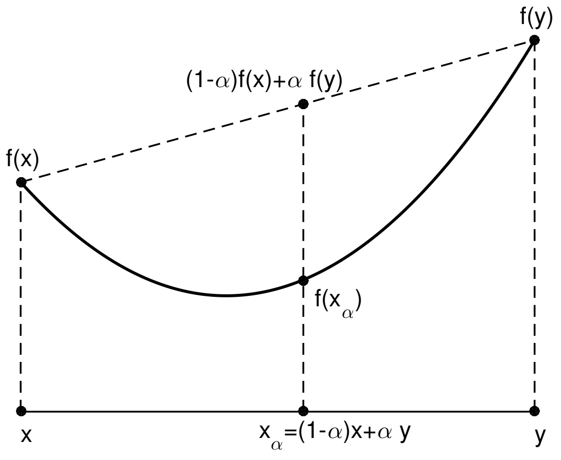

Figure 8.1: Illustration of the concept of a convex function f.

A function $f \colon \mathbb{R}^n \to \mathbb{R}$ is a *strictly convex function* if for all $x, y \in \mathbb{R}^n$, with $x \neq y$, and all $\alpha \in (0, 1)$:

$$f((1-\alpha)x + \alpha y) < (1-\alpha)f(x) + \alpha f(y). \tag{8.2}$$

It is easy to see that a strictly convex function is convex and that not every convex function is strictly convex.

8.4 Some Classes of Convex Functions

Theorem 8.4.1. For any $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the affine function $f(x) = a^T x + b$ is convex.

Proof. Let $\alpha \in [0, 1]$ and $x, y \in \mathbb{R}^n$. Then

$$\begin{aligned} f((1-\alpha)x + \alpha y) &= (1-\alpha)a^T x + \alpha a^T y + b \\ &= (1-\alpha)(a^T x + b) + \alpha(a^T y + b) \\ &= (1-\alpha) f(x) + \alpha f(y). \end{aligned}$$

Hence f is convex. □

Theorem 8.4.2. Every norm on $\mathbb{R}^n$ is a convex function.

Proof. Exercise. □

8.5 Properties of Convex Functions

8.5.1 Combining Convex Functions

Theorem 8.5.1. Let f and g be a convex functions on $\mathbb{R}^m$. Then the following functions are convex:

a) $h(x) = \beta f(x)$ where $\beta \geq 0$.

b) $h(x) = f(x) + g(x)$.

c) $h(x) = \max\{f(x), g(x)\}$.

d) $h(x) = f(Ax + b)$ where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

e) $h(x) = g(f(x))$ when in addition to being convex, g is nondecreasing on the range of f.

Proof. Let $\alpha \in [0, 1]$ and $x, y \in \mathbb{R}^n$. Part a): Exercise.
Part b):

$$\begin{aligned} h((1-\alpha)x + \alpha y) &= f((1-\alpha)x + \alpha y) + g((1-\alpha)x + \alpha y) \\ &\leq (1-\alpha) f(x) + \alpha f(y) + (1-\alpha) g(x) + \alpha g(y) \\ &= (1-\alpha)(f(x) + g(x)) + \alpha(f(y) + g(y) \\ &= (1-\alpha) h(x) + \alpha h(y). \end{aligned}$$

Part c):

$$\begin{aligned} h((1-\alpha)x + \alpha y) &= \max\{f((1-\alpha)x + \alpha y), g((1-\alpha)x + \alpha y)\} \\ &\leq \max\{(1-\alpha) f(x) + \alpha f(y), (1-\alpha) g(x) + \alpha g(y)\} \\ &\leq (1-\alpha) \max\{f(x), g(x)\} + \alpha \max\{f(y), g(y)\} \\ &= (1-\alpha) h(x) + \alpha h(y). \end{aligned}$$

Part d):

$$\begin{aligned} h((1-\alpha)x + \alpha y) &= f((1-\alpha)(Ax) + \alpha(Ay) + b) \\ &= f((1-\alpha)(Ax + b) + \alpha(Ay + b)) \\ &\leq (1-\alpha) f(Ax + b) + \alpha f(Ay + b) \\ &= (1-\alpha) h(x) + \alpha h(y). \end{aligned}$$

Part e): Since f is convex, $f((1-\alpha)x+\alpha y) \le (1-\alpha)f(x)+\alpha f(y)$. Combining this with the properties of g (nondecreasing and convex) yields

$$\begin{aligned} h((1-\alpha)x+\alpha y) &= g(f((1-\alpha)x+\alpha y)) \\ &\le g((1-\alpha)f(x)+\alpha f(y)) \\ &\le (1-\alpha)g(f(x))+\alpha g(f(y)) \\ &= (1-\alpha)h(x)+\alpha h(y). \end{aligned}$$

□

Here is a simple result obtained by using the above properties.

Theorem 8.5.2. If $P \in \mathbb{R}^{n\times n}$ is symmetric positive semidefinite, $q \in \mathbb{R}^n$ and $r \in R$, then the quadratic function $f(w) = {}^1\!/\!_2\, w^T P w + q^T w + r$ is convex.

Proof. The the affine function $q^T w + r$ is convex (Theorem 8.4.1). The norm $\|\cdot\|_2$ is convex and hence the function $\|\sqrt{P}w\|_2$ is convex (Theorem 8.5.1) and clearly takes nonnegative values. The function $g(x) = x^2$ is convex and monotone increasing on the nonnegative reals. Hence $w^T P w = \|\sqrt{P}w\|_2^2$ is a convex function (Theorem 8.5.1). Finally, since nonnegative scaling preserves convexity and the sum of convex functions is convex, $f(w) = {}^1\!/\!_2\, w^T P w + q^T w + r$ is convex. □

8.5.2 Sublevel Sets and Local Minima

Theorem 8.5.3. Let f be a convex function on $\mathbb{R}^n$. Then for each $c \in \mathbb{R}$, the sublevel set $L_c = \{x\colon f(x) \le c\}$ is a convex set.

Proof. If L_c is empty, then it is convex. Otherwise let $x, y \in L_c$. Then $f((1-\alpha x)+\alpha y) \le (1-\alpha)f(x)+\alpha f(y) \le c$. Hence L_c is convex. □

Theorem 8.5.4. If a convex function f has a finite valued local minimum at $x^\star \in \mathbb{R}^n$, then $x^\star$ is a global minimum point of f.

Proof. Let $f(x^\star) = c > -\infty$. Then there exists $r > 0$ such that for all x with $\|x - x^\star\|_2 < r$, $f(x) \ge c$. Suppose that for some $z \in \mathbb{R}^n$, $f(z) < c$. Let $x_\alpha = (1-\alpha)x^\star + \alpha z$ with $\alpha \in (0,1)$. Then for $\alpha > 0$ sufficiently small, $\|x_\alpha - x^\star\|_2 < r$ and $f(x_\alpha) \le (1-\alpha)f(x^\star) + \alpha f(z) < c$; a contradiction. Hence $x^\star$ is a global minima. □

Theorem 8.5.5. The set of all global minima of a convex function is a convex set.

Proof. Let $x^\star$ be a global minima with $f(x^\star) = c > -\infty$. The set of all global minima is the sublevel set $L_c = \{x :: f(x) \le c\}$. Hence by Theorem 8.5.3, the set of all global minima is convex. □

We end out list of properties with the following uniqueness result.

