ELE 535
Machine Learning and Pattern Recognition [1]
Handout #5: Sparse Regression

Peter J. Ramadge

Fall 2016, version 2.1

Chapter 13

Sparse Least Squares

13.1 Introduction

We now consider linear least squares problems with the additional constraint that we seek a sparse solution. By this we mean that many of the entries of the solution $w^\star$ are zero. To motivate why such problems are of interest consider a medical example. The problem is to use a measured set of medical variables $x(1), \dots, x(n)$ to form a prediction of a covariate y of interest. Given training data, we can learn a linear predictor by solving a least squares problem such as ridge regression:

$$\min_{w \in \mathbb{R}^n} \|y - Aw\|_2^2 + \lambda \|w\|_2^2. \tag{13.1}$$

Generally, the solution $\hat{w}$ of this problem has all of its entires non zero. This suggests that all of the predictor variables are important for predicting y. However, in many practical applications this is unlikely to be true. In the limit of an infinite amount of training data, the estimate might reveal that $\hat{w}_\infty(j) = 0$, while for each finite training set $\hat{w}(j)$ is non zero. Hence having obtained $\hat{w}$, one is interested in testing the hypothesis $|\hat{w}_\infty(j)| > 0$. If a suitable test of this hypothesis indicates that with high probability the value $\hat{w}_\infty(j)$ is 0, then we would conclude that the variate $x(j)$ is not important for a linear prediction of y. More generally, we could seek a small subset of the variates that are important for a good linear prediction of y. This is called a subset selection problem. This discussion suggests that a sparse solution should be sought directly in the regression problem. This would use the training data to find a sparse linear predictor $w \in \mathbb{R}^n$. The sparsity of w implies the selection of a subset of predictors from the full set $x(1), \dots, x(n)$.

Let $\|w\|_0$ denote the number of nonzero entries in w. Then sparse linear regression can be posed as

$$\begin{aligned} w^\star = \arg \min_{w \in \mathbb{R}^n} \quad & \|y - Aw\|_2^2 \\ \text{s.t.} \quad & \|w\|_0 \leq k. \end{aligned} \tag{13.2}$$

Problems like (13.2) arise in a variety of contexts. For example, when A has more columns than rows, the problem seeks a sparse solution to an underdetermined least squares problem. Alternatively, it can result from a desire for subset selection. For example, suppose we are provided with a database of labelled face images (f_j, z_j), $j = 1, \dots, p$. Here $f_j \in \mathbb{R}^n$ is the j-th vectorized face image and z_j is its label (person's identity). Form the face examples into matrix $A = [f_1, \dots, f_p]$. Then, given a new (unlabelled) face image y, we want to predict its label. We suspect that the subset of images in the database from the same person will be most important. So we set out to find a sparse representation of y using the columns of A by solving problem (13.2). This finds an approximate representation of y as a linear combination of relatively few columns of A. This is called a *sparse representation* of y. The solution $w^\star$ selects a subset of the columns of A and gives each selected column a nonzero weight. If we extract the subset of selected columns with label z and the corresponding weights, then we can form a class z predictor of y as a linear combination of the columns of A with label z. The class predictor that yields the least RSS error in predicting y, provides the estimated label z of y. This is called *sparse representation classification*.

As another illustration, consider a signal $x \in \mathbb{R}^n$ that has a sparse representation in some basis B. So $x = Bw$, where $\|w\|_0 = k \ll n$. Suppose we take $m \ll n$ measurements of x in the form of random linear combinations of the elements of x to obtain $y = Sx$, where S is the sensing matrix. We say that y is formed by compressive sensing of x using the sensing matrix S. Given $y \in \mathbb{R}^m$ we now want to reconstruct $x \in \mathbb{R}^n$. A natural way to proceed is to first solve the sparse least squares problem (13.2) with $A = SB$. Then set $\hat{x} = Bw^\star$.

The simplest version of problem (13.2) is called *sparse approximation*. Given $y \in \mathbb{R}^n$ we want to find an approximation x to y such that x has at most k nonzero entries. This simple problem can be posed as:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|y - x\| \\ \text{s.t.} \quad & \|x\|_0 \leq k. \end{aligned} \tag{13.3}$$

Sparse vectors and matrices arise naturally in a variety of practical problems, often in variations of the simple problem posed above. For example, since x is close to y but is specified with much fewer (nonzero) coordinates, we can regard x as a compressed form of y. We suspect that such an approximation will be useful if there are relatively few large entries and many small entries in y. In other words, y has some special structure that makes it “compressible”. This is often the case for natural forms of data such as speech and images. For example, after the application of a wavelet transform, most natural images are highly compressible.

In security applications one may need to classify a new face image with respect to a set of previously captured face images. Suppose that the subject is wearing glasses, or sun glasses, or a scarf, in the new image but not in the images in the database. Then the new image will most likely differ significantly from the best match among the previous images by a sparse set of pixels. So the new image y might be modeled as $y \approx y_f + y_s$, where y_f is some image in the database and y_s is a sparse image with most pixels having value 0. One might then pose the problem of finding the best match to y from the database $\mathcal{D}$ as

$$\begin{aligned} \min_{y_f\in\mathcal{D}, y_s\in\mathbb{R}^n} \quad & \|y - (y_f + y_s)\|_2^2 \\ \text{s.t.} \quad & \|y_s\|_0 \leq k. \end{aligned}$$

13.2 Preliminaries

Let $|\alpha|_0$ denote the indicator function of the set $\{\alpha \in \mathbb{R} \colon \alpha \neq 0\}$. So

$$|\alpha|_0 = \begin{cases} 0, & \text{if } \alpha = 0; \\ 1, & \text{otherwise.} \end{cases} \tag{13.4}$$

The number of nonzero entries in a vector $x \in \mathbb{R}^n$ can then be expressed as

$$\|x\|_0 = \sum_{j=1}^{n} |x(j)|_0. \tag{13.5}$$

We show below that the function $\|\cdot\|_0$ is not a norm; it simply counts the number of nonzero entries in x.

The *support* of $x \in \mathbb{R}^n$ is the set

$$S(x) = \{i \colon x(i) \neq 0\}. \tag{13.6}$$

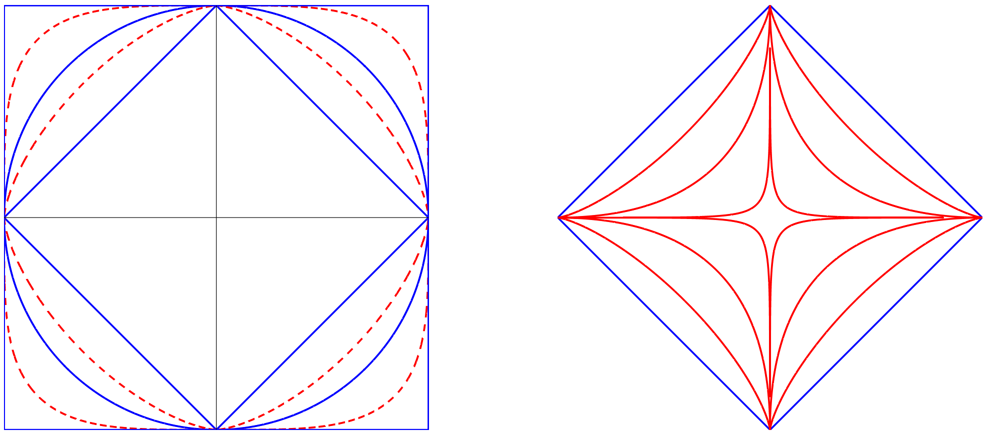

Figure 13.1: sublevel sets $B_p = (\sum_j |x(j)|^p)^{1/p} \leq 1$. Left: $p \geq 1$, the sublevel sets are convex. Right: For $p < 1$, the sublevel sets are not convex.

Clearly $|S(x)| = \|x\|_0$. For an integer $k \geq 0$, we say that x is *k-sparse* if $|S(x)| \leq k$. More generally, we say that a vector $x \in \mathbb{R}^n$ is *sparse* if relatively few of its entries are nonzero.

13.2.1 New Norms

The Euclidean norm (or 2-norm) is one of many norms on $\mathbb{R}^n$. In the development that ensues, it will be useful to have a variety of norms to provide a running set of examples. We introduce these below.

a) $\|x\|_1 = \sum_{j=1}^n |x(j)|$. This is called the *1-norm*.

b) $\|x\|_p = (\sum_{j-1}^n |x(j)|^p)^{1/p}$ for an integer $p \geq 1$. This is called the *p-norm*.

c) $\|x\|_\infty = \max_j\{|x(j)|\}$. This is called the *max norm*.

d) $\|x\| = (x^T P x)^{1/2}$, where $P \in \mathbb{R}^{n \times n}$ is symmetric positive definite.

The verification that each of these functions satisfies the three norm properties (positivity, scaling, and triangle inequality) is left as an exercise (some are easy, some more challenging; see a suitable text). Note that the 1-norm and the 2-norm are instances of the p-norm with $p = 1$ and $p = 2$, respectively.

13.2.2 $\|\cdot\|_0$ is not a Convex Function

Recall that for $c \in \mathbb{R}$, the sublevel set of $f\colon \mathbb{R}^n \to \mathbb{R}$ at c is $S_c = \{x\colon f(x) \leq c\}$. We know that every sublevel set of a convex function is a convex set. The *unit ball* of norm $\|\cdot\|$ is the particular sublevel set $B = \{x\colon \|x\| \leq 1\}$. Since every norm is a convex function, the unit ball of a norm is a convex set. The unit balls of some ℓ_p norms on $\mathbb{R}^2$ are illustrated in the left panel of Figure 13.1. In contrast, the sublevel sets of the function $f_p(x) = (\sum_j |x(j)|^p)^{1/p}$, for $0 < p < 1$, are not convex. This is illustrated for $\mathbb{R}^2$ in the right panel of Figure 13.1. It follows that for $0 < p < 1$, f_p is not a norm.

It is easy to see that $\|\cdot\|_0$ is not a convex function. Simply plot is unit ball in $\mathbb{R}^2$. Here is another way to see this. In the limit as $p \to 0$, the sublevel set $f_p(x) \leq 1$ is the union of the segments of each axis from -1 to 1. This is the intersection of the set of 1-sparse vectors with the unit ball of $\|\cdot\|_\infty$. It is not a convex set. Thus $\|\cdot\|_0$ is not convex. From this analysis it follows that $\|\cdot\|_0$ is not a norm.

