Chapter 6

Approximation and fitting

6.1 Norm approximation

6.1.1 Basic norm approximation problem

The simplest *norm approximation problem* is an unconstrained problem of the form

$$\text{minimize} \quad \|Ax - b\| \tag{6.1}$$

where $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$ are problem data, $x \in \mathbf{R}^n$ is the variable, and $\|\cdot\|$ is a norm on $\mathbf{R}^m$. A solution of the norm approximation problem is sometimes called an *approximate solution* of $Ax \approx b$, in the norm $\|\cdot\|$. The vector

$$r = Ax - b$$

is called the *residual* for the problem; its components are sometimes called the individual *residuals* associated with x.

The norm approximation problem (6.1) is a convex problem, and is solvable, *i.e.*, there is always at least one optimal solution. Its optimal value is zero if and only if $b \in \mathcal{R}(A)$; the problem is more interesting and useful, however, when $b \notin \mathcal{R}(A)$. We can assume without loss of generality that the columns of A are independent; in particular, that $m \geq n$. When $m = n$ the optimal point is simply $A^{-1}b$, so we can assume that $m > n$.

Approximation interpretation

By expressing Ax as

$$Ax = x_1a_1 + \cdots + x_na_n,$$

where $a_1, \ldots, a_n \in \mathbf{R}^m$ are the columns of A, we see that the goal of the norm approximation problem is to fit or approximate the vector b by a linear combination of the columns of A, as closely as possible, with deviation measured in the norm $\|\cdot\|$.

The approximation problem is also called the *regression problem*. In this context the vectors $a_1, \ldots, a_n$ are called the *regressors*, and the vector $x_1a_1 + \cdots + x_na_n$,

where x is an optimal solution of the problem, is called the *regression of* b (onto the regressors).

Estimation interpretation

A closely related interpretation of the norm approximation problem arises in the problem of estimating a parameter vector on the basis of an imperfect linear vector measurement. We consider a linear measurement model

$$y = Ax + v,$$

where $y \in \mathbf{R}^m$ is a vector measurement, $x \in \mathbf{R}^n$ is a vector of parameters to be estimated, and $v \in \mathbf{R}^m$ is some measurement error that is unknown, but presumed to be small (in the norm $\|\cdot\|$). The estimation problem is to make a sensible guess as to what x is, given y.

If we guess that x has the value $\hat{x}$, then we are implicitly making the guess that v has the value $y - A\hat{x}$. Assuming that smaller values of v (measured by $\|\cdot\|$) are more plausible than larger values, the most plausible guess for x is

$$\hat{x} = \operatorname{argmin}_z \|Az - y\|.$$

(These ideas can be expressed more formally in a statistical framework; see chapter 7.)

Geometric interpretation

We consider the subspace $\mathcal{A} = \mathcal{R}(A) \subseteq \mathbf{R}^m$, and a point $b \in \mathbf{R}^m$. A *projection* of the point b onto the subspace $\mathcal{A}$, in the norm $\|\cdot\|$, is any point in $\mathcal{A}$ that is closest to b, *i.e.*, any optimal point for the problem

$$\begin{array}{ll} \text{minimize} & \|u - b\| \\ \text{subject to} & u \in \mathcal{A}. \end{array}$$

Parametrizing an arbitrary element of $\mathcal{R}(A)$ as $u = Ax$, we see that solving the norm approximation problem (6.1) is equivalent to computing a projection of b onto $\mathcal{A}$.

Design interpretation

We can interpret the norm approximation problem (6.1) as a problem of optimal design. The n variables $x_1, \ldots, x_n$ are *design variables* whose values are to be determined. The vector $y = Ax$ gives a vector of m *results*, which we assume to be linear functions of the design variables x. The vector b is a vector of *target* or *desired results*. The goal is to choose a vector of design variables that achieves, as closely as possible, the desired results, *i.e.*, $Ax \approx b$. We can interpret the residual vector r as the deviation between the actual results (*i.e.*, Ax) and the desired or target results (*i.e.*, b). If we measure the quality of a design by the norm of the deviation between the actual results and the desired results, then the norm approximation problem (6.1) is the problem of finding the best design.