Theorem 8.5.6. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be strictly convex. If f has a finite valued local minimum at $x^\star$, then $x^\star$ is the unique global minimum point of f.

Proof. Let $x^\star$ be a local minimum point with $f(x^\star) = c > -\infty$. Then $x^\star$ is a global minimum point. If there are two distinct global minima, say at $x^\star$ and $y^\star$, then all points on the line joining $x^\star$ and $y^\star$ are global minima. But this violates the strict convexity of f. □

8.6 Differentiable Convex Functions

The following theorem shows that a differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex if and only if at every point the best local linear approximation to f is a global lower bound for f.

Theorem 8.6.1. A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for all $u, v \in \mathbb{R}^n$,

$$f(v) - f(u) \geq Df(u)(v-u). \tag{8.3}$$

Proof. (If) Assume that the bound (8.3) holds. Let $x, y \in \mathbb{R}^n$ and consider the point $x_\alpha = (1-\alpha)x + \alpha y$ for some $\alpha \in [0,1]$. We want to show that $(1-\alpha)f(x) + \alpha f(y) \geq f(x_\alpha)$. Applying (8.3) at $u = x_\alpha$ we obtain the lower bound $g(v) = f(x_\alpha) + Df(x_\alpha)(v - x_\alpha)$ to $f(v)$. Evaluating this bound at $v = x$ and $v = y$ gives $f(x) \geq g(x)$, $f(y) \geq g(y)$, and hence $(1-\alpha)f(x) + \alpha f(y) \geq (1-\alpha)g(x) + \alpha g(y)$. Substituting the expressions for $g(x)$ and $g(y)$ into the previous equation yields the desired inequality:

$$\begin{aligned}(1-\alpha)f(x) + \alpha f(y) &\geq (1-\alpha)(f(x_\alpha) + Df(x_\alpha)(x - x_\alpha)) + \alpha(f(x_\alpha) + Df(x_\alpha)(y - x_\alpha))\\ &= f(x_\alpha) + Df(x_\alpha)[(1-\alpha)(x - x_\alpha) + \alpha(y - x_\alpha)]\\ &= f(x_\alpha).\end{aligned}$$

(Only If) Assume f is convex. Then $f(x + \alpha(y-x)) \leq f(x) + \alpha(f(y) - f(y))$. Hence $f(x + \alpha(y - x)) - f(x) \leq \alpha(f(y) - f(x))$. Dividing both sides by α gives

$$\frac{f(x + \alpha(y-x)) - f(x)}{\alpha} \leq f(y) - f(x).$$

Then taking the limit as α approaches 0 yields $Df(x)(y - x) \leq f(y) - f(x)$. □

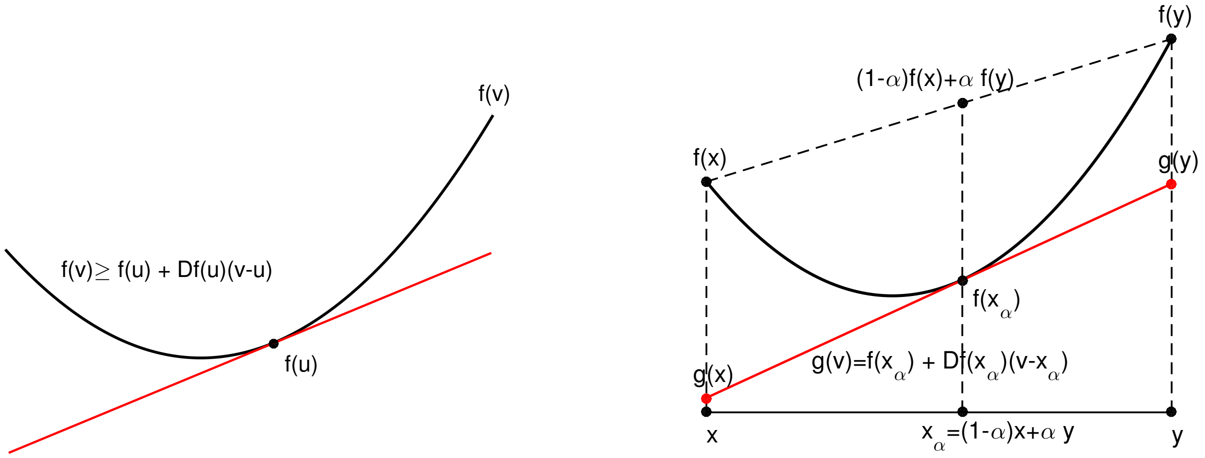

Figure 8.2: Left: Illustration of the bound (8.3). The function $f(v)$ is bounded below by the best linear approximation to the function at any point u. Right: Illustration of the bound (8.3) applied at the point x_α. We see that $f(x) \geq g(x)$ and $f(y) \geq g(y)$. Hence $(1-\alpha)f(x) + \alpha f(y) \geq (1-\alpha)g(x) + \alpha g(y) = f(x_\alpha)$.

8.6.1 Minimization of a Convex Function Over a Convex Set

The basic problem of interest is minimizing a convex function f over a convex set S.

Theorem 8.6.2. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function and S be a convex subset of $\mathbb{R}^n$. A point $x^\star$ minimizes f over S if and only if $x^\star \in S$ and for each $y \in S$

$$Df(x^\star)(y - x^\star) \geq 0. \tag{8.4}$$

Proof. (If) Suppose $x^\star \in S$ and (8.4) holds. Since f is convex, for each $y \in S$, $f(y) - f(x^\star) \geq Df(x^\star)(y - x^\star)$. Hence by (8.4), $f(y) \geq f(x^\star)$.

(Only If) Suppose $x^\star$ minimizes of f over S. Then clearly $x^\star \in S$. If (8.4) does not hold, then for some $y \in S$, $Df(x^\star)(y - x^\star) < 0$. Let $h = y - x^\star$ and note that for $\alpha \in [0, 1]$, $x^\star + \alpha h = (1 - \alpha)x^\star + \alpha y \in S$. Using the definition of the derivative we have

$$Df(x^\star)h = \lim_{\alpha \downarrow 0} \frac{f(x^\star + \alpha h) - f(x^\star)}{\alpha} < 0.$$

By the definition of a limit, there exists $\alpha_0 > 0$ such that for all $0 < \alpha \leq \alpha_0$, $f(x^\star + \alpha h) - f(x^\star) < 0$. For such α, $x^\star + \alpha h \in S$ and $f(x^\star + \alpha h) < f(x^\star)$; a contradiction. □

Sometimes you will see equation (8.4) written as $\nabla f(x^\star)^T (y - x^\star) \geq 0$, where $\nabla f(x)$ is the gradient of f at x. This is equivalent since $\nabla f(x)^T = Df(x)$. The idea behind (8.4) is very simple. Since $y \in S$ and S is convex, $y - x^\star$ is a direction we can move from $x^\star$ without leaving S. Now $x^\star$ is a minimizer of $f(x)$, so $f(x)$ can't decrease as we move from $x^\star$ in the direction $y - x^\star$. Hence $y - x^\star$ must point in a direction of increase of f. Since $\nabla f(x^\star)$ is the direction of greatest increase in f at $x^\star$, that means $\nabla f(x^\star)^T (y - x^\star) \geq 0$.