13.3 Sparse Least Squares Problems

The general sparse least squares problem can be posed in any of the following three forms:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \ & \|y-Ax\|_2^2 \\ \text{s.t. } & \|x\|_0 \leq k \end{aligned} \tag{13.7}$$

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \ & \|x\|_0 \\ \text{s.t. } & \|y-Ax\|_2^2 \leq \epsilon \end{aligned} \tag{13.8}$$

$$\min_{w\in\mathbb{R}^n} \|y-Aw\|_2^2 \ + \lambda\|w\|_0. \tag{13.9}$$

All three formulations share the difficulty that $\|\cdot\|_0$ is not a convex function.

13.3.1 Terminology and Various Simplifications

Let $A=[a_1,\dots,a_p]$. For convenience, we refer to the columns a_i of A as *atoms*. If it is convenient to do so, without loss of generality we can assume that $y\in\mathcal{R}(A)$ and that y and the atoms of A have unit norm. When we say that we can make these assumptions without loss of generality, we mean that the modified problem is equivalent to the original problem.

To show the above claim, first let $y=\hat{y}+\tilde{y}$ with $\hat{y}\in\mathcal{R}(A)$ and $\tilde{y}\in\mathcal{R}(A)^\perp$. Then

$$\begin{aligned} \|y-Aw\|_2^2+\lambda\|w\|_0 &= \|\tilde{y}+\hat{y}-Aw\|_2^2+\lambda\|w\|_0 \\ &= \|\tilde{y}\|_2^2+\|\hat{y}-Aw\|_2^2+\lambda\|w\|_0 \\ &\equiv \|\hat{y}-Aw\|_2^2+\lambda\|w\|_0. \end{aligned}$$

So solving the problem with $\hat{y}$ yields a solution of the original problem. Conversely, if $w^\star$ is a solution of the original problem, then it is also a solution for the problem with $\hat{y}$. Hence without loss of generality we can assume $y\in\mathcal{R}(A)$.

If the atoms do not have unit norm, let $\tilde{A}=AD$ where the atoms of $\tilde{A}$ have unit norm and D is diagonal with positive diagonal entries. If $z=Dw$, then $\|w\|_0=\|D^{-1}z\|_0=\|z\|_0$ and

$$\begin{aligned} \|y-Aw\|_2^2+\lambda\|w\|_0 &= \|y-\tilde{A}Dw\|_2^2+\lambda\|w\|_0 \\ &= \|y-\tilde{A}z\|_2^2+\lambda\|D^{-1}z\|_0 \\ &= \|y-\tilde{A}z\|_2^2+\lambda\|z\|_0. \end{aligned}$$

So if we solve the sparse regression problem using $\tilde{A}$ (with unit norm atoms) to obtain $z^\star$, then $w^\star = D^{-1}z^\star$ is a solution to the original problem. Conversely, if $w^\star$ is a solution of the original problem, then $Dw^\star$ is a solution for the problem using $\tilde{A}$. Hence without loss of generality we can assume that A has unit norm atoms.

Now multiply the objective function by $c^2>0$ and set $u=cy$, $z=cw$ and $\tilde{\lambda}=c^2\lambda$. Noting that $\|cw\|_0=\|w\|_0$ this gives

$$\begin{aligned} c^2\left(\|y-Aw\|_2^2+\lambda\|w\|_0\right) &= \|cy-Acw\|_2^2+\lambda c^2\|cw\|_0 \\ &= \|u-Az\|_2^2+\lambda c^2\|z\|_0 \\ &= \|u-Az\|_2^2+\tilde{\lambda}\|z\|_0. \end{aligned}$$

If $z^\star$ solves the modified problem using $u=cy$ and $\tilde{\lambda}=c^2\lambda$, then $w^\star=z^\star/c$ solves the origin problem using y and λ. Conversely, if $w^\star$ is a solution of the original problem, then $z^\star=cw^\star$ is a solution for the modified problem using u and $\tilde{\lambda}$. Now note that choosing $c=1/\|y\|_2$ ensures u has unit norm. Hence without loss of generality we can assume y has unit norm.

13.3.2 Special Cases

Problem (13.9) is easily solved when $A \in \mathbb{R}^{n\times k}$, with $k \leq n$, has an SVD factorization $A = U\Sigma P^T$, where $U \in \mathcal{O}_{n,k}$, $\Sigma \in \mathbb{R}^{k\times k}$ is diagonal with a positive diagonal, and $P \in \mathbb{R}^{k\times k}$ is a generalized permutation matrix. Special instances of this include:

(a) Sparse approximation:

$$\min_{x\in\mathbb{R}^n} \|y - x\|_2^2 + \lambda\|x\|_0.$$

(b) Sparse weighted approximation: For $D = \text{diag}(d)$ with $d \in \mathbb{R}^n$ with $d > 0$,

$$\min_{x\in\mathbb{R}^n} \|y - Dx\|_2^2 + \lambda\|x\|_0,$$

(c) Sparse representation in an orthonormal basis: For $Q \in \mathcal{O}_{n,k}$,

$$\min_{x\in\mathbb{R}^k} \|y - Qx\|_2^2 + \lambda\|x\|_0,$$

Problem (a) is an easily solved sparse approximation problem. We examine this problem in the next section. The other special cases are all reducible to problem (a) and hence are also easily solved.

Beyond special cases like these, the general sparse least squares problem presents a computational bottleneck. For example, problem (13.8) is known to be NP-hard. In light of the difficulty of finding an efficient general solution method, a number of greedy algorithms have been proposed for efficiently finding an approximate solution. Examples of such methods are examined in §13.5.

13.4 k-Sparse Approximation

Given $y \in \mathbb{R}^n$ and $k < n$, we consider the *sparse approximation problem* of finding a k-sparse vector $x \in \mathbb{R}^n$ that is "closest" to y. The problem can be stated as

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(y - x) \\ \text{s.t.} \quad & \|x\|_0 \leq k, \end{aligned} \tag{13.10}$$

where $f(z) = g(\|z\|)$ for some norm $\|\cdot\|$ on $\mathbb{R}^n$ and g is a strictly monotone increasing function $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $g(0) = 0$.

Example 13.4.1. Examples of possible functions f include:

a) $f(z) = \|z\|_1 = \sum_{j=1}^n |z(j)|$.

b) $f(z) = (\|z\|_2)^2 = \sum_{j=1}^n |z(j)|^2$.

c) $f(z) = (\|z\|_p)^p = \sum_{j=1}^n |z(j)|^p$.

d) $f(z) = \|z\|_\infty = \max_j\{|z(j)|\}$.

e) $f(x) = x^T Px$, where $P \in \mathbb{R}^{n\times n}$ is symmetric postive definite.

Alternatively, we can bring in a parameter $\lambda > 0$ and add the penalty $\lambda\|x\|_0$ to the objective function in (13.10) to form the unconstrained problem:

$$\min_{x\in\mathbb{R}^n} \quad f(y-x) + \lambda\|x\|_0. \tag{13.11}$$

Here $f(y-x)$ and $\lambda\|x\|_0$ are competing objectives. Selecting a small value for λ encourages less sparsity and a better match between y and x. A large value of λ encourages greater sparsity, but potentially a worse match between x and y. So a solution $x^\star$ of (13.11) will be generally sparse, but we can't guarantee it will be k-sparse. Nevertheless, it will be convenient to first solve (13.11).

13.4.1 Separable Objective Function

Problem (13.11) is easy to solve when $f(z)$ is a separable function. To be specific, assume that $f(z) = \sum_{j=1}^n h_j(|z(j)|)$, where the functions $h_j\colon \mathbb{R}_+ \to \mathbb{R}_+$ are strictly monotone increasing functions with $h_j(0) = 0$. In this case, we minimize

$$\sum_{j=1}^n \left(\, h_j(|y(j) - x(j)|) + \lambda|x(j)|_0 \,\right). \tag{13.12}$$

This is the sum of n subproblems each of which can be solved independently. Subproblem j has the form

$$\min_{\alpha} \quad h_j(|y(j) - \alpha|) + \lambda|\alpha|_0.$$

This subproblem can be solved by considering three cases: (1) If $y(j) = 0$, we set $\alpha = 0$; (2) If $y(j) \neq 0$, there are two options to consider: either we set $\alpha = y(j)$ and incur a cost λ, or we set $\alpha = 0$ and incur a cost $h_j(|y(j)|)$. This yields the following result.

Theorem 13.4.1. Assume $f(z) = \sum_{j=1}^n h_j(|z(j)|)$, where the functions $h_j\colon \mathbb{R}_+ \to \mathbb{R}_+$ are strictly monotone increasing functions with $h_j(0) = 0$. Then the solution of problem (13.11) is

$$x^\star(j) = \begin{cases} y(j), & \text{if } h_j(|y(j)|) \geq \lambda; \\ 0, & \text{otherwise}\ . \end{cases} \tag{13.13}$$

So for separable functions f, the solution to the sparse approximation problem is obtained by *hard thresholding* $y(j)$ based on a comparison of $h_j(|y(j)|)$ and λ. Because $h_j(\cdot)$ is strictly monotone increasing, we can also write the thresholding condition as $|y(j)| \geq t_j(\lambda) = h_j^{-1}(\lambda)$. Bring in the generic scalar hard thresholding operator defined for $z \in \mathbb{R}$ by

$$H_t(z) = \begin{cases} z, & \text{if } |z| \geq t; \\ 0, & \text{otherwise}\ . \end{cases} \tag{13.14}$$

Then in terms of $H_t(z)$ and $t_j(\lambda) = h_j^{-1}(\lambda)$ we can write

$$x^\star = \Big[H_{t_j(\lambda)}(y(j)) \Big] = \begin{bmatrix} y(j), & \text{if } |y(j)| \geq t_j(\lambda) \\ 0, & \text{otherwise} \end{bmatrix}.$$

Example 13.4.2. Consider the special case: $\min_{x\in\mathbb{R}^n} \|y-x\|_2^2 + \lambda\|x\|_0$. For this problem, the appropriate hard threshold is $t = \sqrt{\lambda}$. This is applied to each entry to y to obtain:

$$x^\star = H_{\sqrt{\lambda}}(y) = \begin{bmatrix} y(j), & \text{if } |y(j)| \geq \sqrt{\lambda} \\ 0, & \text{otherwise} \end{bmatrix}.$$

□

A key observation from (13.13) is that $x^\star$ is a function of λ and as λ decreases smoothly, $\|x^\star(\lambda)\|_0$ increases monotonically in a staircase-like fashion. Depending on the value of λ, we may obtain $\|x^\star\|_0 > k$, or $\|x^\star\|_0 < k$. As a thought experiment, imagine starting with a large value of λ and computing $x^\star(\lambda)$ as λ smoothly decreases. For simplicity, suppose that at least k of the values $f_j(|y(j)|)$ are nonzero and that the nonzero values are distinct. For each value of λ let $S(\lambda)$ denote the set of indices j for which $x^\star(j) \neq 0$. So $S(\lambda)$ is the support set of $x^\star(\lambda)$. As we decrease λ the following things happen:

0) Start with $\lambda = \lambda_0 > \max_j f_j(|y(j)|)$, then $x^\star(\lambda_0) = \mathbf{0}$ and $S(\lambda_0) = \emptyset$.