Weighted norm approximation problems

An extension of the norm approximation problem is the *weighted norm approximation problem*

$$\text{minimize} \quad \|W(Ax-b)\|$$

where the problem data $W \in \mathbf{R}^{m\times m}$ is called the *weighting matrix*. The weighting matrix is often diagonal, in which case it gives different relative emphasis to different components of the residual vector $r = Ax - b$.

The weighted norm problem can be considered as a norm approximation problem with norm $\|\cdot\|$, and data $\tilde{A} = WA$, $\tilde{b} = Wb$, and therefore treated as a standard norm approximation problem (6.1). Alternatively, the weighted norm approximation problem can be considered a norm approximation problem with data A and b, and the *W-weighted norm* defined by

$$\|z\|_W = \|Wz\|$$

(assuming here that W is nonsingular).

Least-squares approximation

The most common norm approximation problem involves the Euclidean or ℓ_2-norm. By squaring the objective, we obtain an equivalent problem which is called the *least-squares approximation problem*,

$$\text{minimize} \quad \|Ax-b\|_2^2 = r_1^2 + r_2^2 + \cdots + r_m^2,$$

where the objective is the sum of squares of the residuals. This problem can be solved analytically by expressing the objective as the convex quadratic function

$$f(x) = x^TA^TAx - 2b^TAx + b^Tb.$$

A point x minimizes f if and only if

$$\nabla f(x) = 2A^TAx - 2A^Tb = 0,$$

i.e., if and only if x satisfies the so-called *normal equations*

$$A^TAx = A^Tb,$$

which always have a solution. Since we assume the columns of A are independent, the least-squares approximation problem has the unique solution $x = (A^TA)^{-1}A^Tb$.

Chebyshev or minimax approximation

When the ℓ_∞-norm is used, the norm approximation problem

$$\text{minimize} \quad \|Ax-b\|_\infty = \max\{|r_1|,\ldots,|r_m|\}$$

is called the *Chebyshev approximation problem*, or *minimax approximation problem*, since we are to minimize the maximum (absolute value) residual. The Chebyshev approximation problem can be cast as an LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1}, \end{array}$$

with variables $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

Sum of absolute residuals approximation

When the ℓ_1-norm is used, the norm approximation problem

$$\text{minimize} \quad \|Ax - b\|_1 = |r_1| + \cdots + |r_m|$$

is called the sum of (absolute) residuals approximation problem, or, in the context of estimation, a *robust estimator* (for reasons that will be clear soon). Like the Chebyshev approximation problem, the ℓ_1-norm approximation problem can be cast as an LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T t \\ \text{subject to} & -t \preceq Ax - b \preceq t, \end{array}$$

with variables $x \in \mathbf{R}^n$ and $t \in \mathbf{R}^m$.

6.1.2 Penalty function approximation

In ℓ_p-norm approximation, for $1 \leq p < \infty$, the objective is

$$(|r_1|^p + \cdots + |r_m|^p)^{1/p}.$$

As in least-squares problems, we can consider the equivalent problem with objective

$$|r_1|^p + \cdots + |r_m|^p,$$

which is a separable and symmetric function of the residuals. In particular, the objective depends only on the *amplitude distribution* of the residuals, *i.e.*, the residuals in sorted order.

We will consider a useful generalization of the ℓ_p-norm approximation problem, in which the objective depends only on the amplitude distribution of the residuals. The *penalty function approximation problem* has the form

$$\begin{array}{ll} \text{minimize} & \phi(r_1) + \cdots + \phi(r_m) \\ \text{subject to} & r = Ax - b, \end{array} \tag{6.2}$$

where $\phi : \mathbf{R} \to \mathbf{R}$ is called the (residual) *penalty function*. We assume that ϕ is convex, so the penalty function approximation problem is a convex optimization problem. In many cases, the penalty function ϕ is symmetric, nonnegative, and satisfies $\phi(0) = 0$, but we will not use these properties in our analysis.