8.7 Appendix: Convex Functions $f\colon \mathbb{R} \to \mathbb{R}$

We give a brief review of convex sets in $\mathbb{R}$ and convex functions mapping an interval of $\mathbb{R}$ into $\mathbb{R}$.

A convex subset of $\mathbb{R}$ must be an interval. If it is bounded, then it takes one the forms: (a, b), $(a, b]$, $[a, b)$, or $[a, b]$, where $a, b \in \mathbb{R}$ and $a < b$. If it is unbounded, then it takes one of the forms (a, ∞), $[a, \infty)$, $(-\infty, a]$, $(-\infty, a)$, or $\mathbb{R}$. In all cases, the interior of the interval has the form (a, b), where a could be $-\infty$ and b could be ∞.

Here are some useful results on convex functions $f\colon I \to \mathbb{R}$, where I is an interval.

Lemma 8.7.1 (Jenson's Inequality)**.** The value of a convex function at a (weighted) average of a set of points is less than the same average of the function values at these points.

Proof. Exercise. □

For example, if f is a convex function on an interval I, then for nonnegative α, β with $\alpha + \beta = 1$, $f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y)$. In particular,

$$f\left(\frac{x + y}{2}\right) \leq \frac{f(x) + f(y)}{2}.$$

More generally, if $\{x_i\}_{i=1}^k \subset I$, $\alpha_i > 0$ with $\sum_{i=1}^k \alpha_i = 1$, then

$$f\left(\sum_{i=1}^k \alpha_i x_i\right) \leq \sum_{i=1}^k \alpha_i f(x_i).$$

Lemma 8.7.2. Let I be an interval of $\mathbb{R}$. If $f\colon I \to \mathbb{R}$ is convex, then f is continuous on the interior of I.

Proof. See any standard text covering convex functions. □

Lemma 8.7.3. Let $f\colon I \to \mathbb{R}$ be twice differentiable on an open interval I. Then f is convex on I if and only if $f''(x) \geq 0$ at each point $x \in I$.

Proof. (IF) By Taylor's theorem $f(x+h) = f(x) + f'(x)h + f''(z)h^2$ where z is a point between x and $x+h$. Hence $f(x+h) \geq f(x) + f'(x)h$. Since f is bounded below by its derivative at any point, it is convex.

(ONLY IF) $f(x) = f(\frac{x+h+x-h}{2}) \leq \frac{f(x+h)+f(x-h)}{2}$. Hence $f(x+h) - 2f(x) + f(x-h) \geq 0$. The second derivative of f at x is found by taking the limit as $h \downarrow 0$ of

$$\frac{1}{h}\left(\frac{f(x+h)-f(x)}{h} - \frac{f(x)-f(x-h)}{h}\right) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \geq 0$$

Hence $f''(x) \geq 0$. □

Example 8.7.1. Some examples:

a) $f(x) = x^2$ has $f^{''}(x) = 2 > 0$. Hence f is convex on every open interval. It follows that it is also convex on every closed interval.

b) $f(x) = \sqrt{x}$ on the interval $[0,\infty)$. Since $f^{''}(x) = -\frac{1}{4}x^{-3/2}$ is negative for $x > 0$, f is not convex on $[0,\infty)$. For example, $f(1) = 1$, $f(4) = 2$ and $(f(1) + f(4))/2 = 1.5$. On the other hand, $f((1+4)/2) = \sqrt{5/2} \approx 1.581 > 1.5$. However, $g(x) = -\sqrt{x}$ is convex on $[0,\infty)$.

c) $f(x) = x \ln x$ for $x \in (0,\infty)$. This has $f^{'}(x) = \ln x + 1$ and $f^{''}(x) = 1/x$. Since $f^{''}(x) > 0$ at each $x \in (0,\infty)$, f is convex on $(0,\infty)$.

8.8 Exercises

Exercise 8.1. Show that each of the examples in Example 8.2.1 is indeed convex.

Exercise 8.2. Prove Theorem 8.2.1.

Exercise 8.3. Prove that every norm is a convex function.

Exercise 8.4. Determine general sufficient conditions (if any exist) under which the indicated function is convex.
(a) $f\colon [0,\infty) \to \mathbb{R}$ with $f(x) = x^r$.
(b) $f\colon \mathbb{R} \to \mathbb{R}$ with $f(x) = |x|$.
(c) $f\colon \mathbb{R} \to \mathbb{R}$ with $f(x) = |x|^r$.
(d) $f\colon (0,\infty) \to \mathbb{R}$ with $f(x) = 1/x^r$.
(e) $f\colon [d,\infty) \to \mathbb{R}$ with $f(x) = ax^3 + bx^2 + c$.
(f) $f\colon \mathbb{R}^n \to \mathbb{R}$ with $f(x) = (x^TQx)^r$. Here $Q \in \mathbb{R}^{n\times n}$ is symmetric PSD.
(g) $f\colon \mathbb{R}^n \to \mathbb{R}$ with $f(x) = 1 + e^{(\sum_{i=1}^n |x(i)|)^r}$.

8.9 Notes and References

The material is this chapter is standard and can be found in any good book on optimization. See, for example, the books by Chong and Zak [3], Boyd and Vandenberghe [2], and Bertsekas [1].

Chapter 9

Least Squares Regression

9.1 Learning a Linear Function

Let $\mathcal{G}$ be the family of linear functions $\{g\colon g(x) = w^T x, w \in \mathbb{R}^n\}$. This set is parameterized by the vector variable $w \in \mathbb{R}^n$. Our objective is to use a finite set of training examples $\{(x_j, y_j) \in \mathbb{R}^n \times \mathbb{R}\}_{j=1}^m$ to learn $\hat{w} \in \mathbb{R}^n$ so that the linear function $\hat{g}(x) = \hat{w}^T x$ best approximates the hidden relationship between x and y.

Let $X = [x_1, \dots, x_m] \in \mathbb{R}^{n \times m}$ be the matrix of input training data and $y = [y_j] \in \mathbb{R}^m$ be the vector of corresponding output values. For given $w \in \mathbb{R}^n$, the corresponding vector of predicted values $\hat{y} \in \mathbb{R}^m$ and prediction errors $\varepsilon = y - \hat{y}$ on the training data are

$$\hat{y} = X^T w \qquad \varepsilon = y - X^T w. \tag{9.1}$$

The error vector ε is often called the *residual*. It is the part of y that is "unexplained" by $X^T w$. Let $A = X^T$; this is simply a rearrangement of the input data. Each input example is now a row in A and each column of A corresponds to a measured feature. Hence the columns of A are often called *features*.

The objective is to learn the "best" value for the parameter vector w. These best value should some how be linked to good prediction performs. Hence we need a means of assessing prediction performance of linear predictors over $w \in \mathbb{R}^n$. Problems of this form are often called *linear regression problems*. Linear regression finds the "best approximation" to the vector y as a linear combination of the columns $a_1, \dots, a_n$ of A by minimizing some cost function of the residual $\varepsilon = y - Aw$ on the training data. In this context, the matrix A is often called the *regression matrix* or *design matrix* and a column of A is called a *regressor*.