1) When λ decreases to $\lambda = \lambda_1 = \max_j\{f_j(|y(j)|)\}$ with $j_1 = \arg\max_j\{f_j(|y(j)|)\}$, there is a jump change. At this point $A(\lambda_1) = \{j_1\}$, $x^\star(\lambda_1)(j_1) = y(j_1)$, and $x^\star(\lambda_1)$ is the unique optimal 1-sparse approximation to y.

2) Continuing to decrease λ we reach a value λ_2 equal to the second largest value of $\{h_j(|y(j)|)\}$. Suppose $\lambda_2 = h_{j_2}(|y(j_2)|)$. At this point, j_2 is added to S so that $S(\lambda_2) = \{j_1, j_2\}$, and $x^\star$ is modified so that $x^\star(j_2) = y(j_2)$. Then $x^\star(\lambda_2)$ is the unique optimal 2-sparse approximation to y.

3) Continuing in this fashion, we see that the optimal k-sparse approximation $x^\star$ to y is obtained by letting S be the set of indices of the first k largest values of $h_j(y(j))$, and setting

$$x^\star(j) = \begin{cases} y(j), & \text{if } j \in S; \\ 0, & \text{otherwise.} \end{cases} \tag{13.15}$$

The simplifying assumption that all of the nonzero values $f_j(|y(j)|)$ are distinct was made to simplify the explanation. More generally, one can prove the following result.

Theorem 13.4.2. Let S be the indices of any set of k largest values of $h_j(|y(j)|)$. Then $x^\star$ defined by (13.15) is a solution to problem (13.10). This solution is unique if and only if the k-th largest value of $h_j(|y(j)|)$ is strictly larger than the $(k+1)$-st value.

Proof. Exercise. □

For a fixed value of λ, solving Problem (13.11) under the assumption of separability is very efficient. One just needs to threshold the values of $y(j)$ based on the corresponding values of $h_j(|y(j)|)$ and λ. The downside is that solving problem (13.11) doesn't give precise control of the resulting value of $\|x^\star(\lambda)\|_0$.

But we now see how to solve problem (13.10) and this gives precise control over $\|x^\star\|_0$. We simply need to find the indices $j_1, \dots, j_k$ for any k largest values of $h_i(|y(i)|)$. This is not difficult. For example, we can use the following algorithm:

1) Scan the entries $y(j)$ in order from $j = 1, \dots, n$.

2) Maintain a sorted list of at most k pairs $(j, h_j(|y(j)|))$ with the k largest values of $h_j(|y(j)|)$ seen so far.

3) When entry j of $y(j)$ is scanned, compute $h_j(|y(j)|)$. If the number of table entries is less than k, add $(j, h_j(|y(j)|))$. Otherwise, if $h_j(|y(j)|)$ is larger than the smallest corresponding value among in the table, add $(j, h_j(|y(j)|))$ to the sorted table and remove the entry $(i, h_i(|y(i)|))$ with the smallest value $h_i(|y(i)|)$. Otherwise, read the next value.

The overall complexity of this algorithm $O(n)$ for computing $h_j(|y(j)|)$ and making a comparison and an additional $O(k \log k)$ overhead for keeping an ordered list of the k largest values. If a predetermined value of k is required, then the second solution method is probably more efficient. But if either k or λ is to be determined by cross-validation (checking performance on held out testing data), then the solution method based on thresholding using λ may have an efficiency advantage.

13.4.2 A Non-Separable Objective Function

As an example, we examine k-sparse approximation under the non-separable max norm:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|y-x\|_\infty \\ \text{s.t.} \quad & \|x\|_0 \leq k. \end{aligned} \tag{13.16}$$

For simplicity, initially assume that the entries of y are nonzero with distinct absolute values, and that the entries are arranged from largest to smallest by absolute value. Hence $\|y\|_\infty = |y(1)| > |y(2)| > \cdots > |y(n)| > 0$.

We can use up to k nonzero elements in x to minimize $\|y-x\|_\infty$. The optimal allocation is to use $x(1), \dots x(k)$ to make $|y(j)-x(j)|$ no larger than $|y(k+1)|$, $j = 1, \dots, k$. This yields $\|y-x^\star\|_\infty = |y(k+1)|$. This is the smallest achievable value of $\|y-x\|_\infty$ using a k-sparse x. Any $x^\star$ with

$$|x^\star(j) - y(j)| \leq |y(k+1)|, \quad j = 1, \dots, k,$$

is an optimal solution. So in general, problem (13.16) doesn't have a unique solution.

One should now see how to obtain the general solution without the simplifying assumptions. Sequentially scan the elements of y and determine an ordered list of the k largest values of $|y(j)|$. We also need to record the indices j_p of these elements (and the corresponding values $y(j_p)$, $p = 1, \dots, k$). The entries of the list determine which components of $x^\star$ will be nonzero and the required values of these components. This solution (and the algorithm for obtaining it) is remarkably similar to that discussed in the previous section for problem (13.10) under a separability assumption. This suggests that separability is not the critical feature defining this solution. This is explored further in the appendices and exercises.

13.5 Greedy Solution Methods for Sparse Least Squares

Greedy solution methods generally operate by iteratively alternating between two actions: (1) adding an atom (or atoms) to the support set of the estimated solution $w^\star$ and (2) updating the weights assigned to atoms indexed by this support set. Let $c = 1, 2, 4 \dots$ count the iterations used; $\hat{S}$ denote the indices of the currently selected atoms, $\hat{y}$ denote the associated sparse approximation to y and $r = y - \hat{y}$ denote the associated residual.

13.5.1 Matching Pursuit

The *Matching Pursuit* (MP) algorithm iteratively either adds one new atom to the estimated support set and assigns a new weight to this atom or updates the weight of an atom already in the estimated support set.

```
(0) Initialize $c = 0$, $\hat{S} = \emptyset$, $\hat{y} = \mathbf{0}$, $r = y$.
(1) Update the iteration count $c = c + 1$
    Select an atom $a_{p_c}$ with index $p_c \in \arg\max_i |a_i^T r|$
    Update the indices of the selected atoms $\hat{S} = \hat{S} \cup \{p_c\}$
    Set $w(p_c) = a_{p_c}^T r$
    The projection of $r$ onto $a_{p_c}$ is $\hat{r} = w(p_c) a_{p_c}$
    Update $\hat{y} = \hat{y} + w(p_c) a_{p_c}$ and $r = y - \hat{y}$.
(2) Check if a termination condition is satisfied (see below). If not, go to step (1).
```

The construction terminates after a desired number of distinct atoms have been selected (problem (13.7)) or the size of the residual falls below some threshold (problem (13.8)). On termination, the algorithm results in a list of the indices of the selected atoms $p_1, \dots, p_k$ and weights $w(p_1), \dots, w(p_k)$. These give the sparse approximation $\hat{y} = \sum_{i=1}^{k} w(p_i) a_{p_i}$ to y with residual $r = y - \hat{y}$.

13.5.2 Othogonal Matching Pursuit

Orthogonal Matching Pursuit (OMP) is a similar algorithm to MP except that at each iteration the weights are updated jointly by orthogonal projection of y onto the span of the atoms selected so far. Let $A_{\hat{S}}$ denote the matrix consisting of the columns of A with indices in $\hat{S}$.

```
(0) Initialize the set of selected atoms $\hat{S} = \emptyset$, the residual $r = y$ and iteration count $c = 0$.

(1) (a) Update the iteration count $c = c + 1$
    (b) Select $p_c \in \arg\max_i |a_i^T r|$
    (c) Update the indices of the selected atoms $\hat{S} = \hat{S} \cup \{p_c\}$
    (d) Determine the orthogonal projection $\hat{y}$ of $y$ onto the range of $A_{\hat{S}}$
    (e) Update the residual $r = y - \hat{y}$

(2) Check if a termination condition is satisfied (see below). If not, go to step (1).
```

Step (1)(d) requires solving a least squares problem with one additional column than at the previous iteration. Note that after step (1) is completed, $r \perp \text{span}(A_{\hat{S}})$. So once an atom has been selected it can't be selected a second time. The construction terminates after a desired number of atoms have been selected (for (13.7)), the size of the residual falls below some threshold (for (13.8)), or no atom can be found that has a nonzero correlation with the residual. On termination, the algorithm results in a set of selected atoms $a_{p_1}, \dots, a_{p_k}$ and weights $w(p_1), \dots, w(p_k)$. These give $\hat{y} = \sum_{i=1}^{k} w(p_i) a_{p_i}$ as a sparse approximation to y with residual $r = y - \hat{y}$.

OMP (and its extensions) can quickly produce a solution with a specified sparsity or accuracy, and has been found useful in a variety of applications.

13.6 Exercises

Sparsity

Exercise 13.1. Which properties of a norm does $\|\cdot\|_0$ fail to satisfy?

Exercise 13.2. Show that $\|\cdot\|_0$ is a symmetric function in the sense that it is invariant under the group of generalized permutations on $\mathbb{R}^n$. Show that it is also invariant under the action of any diagonal matrix with nonzero diagonal entries. Finally, show that it is not invariant under the orthogonal group $\mathcal{O}_n$.

Separability

Exercise 13.3. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a separable convex function with $f(z) = \sum_{j=1}^n h_j(z(j))$ for functions $h_j\colon \mathbb{R} \to \mathbb{R}$ with $h_j(0) = 0$. Show that each h_j is a convex function.

Sparse Approximation

Exercise 13.4. **(Approximation with an ℓ_1 penalty)** Let $x, y \in \mathbb{R}^n$. In problem (13.11), suppose we replace the sparsity penalty $\|x\|_0$ by a 1-norm penalty $\|x\|_1$. This gives the approximation problem:

$$\min_{x\in\mathbb{R}^n} \ \|y - x\|_2^2 + \lambda\|x\|_1.$$

Solving this problem can also give a sparse solution.

a) Find a solution to the above the problem, and determine if the solution is unique.

b) Interpret the solution in terms of an operation on the components of y.

c) Assume the components of y are permuted so that $|y(1)| \geq |y(2)| \geq \cdots \geq |y(n)|$. Show that the solution is k-sparse if and only if $|y_{k+1}| \leq \lambda < |y_k|$.

d) Does this solution give a better k-sparse approximation to y than the solution of problem (13.11)?

