Chapter 2
Sampling and linear models

2.1 Positive spanning sets and positive bases

Positive spanning sets and positive bases are used in directional direct-search methods. As we will see later, the main motivation to look at a positive spanning set $D \subset \mathbb{R}^n$ is the guarantee that, given any nonzero vector v in $\mathbb{R}^n$, there is at least one vector d in D such that v and d form an acute angle. The implication in optimization is then obvious. Suppose that the nonzero vector v is the negative gradient, $-\nabla f(x)$, of a continuously differentiable function f at a given point x. Any vector d that forms an acute angle with $-\nabla f(x)$ is a descent direction.[1] In order to decrease $f(x)$, it might be required to evaluate the points $x + \alpha d$ (for all $d \in D$), where $\alpha > 0$, and to repeat this evaluation for smaller positive values of α. But, since the gradient $\nabla f(x)$ is nonzero, there exist a positive value for α and a vector d in D for which $f(x + \alpha d) < f(x)$, which shows that such a scheme should be terminated after a finite number of reductions of the parameter α.

In this section, we will review some of the basic properties of positive spanning sets and positive bases and show how to construct simple positive bases. Most of the basic properties about positive spanning sets are extracted from the theory of positive linear dependence developed by Davis [74] (see also the paper by Lewis and Torczon [153]).

Definitions and properties

The *positive span*[2] of a set of vectors $[v_1 \cdots v_r]$ in $\mathbb{R}^n$ is the convex cone

$$\left\{ v \in \mathbb{R}^n : \; v = \alpha_1 v_1 + \cdots + \alpha_r v_r, \quad \alpha_i \geq 0, \, i = 1, \ldots, r \right\}.$$

(Many times it will be convenient for us in this book to regard a set of vectors as a matrix whose columns are the vectors in the set.)

[1] By a descent direction for f at x we mean a direction d for which there exists an $\bar{\alpha} > 0$ such that $f(x + \alpha d) < f(x)$ for all $\alpha \in (0, \bar{\alpha}]$.

[2] Strictly speaking we should have written *nonnegative* instead of positive, but we decided to follow the notation in [74, 153]. We also note that by *span* we mean *linear span*.

Definition 2.1. *A positive spanning set in $\mathbb{R}^n$ is a set of vectors whose positive span is $\mathbb{R}^n$.*

The set $[v_1 \cdots v_r]$ is said to be positively dependent if one of the vectors is in the convex cone positively spanned by the remaining vectors, i.e., if one of the vectors is a positive combination of the others; otherwise, the set is positively independent.

A positive basis in $\mathbb{R}^n$ is a positively independent set whose positive span is $\mathbb{R}^n$.

Equivalently, a positive basis for $\mathbb{R}^n$ can be defined as a set of nonzero vectors of $\mathbb{R}^n$ whose positive combinations span $\mathbb{R}^n$ but for which no proper set exhibits the same property.

The following theorem due to [74] indicates that a positive spanning set contains at least $n+1$ vectors in $\mathbb{R}^n$.

Theorem 2.2. *If $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$ positively, then it contains a subset with $r-1$ elements that spans $\mathbb{R}^n$.*

Proof. The set $[v_1 \cdots v_r]$ is necessarily linearly dependent (otherwise, it would be possible to construct a basis for $\mathbb{R}^n$ that would span $\mathbb{R}^n$ positively). As a result, there are scalars $\bar{a}_1, \ldots, \bar{a}_r$ (not all zero) such that $\bar{a}_1 v_1 + \cdots + \bar{a}_r v_r = 0$. Thus, there exists an $i \in \{1, \ldots, r\}$ for which $\bar{a}_i \neq 0$.

Now let v be an arbitrary vector in $\mathbb{R}^n$. Since $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$ positively, there exist nonnegative scalars $a_1, \ldots, a_r$ such that $v = a_1 v_1 + \cdots + a_r v_r$.

As a result, we get

$$v = \sum_{j=1}^{r} a_j v_j = \sum_{\substack{j=1 \\ j \neq i}}^{r} \left(a_j - \frac{\bar{a}_j}{\bar{a}_i} a_i \right) v_j.$$

Since v is arbitrary, we have proved that $\{v_1, \ldots, v_r\} \backslash \{v_i\}$ spans $\mathbb{R}^n$. □

It can also be shown that a positive basis cannot contain more than $2n$ elements (see [74]). Positive bases with $n+1$ and $2n$ elements are referred to as *minimal* and *maximal* positive bases, respectively.

The positive basis formed by the vectors of the canonical basis and their negative counterparts is the most simple maximal positive basis one can think of. In $\mathbb{R}^2$, this positive basis is defined by the columns of the matrix

$$D_1 = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.$$

Later in the book we will designate this basis by $D_\oplus$.