Interpretation

We can interpret the penalty function approximation problem (6.2) as follows. For the choice x, we obtain the approximation Ax of b, which has the associated residual vector r. A penalty function assesses a cost or penalty for each component of residual, given by $\phi(r_i)$; the total penalty is the sum of the penalties for each residual, *i.e.*, $\phi(r_1) + \cdots + \phi(r_m)$. Different choices of x lead to different resulting residuals, and therefore, different total penalties. In the penalty function approximation problem, we minimize the total penalty incurred by the residuals.

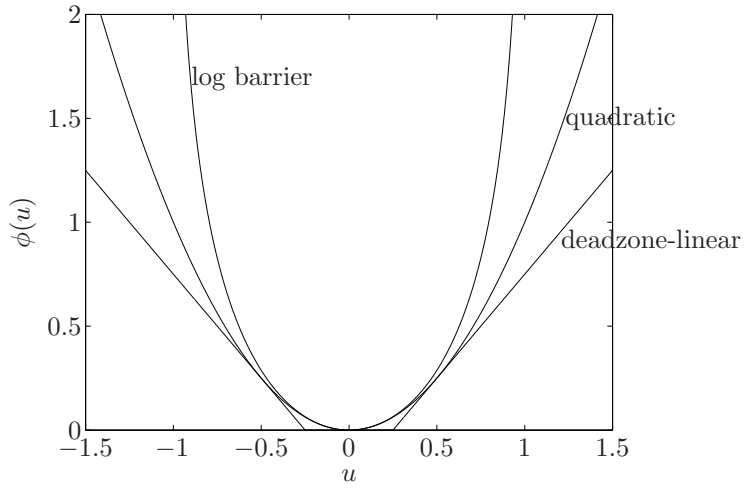

Figure 6.1 Some common penalty functions: the quadratic penalty function $\phi(u) = u^2$, the deadzone-linear penalty function with deadzone width $a = 1/4$, and the log barrier penalty function with limit $a = 1$.

Example 6.1 *Some common penalty functions and associated approximation problems.*

- By taking $\phi(u) = |u|^p$, where $p \geq 1$, the penalty function approximation problem is equivalent to the ℓ_p-norm approximation problem. In particular, the quadratic penalty function $\phi(u) = u^2$ yields least-squares or Euclidean norm approximation, and the absolute value penalty function $\phi(u) = |u|$ yields ℓ_1-norm approximation.
- The *deadzone-linear* penalty function (with deadzone width $a > 0$) is given by

$$\phi(u) = \begin{cases} 0 & |u| \leq a \\ |u| - a & |u| > a. \end{cases}$$

  The deadzone-linear function assesses no penalty for residuals smaller than a.
- The *log barrier* penalty function (with limit $a > 0$) has the form

$$\phi(u) = \begin{cases} -a^2 \log(1 - (u/a)^2) & |u| < a \\ \infty & |u| \geq a. \end{cases}$$

  The log barrier penalty function assesses an infinite penalty for residuals larger than a.

A deadzone-linear, log barrier, and quadratic penalty function are plotted in figure 6.1. Note that the log barrier function is very close to the quadratic penalty for $|u/a| \leq 0.25$ (see exercise 6.1).

Scaling the penalty function by a positive number does not affect the solution of the penalty function approximation problem, since this merely scales the objective

function. But the *shape* of the penalty function has a large effect on the solution of the penalty function approximation problem. Roughly speaking, $\phi(u)$ is a measure of our dislike of a residual of value u. If ϕ is very small (or even zero) for small values of u, it means we care very little (or not at all) if residuals have these values. If $\phi(u)$ grows rapidly as u becomes large, it means we have a strong dislike for large residuals; if ϕ becomes infinite outside some interval, it means that residuals outside the interval are unacceptable. This simple interpretation gives insight into the solution of a penalty function approximation problem, as well as guidelines for choosing a penalty function.