Exercise 13.5. **(Approximation with an ℓ_1 constraint)** Let $x, y \in \mathbb{R}^n$.

a) Use the solution of part a) in exercise (13.4) to solve the following problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|y - x\|_2^2 \\ \text{subject to:} \quad & \|x\|_1 \leq \alpha. \end{aligned}$$

Determine if the solution is unique.

b) Assume the components of y are permuted so that $|y(1)| \geq |y(2)| \geq \cdots \geq |y(n)|$. Show that the solution of part a) is k-sparse if and only if $\sum_{j=1}^k |y(j)| - k|y(k)| \leq \alpha < \sum_{j=1}^k |y(j)| - k|y(k+1)|$.

c) Give an algorithm for directly solving the problem in part a).

Exercise 13.6. Let $D \in \mathbb{R}^{n\times n}$ be diagonal with nonnegative diagonal entries. The following problem seeks the best approximation x to y with a non-symmetric ℓ_1 penalty on x. Solve the problem.

$$\min_{x\in\mathbb{R}^n} \quad \|y - x\|_2^2 + \lambda\|Dx\|_1.$$

Exercise 13.7. (**Sparse approximation in a non-symmetric norm**) Let $D \in \mathbb{R}^{n\times n}$ be diagonal and positive definite. Then $\|x\|_D = (x^T Dx)^{1/2}$ is a norm. Find a solution of the following problem and determine if the solution is unique.

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|y-x\|_D^2 \\ \text{subject to:} \quad & \|x\|_0 \leq k. \end{aligned}$$

The similar approach also provides a solution for: $\min_{x\in\mathbb{R}^n} \|y-x\|_D^2 + \lambda\|x\|_0$. For $D \neq \alpha I_n$, $\|\cdot\|_D$ is non-symmetric. So sparse approximation is easily solved for some non-symmetric norms.

Special Cases of Sparse Regression

Exercise 13.8. (**Sparse representation in an ON basis**) Let $r \leq n$ and $Q \in \mathbb{R}^{n\times r}$ have orthonormal columns. Find a solution of the following problem and determine if the solution is unique.

$$\begin{aligned} \min_{x\in\mathbb{R}^r} \quad & \|y-Qx\|_2^2 \\ \text{subject to:} \quad & \|x\|_0 \leq k. \end{aligned}$$

The same method also gives solutions for:

$$\begin{aligned} &\min_{x\in\mathbb{R}^r} \quad \|y-Qx\|_2^2 + \lambda\|x\|_0; \\ &\min_{x\in\mathbb{R}^r} \quad \|y-Qx\|_2^2, \text{ subject to } \|x\|_1 \leq c; \text{ and} \\ &\min_{x\in\mathbb{R}^r} \quad \|y-Qx\|_2^2 + \lambda\|x\|_1. \end{aligned}$$

In each case, briefly give the corresponding solution.

Exercise 13.9. (**Sparse representation when $\mathbf{A} = \mathbf{U\Sigma P^T}$**) Let $r \leq n$, $U \in \mathbb{R}^{n\times r}$ have orthonormal columns, $\Sigma \in \mathbb{R}^{r\times r}$ be diagonal with positive diagonal entries, and $P \in \mathbb{R}^{r\times r}$ be a generalized permutation. Set $A = U\Sigma P^T$. Find a solution of the following problem and determine if the solution is unique.

$$\begin{aligned} \min_{x\in\mathbb{R}^r} \quad & \|y-Ax\|_2^2 \\ \text{subject to:} \quad & \|x\|_0 \leq k. \end{aligned}$$

A similar approach gives the solution for: $\min_{x\in\mathbb{R}^r} \|y-Ax\|_2^2 + \lambda\|x\|_0$.

Exercise 13.10. (**Sparse approximation in a quadratic norm**) Let $P, D \in \mathcal{S}_n^{++}$ with D diagonal, and let $Q \in O(n)$. Then for $x \in \mathbb{R}^n$, $\|x\|_P = (x^T Px)^{1/2}$ and $\|x\|_D = (x^T Dx)^{1/2}$ are norms.

Show that the following two problems are equivalent, in the sense that an instance of one can be transformed into an instance of the other. Thus a solution method for either, gives a solution method for the other.

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|y-x\|_P^2, \text{ subject to } \|x\|_0 \leq k \qquad \min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|z-Qx\|_D^2, \text{ subject to } \|x\|_0 \leq k.$$

For general P, the first problem is a sparse approximation problem in a non-symmetric norm. The second problem is a sparse representation problem with respect to an orthonormal basis and a simpler non symmetric norm.

13.7 Appendix: When is $\mathbf{f(x) = x^T P x}$ Separable?

Lemma 13.7.1. For symmetric $P \in \mathbb{R}^{n\times n}$, the function $f(x) = x^T P x$ is separable if and only if P is diagonal.

Proof. If P is diagonal, $f(x) = \sum_{i=1}^n p_{ii} x(i)^2$ which is clearly separable. For general P write $f(x) = \sum_i \sum_j p_{ij} x(i)x(j)$. Suppose $f(x) = \sum_{i=1}^n h_i(x(i))$. Setting $x = \mathbf{0}$ we must have $0 = \sum_{i=1}^n h_i(0)$. Hence $\sum_{i\neq k} h_i(0) = -h_k(0)$. If $x = x(k)e_k$, where e_k is the k-th standard basis vector, then

$$x(k)^2 p_{kk} = h_k(x(k)) + \textstyle\sum_{i\neq k} h_i(0) = h_k(x(k)) - h_k(0).$$

Hence for every x

$$\textstyle\sum_{k=1}^n p_{kk} x(k)^2 = \sum_{k=1}^n h_k(x(k)) - \sum_{k=1}^n h_k(0) = \sum_{k=1}^n h_k(x(k)) = f(x). \tag{13.17}$$

Let $J = P - D$ where D is the diagonal matrix with the diagonal entries of P down the diagonal. Since $J = P - D$ is symmetric, J has real eigenvalues and n orthonormal eigenvectors. Let these eigenvectors be the columns of $V = [v_1, \dots, v_n]$ and Λ be the diagonal matrix with the corresponding eigenvalues down the diagonal. Then $J = V\Lambda V^T$.

By (13.17), $x^T J x = x^T(P - D)x = f(x) - \sum_{k=1}^n p_{kk} x(k)^2 = 0$ for each $x \in \mathbb{R}^n$. This implies that all of the eigenvalues of J are zero. Thus $J = V\Lambda V^T = \mathbf{0}$. Thus $P = D$. □

13.8 Appendix: Sparse Approximation Under a Symmetric Norm

13.8.1 Symmetric Norms

A norm on $\mathbb{R}^n$ is said to be *symmetric* if it has the following two properties:
(a) for any permutation matrix P and all $x \in \mathbb{R}^n$, $\|Px\| = \|x\|$; and
(b) for any diagonal matrix D with diagonal entries in $\{\pm 1\}$, and all $x \in \mathbb{R}^n$, $\|Dx\| = \|x\|$.
A norm that satisfies the first property is called a *permutation invariant norm*, and one that satisfies the second property is called an *absolute norm* (since $\|x\| = \||x|\|$). A square matrix of the form DP where $D = \text{diag}[\pm 1]$ and P is a permutation is called a *generalized permutation*. Let $\text{GP}(n)$ denote the set of $n \times n$ generalized permutation matrices. Then a norm is symmetric if and only if it is invariant under all generalized permutations.

Lemma 13.8.1. $\text{GP}(n)$ is closed under matrix multiplication, contains the identity, and every $P \in \text{GP}(n)$ has an inverse $P^{-1} \in \text{GP}(n)$.

Proof. Let $\mathcal{D}_n$ denote the family of $n \times n$ diagonal matrices with diagonal entries in $\{\pm 1\}$. Then $I \in \mathcal{D}_n$ and if $D \in \mathcal{D}_n$, then $D^2 = I$. So $D^{-1} = D \in \mathcal{D}_n$. Finally, if $D_1, D_2 \in \mathcal{D}_n$, then $D = D_1 D_2 \in \mathcal{D}_n$.

If $P = DQ \in \text{GP}(n)$, then $P = QD'$ for some $D' \in \mathcal{D}_n$ and hence $P^{-1} = D'Q^T \in \text{GP}(n)$. Now let $P_i = D_i Q_i \in \text{GP}(n)$, $i = 1, 2$. Then $P_1 P_2 = D_1 Q_1 D_2 Q_2 = DQ$ with $Q_1 D_2 = D_2' Q_1$, $D = D_1 D_2'$ and $Q = Q_1 Q_2$. So $P_1 P_2 \in \text{GP}(n)$. □

Example 13.8.1. Some examples of symmetric norms are given below. In the accompanying verifications, $Q \in \text{GP}(n)$.

a) Every p-norm: $\|Qx\|_p = (\sum_{j=1}^n |(Qx)(j)|^p)^{1/p} = (\sum_{j=1}^n |x(j)|^p)^{1/p} = \|x\|_p$.

b) The max norm: $\|Qx\|_\infty = \max_j\{|(Qx)(j)|\}_{j=1}^n = \max_j\{|x(j)|\}_{j=1}^n = \|x\|_\infty$.

c) The *c-norm* defined by $\|x\|_c = \max_{P\in \mathrm{GP}(n)}\{x^T Pc\}$, where $c \in \mathbb{R}^n$ is nonzero:

$$\|Qx\|_c = \max_{P\in \mathrm{GP}(n)}\{(Qx)^T Pc\} = \max_{P\in \mathrm{GP}(n)}\{x^T Pc\} = \|x\|_c.$$

Example 13.8.2. Let

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix},$$

and consider the norm $\|x\|_D = (x^T Dx)^{1/2}$. We have $e_1^T De_1 = 2$, $e_2^T De_2 = 1$, with e_1 and e_2 related by a permutation $e_2 = Pe_1$. Hence $\|x\|_D$ is not a symmetric norm.

13.8.2 Properties of Symmetric Norms

It will be convenient to use the following notation. If the entries of a vector $x \in \mathbb{R}^n$ are positive (resp. nonnegative) we write $x > \mathbf{0}$ (resp. $x \geq \mathbf{0}$). For $x, y \in \mathbb{R}^n$, $x \leq y$ means $y - x \geq \mathbf{0}$. The following property of symmetric norms will be useful.

Lemma 13.8.2. For any symmetric norm $\|\cdot\|$ on $\mathbb{R}^n$, if $\mathbf{0} \leq x \leq y$, then $\|x\| \leq \|y\|$.

Proof. This proof will draw on some advanced aspects of norms and convex sets. If $x = \mathbf{0}$, then $\|x\| = 0 \leq \|y\|$. So we can assume $x \neq \mathbf{0}$. The set $S = \{u\colon \|u\| \leq \|x\|\}$ is called the sublevel set of $\|\cdot\|$ at the value $\|x\| > 0$. Clearly $x \in S$. The sublevel sets of a convex function are convex sets. Since a norm is a convex function, S is a convex set.