A simple minimal basis in $\mathbb{R}^2$ is formed by the vectors of the canonical basis and the negative of their sum:

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

For convenience, we normalize the third vector and write

$$D_2 = \begin{bmatrix} 1 & 0 & -\sqrt{2}/2 \\ 0 & 1 & -\sqrt{2}/2 \end{bmatrix}.$$

If we add one more vector to this positive basis, we get a positive spanning set that is not a positive basis:

$$D_3 = \begin{bmatrix} 1 & 0 & -\sqrt{2}/2 & 0 \\ 0 & 1 & -\sqrt{2}/2 & -1 \end{bmatrix}.$$

In Figure 2.1, we plot the positive bases D_1 and D_2 and the positive spanning set D_3.

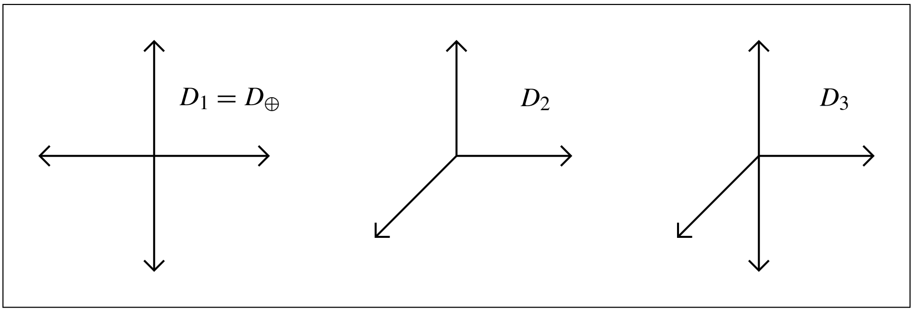


Figure 2.1. *A maximal positive basis (left), a minimal positive basis (center), and a positive spanning set that is not a positive basis (right).*

We now present three necessary and sufficient characterizations for a set that spans $\mathbb{R}^n$ to also span $\mathbb{R}^n$ positively (see also [74]).

Theorem 2.3. *Let $[v_1 \cdots v_r]$, with $v_i \neq 0$ for all $i \in \{1,\ldots,r\}$, span $\mathbb{R}^n$. Then the following are equivalent:*

(i) $[v_1 \cdots v_r]$ *spans $\mathbb{R}^n$ positively.*

(ii) *For every $i = 1,\ldots,r$, the vector $-v_i$ is in the convex cone positively spanned by the remaining $r-1$ vectors.*

(iii) *There exist real scalars $\alpha_1,\ldots,\alpha_r$ with $\alpha_i > 0$, $i \in \{1,\ldots,r\}$, such that $\sum_{i=1}^r \alpha_i v_i = 0$.*

(iv) *For every nonzero vector $w \in \mathbb{R}^n$, there exists an index i in $\{1,\ldots,r\}$ for which $w^\top v_i > 0$.*

Proof. The proof is made by showing the following implications: (i) $\Rightarrow$ (ii), (ii) $\Rightarrow$ (iii), (iii) $\Rightarrow$ (i), (i) $\Rightarrow$ (iv), and (iv) $\Rightarrow$ (i).

(i) $\Rightarrow$ (ii) Since $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$ positively, the vector $-v_i$, with i in $\{1,\ldots,r\}$, can be written as

$$-v_i = \sum_{j=1}^{r} \lambda_{ij} v_j,$$

where the scalars $\lambda_{i1},\ldots,\lambda_{ir}$ are nonnegative. As a consequence, we obtain

$$-v_i - \lambda_{ii} v_i \;=\; \sum_{\substack{j=1\\ j\neq i}}^{r} \lambda_{ij} v_j$$

and

$$-v_i \;=\; \sum_{\substack{j=1\\ j\neq i}}^{r} \frac{\lambda_{ij}}{1+\lambda_{ii}} v_j \;=\; \sum_{\substack{j=1\\ j\neq i}}^{r} \tilde{\lambda}_{ij} v_j,$$

where $\tilde{\lambda}_{ij} = \frac{\lambda_{ij}}{1+\lambda_{ii}} \geq 0$ for all $j \in \{1,\ldots,r\}\backslash\{i\}$. This shows that $-v_i$ is in the convex cone positively spanned by the remaining $r-1$ vectors.