9.2 Ordinary Least Squares

Least squares is a linear regression method based selecting the parameter w to minimize the Euclidean norm of the residual $\|\varepsilon\|_2^2 = \|y - Aw\|_2^2$. Since this metric is the sum of squared errors: $\sum_j \varepsilon(j)^2$, it is called the *least squares* or *residual sum of squares* (RSS) objective.

This gives rise to the standard *least squares problem*:

$$w^\star = \arg\min_{w \in \mathbb{R}^n} \|y - Aw\|_2^2. \tag{9.2}$$

We do not claim that this method of learning is "optimal" beyond that $w^\star$ minimizes the least squares objective function on the training data.

9.2.1 Some Simple Variations

Once we have solved problem (9.2) and determined the properties of the solution, we can often transfer these results and insights to related forms of least squares. Here are some examples of closely related problems.

(1) Learning an affine function

Instead of learning a linear function suppose we want to learn an affine function of the form $g(x) = w^T x + \beta$. Letting A and y be defined as before, you then set out to solve:

$$w^\star, \beta^\star = \arg \min_{w \in \mathbb{R}^n, \beta \in \mathbb{R}} \|y - Aw - \beta\|_2^2.$$

By a simple reorganization this can be recast as a standard least squares problem. Let $\tilde{A}$ be formed from A by adding a column of all 1's at the right. So $\tilde{A} = \begin{bmatrix} A & \mathbf{1} \end{bmatrix}$. Similarly, let $z = \begin{bmatrix} w^T & \beta \end{bmatrix}^T$. Then the above problem is equivalent to the standard problem:

$$z^\star = \arg \min_{z \in \mathbb{R}^{n+1}} \|y - \tilde{A}z\|_2^2,$$

with $z^\star = \begin{bmatrix} w^{\star T} & \beta^\star \end{bmatrix}^T$.

(2) Quadratic performance metric

Instead of the Euclidean norm suppose we can use the quadratic norm $\|x\|_P = \sqrt{x^T P x}$ where $P \in \mathbb{R}^{n \times n}$ is symmetric positive definite. This is sometimes called a *Mahalanobis norm*. In this case the least squares problem becomes

$$w^\star = \arg \min_{w \in \mathbb{R}^n} \|Fw - g\|_P^2, \tag{9.3}$$

This form uses a modified norm (akin to a Mahalanobis norm) to measure the residual error.

By a simple transformation, the problem (9.3) can be written in the standard form (9.2). To see this, we introduce the symmetric positive definite square root of the matrix P. First use an eigendecomposition of P to write

$$P = U \Lambda U^T = U \Lambda^{1/2} U^T U \Lambda U^T.$$

Denote the diagonal entries of Λ by λ_i, $i = 1, \dots, m$, and let $\sqrt{\Lambda}$ be the diagonal matrix with ii-th entry $\sqrt{\lambda_i}$. Let $\sqrt{P} = U \Lambda^{1/2} U^T$. Clearly $P = \sqrt{P}\sqrt{P}$. The matrix $\sqrt{P}$ is the unique symmetric positive definite square root of P.

Using the square root of P we can write

$$\|Fw - g\|_P^2 = (Fw - g)^T P (Fw - g) = \|\sqrt{P}Fw - \sqrt{P}g\|_2^2 = \|\tilde{F}w - \tilde{g}\|_2^2.$$

This reduces (9.3) to a standard least squares problem with a modified regressor matrix $\tilde{F} = \sqrt{P}F$ and output vector $\tilde{g} = \sqrt{P}g$.

(3) Multiple least squares objectives

As a second example consider a problem with multiple least squares objectives:

$$w^\star = \arg \min_{w \in \mathbb{R}^n} \|F_1 w - g_1\|_2^2 + \|F_2 w - g_2\|_2^2. \tag{9.4}$$

Noting that the sum $\|F_1w - g_1\|_2^2 + \|F_2w - g_2\|_2^2$ is just a sum of squares we can write:

$$\|F_1w - g_1\|_2^2 + \|F_2w - g_2\|_2^2 = \left\| \begin{bmatrix} F_1w - g_1 \\ F_2w - g_2 \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} w - \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \right\|_2^2 = \|\tilde{F}w - \tilde{g}\|_2^2,$$

where

$$\tilde{F} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \in \mathbb{R}^{(m_1+m_2)\times n} \quad \text{and} \quad \tilde{y} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \in \mathbb{R}^{m_1+m_2}.$$

This reduces (9.4) to a standard least squares problem with an augmented regression matrix $\tilde{F}$ and output vector $\tilde{g}$. A similar transformation can be applied if the objective is a finite sum of quadratic terms: $\sum_{j=1}^{k} \|F_jw - g_j\|_2^2$.

(4) Ridge regression

The third example is called *ridge regression*:

$$w^\star = \arg\min_{w\in\mathbb{R}^n} \|Fw - g\|_2^2 + \lambda\|w\|_2^2. \tag{9.5}$$

Here the scalar parameter $\lambda > 0$ is selected to appropriately balance the competing objectives of minimizing the residual squared error while keeping w small. Notice that (9.5) is a special case of (9.4) with $F_1 = F$, $g_1 = g$, $F_2 = \sqrt{\lambda}I_n$ and $g_2 = \mathbf{0}$. Hence (9.5) can be transformed into a standard least squares problem with the objective $\|\tilde{F}w - \tilde{g}\|_2^2$ where

$$\tilde{F} = \begin{bmatrix} F \\ \sqrt{\lambda}I_n \end{bmatrix} \in \mathbb{R}^{(m+n)\times n} \quad \text{and} \quad \tilde{y} = \begin{bmatrix} g \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m+n}.$$

9.3 The Least Squares Solution

The objective function in (9.2) is the square of a norm of an affine function of w. Hence the objective function is convex. Thus any local minimum is a global minimum. The objective function is also twice differentiable with respect to w. We show below that any stationary point of the derivative is a solution of (9.2).

Expanding the objective function in (9.2) yields

$$\begin{aligned} J(w) &= \|y - Aw\|_2^2 \\ &= (y - Aw)^T(y - Aw) \\ &= y^Ty - y^TAw - w^TA^Ty + w^TA^TAw. \end{aligned}$$

Computing $DJ(w)h$ (the derivative acting on $h \in \mathbb{R}^n$) with respect to w and setting this equal to zero, gives

$$\begin{aligned} DJ(w)h &= -y^TAh - h^TA^Ty + w^TA^TAh + h^TA^TAw \\ &= 2(-y^TA + w^TA^TA)h \\ &= 0. \end{aligned}$$

At a stationary point, equality must hold for all $h \in \mathbb{R}^n$. Hence we deduce that a necessary condition for $w^\star$ to be a solution of the least squares problem is

$$A^TAw^\star = A^Ty\,. \tag{9.6}$$

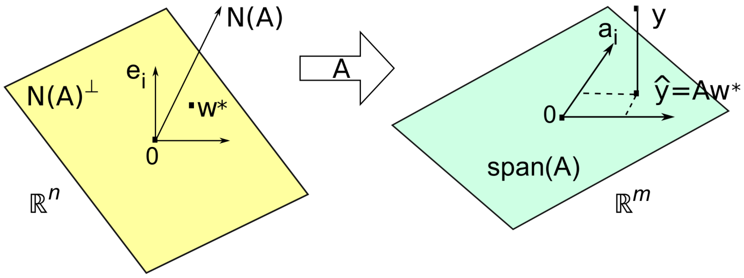

Figure 9.1: Least squares regression. Find $\mathbf{w}^\star$ such that $A\mathbf{w}^\star = \hat{y}$.