Consider the two convex sets S and $C = \{x\}$. S contains interior points, C is nonempty and contains no interior points of S. Hence by the separation theorem for convex sets, there exists a hyperplane $w^T z = c$, with $w \neq 0$, such that for all $z \in S$, $w^T z \leq c$, and for all $z \in C$, $w^T z \geq c$. Since x is in both sets, $w^T x = c$.

Under the assumption that $x > \mathbf{0}$, we show that $w \geq \mathbf{0}$. Suppose to the contrary that $w(i) < 0$. Form $\hat{x}$ from x by setting $\hat{x}(i) = -x(i)$. By the symmetry of the norm we have $\|\hat{x}\| = \|x\|$ and $\hat{x} \in S$. Hence we must have $w^T\hat{x} \leq c$. On the other hand, $w(i) < 0$ and $\hat{x}_i = -x(i) < 0$ imply that

$$w^T\hat{x} = \textstyle\sum_{j=1}^n w(j)\hat{x}(j) = \sum_{j\neq i} w(j)x(j) + (-x(i)w(i)) > w^T x = c.$$

This is a contradiction. Thus $w \geq \mathbf{0}$.

Continuing with the assumption $x > \mathbf{0}$, we now show that $\|x\| \leq \|y\|$. Since $x \leq y$ and $x > \mathbf{0}$ we have $y > \mathbf{0}$ and $z = y - x \geq \mathbf{0}$. If $w^T y > c$, then $\|y\| > \|x\|$ and we are done. On the other hand, if $w^T y = w^T(x + z) \leq c$, then using $w^T x = c$ we conclude that $w^T z \leq 0$. But $w, z \geq \mathbf{0}$. So we must have $w^T z = 0$. Thus $w^T y = w^T(x + z) = w^T x = c$. This still leaves three possibilities:

(1) $y \notin S$. In this case $\|x\| < \|y\|$ and we are done.
(2) y is a boundary point of S. In this case $\|x\| = \|y\|$ and we are done.
(3) y is an interior point of S.

If y is an interior point of S, then there exists a ball $B(y, \epsilon) = \{a\colon \|a - y\| \leq \epsilon\}$ centered at y with radius $\epsilon > 0$ such that $B(y, \epsilon) \subset S$. Hence for some small $\delta > 0$, $(1 + \delta)y \in S$. But this means that $w^T(1 + \delta)y = (1 + \delta)w^T y = (1 + \delta)c > c$. A contradiction. Thus $\|x\| \leq \|y\|$.

The final step is to consider $x \geq \mathbf{0}$ with $x \leq y$. Form

$$\tilde{y}(j) = \begin{cases} y(j), & \text{if } y(j) > 0; \\ 1, & \text{if } y(j) = 0, \end{cases} \quad \text{and} \quad \tilde{x}(j) = \begin{cases} x(j), & \text{if } x(j) > 0; \\ \tilde{y}(j), & \text{if } x(j) = 0. \end{cases}$$

Then $\mathbf{0} < \tilde{x} \leq \tilde{y}$. Let $x_\alpha = (1 - \alpha)x + \alpha\tilde{x}$ and $y_\alpha = (1 - \alpha)y + \alpha\tilde{y}$ for $\alpha \in [0, 1]$. For $\alpha > 0$,

$$\mathbf{0} < x_\alpha = (1 - \alpha)x + \alpha\tilde{x} \leq (1 - \alpha)y + \alpha\tilde{y} = y_\alpha.$$

Hence by the result proved above, $\|x_\alpha\| \leq \|y_\alpha\|$. Now take the limit as $\alpha \to 0$ and use the continuity of the norm to conclude that $\|x\| \leq \|y\|$. □

13.8.3 Sparse Approximation under a Symmetric Norm

Under a symmetric norm, the sparse approximation problem (13.3) has the following simple solution.

Theorem 13.8.1. Let $\|\cdot\|$ be a symmetric norm, $y \in \mathbb{R}^n$, and S be the indices of k largest values of $|y(j)|$, $j = 1, \dots, n$. Then (13.15) gives a solution to problem (13.3).

Proof. For any $P \in GP(n)$, we have $\|y - x\| = \|Py - Px\|$ and $\|x\|_0 = \|Px\|_0$. So we can select a P to ensure $(Py)(1) \geq (Py)(2) \geq \cdots \geq (Py)(n) \geq 0$. Hence from this point forward we simply assume that $y(1) \geq y(2) \geq \cdots \geq y(n) \geq 0$. To simplify the proof, we will also assume the largest k values in y are distinct, but this is not required.

Let $z = y - x$. The possibilities for the best k-sparse solution fall into two forms: either $z(j) = 0$, for all $j = 1, \dots, k$, or there exist integers p, q with $1 \leq p \leq k$ and $k + 1 \leq q \leq n$, such that $z(p) \neq 0$, and $z(q) = 0$ with $y(q) < y(p)$. In the first case, let $z_1 = y - x$, and the second, let $z_2 = y - x$. We can permute the entries of z_2 to form z_2' by swapping the zero values outside the range $1, \dots, k$ with the locations with nonzero values in the range $1, \dots, k$. This is visualized below.

$$\begin{array}{lcccccccc}
z_1: & 0 & \dots & 0 & 0 & y(k+1) & y(k+2) & \cdots & y(n) \\
z_2: & 0 & \dots & y(k-1) & y(k) & 0 & y(k+2) & \cdots & 0 \\
z_2': & 0 & \dots & 0 & 0 & y(k-1) & y(k+2) & \cdots & y(k)
\end{array}$$

So $\|z_2\| = \|z_2'\|$ and $\mathbf{0} \leq z_1 \leq z_2'$. Hence by Lemma 13.8.2, $\|z_1\| \leq \|z_2'\| = \|z_2\|$. Thus $x^\star(j) = y(j)$, $j = 1, \dots, k$, achieves an objective value at least as good as any other x. □

The solution under a symmetric norm need not be unique, even when $|y(k)| > |y(k+1)|$ in an ordered list of these values. For example, it is not unique under the max norm.

Chapter 14

Convex Relaxation: the LASSO

14.1 Introduction

An alternative approach to problem (13.9) is to relax the sparsity penalty $\|w\|_0$ to the convex function $\|w\|_1$. This leads to the convex *ℓ_1-regularized least squares problem*:

$$\min_{w\in\mathbb{R}^m} \quad \|y - Aw\|_2^2 + \lambda\|w\|_1 \ . \tag{14.1}$$

This is often called the *lasso problem*.

The objective function in (14.1) is the sum of two competing convex terms: $\|Aw - b\|_2^2$ and $\lambda\|w\|_1$. It is hence a convex function. Example sublevel sets, $\{w\colon \|Aw - y\|_2^2 \leq \alpha\}$ and $\{w\colon \|w\|_1 \leq \beta\}$, of these two terms are illustrated in Figure 14.1. As shown in the figure, usually we can't minimize both terms at the same time.

Consideration of the sublevel sets leads to two equivalent formulations of ℓ_1 regularized regression. The first is

$$\begin{aligned} \min_{w\in\mathbb{R}^n} \quad & \|Aw - y\|_2^2 \\ \text{s.t.} \quad & \|w\|_1 \leq \epsilon \ . \end{aligned} \tag{14.2}$$

In this formulation we minimize $\|Aw - y\|_2^2$ over a fixed sublevel set of the ℓ_1 norm. The optimal solution occurs at the point $w_\epsilon^\star$ where a level set of $\|Aw - y\|_2^2$ first intersects the ϵ-ball of $\|\cdot\|_1$. This is illustrated in Figure 14.1. We also illustrate the ℓ_2-ball that first intersects the same level set of $\|Aw - y\|_2^2$. Notice the difference in the sparsity of the two intersection points.

The second formulation is

$$\begin{aligned} \min_{w\in\mathbb{R}^n} \quad & \|w\|_1 \\ \text{s.t.} \quad & \|Aw - y\|_2^2 \leq \delta \ . \end{aligned} \tag{14.3}$$

Here we minimize $\|w\|_1$ over a fixed sublevel set of $\|Aw - y\|_2^2$. The optimal solution occurs at the point $w_\delta^\star$ where a level set of $\|\cdot\|_1$ first intersects the δ-sublevel set of $\|Aw - y\|^2$. For appropriate choice of δ and ϵ, the two problems have the same solution. See Figure 14.1.

Multiplying the objective of (14.1) by α^2, with $\alpha > 0$, yields the equivalent problem:

$$\min_{w\in\mathbb{R}^m} \|\bar{y} - \bar{A}w\|_2^2 + \bar{\lambda}\|w\|_1,$$

where $\bar{y} = \alpha y$, $\bar{A} = \alpha A$, and $\bar{\lambda} = \alpha^2\lambda$. As a result, it is meaningless to talk about the value of λ employed when solving (14.1) without accounting for possible scaling. One way to do this is to let a_j denote the j-th column of A and define $\lambda_{\max} = \max_{j=1}^m |a_j^T y|$. Then the ratio $\lambda/\lambda_{\max}$ is invariant to scaling.

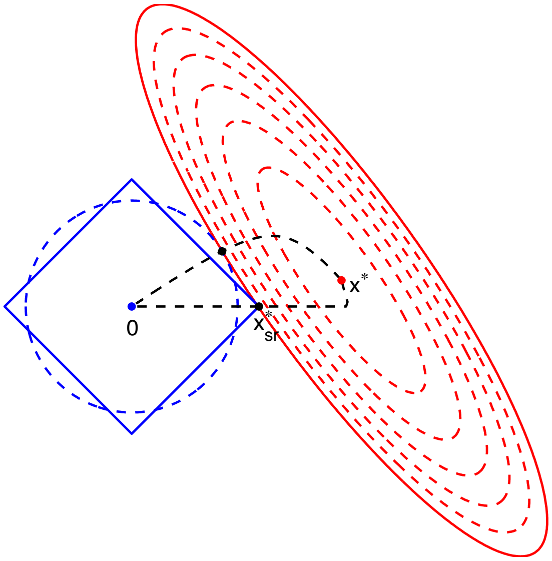

Figure 14.1: Sublevel sets of $\|Aw - y\|_2^2$ (red) and the $\|w\|_1$ regularizer (blue) together with the regularization path of the lasso solution $x^\star_{\text{sr}}$. For comparison, a sublevel set of the ℓ_2 norm (dashed blue) and the corresponding ridge regression solution are also shown.

In general, no closed form expressions for the solutions of (14.1), (14.2) and (14.3) are known. This is the usual situation for many interesting optimization problems. The important point is that ℓ_1-regularized least squares is a convex optimization problem and is amenable to solution via efficient numerical algorithms. Moreover, many such algorithms are readily available. The special case of (14.1) for sparse approximation is particularly easy to solve and this connects to the sparse approximation problems studied under $\|\cdot\|_0$. We discuss this in the following section.