(ii) $\Rightarrow$ (iii) From the assumption (ii), there exist nonnegative scalars $\bar{\lambda}_{ij}$, $i,j=1,\ldots,r$, such that

$$\begin{aligned}
v_1 \;+\; \bar{\lambda}_{12} v_2 \;+\cdots+\; \bar{\lambda}_{1r} v_r \;&=\; 0,\\
\bar{\lambda}_{21} v_1 \;+\; v_2 \;+\cdots+\; \bar{\lambda}_{2r} v_r \;&=\; 0,\\
&\;\;\vdots\\
\bar{\lambda}_{r1} v_1 \;+\; \bar{\lambda}_{r2} v_2 \;+\cdots+ v_r \;&=\; 0.
\end{aligned}$$

By adding these r equalities, we get

$$\left(1+\sum_{i=2}^{r}\bar{\lambda}_{i1}\right) v_1 + \cdots + \left(1+\sum_{i=1}^{r-1}\bar{\lambda}_{ir}\right) v_r \;=\; 0,$$

which can be rewritten as

$$\alpha_1 v_1 + \cdots + \alpha_r v_r \;=\; 0,$$

with $\alpha_j = 1 + \sum_{\substack{i=1\\ i\neq j}}^{r} \bar{\lambda}_{ij} > 0$, $j \in \{1,\ldots,r\}$.

(iii) $\Rightarrow$ (i) Let $\alpha_1,\ldots,\alpha_r$ be positive scalars such that $\alpha_1 v_1 + \cdots + \alpha_r v_r = 0$, and let v be an arbitrary vector in $\mathbb{R}^n$. Since $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$, there exist scalars $\lambda_1,\ldots,\lambda_r$ such that $v = \lambda_1 v_1 + \cdots + \lambda_r v_r$. By adding to the right-hand side of this equality a sufficiently large multiple of $\alpha_1 v_1 + \cdots + \alpha_r v_r$, one can show that v can be expressed as a positive linear combination of $v_1,\ldots,v_r$. Thus, $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$ positively.

(i) $\Rightarrow$ (iv) Let w be a nonzero vector in $\mathbb{R}^n$. From the assumption (i), there exist nonnegative scalars $\lambda_1,\ldots,\lambda_r$ such that

$$w \;=\; \lambda_1 v_1 + \cdots + \lambda_r v_r.$$

Since $w \neq 0$, we get that

$$\begin{aligned}
0 \;<\; w^\top w &= (\lambda_1 v_1 + \cdots + \lambda_r v_r)^\top w\\
&= \lambda_1 v_1^\top w + \cdots + \lambda_r v_r^\top w,
\end{aligned}$$

from which we conclude that at least one of the scalars $w^\top v_1,\ldots,w^\top v_r$ has to be positive.

(iv) $\Rightarrow$ (i) If the convex cone positively spanned by $v_1,\ldots,v_r$ is not $\mathbb{R}^n$, then there exists a hyperplane $H = \{v \in \mathbb{R}^n : v^\top h = 0\}$, with $h \neq 0$, such that this convex cone (and so all of its generators) is contained in either $\{v \in \mathbb{R}^n : v^\top h \geq 0\}$ or $\{v \in \mathbb{R}^n : v^\top h \leq 0\}$; see [200, Corollary 11.7.3]. The assumption would then be contradicted with either $w = h$ or $w = -h$. $\square$

As mentioned before, the characterization (iv) of Theorem 2.3 is at the heart of directional direct-search methods. It implies that, given a continuously differentiable function f at some given point x where $\nabla f(x) \neq 0$, there must always exist a vector d in a given positive spanning set (or in a positive basis) such that

$$-\nabla f(x)^\top d \; > \; 0.$$

In other words, there must always exist a direction of descent in such a set. In Figure 2.2, we identify such a vector d for the three spanning sets D_1, D_2, and D_3 given before.

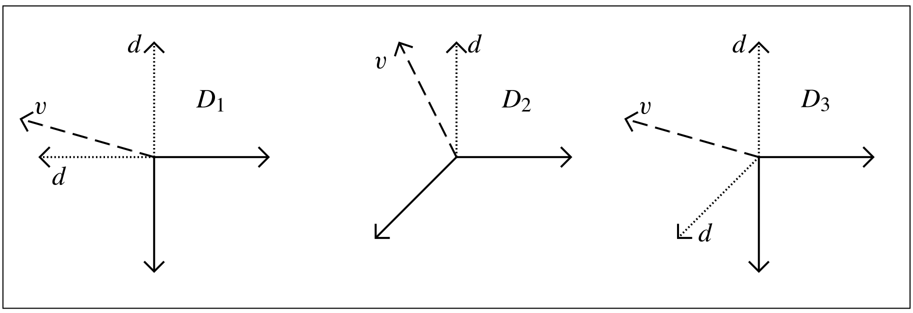


Figure 2.2. *Given a positive spanning set and a vector $v = -\nabla f(x)$ (dashed), there exists at least one element d (dotted) of the set such that $v^\top d > 0$.*

Simple positive bases

Now we turn our attention to the construction of positive bases. The following result (given in [153]) provides a simple mechanism for generating different positive bases.