These are called the *normal equations*.

Because the objective function is convex, a solution of the normal equations is in fact a solution of the least squares problem. This follows by Theorem 8.6.2. Alternatively, we can verify that the second derivative is positive semidefinite at $w^\star$. The second derivative is

$$D^2J(w)(h) = 2\, h^T(A^TA)h.$$

Since A^TA is symmetric positive semidefinite at every stationary point of the derivative, every solution of the normal equations is a local (and hence global) minimum.

Let $w^\star$ be a solution of (9.6) and $\hat{y} = Aw^\star$. Then $\hat{y} \in \mathcal{R}(A)$ and by (9.6),

$$A^T(y - \hat{y}) = A^Ty - A^TAw^\star = 0.$$

So the residual $\varepsilon = y - \hat{y}$ is in the null space of A^T. Since the rows of A^T are the transposed regressors, this indicates that the residual is orthogonal to each regressor and hence to $\mathcal{R}(A)$. So

$$y = \hat{y} + \varepsilon \quad \text{with} \quad \hat{y} \in \mathcal{R}(A) \text{ and } \varepsilon \in \mathcal{R}(A)^\perp.$$

Thus $\hat{y}$ is the unique orthogonal projection of y onto $\mathcal{R}(A)$ and ε is the orthogonal residual.

Any a vector $w^\star \in \mathbb{R}^n$ with $\hat{y} = Aw^\star$, is called a *least squares solution* of (9.2). Every least squares solution $w^\star$ gives an exact representation of $\hat{y}$ as a linear combination of the columns of A. Hence $w^\star$ is unique if and only if the columns of A are linearly independent ($\text{rank}(A) = n \leq m$). One can readily show that the columns of A are linearly independent if and only if A^TA is invertible. In that case,

$$w^\star = (A^TA)^{-1}A^Ty.$$

On the other hand, if the columns of A are linearly dependent ($\text{rank}(A) < n$), then $\mathcal{N}(A)$ is nontrivial and there are infinitely many solutions $w^\star$, each giving a different representation of the same point $\hat{y}$.

Finding the solution of a standard least squares problem thus involves two operations:

a) **Orthogonal Projection:** y is orthogonally projected onto $\mathcal{R}(A)$ to yield the unique vector $\hat{y}$.

b) **Representation:** $\hat{y}$ is exactly represented as a linear combination of the columns of A: $\hat{y} = Aw^\star$. However, the uniqueness of $w^\star$ depends on the rank of A. If $r = \text{rank}(A) = n$, then the columns of A are linearly independent and the solution $w^\star$ is unique. But if $\text{rank}(A) < n$ the columns of A are linearly dependent, $\mathcal{N}(A)$ is nontrivial, and $w^\star$ is not unique.

9.3.1 Computing a Solution

Now we examine one way to compute a solution. Let $A = U\Sigma V^T$ be a compact SVD of A. Substituting this into the normal equations gives $V\Sigma U^T U\Sigma V^T w = V\Sigma U^T y$. Using the properties of U, V and Σ this can be simplified to

$$VV^T w^\star = V\Sigma^{-1}U^T y. \tag{9.7}$$

Recall that the columns of V span $\mathcal{N}(A)^\perp$. When the columns of A are linearly independent, $\mathcal{N}(A) = \mathbf{0}$ and $\mathcal{N}(A)^\perp = \mathbb{R}^n$. In this case, $V \in \mathcal{O}_n$, $VV^T w^\star = w^\star$ and the unique least squares solution is given by

$$w^\star = V\Sigma^{-1}U^T y. \tag{9.8}$$

Suppose that the columns of A are linearly dependent and hence that $\mathcal{N}(A)$ is nontrivial. It follows that if $w^\star$ is a solution of the normal equations, then so is $w^\star + v$ for every $v \in \mathcal{N}(A)$. Conversely, if w is a solution of the normal equations, then $A^T A(w - w^\star) = 0$. So $w = w^\star + v$ with $v \in \mathcal{N}(A)$. Hence the set of solutions is a linear manifold formed by a translation of the subspace $\mathcal{N}(A)$ by a particular solution: $w^\star + \mathcal{N}(A)$.

It is intuitively clear that among the solutions there is always a unique solution of least norm. Indeed if $\mathbf{0}$ is a solution, then it is the least norm solution. Otherwise, the point on the solution manifold closest to $\mathbf{0}$ is the least norm solution. What is less obvious is that the point given by (9.8) is the least norm solution.

Lemma 9.3.1. Let A have compact SVD $U\Sigma V^T$. Then $w^\star_{\text{ln}} = V\Sigma^{-1}U^T y$ is the unique least norm solution of (9.6).

Proof. We first show that $w^\star_{\text{ln}}$ is a solution. This follows by noting that $\|y - Aw\|_2^2 = \|(I - UU^T)y + UU^T y - U(\Sigma V^T w)\|_2^2 = \|(I - UU^T)y\|_2^2 + \|U^T y - \Sigma V^T w\|_2^2$. The first term is a constant and the second is made zero by setting $w = w^\star_{\text{ln}}$. Thus $w^\star_{\text{ln}}$ is a solution. If $w^\star$ is any solution, then we can write $w^\star = w^\star_{\text{ln}} + w_0$ for some $w_0 \in \mathcal{N}(A)$. Since $w^\star_{\text{ln}} \in \mathcal{N}(A)^\perp$, we have $w^\star_{\text{ln}} \perp w_0$ and hence by Pythagorous, $\|w^\star\|_2^2 = \|w^\star_{\text{ln}}\|_2^2 + \|w_0\|_2^2$. So $\|w^\star\|_2^2 \geq \|w^\star_{\text{ln}}\|_2^2$ with equality if and only if $w^\star = w^\star_{\text{ln}}$. Thus $w^\star_{\text{ln}}$ is the unique least norm solution. □

So when the columns of A are linearly independent, (9.8) gives the unique solution and when the columns are linearly dependent, it gives the least norm solution.

9.3.2 Linear Independence of the Columns of A

In terms of the data, what does it mean if the columns of A are linearly dependent? Recall that the columns of A are the features of the data. The first column of A is the vector of measurements of the first feature (say, heart rate), the second column is the vector of measurements of the second features (say, blood pressure), and so on. If the third column is linearly dependent on the first two, then the third feature can be exactly predicted as a linear function of the first two features. In this sense, the selected features are redundant. As a result, the feature vectors lie in a proper subspace $\mathcal{R}(A) \subset \mathbb{R}^n$ of dimension $r = \text{rank}(A)$.

Consider the least norm solution $w^\star_{\text{ln}} = V\Sigma^{-1}U^T y$. The columns of U are the principal components of the uncentered feature vectors in A. So $w^\star_{\text{ln}}$ is obtained by projecting y onto the r principal components of the features: $U^T y$, then using $V\Sigma^{-1}$ to map these coordinates to the vector $w^\star_{\text{ln}} \in \mathcal{N}(A)^\perp$.