14.2 ℓ_1-Sparse Approximation and Soft Thresholding

A special case of (14.1) is the ℓ_1-sparse approximation problem:

$$x^\star = \arg\min_{x\in\mathbb{R}^n} \quad \|y - x\|_2^2 + \lambda\|x\|_1. \tag{14.4}$$

This requires selecting x to approximate y, as measured by $\|y - x\|_2^2$, but with a convex ℓ_1-penalty on x to encourage a sparse approximation. The objective function is convex and separable:

$$\|y - x\|_2^2 + \lambda\|x\|_1 = \textstyle\sum_{j=1}^n (y(j) - x(j))^2 + \lambda|x(j)| \ .$$

Each term in the sum can be optimized separately, leading to n scalar problems of the form:

$$\min_{z\in\mathbb{R}} \quad (y(j) - z)^2 + \lambda|z| \ . \tag{14.5}$$

We would like to use differential calculus to solve the above problem, but we see that $|z|$ is not differentiable at $z = 0$. This can be handled as follows. First consider $z > 0$ and set the derivative w.r.t. z of the objective in (14.5) equal to 0. This yields $z - y(j) + \lambda/2 = 0$ which implies that $z = y(j) - \lambda/2$ provided $y(j) > \lambda/2$. Doing the same for $z < 0$ yields $z - y(j) - \lambda/2 = 0$ which implies $z = y(j) + \lambda/2$ provided

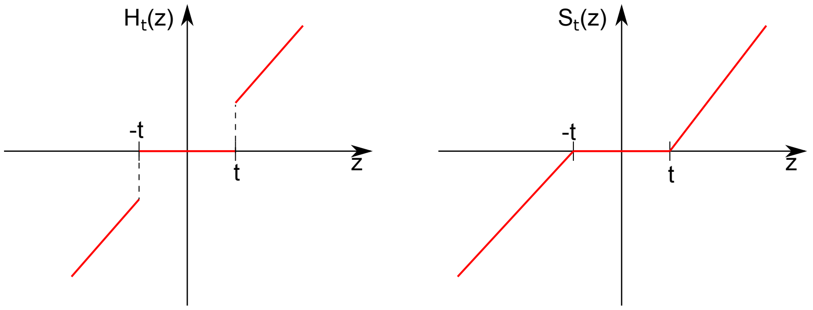

Figure 14.2: The hard (left) and soft (right) scalar thresholding functions $H_t(z)$ and $S_t(z)$.

$y(j) < -\lambda/2$. The only case that remains is $z = 0$, with objective value $y(j)^2$. This must be solution for $-\lambda/2 \leq y(j) \leq \lambda/2$. So the optimal solution of each scalar problem is:

$$x^\star(j) = \begin{cases} y(j) - \lambda/2, & \text{if } y(j) > \lambda/2; \\ 0, & \text{if } -\lambda/2 \leq y(j) \leq \lambda/2; \\ y(j) + \lambda/2, & \text{if } y(j) < -\lambda/2. \end{cases}$$

The component $x^\star(j)$ is set to 0 if $y(j)$ is smaller in magnitude than $\lambda/2$. Doing so introduces sparsity in $x^\star$. The remaining nonzero components of $x^\star$ are formed by reducing the magnitude of the corresponding values in y by $\lambda/2$. For this reason, this operation is also called *shrinkage*.

Bring in the scalar *soft thresholding function*:

$$S_t(z) = \begin{cases} z - t, & \text{if } z \geq t; \\ 0, & \text{if } -t < z < t; \\ z + t, & \text{if } z \leq -t. \end{cases}$$

This is illustrated in Figure 14.2. We have shown above that $x^\star(j) = S_{\lambda/2}(y(j))$. The optimal solution of (14.4) can then be written in vector form as:

$$x^\star = S_{\lambda/2}(y), \tag{14.6}$$

where the vector function $S_t \colon \mathbb{R}^n \to \mathbb{R}^n$ acts componentwise.

More generally, the same approach can be used to solve the problem $\min_{x\in\mathbb{R}^n} \|y - x\|_2^2 + \lambda\|Dx\|_1$, where $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix with positive diagonal. The solution is again a soft thresholding operation on y except that the threshold for component j is $t_j = \lambda d(j)/2$. This can be generalized to solve any problem of the form $\min_{x\in R^k} \|y - UDP^T x\|_2^2 + \|x\|_1$, where $U \in \mathcal{O}_{n,k}$ has orthonormal columns, $D \in \mathbb{R}^{k\times k}$ is diagonal with positive diagonal entries, and $P \in \mathbb{R}^{k\times k}$ is a generalized permutation matrix.

Example 14.2.1. The sparse approximation problems:

$$\min_{x\in\mathbb{R}^n} \|y - x\|_2^2 + \lambda\|x\|_0 \qquad \text{and} \qquad \min_{x\in\mathbb{R}^n} \|y - x\|_2^2 + \lambda\|x\|_1,$$

have the solutions

$$x_0^\star = H_{\sqrt{\lambda}}(y) \qquad \text{and} \qquad x_1^\star = S_{\lambda/2}(y).$$

The two solutions use similar but distinct thresholding functions (hard versus soft) but also use distinct threshold values: $\sqrt{\lambda}$ versus $\lambda/2$. These functions are compared in Figure 14.3. In both cases, components

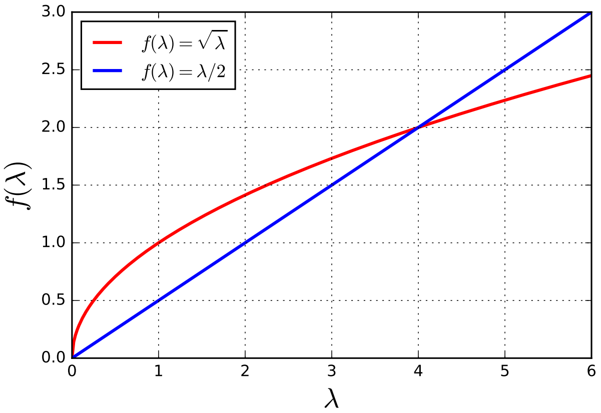

Figure 14.3: Plots of $\sqrt{\lambda}$ and $\lambda/2$ versus λ.

of y with magnitudes above the threshold will remain nonzero, otherwise the component value is set to zero. So a higher value of the threshold creates the opportunity for more sparsity in the solution. We see that for $\lambda < 4$, sparse approximation using $\|x\|_0$ has the opportunity to yield a sparser solution. We say opportunity because the sparsity of the solution will also depend on the values of the components of y. For example, if y has a few large values and all other values equal to zero, then for λ sufficiently small, $x_0^\star$ and $x_1^\star$ will have the same support set. On the other hand, if y has a few large magnitude components and many components with magnitudes clustered in a neighborhood of 0, then it is expected that $x_0^\star$ will be more sparse that $x_1^\star$.

14.3 Subgradients and the Subdifferential

We have already seen that the scalar function $g(z) = |z|$ is not differentiable at $z = 0$. Since $\|w\|_1 = \sum_{j=1}^n |w(j)|$, it follows that $\|w\|_1$ is not differentiable at any point at which a component of w is zero. But (14.4) seeks a sparse solution; so it's unlikely that $\|w\|_1$ will be differentiable at a solution $w^\star$.

Consider a scalar valued differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$. The derivative of f at x is a linear function from $\mathbb{R}^n$ into $\mathbb{R}$. Hence it can be represented by a row vector g_x^T where $g_x \in \mathbb{R}^n$ is the gradient $\nabla f(x)$. Let $\alpha > 0$ be small and $y \in \mathbb{R}^n$. A first order Taylor series expansion of f at x gives

$$f(x+\alpha y) = f(x) + \alpha g_x^T y + \text{H.O.T.} \tag{14.7}$$

If f is also convex, then for any $z \in \mathbb{R}^n$,

$$f((1-\alpha)x + \alpha z) \leq (1-\alpha)f(x) + \alpha f(z)\ . \tag{14.8}$$

Putting (14.7) and (14.8) together gives

$$(1-\alpha)f(x) + \alpha f(z) \geq f(x + \alpha(z-x)) = f(x) + \alpha g_x^T(z-x) + \text{H.O.T.}$$

Cancelling common terms, dividing both sides by α, and letting $\alpha \downarrow 0$, yields:

$$f(z) \geq f(x) + g_x^T(z-x). \tag{14.9}$$

This lower bounds the values $f(z)$ of the convex function f by the sum of its value at x and a linear function g_x^T of the deviation of z from x. This result was given previously when we discussed differentiable convex functions.

The bound (14.9) gives a useful way to define a "generalized gradient" for nondifferentiable convex functions. A vector $g \in \mathbb{R}^n$ is called a *subgradient* of f at x if for all $z \in \mathbb{R}^n$,

$$f(z) \geq f(x) + g^T(z - x) \ .$$

The function f is called *subdifferentiable* at x if it has a nonempty set of subgradients. When this holds, the set of subgradients, denoted by $\partial f(x)$, is called the *subdifferential* of f at x.

Example 14.3.1. Consider the scalar function $|z|$. For $z \neq 0$, this is differentiable with $\partial|z| = 1$ if $z > 0$ and $\partial|z| = -1$ if $z < 0$. At $z = 0$, we have $\partial|z| = [-1, 1]$. We can write this as

$$g \in \partial|z| \iff g = \begin{cases} 1, & \text{if } z > 0; \\ \gamma \in [-1,1], & \text{if } z = 0; \\ -1, & \text{if } z < 0. \end{cases}$$

Similarly,

$$g \in \partial\|w\|_1 \iff g(j) = \begin{cases} 1, & \text{if } w(j) > 0; \\ \gamma \in [-1,1], & \text{if } w(j) = 0; \\ -1, & \text{if } w(j) < 0. \end{cases} \tag{14.10}$$

We will use the following important property of subdifferentials.

Lemma 14.3.1. The point $w^\star$ minimizes the convex function $f \colon \mathbb{R}^n \to \mathbb{R}$ if and only if $\mathbf{0} \in \partial f(w^\star)$.

Proof. If $w^\star$ minimizes f, then for every z, $f(z) \geq f(w^\star) = f(w^\star) + \mathbf{0}^T(z - w^\star)$. Hence $\mathbf{0} \in \partial f(w^\star)$. Conversely, if $\mathbf{0} \in \partial f(w^\star)$, then for all z, $f(z) \geq f(w^\star) + \mathbf{0}^T(z - w^\star) = f(w^\star)$. So $w^\star$ minimizes f. □

14.4 Application to ℓ_1-Regularized Least Squares

We can use Lemma 14.3.1 to find necessary and sufficient conditions for $w^\star$ to be a minimizer of the function $f(w) = \|y - Aw\|_2^2 + \lambda\|w\|_1$. The first term in f can be expanded as $w^T A^T A w - 2w^T A^T y + y^T y$. This term has a gradient given by $2(A^T A w - A^T y)$. So the subdifferential of $f(w)$ is:

$$\partial f(w) = 2A^T(Aw - y) + \lambda\partial\|w\|_1 \ . \tag{14.11}$$

Using Lemma 14.3.1, we have the following result.