Theorem 2.4. *Suppose $[v_1 \cdots v_r]$ is a positive basis for $\mathbb{R}^n$ and $W \in \mathbb{R}^{n\times n}$ is a nonsingular matrix. Then $[Wv_1 \cdots Wv_r]$ is also a positive basis for $\mathbb{R}^n$.*

Proof. It is obvious that $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$ since it does it positively. Since W is nonsingular, $[Wv_1 \cdots Wv_r]$ also spans $\mathbb{R}^n$. Thus we can apply Theorem 2.3 for both $[v_1 \cdots v_r]$ and $[Wv_1 \cdots Wn_r]$.

Now let w be a nonzero vector in $\mathbb{R}^n$. Since $[v_1 \cdots v_r]$ spans $\mathbb{R}^n$ positively and W is nonsingular, we get from (iv) in Theorem 2.3 that

$$(W^\top w)^\top v_i \; > \; 0$$

for some i in $\{1,\ldots,r\}$. In other words,

$$w^\top(Wv_i) > 0$$

for some i in $\{1,\ldots,r\}$, from which we conclude that $[Wv_1 \cdots Wv_r]$ also spans $\mathbb{R}^n$ positively.

It is a direct consequence of the definition of positive dependence that if $[Wv_1 \cdots Wv_r]$ was positively dependent, then $[v_1 \cdots v_r]$ would also be positively dependent, which concludes the proof of the theorem. □

One can easily prove that $D_\oplus = [I \ -I]$ is a (maximal) positive basis. The result just stated in Theorem 2.4 allows us to say that $[W \ -W]$ is also a (maximal) positive basis for any choice of the nonsingular matrix $W \in \mathbb{R}^{n\times n}$.

From Theorems 2.3 and 2.4, we can easily deduce the following corollary. The proof is left as an exercise.

Corollary 2.5.

(i) $[I \ -e]$ *is a (minimal) positive basis.*

(ii) *Let* $W = [w_1 \cdots w_n] \in \mathbb{R}^{n\times n}$ *be a nonsingular matrix. Then* $[W \ -\sum_{i=1}^n w_i]$ *is a (minimal) positive basis for* $\mathbb{R}^n$.

Positive basis with uniform angles

Consider $n+1$ vectors $v_1,\ldots,v_{n+1}$ in $\mathbb{R}^n$ for which all the angles between pairs v_i, v_j $(i \neq j)$ have the same amplitude α. Assuming that the $n+1$ vectors are normalized, this requirement is expressed as

$$a = \cos(\alpha) = v_i^\top v_j, \qquad i,j \in \{1,\ldots,n+1\},\ i \neq j, \tag{2.1}$$

where $a \neq 1$. One can show that $a = -1/n$ (see the exercises).

Now we seek a set of $n+1$ normalized vectors $[v_1 \cdots v_{n+1}]$ satisfying the property (2.1) with $a = -1/n$. Let us first compute $v_1,\ldots,v_n$; i.e., let us compute a matrix $V = [v_1 \cdots v_n]$ such that

$$V^\top V = A,$$

where A is the matrix given by

$$A = \begin{bmatrix} 1 & -1/n & -1/n & \cdots & -1/n \\ -1/n & 1 & -1/n & \cdots & -1/n \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ -1/n & -1/n & -1/n & \cdots & 1 \end{bmatrix}. \tag{2.2}$$

The matrix A is symmetric and diagonally dominant with positive diagonal entries, and, therefore, it is positive definite [109]. Thus, we can make use of its Cholesky decomposition

$$A = CC^\top,$$

where $C \in \mathbb{R}^{n\times n}$ is a lower triangular matrix of order n with positive diagonal entries. Given this decomposition, one can easily see that a choice for V is determined by

$$V = [v_1 \cdots v_n] = C^\top.$$

The vector v_{n+1} is then computed by

$$v_{n+1} = -\sum_{i=1}^{n} v_i. \tag{2.3}$$

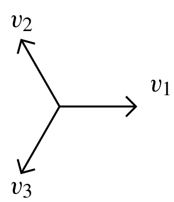
One can easily show that $v_i^\top v_{n+1} = -1/n$, $i = 1,\ldots,n$, and $v_{n+1}^\top v_{n+1} = 1$.