As an example, let us compare ℓ_1-norm and ℓ_2-norm approximation, associated with the penalty functions $\phi_1(u) = |u|$ and $\phi_2(u) = u^2$, respectively. For $|u| = 1$, the two penalty functions assign the same penalty. For small u we have $\phi_1(u) \gg \phi_2(u)$, so ℓ_1-norm approximation puts relatively larger emphasis on small residuals compared to ℓ_2-norm approximation. For large u we have $\phi_2(u) \gg \phi_1(u)$, so ℓ_1-norm approximation puts less weight on large residuals, compared to ℓ_2-norm approximation. This difference in relative weightings for small and large residuals is reflected in the solutions of the associated approximation problems. The amplitude distribution of the optimal residual for the ℓ_1-norm approximation problem will tend to have more zero and very small residuals, compared to the ℓ_2-norm approximation solution. In contrast, the ℓ_2-norm solution will tend to have relatively fewer large residuals (since large residuals incur a much larger penalty in ℓ_2-norm approximation than in ℓ_1-norm approximation).

Example

An example will illustrate these ideas. We take a matrix $A \in \mathbf{R}^{100\times 30}$ and vector $b \in \mathbf{R}^{100}$ (chosen at random, but the results are typical), and compute the ℓ_1-norm and ℓ_2-norm approximate solutions of $Ax \approx b$, as well as the penalty function approximations with a deadzone-linear penalty (with $a = 0.5$) and log barrier penalty (with $a = 1$). Figure 6.2 shows the four associated penalty functions, and the amplitude distributions of the optimal residuals for these four penalty approximations. From the plots of the penalty functions we note that

- The ℓ_1-norm penalty puts the most weight on small residuals and the least weight on large residuals.
- The ℓ_2-norm penalty puts very small weight on small residuals, but strong weight on large residuals.
- The deadzone-linear penalty function puts no weight on residuals smaller than 0.5, and relatively little weight on large residuals.
- The log barrier penalty puts weight very much like the ℓ_2-norm penalty for small residuals, but puts very strong weight on residuals larger than around 0.8, and infinite weight on residuals larger than 1.

Several features are clear from the amplitude distributions:

- For the ℓ_1-optimal solution, many residuals are either zero or very small. The ℓ_1-optimal solution also has relatively more large residuals.

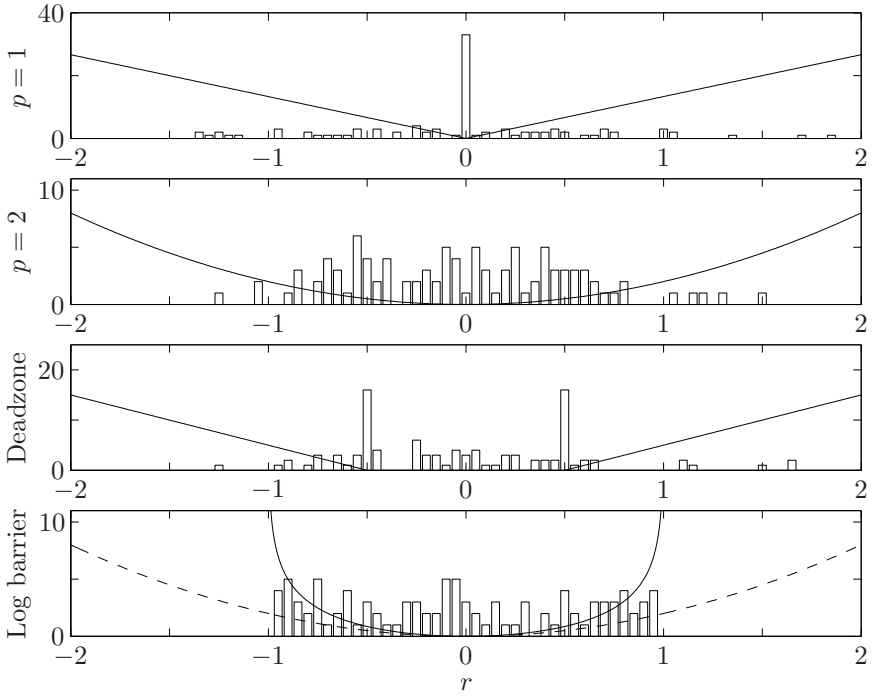

Figure 6.2 Histogram of residual amplitudes for four penalty functions, with the (scaled) penalty functions also shown for reference. For the log barrier plot, the quadratic penalty is also shown, in dashed curve.