Theorem 14.4.1. $w^\star$ is a solution of (14.1) if and only if

$$A^T(y - Aw^\star) = \tfrac{\lambda}{2} g \quad \text{for some } g \in \partial\|w^\star\|_1. \tag{14.12}$$

Proof. By Lemma 14.3.1, $w^\star$ is a solution of (14.1) if and only if $\mathbf{0} \in A^T(Aw - y) + \frac{\lambda}{2}\partial\|w\|_1$. This is equivalent to (14.12). □

The term $y - Aw^\star$ is the residual, and the rows of A^T are the atoms (columns of A). The j-th entry in g depends on the sign of the j-th entry of $w^\star$ and by the above this must equal the inner product of the residual with the j-th atom. Specifically,

$$a_j^T(y - Aw^\star) = \begin{cases} \lambda/2, & \text{if } w^\star(j) > 0; \\ \gamma \in [-\lambda/2, \lambda/2], & \text{if } w^\star = 0; \\ -\lambda/2, & \text{if } w^\star(j) < 0. \end{cases}$$

So when $w^\star$ uses atom a_j (i.e., $w^\star(j) \neq 0$), the inner product of a_j and the residual is $\frac{\lambda}{2}\operatorname{sign}(w^\star(j))$; but if $w^\star$ does not use atom a_j, then the inner product of a_j and the residual lies in the interval $[-\frac{\lambda}{2}, \frac{\lambda}{2}]$. Notice that the residual is never orthogonal to an atom used by $w^\star$.

We have previously seen corresponding conditions for ridge and least squares regression. For ridge regression the corresponding condition is $A^T(y - Aw^\star) = \lambda w^\star$. The term $y - Aw^\star$ is the residual and for any atom a_i, $a_i^T(y - A^T w^\star) = \lambda w^\star(i)$. So an atom is only orthogonal to the residual if it has zero weight in $w^\star$. Since $A^TA + \lambda I$ is invertible for $\lambda > 0$, the ridge solution is $w^\star = (A^TA + \lambda I)^{-1}A^Ty$, which is linear in y. For least squares the corresponding condition is $A^T(y - Aw) = 0$ (the normal equations). In this case, $a_i^T(y - Aw^\star) = 0$. So every atom is orthogonal to the residual. If A^TA is invertible, the solution is $w^\star = (A^TA)^{-1}A^Ty$, which is also linear in y.

Equation (14.12) is of the same form as these two examples but g is a more complex function of y and $w^\star$. In particular, $w^\star$ is not a linear function of y. This is shown in the examples below.

Example 14.4.1. We can easily apply the necessary and sufficient conditions (14.12) to the ℓ_1-sparse approximation problem to obtain the known solution. Applying (14.12) yields $y - x^\star \in {}^{\lambda}\!/_2\partial\|x^\star\|_1$. Writing this out componentwise we obtain

$$\begin{aligned} x^\star(i) &= y(i) - \begin{cases} \lambda/2 & \text{if } x^\star(i) > 0 \quad \equiv \quad y(i) > \lambda/2; \\ \gamma \in [-\lambda/2, \lambda/2] & \text{if } x^\star(i) = 0 \quad \equiv \quad y(i) \in [-\lambda/2, \lambda/2] \text{ with } \gamma = y(i); \\ -\lambda/2 & \text{if } x^\star(i) < 0 \quad \equiv \quad y(i) < -\lambda/2. \end{cases} \\ &= S_{\lambda/2}(y(i)). \end{aligned}$$

Example 14.4.2. In this example we hand compute solutions of the ℓ_1-sparse regression problem. To simplify the computation we consider the slightly modified problem

$$\arg\min_{w\in\mathbb{R}^n} {}^1\!/_2\|y - Aw\|_2^2 + \lambda\|w\|_1,$$

where

$$A = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \end{bmatrix}, \qquad \text{and} \quad y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

For $\lambda > 0$ small, we suspect the solution of the sparse regression problem takes the form $w^\star = \alpha e_1 \in \mathbb{R}^3$ for some $\alpha > 0$. To check this we test the condition (14.12) adapted to this version of the problem:

$$A^Ty - A^TAw^\star = \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} - \alpha \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1-\alpha \\ 0 \\ 0.5(1-\alpha) \end{bmatrix} = \lambda \begin{bmatrix} (1-\alpha)/\lambda \\ 0 \\ 0.5(1-\alpha)/\lambda \end{bmatrix}$$

To ensure the vector on the right is in $\partial\|w^\star\|_1$ we need:

$$\begin{aligned} \alpha &= 1 - \lambda \\ {}^1\!/_2(1-\alpha)/\lambda &\in [-1, 1]\,. \end{aligned}$$

Once α is chosen to satisfy the first condition, the second condition holds. So $w^\star = (1-\lambda)e_1$ is indeed a solution for $0 < \lambda < 1$. This also suggests that at $\lambda = 1$, $w^\star = \mathbf{0}$ is a solution. To check this we again examine the appropriately modified version of (14.12):

$$A^Ty - A^TAw^\star = \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}$$

The last vector is clearly in $\partial\|w^\star\|_1$. Hence $w^\star = (1-\lambda)e_1$ is a solution for $0 < \lambda \leq 1$. A symmetric argument shows that for $y = e_2 = (0,1)^T$, $w^\star = (1-\lambda)e_2$ is a solution for $0 < \lambda \leq 1$.

Example 14.4.3. Consider the same problem formulation and dictionary from Example 14.4.2. This time we seek the solution for $y_\gamma = (1-\gamma)e_1 + \gamma e_2$ for $\gamma \in [0,1]$. For simplicity we will assume $\lambda < 1/2$.

We first examine if the solution takes the form $w^\star = \alpha e_1$ with $\alpha > 0$. To check (14.12) we compute:

$$A^T y_\gamma - A^T A w^\star = \begin{bmatrix} 1-\gamma \\ \gamma \\ 0.5 \end{bmatrix} - \alpha \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1-\gamma-\alpha \\ \gamma \\ 0.5(1-\alpha) \end{bmatrix} = \lambda \begin{bmatrix} (1-\gamma-\alpha)/\lambda \\ \gamma/\lambda \\ 0.5(1-\alpha)/\lambda \end{bmatrix}$$

To ensure the last vector is in $\partial\|w^\star\|_1$ we need

$$\begin{aligned} \alpha + \gamma &= 1-\lambda \\ \gamma/\lambda &\in [-1,1] \\ 1/2(1-\alpha)/\lambda &\in [-1,1]\,. \end{aligned}$$

To satisfy the first condition set $\alpha = 1-\gamma-\lambda$ and require $0 \le \gamma < 1-\lambda$. Since γ and λ are positive, the second condition requires $0 \le \gamma \le \lambda$. For the third condition, $\frac{1}{2}(1-\alpha)/\lambda = \frac{1}{2}(\gamma/\lambda + 1)$. So if the second condition holds, so does the third. We require $0 \le \gamma < 1-\lambda$ and $0 \le \gamma \le \lambda$. Since we have assumed $\lambda < 1/2$, $\lambda < 1-\lambda$. Hence we only need the condition $0 \le \gamma \le \lambda$.

Thus we have

$$w^\star = (1-\gamma-\lambda)e_1 \text{ is a solution for } 0 \le \gamma \le \lambda\,. \tag{14.13}$$

At $\gamma = 0$, $w^\star = (1-\lambda)e_1$. Since this has only one nonzero weight, we say it is 1-sparse. As γ increases from 0, the solution (14.13) shrinks towards the origin and remains 1-sparse over the range $0 \le \gamma \le \lambda$.

For $\gamma > \lambda$, we suspect that $w^\star = \alpha e_1 + \beta e_2$, with $\alpha, \beta > 0$. To test (14.12) we compute:

$$A^T y_\gamma - A^T A w^\star = \begin{bmatrix} 1-\gamma \\ \gamma \\ 0.5 \end{bmatrix} - \alpha \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} - \beta \begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1-\gamma-\alpha \\ \gamma-\beta \\ 0.5(1-\alpha-\beta) \end{bmatrix} = \lambda \begin{bmatrix} (1-\gamma-\alpha)/\lambda \\ (\gamma-\beta)/\lambda \\ 0.5(1-\alpha-\beta)/\lambda \end{bmatrix}.$$

To ensure the last vector is in $\partial\|w^\star\|_1$ we need

$$\begin{aligned} \alpha + \gamma &= 1-\lambda \\ \gamma - \beta &= \lambda \\ 1/2(1-(\alpha+\beta)) &\in [-\lambda,\lambda]\,. \end{aligned}$$

Subtracting the second equation from the first, gives $\alpha + \beta = 1-2\lambda$. So if the first two conditions are satisfied, so is the third. To satisfy the first equation we set $\alpha = 1-\lambda-\gamma$ and require $\gamma < 1-\lambda$. To satisfy the second equation we set $\beta = \gamma - \lambda$ and require $\gamma > \lambda$. Hence

$$w^\star = (1-\gamma-\lambda)e_1 + (\gamma-\lambda)e_2 \text{ is a solution for } \lambda < \gamma < 1-\lambda\,. \tag{14.14}$$

Symmetry demands that

$$w^\star = (1-(1-\gamma)-\lambda)e_2 \text{ is a solution for } 1-\lambda \le \gamma \le 1\,. \tag{14.15}$$

So we have found a complete solution path $w^\star_\gamma$ as γ increases from 0 to 1 and y_γ moves in a straight line from $y_0 = e_1$ to $y_1 = e_2$.