Since V is nonsingular and v_{n+1} is determined by (2.3), we can apply Corollary 2.5(ii) to establish that $[v_1 \cdots v_{n+1}]$ is a (minimal) positive basis. The angles between any two vectors in this positive basis exhibit the same amplitude. We summarize this result below.

Corollary 2.6. *Let* $V = C^\top = [v_1 \cdots v_n] \in \mathbb{R}^{n\times n}$, *where* $A = CC^\top$ *and* A *is given by* (2.2). *Let* $v_{n+1} = -\sum_{i=1}^{n} v_i$.

Then $[v_1 \cdots v_{n+1}]$ *is a (minimal) positive basis for* $\mathbb{R}^n$ *satisfying* $v_i^\top v_j = -1/n$, $i, j \in \{1,\ldots,n+1\}$, $i \neq j$, *and* $\|v_i\| = 1$, $i = 1,\ldots,n+1$.

A minimal positive basis with uniform angles is depicted in Figure 2.3, computed as in Corollary 2.6.



$$[\, v_1 \; v_2 \; v_3 \,] = \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3/4} & -\sqrt{3/4} \end{bmatrix}$$

Figure 2.3. *A minimal positive basis with uniform angles.*

2.2 Gradient estimates used in direct search

What we now show is that if we sample $n+1$ points of the form $x+\alpha d$ defined by a positive basis D, and their function values are no better than the function value at x, then the size of the gradient (considered Lipschitz continuous) of the function at x is of the order of the distance between x and the sample points $x+\alpha d$ and, furthermore, the order constant depends only upon the nonlinearity of f and the geometry of the sample set.

To prove this result, used in the convergence theory of directional direct-search methods, we must first introduce the notion of cosine measure for positive spanning sets.

Definition 2.7. *The cosine measure of a positive spanning set (with nonzero vectors) or of a positive basis D is defined by*

$$\mathrm{cm}(D) \;=\; \min_{0\neq v\in\mathbb{R}^n} \max_{d\in D} \frac{v^\top d}{\|v\|\|d\|}.$$

Given any positive spanning set, it necessarily happens that

$$\mathrm{cm}(D) \;>\; 0.$$

Values of the cosine measure close to zero indicate a deterioration of the positive spanning property. For example, the maximal positive basis $D_\oplus = [I \; -I]$ has cosine measure equal to $1/\sqrt{n}$. When $n = 2$ we have $\mathrm{cm}(D_\oplus) = \sqrt{2}/2$.

Now let us consider the following example. Let θ be an angle in $(0, \pi/4]$ and D_θ be a positive basis defined by

$$D_\theta \;=\; \begin{bmatrix} 1 & 0 & -\cos(\theta) \\ 0 & 1 & -\sin(\theta) \end{bmatrix}.$$

Observe that $D_{\frac{\pi}{4}}$ is just the positive basis D_2 considered before. The cosine measure of D_θ is given by $\cos((\pi-\theta)/2)$, and it converges to zero when θ tends to zero. Figure 2.4 depicts this situation.

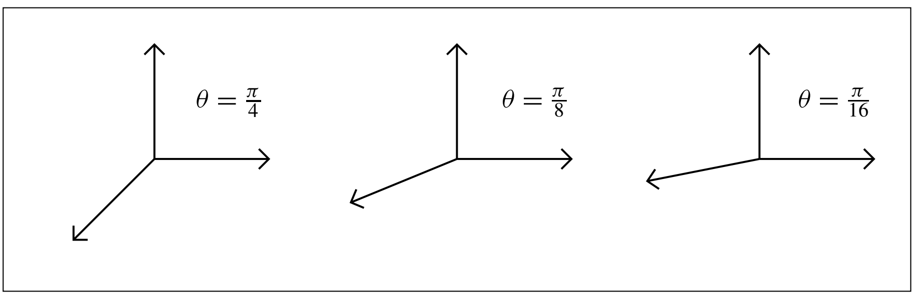


Figure 2.4. *Positive bases D_θ for three values of θ. The cosine measure is approaching zero.*

Another key point, related to the descent properties of positive spanning sets, is that, given any vector $v \neq 0$, we have

$$\mathrm{cm}(D) \;\leq\; \max_{d\in D} \frac{v^\top d}{\|v\|\|d\|}.$$

Thus, there must exist a $d \in D$ such that

$$\mathrm{cm}(D) \;\leq\; \frac{v^\top d}{\|v\|\|d\|}$$

or, equivalently,

$$\mathrm{cm}(D)\|v\|\|d\| \;\leq\; v^\top d.$$

Given a positive spanning set D, a point x, and a positive value for the parameter α, we are interested in looking at the points of the form $x+\alpha d$ for all $d \in D$. These points are in a ball centered at x, of radius Δ, defined by

$$\Delta \;=\; \alpha \max_{d\in D} \|d\|.$$

We point out that if only a finite number of positive spanning sets are used in an algorithm, then Δ tends to zero if and only if α tends to zero. The following result is taken from [80, 145].