If the columns of A are almost linearly dependent, we expect the features to be very close to a subspace of dimension $k < m$ in $\mathbb{R}^m$. The natural candidates for this subspace are the PCA approximation subspaces. Let U_k, V_k denote the matrices consisting of the first k columns of U and V respectively, and Σ_k

denote the top left $k \times k$ submatrix of Σ. Then the least squares solution based on the rank k approximation $U_k \Sigma_k V_k^T$ to A is

$$w_k^\star = V_k \Sigma_k^{-1} U_k^T y = \sum_{j=1}^{k} \frac{1}{\sigma_j} v_j u_j^T y.$$

This gives a sequence of solutions $w_k^\star$, $k = 1, \dots, r$, with $w_r^\star = w_{\text{ln}}^\star$.

9.4 Tikhonov and Ridge Regularization

We have seen that one way to deal with non unique solutions is to select the least norm solution $w_{\text{ln}}^\star$. Another approach is to add a new term to the objective function that reflects prior knowledge and ensures a unique solution. For example, one can modify the least squares problem to $\min_{w \in \mathbb{R}^n} \|y - Aw\|^2 + \lambda \|w\|^2$. Here the selectable positive parameter λ balances the two competing components, $\|y - Aw\|_2^2$ and $\|w\|_2^2$, of the new objective. We will show below that this modified problem always has a unique solution. More generally, prior information may indicate that w should be close to a given vector $b \in \mathbb{R}^n$. This information can be incorporated by selecting w to minimize $\|y - Aw\|_2^2 + \lambda \|w - b\|_2^2$. This problem also has a unique solution.

The imposition of an additional component into the least squares objective, as illustrated above, is called *regularization*. It was first investigated in the context of underdetermined problems by the Russian mathematician Tikhonov (1943). He studied regularization terms of the form $\|Fx - g\|_2^2$ for a specified matrix F and vector g. Hence this form of regularization is sometimes referred to as *Tikhonov regularization*. Somewhat later, regularized linear regression was studied by Hoerl (1962) and Hoerl and Kennard (1970) using the regularization term $\|w\|_2^2$. This approach is widely known as *ridge regression*.

Tikhonov regularized least squares can be posed as:

$$\min_{w \in \mathbb{R}^n} \quad \|Aw - y\|_2^2 + \lambda \|Fw - g\|_2^2, \tag{9.9}$$

where $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$ are the usual elements of the least squares problem, and $F \in \mathbb{R}^{k \times n}$ and $g \in \mathbb{R}^k$ are the new elements of the regularization penalty. The selectable parameter $\lambda > 0$ balances the two competing objectives. Ridge regularization is a special case with $F = I_n$ and $g = \mathbf{0}$.

We have already seen that (9.9) can be readily transformed into a standard least squares problem. Since the objective of (9.9) is a sum of squares, we can write it as:

$$\|Aw - y\|^2 + \lambda \|Fw - g\|_2^2 = \left\| \begin{bmatrix} Aw - y \\ \sqrt{\lambda}(Fw - g) \end{bmatrix} \right\|_2^2 = \|\tilde{A}w - \tilde{y}\|_2^2,$$

where

$$\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} F \end{bmatrix} \in \mathbb{R}^{(m+k) \times n} \quad \text{and} \quad \tilde{y} = \begin{bmatrix} y \\ \sqrt{\lambda} g \end{bmatrix} \in \mathbb{R}^{m+k}.$$

This reduces problem (9.9) to a basic least squares problem with an augmented matrix $\tilde{A}$ and vector $\tilde{y}$. The corresponding augmented normal equations are

$$(A^T A + \lambda F^T F) w = A^T y + \lambda F^T g.$$

If $\operatorname{rank}(\tilde{A}) = n$, then $\tilde{A}$ has n linearly independent columns and the augmented problem has the unique solution:

$$w^\star(\lambda) = (A^T A + \lambda F^T F)^{-1} (A^T y + \lambda F^T g)\ . \tag{9.10}$$

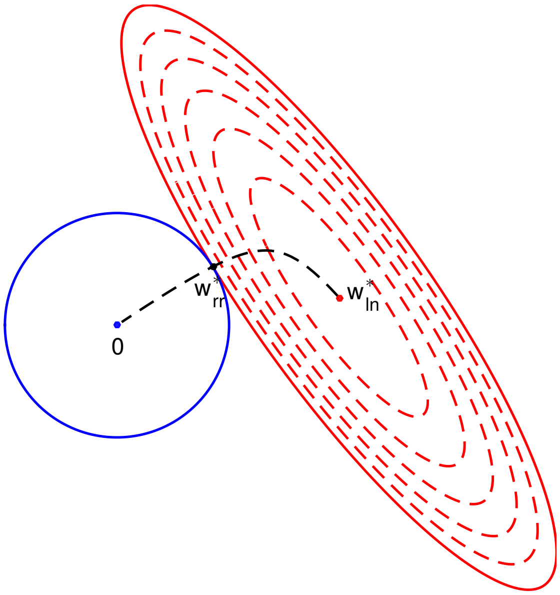

Figure 9.2: An example of a ridge regression regularization path as λ varies from 0 to ∞.

A sufficient condition ensuring that $\operatorname{rank}(\tilde{A}) = n$ is $\operatorname{rank}(F) = n$ (Exercise 9.7). In particular, ridge regularization has $k = n$, $F = I_n$ and $g = \mathbf{0}$ and hence always has the unique solution:

$$w^\star_{\mathrm{rr}}(\lambda) = (A^T A + \lambda I_n)^{-1} A^T y\ . \tag{9.11}$$

$A^T A$ is positive semidefinite and adding λI_n ensures the sum is positive definite and hence invertible. Note that these solutions are functions of the regularization parameter λ. As λ varies the solutions trace out a curve in $\mathbb{R}^n$ called the *regularization path*.

The residual $\varepsilon = y - Aw^\star$ under Tikhonov and ridge regularization is generally not orthogonal to the columns of A. You can see this for ridge regularization, for example, by using the normal equations to write $A^T(y - Aw^\star_{\mathrm{rr}}) = \lambda w^\star_{\mathrm{rr}}$. However, for all points on the ridge regularization path, $w^\star_{\mathrm{rr}}(\lambda) \in \mathcal{N}(A)^\perp$.

Lemma 9.4.1. For $\lambda > 0$, $w^\star_{\mathrm{rr}}(\lambda) \in \mathcal{N}(A)^\perp$.