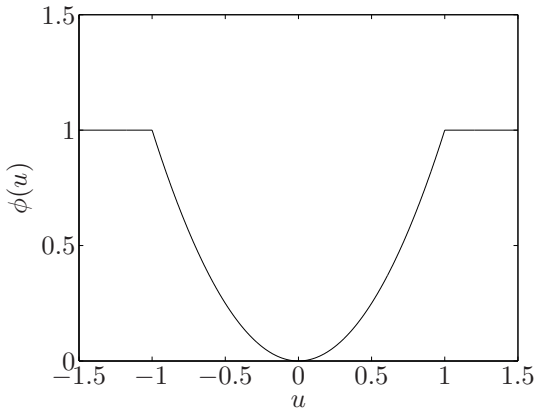

Figure 6.3 A (nonconvex) penalty function that assesses a fixed penalty to residuals larger than a threshold (which in this example is one): $\phi(u) = u^2$ if $|u| \leq 1$ and $\phi(u) = 1$ if $|u| > 1$. As a result, penalty approximation with this function would be relatively insensitive to outliers.

- The ℓ_2-norm approximation has many modest residuals, and relatively few larger ones.
- For the deadzone-linear penalty, we see that many residuals have the value ± 0.5, right at the edge of the 'free' zone, for which no penalty is assessed.
- For the log barrier penalty, we see that no residuals have a magnitude larger than 1, but otherwise the residual distribution is similar to the residual distribution for ℓ_2-norm approximation.

Sensitivity to outliers or large errors

In the estimation or regression context, an *outlier* is a measurement $y_i = a_i^T x + v_i$ for which the noise v_i is relatively large. This is often associated with faulty data or a flawed measurement. When outliers occur, any estimate of x will be associated with a residual vector with some large components. Ideally we would like to guess which measurements are outliers, and either remove them from the estimation process or greatly lower their weight in forming the estimate. (We cannot, however, assign zero penalty for very large residuals, because then the optimal point would likely make all residuals large, which yields a total penalty of zero.) This could be accomplished using penalty function approximation, with a penalty function such as

$$\phi(u) = \begin{cases} u^2 & |u| \leq M \\ M^2 & |u| > M, \end{cases} \tag{6.3}$$

shown in figure 6.3. This penalty function agrees with least-squares for any residual smaller than M, but puts a fixed weight on any residual larger than M, no matter how much larger it is. In other words, residuals larger than M are ignored; they are assumed to be associated with outliers or bad data. Unfortunately, the penalty

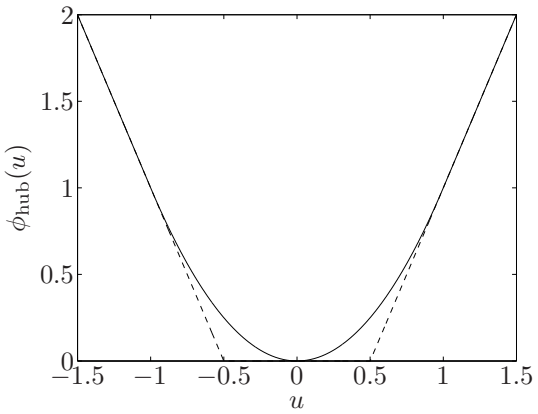

Figure 6.4 The solid line is the robust least-squares or Huber penalty function $\phi_{\rm hub}$, with $M = 1$. For $|u| \leq M$ it is quadratic, and for $|u| > M$ it grows linearly.

function (6.3) is not convex, and the associated penalty function approximation problem becomes a hard combinatorial optimization problem.