Theorem 2.8. *Let D be a positive spanning set and $\alpha > 0$ be given. Assume that ∇f is Lipschitz continuous (with constant $\nu > 0$) in an open set containing the ball $B(x;\Delta)$. If $f(x) \leq f(x+\alpha d)$, for all $d \in D$, then*

$$\|\nabla f(x)\| \;\leq\; \frac{\nu}{2}\,\mathrm{cm}(D)^{-1} \max_{d\in D}\|d\|\,\alpha.$$

Proof. Let d be a vector in D for which

$$\mathrm{cm}(D)\|\nabla f(x)\|\|d\| \;\leq\; -\nabla f(x)^\top d.$$

Now, from the integral form of the mean value theorem and the fact that $f(x) \leq f(x+\alpha d)$, we get, for all $d \in D$, that

$$0 \;\leq\; f(x+\alpha d) - f(x) \;=\; \int_0^1 \nabla f(x+t\alpha d)^\top(\alpha d)dt.$$

By multiplying the first inequality by α and by adding it to the second one, we obtain

$$\mathrm{cm}(D)\|\nabla f(x)\|\|d\|\alpha \;\leq\; \int_0^1 \left(\nabla f(x+t\alpha d) - \nabla f(x)\right)^\top(\alpha d)dt \;\leq\; \frac{\nu}{2}\|d\|^2\alpha^2,$$

and the proof is completed. $\square$

If a directional direct-search method is able to generate a sequence of points x satisfying the conditions of Theorem 2.8 for which α (and thus Δ) tends to zero, then clearly the gradient of the objective function converges to zero along this sequence.

The bound proved in Theorem 2.8 can be rewritten in the form

$$\|\nabla f(x)\| \;\leq\; \kappa_{eg}\Delta,$$

where $\kappa_{eg} = \nu\,\mathrm{cm}(D)^{-1}/2$. We point out that this bound has the same structure as other bounds used in different derivative-free methods. The bound is basically given by Δ times a constant that depends on the nonlinearity of the function (expressed by the Lipschitz constant ν) and on the geometry of the sample set (measured by $\mathrm{cm}(D)^{-1}$).

2.3 Linear interpolation and regression models

Now we turn our attention to sample sets not necessarily formed by a predefined set of directions. Consider a sample set $Y = \{y^0, y^1, \ldots, y^p\}$ in $\mathbb{R}^n$. The simplest model based on $n+1$ sample points ($p = n$) that we can think of is an interpolation model.

Linear interpolation

Let $m(x)$ denote a polynomial of degree $d = 1$ interpolating f at the points in Y, i.e., satisfying the interpolation conditions

$$m(y^i) \;=\; f(y^i), \quad i = 0, \ldots, n. \tag{2.4}$$

We can express $m(x)$ in the form

$$m(x) \;=\; \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n,$$

using, as a basis for the space $\mathcal{P}_n^1$ of linear polynomials of degree 1, the polynomial basis $\bar{\phi} = \{1, x_1, \ldots, x_n\}$. However, we point out that other bases could be used, e.g., $\{1, 1 + x_1, 1 + x_1 + x_2, \ldots, 1 + x_1 + x_2 + \cdots + x_n\}$. We can then rewrite (2.4) as

$$\begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 \\ 1 & y_1^1 & \cdots & y_n^1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_1^n & \cdots & y_n^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} f(y^0) \\ f(y^1) \\ \vdots \\ f(y^n) \end{bmatrix}.$$

The matrix of this linear system is denoted by

$$M \;=\; M(\bar{\phi}, Y) \;=\; \begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 \\ 1 & y_1^1 & \cdots & y_n^1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_1^n & \cdots & y_n^n \end{bmatrix}. \tag{2.5}$$

In this book we write M as $M(\bar{\phi}, Y)$ to highlight the dependence of M on the basis $\bar{\phi}$ and on the sample set Y.

Definition 2.9. *The set $Y = \{y^0, y^1, \ldots, y^n\}$ is poised for linear interpolation in $\mathbb{R}^n$ if the corresponding matrix $M(\bar{\phi}, Y)$ is nonsingular.*

The definition of poisedness is independent of the basis chosen. In other words, if Y is poised for a basis ϕ, then it is poised for any other basis in $\mathcal{P}_n^1$. The definition of $m(x)$ is also independent of the basis chosen. These issues are covered in detail in Chapter 3. It is straightforward to see that the sample set Y is poised for linear interpolation if and only if the linear interpolating polynomial $m(x) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n$ is uniquely defined.