Proof. Abbreviate $w^\star_{\mathrm{rr}}(\lambda)$ to $w^\star_{\mathrm{rr}}$, and let $w^\star_{\mathrm{rr}} = v + w$ with $v \in \mathcal{N}(A)^\perp$ and $w \in \mathcal{N}(A)$. Then $Aw^\star_{\mathrm{rr}} = A(v + w) = Av$ and $\|Aw^\star_{\mathrm{rr}} - y\|_2^2 = \|Av - y\|_2^2$. So $w^\star_{\mathrm{rr}}$ and v have the same least squares cost. By Pythagorous, $\|w^\star_{\mathrm{rr}}\|_2^2 = \|v\|_2^2 + \|w\|_2^2$. If $w \neq 0$, $\|v\|_2^2 < \|w^\star_{\mathrm{rr}}\|_2^2$; a contradiction. Hence $w^\star_{\mathrm{rr}} \in \mathcal{N}(A)^\perp$. □

We can say more by bringing in a compact SVD $A = U\Sigma V^T$. Recall that the columns of U form an ON basis for $\mathcal{R}(A)$ and the columns of V form an ON basis for $\mathcal{N}(A)^\perp$. Using $A = U\Sigma V^T$ yields:

$$\begin{aligned}(A^T A + \lambda I_n) w^\star_{\mathrm{rr}} &= (V\Sigma^2 V^T + \lambda I_n) w^\star_{\mathrm{rr}}, \quad \text{and} \\ A^T y &= V\Sigma U^T y\ .\end{aligned}$$

By Lemma 9.4.1, $VV^T w^\star_{\mathrm{rr}} = w^\star_{\mathrm{rr}}$. Hence

$$(V\Sigma^2 V^T + \lambda I_n) w^\star_{\mathrm{rr}} = (V\Sigma^2 V^T + \lambda I_n) VV^T w^\star_{\mathrm{rr}} = V(\Sigma^2 + \lambda I_r) V^T w^\star_{\mathrm{rr}}.$$

This allows us to write the normal equations as $V(\Sigma^2 + \lambda I_r) V^T w^\star_{\mathrm{rr}} = V\Sigma U^T y$. Multiplying both sides of this equation by $V(\Sigma^2 + \lambda I_r)^{-1} V^T$ yields:

$$w^\star_{\mathrm{rr}}(\lambda) = V(\Sigma^2 + \lambda I_r)^{-1} \Sigma U^T y = V \operatorname{diag}\left[\frac{\sigma_j}{\lambda + \sigma_j^2}\right] U^T y = \sum_{j=1}^r \frac{\sigma_j}{\lambda + \sigma_j^2} v_j u_j^T y. \tag{9.12}$$

By varying λ in (9.12), the entire regularization path is easily computed. In addition, (9.12) indicates the limit of the ridge solution as $\lambda \to 0$.

Lemma 9.4.2. $\lim_{\lambda \to 0} w^\star_{\text{rr}}(\lambda) = w^\star_{\text{ln}}$.

Proof. By (9.12), $\lim_{\lambda \to 0} w^\star_{\text{rr}}(\lambda) = V\Sigma^{-1}U^Ty$. Thus by Lemma 9.3.1, $\lim_{\lambda \to 0} w^\star_{\text{rr}}(\lambda) = w^\star_{\text{ln}}$. □

So as λ increases from 0, the ridge solution $w^\star_{\text{rr}}(\lambda)$ starts at $w^\star_{\text{ln}}$, moves along the regularization path within the subspace $\mathcal{N}(A)^\perp$, and gradually shrinks to $\mathbf{0}$ as $\lambda \uparrow \infty$. An example is shown in Figure 9.2.

9.5 On-Line Least Squares

An important variation of least squares is when the training examples are collected (or sampled) in a sequential fashion and we update an existing least squares solution after each new training example is available. In this situation, we would like to add the contribution of each new training example as efficiently as possible. We examine an efficient update algorithm known as *recursive least squares*.

If one continues to add training examples, then eventually $n < m$ and the regression becomes overdetermined. In this situation, let $A_m \in \mathbb{R}^{m\times n}$ denote the design matrix, $y_m \in \mathbb{R}^m$ denote the data vector, and assume that A_m has n linearly independent columns. Since we will be adding rows to a narrow regression matrix ($n < m$), it is natural to write the normal equations in terms of the rows of A_m. To this end, let

$$P_m = A_m^T A_m = \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix}^T = \sum_{j=1}^m x_j x_j^T \ , \tag{9.13}$$

and

$$s_m = A_m^T y_m = \sum_{j=1}^m x_j y(j) \ . \tag{9.14}$$

In terms of P_m and s_m, the normal equations become $P_m w^\star_m = s_m$. Since the columns of A_m are linearly independent, P_m is nonsingular and the unique least squares solution is $w^\star_m = P_m^{-1} s_m$.

Adding an $(m+1)$-th training example adds a new row to A_m and a new entry to y_m and this yields

$$P_{m+1} = A_{m+1}^T A_{m+1} = P_m + x_{m+1}x_{m+1}^T, \tag{9.15}$$

$$s_{m+1} = A_{m+1}^T y_{m+1} = s_m + y(m+1)x_{m+1} \ . \tag{9.16}$$

Hence to find $w^\star_{m+1}$ we need to solve the updated normal equations

$$(P_m + x_{m+1}x_{m+1}^T)w^\star_{m+1} = s_m + y(m+1)x_{m+1}. \tag{9.17}$$

Since P_{m+1} and s_{m+1} have the same dimensions as $P_m \in \mathbb{R}^{m\times n}$ and $s_m \in \mathbb{R}^n$, respectively, directly solving the updated normal equations requires $O(n^3)$ time. Assuming we have previously computed P_m^{-1}, can we use the connection between P_m and P_{m+1} is update the least squares solution more efficiently?

The problem we face can be stated as follows. Given an invertible symmetric matrix $P \in \mathbb{R}^{k\times k}$ and a vector $d \in \mathbb{R}^k$, efficiently find the inverse of the matrix $P + dd^T$, assuming it exists. Remarkably, this problem has a very simple solution.

Lemma 9.5.1. Let $P \in \mathbb{R}^{k\times k}$ be an invertible symmetric matrix and $d \in \mathbb{R}^k$. If $1 + d^TP^{-1}d \neq 0$, then the matrix $P + dd^T$ has the inverse

$$\left(P + dd^T\right)^{-1} = P^{-1} - \frac{1}{1 + d^TP^{-1}d}(P^{-1}d)(P^{-1}d)^T \tag{9.18}$$

Proof. Here is the standard proof. Since $1 + d^T P^{-1} d \neq 0$, the RHS of (9.18) is a finite valued $k \times k$ real matrix. To prove the claim simply multiply the RHS by $(P + dd^T)$:

$$\begin{aligned}
& (P + dd^T)\left(P^{-1} - \frac{1}{1 + d^T P^{-1} d}(P^{-1}d)(P^{-1}d)^T\right) \\
= \quad & I_n + dd^T P^{-1} - \frac{1}{1 + d^T P^{-1} d} dd^T P^{-1} - \frac{1}{1 + d^T P^{-1} d} d(d^T P^{-1} d) d^T P^{-1} \\
= \quad & I_n + \left(1 - \frac{1}{1 + d^T P^{-1} d} - \frac{d^T P^{-1} d}{1 + d^T P^{-1} d}\right) dd^T P^{-1} \\
= \quad & I_n
\end{aligned}$$