The sensitivity of a penalty function based estimation method to outliers depends on the (relative) value of the penalty function for large residuals. If we restrict ourselves to convex penalty functions (which result in convex optimization problems), the ones that are least sensitive are those for which $\phi(u)$ grows linearly, *i.e.*, like $|u|$, for large u. Penalty functions with this property are sometimes called *robust*, since the associated penalty function approximation methods are much less sensitive to outliers or large errors than, for example, least-squares.

One obvious example of a robust penalty function is $\phi(u) = |u|$, corresponding to ℓ_1-norm approximation. Another example is the *robust least-squares* or *Huber penalty function*, given by

$$\phi_{\rm hub}(u) = \begin{cases} u^2 & |u| \leq M \\ M(2|u| - M) & |u| > M, \end{cases} \tag{6.4}$$

shown in figure 6.4. This penalty function agrees with the least-squares penalty function for residuals smaller than M, and then reverts to ℓ_1-like linear growth for larger residuals. The Huber penalty function can be considered a convex approximation of the outlier penalty function (6.3), in the following sense: They agree for $|u| \leq M$, and for $|u| > M$, the Huber penalty function is the convex function closest to the outlier penalty function (6.3).

Example 6.2 *Robust regression.* Figure 6.5 shows 42 points (t_i, y_i) in a plane, with two obvious outliers (one at the upper left, and one at lower right). The dashed line shows the least-squares approximation of the points by a straight line $f(t) = \alpha + \beta t$. The coefficients α and β are obtained by solving the least-squares problem

$$\text{minimize} \quad \sum_{i=1}^{42} (y_i - \alpha - \beta t_i)^2,$$

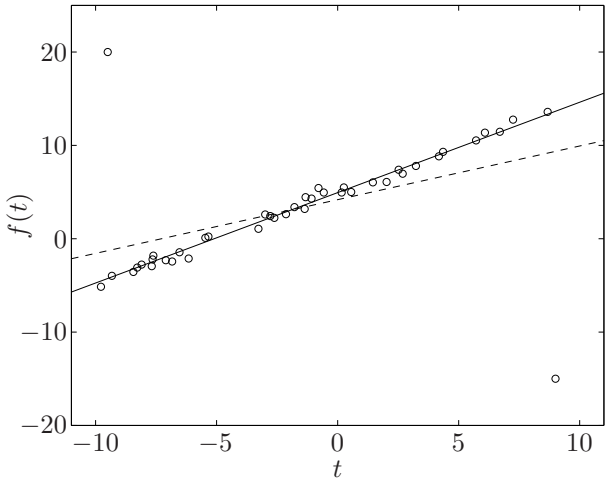

Figure 6.5 The 42 circles show points that can be well approximated by an affine function, except for the two outliers at upper left and lower right. The dashed line is the least-squares fit of a straight line $f(t) = \alpha + \beta t$ to the points, and is rotated away from the main locus of points, toward the outliers. The solid line shows the robust least-squares fit, obtained by minimizing Huber's penalty function with $M = 1$. This gives a far better fit to the non-outlier data.

with variables α and β. The least-squares approximation is clearly rotated away from the main locus of the points, toward the two outliers.

The solid line shows the robust least-squares approximation, obtained by minimizing the Huber penalty function

$$\text{minimize} \quad \sum_{i=1}^{42} \phi_{\text{hub}}(y_i - \alpha - \beta t_i),$$

with $M = 1$. This approximation is far less affected by the outliers.

Since ℓ_1-norm approximation is among the (convex) penalty function approximation methods that are most robust to outliers, ℓ_1-norm approximation is sometimes called *robust estimation* or *robust regression*. The robustness property of ℓ_1-norm estimation can also be understood in a statistical framework; see page 353.

Small residuals and ℓ_1-norm approximation

We can also focus on small residuals. Least-squares approximation puts very small weight on small residuals, since $\phi(u) = u^2$ is very small when u is small. Penalty functions such as the deadzone-linear penalty function put *zero* weight on small residuals. For penalty functions that are very small for small residuals, we expect the optimal residuals to be small, but not very small. Roughly speaking, there is little or no incentive to drive small residuals smaller.