If the solution $w^\star_\gamma$ was linear in y_γ, then as y_γ moves from y_0 to y_1, the solution path would be $w^\star = (1-\gamma)w^\star_0 + \gamma w^\star_1$, where $w^\star_0$ is the solution obtained for y_0 and $w^\star_1$ for y_1. So the solution path would move in a straight line from $w^\star_0$ at $\gamma = 0$ to $w^\star_1$ at $\gamma = 1$. But it does not do so. From $w^\star_0 = (1-\lambda)e_1$, as γ increases the solution first shrinks until $\gamma = \lambda < 1/2$. At this point $w^\star_\lambda = (1-2\lambda)e_1$. Symmetrically,

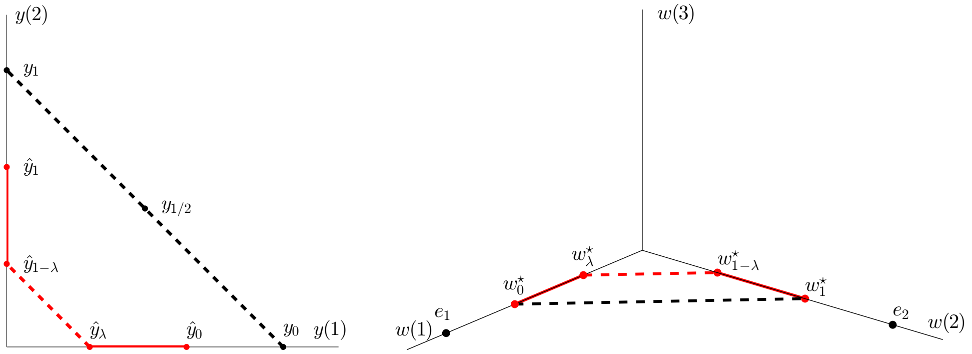

Figure 14.4: Example 14.4.3. A solution path of sparse regression. Left: The black dashed line indicates the path of test vectors y_γ. The red lines show the resulting approximation $\hat{y}_\gamma = Dw^\star_\gamma$. Right: The dashed line indicates linear interpolation between $w^\star_0$ and $w^\star_1$. The red lines indicate the solution path $w^\star_\gamma$ as γ varies from 0 to 1.

as γ decreases from 1, the solution $w^\star_1 = (1-\lambda)e_2$ shrinks until $\gamma = 1-\lambda > 1/2$ at which point $w^\star_{1-\lambda} = (1-2\lambda)e_2$. As γ transits from λ to $1-\lambda$, the solution path linearly interpolates between the solution $w^\star_\lambda$ and $w^\star_{1-\lambda}$. To see this set $\delta = (\gamma-\lambda)/(1-2\gamma)$. Then as γ increases from λ to $1-\lambda$, δ increases from 0 to 1. So the linear interpolation between $w^\star_\lambda$ and $w^\star_{1-\lambda}$ is

$$\begin{aligned}(1-\delta)w^\star_\lambda + \delta w^\star_{1-\lambda} &= (1-\frac{\gamma-\lambda}{1-2\lambda})(1-2\lambda)e_1 + \frac{\gamma-\lambda}{1-2\gamma}(1-2\lambda)e_2\\ &= (1-\gamma-\lambda)e_1 + (\gamma-\lambda)e_2\\ &= w^\star_\gamma \quad \text{for } \lambda < \gamma < 1-\lambda\,.\end{aligned}$$

The solution path when $\lambda < 1/2$ is illustrated in Fig. 14.4.

14.5 Appendix: Related Problems and Concepts

Let $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$ and consider the set of linear equations $Aw = y$. Assume that A has rank m. Hence $Aw = b$ has a solution but in general this need not be unique. To resolve this one could find the unique least norm solution. However, we now have another option: find the sparsest solution. It is clear that we can always find a solution that minimizes $\|w\|_0$. The question is whether this solution unique. In other words, does the following problem have a unique solution:

$$\begin{aligned}\min_{w\in\mathbb{R}^n} \quad & \|w\|_0\\ \text{s.t.} \quad & Aw = y.\end{aligned} \tag{14.16}$$

14.5.1 Spark

Let $A \in \mathbb{R}^{m\times n}$ have rank $r \leq \min(m,n)$. Let $\text{spark}(A)$ be the least number of linearly dependent cols of A. It must hold that

$$2 \leq \text{spark}(A) \leq r+1.$$

$\text{spark}(A)$ gives a condition for (14.16) to have a unique solution. Roughly, if a solution is sufficiently sparse, then it is the unique sparsest solution.

Theorem 14.5.1. If $Ax^\star = y$ and $\|x^\star\|_0 < \text{spark}(A)/2$, then $x^\star$ is the unique sparsest solution of $Ax = y$.

Proof. Suppose z satisfies $Az = y$. Then $A(x^\star - z) = 0$. Hence $\|x^\star - z\|_0 \geq \text{spark}(A)$. This gives

$$\|x^\star\|_0 + \|z\|_0 \geq \|x^\star - z\|_0 \geq \text{spark}(A).$$

Since $\|x^\star\|_0 < \text{spark}(A)/2$, we must have $\|z\|_0 > \text{spark}(A)/2$. So z is not as sparse as $x^\star$. Thus $x^\star$ is the unique sparsest solution. □

The theorem is interesting because it indicates that there are instances in which a unique sparsest solution exists. Unfortunately, computing $\text{spark}(A)$ is expensive. So the theorem does not give a useful computational test for the uniqueness of a sparse solution.

14.5.2 Coherence

For $A \in \mathbb{R}^{m\times n}$, the *coherence* of A is

$$\mu(A) = \max_{i\neq j} \frac{|a_i^T a_j|}{\|a_i\|_2\ \|a_j\|_2}.$$

The coherence of A is relatively easy to compute and gives a measure of dependence among the columns of A. For example, if $Q \in \mathcal{O}_{n,k}$, then the columns of Q point in orthogonal directions in $\mathbb{R}^m$. In this case, $\mu(A) = 0$ and we say that the columns are *incoherent*.

Let A have unit norm columns and consider the Gram matrix $G = A^TA$ of A. The off diagonal entires of the Gram matrix indicates the similarity of pairs of distinct unit norm columns of A. Moreover, $\mu(A) = max_{i\neq j}|G_{ij}|$. So the coherence of A is determined by the two most similar columns of A without regard to the sign of the inner product.

The coherence of A can be used to lower bound $spark(A)$.

Lemma 14.5.1. $\text{spark}(A) \geq 1 + \frac{1}{\mu(A)}$.

Proof. Without loss of generality, assume A has cols of unit 2-norm. Consider the Gram matrix $G = A^TA$. Then $G_{ij} = 1$ if $i = j$ and is $\leq \mu(A)$ otherwise.

Select p cols of A and consider the corresponding $p \times p$ Gram matrix. If $\sum_{i\neq j}|G_{ij}| < |G_{jj}|$ for each j, then by the Gershgorin disk theorem this $p \times p$ matrix is positive definite and the p selected cols of A are linearly independent. The above condition holds if $(p-1)\mu(A) < 1$, i.e., $p < 1 + 1/\mu(A)$. Hence $\text{spark}(A) \geq p + 1$. Selecting p to be the largest integer less that $1 + 1/\mu$, gives $\text{spark}(A) \geq p + 1 \geq 1 + 1/\mu(A)$. □

Now we can state the following result.

Theorem 14.5.2. If $Ax^\star = y$ and $\|x^\star\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right)$, then $x^\star$ is the unique solution of (14.16).

Proof. Exercise. □

This test is weaker than the result for spark, but the sufficient condition is easily verifiable. Would we like $\mu(A)$ to be small. This means we want to spread the directions of the columns of A out uniformly on the unit sphere.

14.5.3 Basis Pursuit

When $\operatorname{rank}(A) = m$, the following relaxed version of (14.16) is called *Basis Pursuit*:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|_1 \\ \text{s.t.} \quad & Ax = y \end{aligned} \tag{14.17}$$

Problem (14.17) is a convex program. In fact, it is equivalent to the (large) linear program:

$$\begin{array}{lrcl} \min_{x,z \in \mathbb{R}^n} & \mathbf{1}^T z & & \\ \text{s.t.} & Ax & = & y \\ & x - z & \leq & \mathbf{0} \\ & -x - z & \leq & \mathbf{0} \\ & -z & \leq & \mathbf{0} \end{array}$$

It is interesting that there are instances where solving the relaxed problem (14.17) will give the unique solution of (14.16). Roughly, if the solution of (14.16) is sufficient sparse, then it is unique and solving the relaxed problem (14.17) will find it.

Theorem 14.5.3 (Donoho and Elad, 2003)**.** If $Ax^\star = y$ and $\|x^\star\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right)$, then $x^\star$ is the unique solution of both (14.16) and (14.17).

Proof. The first part is the result of Theorem 14.5.2. For the proof of the second part see See Donoho and Elad, 2003. □

14.6 Exercises

Exercise 14.1. Solve each of the following problems:

a) $\min_{x\in\mathbb{R}^n} \|y-x\|_2^2 + \lambda\|Dx\|_1$.

b) $\min_{x\in\mathbb{R}^k} \|y-Ux\|_2^2 + \lambda\|x\|_1$, where $U \in \mathcal{O}_{n,k}$.

c) $\min_{x\in\mathbb{R}^k} \|y-UDP^Tx\|_2^2 + \lambda\|x\|_1$, where $U \in \mathbb{R}^{n\times k}$ has ON columns, $D \in \mathbb{R}^{k\times k}$ is diagonal with a positive diagonal, and $P \in \mathbb{R}^{k\times k}$ is a generalized permutation.

Exercise 14.2. Consider the dictionary

$$A = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \end{bmatrix} .$$

Suppose $\lambda > 1/2$ and $y_\gamma = (1-\gamma)e_1 + \gamma e_2$ where $\gamma \in [0,1]$. Find and plot the solution $w_\gamma^\star$ of the corresponding lasso regression problem as a function of γ. Also plot $\hat{y}_\gamma = Aw_\gamma^\star$.

Exercise 14.3. Consider the dictionary

$$A = \begin{bmatrix} 1 & 0 & 1/\sqrt{2} \\ 0 & 1 & 1/\sqrt{2} \end{bmatrix} .$$

Let $y_\gamma = (1-\gamma)e_1 + \gamma e_2$ where $\gamma \in [0,1]$. By imposing an upper bound on λ, find and plot the solution $w_\gamma^\star$ of the corresponding lasso regression problem as a function of γ. Also plot $\hat{y}_\gamma = Aw_\gamma^\star$.

Exercise 14.4. Let $D = [d_1, \dots, d_p]$ be a dictionary of unit norm atoms and consider the sparse representation problem

$$\min_{w\in\mathbb{R}^p} \|y - Dw\|_2^2 + \lambda\|w\|_1.$$

a) Let $y = d_j$. Show that $w^\star = (1-\lambda)e_j$ and $Dw^\star = (1-\lambda)d_j$, for $0 < \lambda < 1$.

b) Let $y = d_j$. This time we pick an atom d_k, $k \neq j$, that maximizes $c_{ij} = |d_i^T d_j|$ over $i \neq j$. Furthermore, assume that $d_k^T d_j > 0$. Constrain λ to be small in the sense that $\lambda < (1 + c_{jk})/2$. Now let $y_\gamma = (1-\gamma)d_j + \gamma d_k$, for $\gamma \in [0,1]$. So y_γ traces out line segment from d_j to d_k. Determine the corresponding solution of $w_\gamma^\star$ of the ℓ_1-sparse regression problem as a function of γ.

14.7 Notes and References

Not done yet.

Bibliography