For an alternative proof see Appendix 1. □

Recall that by assumption the columns of A_m are linearly independent. Hence P_m is nonsingular and the least squares solution is $w^\star_m = P_m^{-1} s_m$. After obtaining $w^\star_m$ in this fashion, We assume that P_m^{-1} and s_m remain available. When a new training example is added, A_{m+1} still has n linearly independent columns (A_m has linearly independent columns and adding a row to these vectors does not change this). Hence P_{m+1} is also invertible. Application of Lemma 9.5.1 to equation (9.17) yields the following set of recursive equations for computing P_{m+1}^{-1} from P_m^{-1} and x_{m+1}, and hence for obtaining $w^\star_{m+1}$:

$$P_{m+1}^{-1} = P_m^{-1} - \frac{(P_m^{-1} x_{m+1})(P_m^{-1} x_{m+1})^T}{1 + x_{m+1}^T P_m^{-1} x_{m+1}} \ , \tag{9.19}$$

$$w^\star_{m+1} = P_{m+1}^{-1} s_{m+1} \ . \tag{9.20}$$

If we substitute (9.19) and (9.16) into (9.20) and simplify we obtain

$$\hat{y}(m+1) = x_{m+1}^T w^\star_m \ , \tag{9.21}$$

$$w^\star_{m+1} = w^\star_m + \frac{P_m^{-1} x_{m+1}}{1 + x_{m+1}^T P_m^{-1} x_{m+1}} \left(y(m+1) - \hat{y}(m+1)\right) \ . \tag{9.22}$$

This update procedure is known as *recursive least squares*. It gives a concise and efficient update formula for $w^\star_{m+1}$ in terms of $w^\star_m$ and each new training example. The update is driven by the prediction error $y(m+1) - \hat{y}(m+1)$ with no update required if the prediction error is zero. Inverting $P_{m+1} \in \mathbb{R}^{n \times n}$ requires $O(n^3)$ operations. On the other hand, the recursive equations above require $O(n^2)$ operations. Hence RLS is an efficient procedure when new examples are presented sequentially and a new least squares solution is needed after each new example is presented.

9.6 Exercises

Preliminaries

Exercise 9.1. Let $A \in \mathbb{R}^{m \times n}$. Show each of the following claims.

a) $A^T A$ is positive definite (hence invertible) if and only if the columns of A are linearly independent.

b) AA^T is positive definite if and only if the rows of A are linearly independent.

c) $\mathcal{N}(A^T A) = \mathcal{N}(A)$ and $\mathcal{N}(AA^T) = \mathcal{N}(A^T)$.

d) $\mathcal{R}(A^T A) = \mathcal{R}(A^T)$ and $\mathcal{R}(AA^T) = \mathcal{R}(A)$.

Exercise 9.2. Let $P \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix.

a) Show that there exists a unique symmetric positive definite matrix $P^{1/2}$ such that $P = P^{1/2}P^{1/2}$.

b) Let $\|x\|_P = (x^T P x)^{1/2}$. Show that $\|\cdot\|_P$ is a norm on $\mathbb{R}^n$.

Exercise 9.3. (**Vandermonde matrix**) Show that if the real numbers $\{t_j\}_{j=1}^m$ are distinct and $m > n-1$, then the Vandermonde matrix:

$$V = \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \dots & t_2^{n-1} \\ \vdots & & & & \vdots \\ 1 & t_m & t_m^2 & \dots & t_m^{n-1} \end{bmatrix}$$

has n linearly independent columns (Hint: a polynomial of degree $n-1$ has at most $n-1$ roots).

Least Squares Regression

Exercise 9.4. You are given k points $\{y_i\}_{i=1}^k$ in $\mathbb{R}^n$ and you want to find the points $x \in \mathbb{R}^n$ that minimize the sum of squared distances to these fixed points.
(a) Solve this directly using calculus.
(b) Now pose it as a regression problem and solve it using your knowledge of least squares and the SVD.

Exercise 9.5. (**SVD/PCA Regression**). Let $A \in \mathbb{R}^{m \times n}$ have rank r and compact SVD $U\Sigma V^T$. Let $y \in \mathbb{R}^m$ and $w_{\text{ln}}^\star$ denote the least norm solution to the regression problem $\min_{w \in \mathbb{R}^n} \|y - Aw\|^2$. For $1 \leq k \leq r$, let $A_k = U_k \Sigma_k V_k^T$ be the SVD rank k approximation to A. Here U_k and V_k consist of the first k columns of U and V, respectively, and Σ_k is the upper right $k \times k$ submatrix of Σ. Then let $w_k^\star$ be the *least norm solution* to the problem:

$$\min_{w \in \mathbb{R}^n} \|y - A_k w\|^2.$$

Show that:

a) $w_k^\star = V_k \Sigma_k^{-1} U_k^T y$.

b) $w_k^\star$ is the orthogonal projection of $w_{\text{ln}}^\star$ onto the subspace $\mathcal{R}(V_k)$.

c) $A_k w_k^\star$ yields the orthogonal projection $\hat{y}_k$ of y onto $\mathcal{R}(U_k)$.

Instead of regressing y on A_k, suppose we regress y on the first k left singular vectors U_k. This problem always has a unique solution. In this case we solve:

$$z_k^\star = \arg\min_{z \in \mathbb{R}^k} \|y - U_k z\|^2$$

d) Show that the unique solution is $z_k^\star = U_k^T y$.

e) Show that $z_k^\star = \Sigma_k V_k^T w_k^\star$. So $z_k^\star$ is simply the vector of coordinates of the least norm solution $w_k^\star$ w.r.t. V_k scaled by Σ_k.

f) What can you say about the solutions $\hat{w_k^\star}$ and $\hat{z_k^\star}$ when $\hat{U}_k$ and $\hat{V}_k$ are k corresponding columns of U and V (not necessarily the first k), and $\hat{\Sigma}_k$ is the corresponding submatrix of Σ?

Exercise 9.6. Let $D \in \mathbb{R}^{n\times n}$ be diagonal with nonnegative diagonal entries and consider the problem:

$$\min_{x\in\mathbb{R}^n} \quad \|x-y\|_2^2 + \lambda\|Dx\|_2^2.$$

This problem seeks to best approximate $y \in \mathbb{R}^n$ with a nonuniform penalty for a large entries in x.

a) Solve this problem using the formula for the solution of ridge regression.

b) Show that the objective function is separable into a sum of decoupled terms. Show that this decomposes the problem into n independent scalar problems.

c) Find the solution of each scalar problem.

d) By putting these scalar solutions together, find and interpret the solution to the original problem.

Exercise 9.7. Show that in Tikhonov regularized least squares, if $\operatorname{rank}(F) = n$, then $m + k \geq n$ and $\operatorname{rank}(\tilde{A}) = n$.

Exercise 9.8 (Weighted least squares). Suppose you want to put more weight on matching the labels of some examples and less weight on others. You can do this by bringing in a diagonal matrix $D \in \mathbb{R}^{m\times m}$ with diagonal weights $d_i > 0$ and solving $\arg\min_{w\in\mathbb{R}^n} \|y - Aw\|_{\sqrt{D}}^2$. (a) Determine and interpret the effect of the d_i on the least norm least squares solution, and (b) Do the same for the corresponding form of ridge regression.

On-line Least Squares

Exercise 9.9. (a) Determine the detailed equations for on-line ridge regression. (b) Use these equations to explain how RLS and on-line ridge regression differ. (c) As $m \to \infty$, will the two solutions differ? (d) In light of your answer to (c), what is the role of λ in ridge regression?

Bibliography

[1] Demitri Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Edwin Chong and Stanislaw Zak. *An Introduction to Optimization*. John Wiley and Sons, 2008.