In contrast, penalty functions that put relatively large weight on small residuals, such as $\phi(u) = |u|$, corresponding to ℓ_1-norm approximation, tend to produce

optimal residuals many of which are very small, or even exactly zero. This means that in ℓ_1-norm approximation, we typically find that many of the equations are satisfied exactly, *i.e.*, we have $a_i^T x = b_i$ for many i. This phenomenon can be seen in figure 6.2.

6.1.3 Approximation with constraints

It is possible to add constraints to the basic norm approximation problem (6.1). When these constraints are convex, the resulting problem is convex. Constraints arise for a variety of reasons.

- In an approximation problem, constraints can be used to rule out certain unacceptable approximations of the vector b, or to ensure that the approximator Ax satisfies certain properties.
- In an estimation problem, the constraints arise as prior knowledge of the vector x to be estimated, or from prior knowledge of the estimation error v.
- Constraints arise in a geometric setting in determining the projection of a point b on a set more complicated than a subspace, for example, a cone or polyhedron.

Some examples will make these clear.

Nonnegativity constraints on variables

We can add the constraint $x \succeq 0$ to the basic norm approximation problem:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\| \\ \text{subject to} & x \succeq 0. \end{array}$$

In an estimation setting, nonnegativity constraints arise when we estimate a vector x of parameters known to be nonnegative, *e.g.*, powers, intensities, or rates. The geometric interpretation is that we are determining the projection of a vector b onto the cone generated by the columns of A. We can also interpret this problem as approximating b using a nonnegative linear (*i.e.*, conic) combination of the columns of A.

Variable bounds

Here we add the constraint $l \preceq x \preceq u$, where $l,\ u \in \mathbf{R}^n$ are problem parameters:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\| \\ \text{subject to} & l \preceq x \preceq u. \end{array}$$

In an estimation setting, variable bounds arise as prior knowledge of intervals in which each variable lies. The geometric interpretation is that we are determining the projection of a vector b onto the image of a box under the linear mapping induced by A.

Probability distribution

We can impose the constraint that x satisfy $x \succeq 0$, $\mathbf{1}^T x = 1$:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\| \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

This would arise in the estimation of proportions or relative frequencies, which are nonnegative and sum to one. It can also be interpreted as approximating b by a convex combination of the columns of A. (We will have much more to say about estimating probabilities in §7.2.)

Norm ball constraint

We can add to the basic norm approximation problem the constraint that x lie in a norm ball:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\| \\ \text{subject to} & \|x - x_0\| \leq d, \end{array}$$

where x_0 and d are problem parameters. Such a constraint can be added for several reasons.

- In an estimation setting, x_0 is a prior guess of what the parameter x is, and d is the maximum plausible deviation of our estimate from our prior guess. Our estimate of the parameter x is the value $\hat{x}$ which best matches the measured data (*i.e.*, minimizes $\|Az - b\|$) among all plausible candidates (*i.e.*, z that satisfy $\|z - x_0\| \leq d$).
- The constraint $\|x - x_0\| \leq d$ can denote a *trust region*. Here the linear relation $y = Ax$ is only an approximation of some nonlinear relation $y = f(x)$ that is valid when x is near some point x_0, specifically $\|x - x_0\| \leq d$. The problem is to minimize $\|Ax - b\|$ but only over those x for which the model $y = Ax$ is trusted.

These ideas also come up in the context of regularization; see §6.3.2.

6.2 Least-norm problems

The basic *least-norm problem* has the form

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & Ax = b \end{array} \tag{6.5}$$

where the data are $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$, the variable is $x \in \mathbf{R}^n$, and $\|\cdot\|$ is a norm on $\mathbf{R}^n$. A solution of the problem, which always exists if the linear equations $Ax = b$ have a solution, is called a *least-norm solution* of $Ax = b$. The least-norm problem is, of course, a convex optimization problem.

We can assume without loss of generality that the rows of A are independent, so $m \leq n$. When $m = n$, the only feasible point is $x = A^{-1}b$; the least-norm problem is interesting only when $m < n$, *i.e.*, when the equation $Ax = b$ is underdetermined.