Linear regression

When the number $p+1$ of points in the sample set exceeds by more than 1 the dimension n of the sampling space, it might not be possible to fit a linear polynomial. In this case,

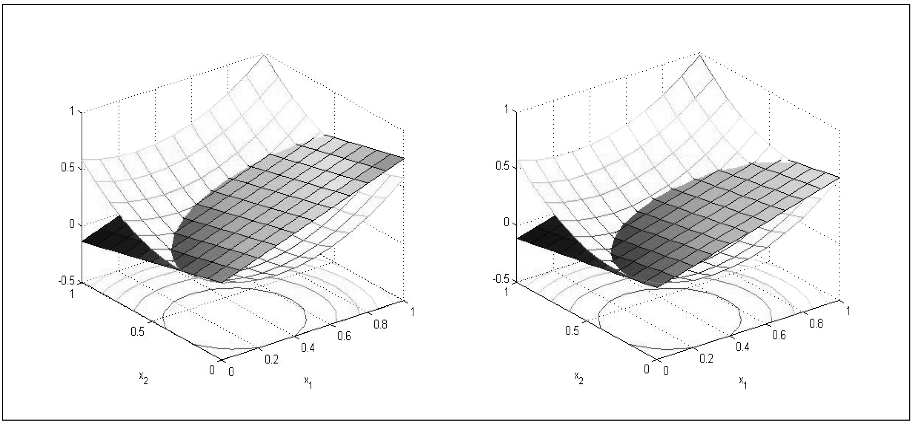


Figure 2.5. *At the left, the linear interpolation model of* $f(x_1,x_2) = (x_1 - 0.3)^2 + (x_2 - 0.3)^2$ *at* $y^0 = (0,0)$*,* $y^1 = (0.2,0.6)$*, and* $y^2 = (0.8,0.7)$*. At the right, the linear regression model of* $f(x_1,x_2) = (x_1 - 0.3)^2 + (x_2 - 0.3)^2$ *at* $y^0 = (0,0)$*,* $y^1 = (0.2,0.6)$*,* $y^2 = (0.8,0.7)$*, and* $y^3 = (0.5,0.5)$*. The inclusion of* y^3 *pushes the model down.*

one option is to use linear regression and to compute the coefficients of the linear (least-squares) regression polynomial $m(x) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n$ as the least-squares solution of the system

$$\begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 \\ 1 & y_1^1 & \cdots & y_n^1 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} f(y^0) \\ f(y^1) \\ \vdots \\ \vdots \\ f(y^p) \end{bmatrix}. \tag{2.6}$$

Again, we denote the matrix of this (possibly overdetermined) system of linear equations by $M = M(\bar{\phi}, Y)$.

The definition of poisedness generalizes easily from linear interpolation to linear regression.

Definition 2.10. *The set* $Y = \{y^0, y^1, \ldots, y^p\}$*, with* $p > n$*, is poised for linear regression in* $\mathbb{R}^n$ *if the corresponding matrix* $M(\bar{\phi}, Y)$ *has full (column) rank.*

It is also possible to prove, in the regression case, that if a set Y is poised for a basis ϕ, then it is also poised for any other basis in $\mathcal{P}_n^1$ and that $m(x)$ is independent of the basis chosen. Finally, we point out that the sample set is poised for linear regression if and only if the linear regression polynomial $m(x)$ is uniquely defined. These issues are covered in detail in Chapter 4.

2.4 Error bounds for linear interpolation and regression

Error bounds for linear interpolation

Let us rewrite the linear interpolating polynomial in the form

$$m(y) = c + g^\top y$$

by setting $\alpha_0 = c$ and $\alpha_i = g_i$, $i = 1,\ldots,n$.

One of the natural questions that arises in interpolation is how to measure the quality of $m(y)$ as an approximation to $f(y)$. We start by looking at the quality of the gradient $g = \nabla m(y)$ of the model as an approximation to $\nabla f(y)$. We consider that the interpolation points $y^0, y^1, \ldots, y^n$ are in a ball of radius Δ centered at y^0. In practice, one might set

$$\Delta = \Delta(Y) = \max_{1\le i\le n} \|y^i - y^0\|.$$

We are interested in the quality of $\nabla m(y)$ and $m(y)$ in the ball of radius Δ centered at y^0. The assumptions needed for this result are summarized below.

Assumption 2.1. *We assume that $Y = \{y^0, y^1, \ldots, y^n\} \subset \mathbb{R}^n$ is a poised set of sample points (in the linear interpolation sense) contained in the ball $B(y^0;\Delta(Y))$ of radius $\Delta = \Delta(Y)$.*

Further, we assume that the function f is continuously differentiable in an open domain Ω containing $B(y^0;\Delta)$ and ∇f is Lipschitz continuous in Ω with constant $\nu > 0$.

As we can observe from the proof of Theorem 2.11 below, the derivation of the error bounds is based on the application of one step of Gaussian elimination to the matrix $M = M(\bar{\phi}, Y)$ in (2.5). After performing such a step we arrive at the matrix

$$\begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 \\ 0 & y_1^1 - y_1^0 & \cdots & y_n^1 - y_n^0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & y_1^n - y_1^0 & \cdots & y_n^n - y_n^0 \end{bmatrix},$$

which can be expressed by blocks as

$$\begin{bmatrix} 1 & y_0^\top \\ 0 & L \end{bmatrix},$$

with

$$L = \left[y^1 - y^0 \cdots y^n - y^0\right]^\top = \begin{bmatrix} (y^1 - y^0)^\top \\ \vdots \\ (y^n - y^0)^\top \end{bmatrix} = \begin{bmatrix} y_1^1 - y_1^0 & \cdots & y_n^1 - y_n^0 \\ \vdots & \vdots & \vdots \\ y_1^n - y_1^0 & \cdots & y_n^n - y_n^0 \end{bmatrix}.$$

It is evident to see that L is nonsingular if and only if M is nonsingular, since $\det(L) = \det(M)$. Notice that the points appear listed in L by rows, which favors factorizations by rows.

It turns out that the error bounds for the approximation which we derive are in terms of the scaled matrix

$$\hat{L} = \frac{1}{\Delta} L = \frac{1}{\Delta}\left[y^1 - y^0 \cdots y^n - y^0\right]^\top = \begin{bmatrix} \frac{y_1^1 - y_1^0}{\Delta} & \cdots & \frac{y_n^1 - y_n^0}{\Delta} \\ \vdots & \vdots & \vdots \\ \frac{y_1^n - y_1^0}{\Delta} & \cdots & \frac{y_n^n - y_n^0}{\Delta} \end{bmatrix}.$$

This matrix $\hat{L}$ corresponds to a scaled sample set

$$\hat{Y} = \{y^0/\Delta, y^1/\Delta, \ldots, y^n/\Delta\} \subset B(y^0/\Delta; 1),$$

which is contained in a ball of radius 1 centered at y^0/Δ.

Theorem 2.11. *Let Assumption* 2.1 *hold. The gradient of the linear interpolation model satisfies, for all points y in $B(y^0;\Delta)$, an error bound of the form*

$$\|\nabla f(y) - \nabla m(y)\| \le \kappa_{eg}\,\Delta, \tag{2.7}$$

where $\kappa_{eg} = \nu(1 + n^{\frac{1}{2}}\|\hat{L}^{-1}\|/2)$ and $\hat{L} = L/\Delta$.

Proof. If the set Y is poised, then the $(n+1)\times(n+1)$ matrix $M = M(\bar{\phi}, Y)$ is nonsingular and so is the $n \times n$ matrix L.

We look initially at the gradient of f at the point y^0. Subtracting the first interpolating condition from the remaining n, we obtain

$$(y^i - y^0)^\top g = f(y^i) - f(y^0), \quad i = 1, \ldots, n.$$

Then, if we use the integral form of the mean value theorem

$$f(y^i) - f(y^0) = \int_0^1 (y^i - y^0)^\top \nabla f(y^0 + t(y^i - y^0))dt,$$

we obtain, from the Lipschitz continuity of ∇f, that

$$(y^i - y^0)^\top(\nabla f(y^0) - g) \le \frac{\nu}{2}\|y^i - y^0\|^2, \quad i = 1, \ldots, n.$$

Then, from these last n inequalities, we derive

$$\|L(\nabla f(y^0) - g)\| \le (n^{\frac{1}{2}}\nu/2)\,\Delta^2,$$

from which we conclude that

$$\|\nabla f(y^0) - g\| \le (n^{\frac{1}{2}}\|\hat{L}^{-1}\|\nu/2)\,\Delta.$$

The error bound for any point y in the ball $B(y^0;\Delta)$ is easily derived from the Lipschitz continuity of the gradient of f:

$$\|\nabla f(y) - g\| \le \|\nabla f(y) - \nabla f(y^0)\| + \|\nabla f(y^0) - g\| \le (\nu + n^{\frac{1}{2}}\|\hat{L}^{-1}\|\nu/2)\,\Delta. \quad \square$$

Assuming a uniform bound on $\|\hat{L}^{-1}\|$ independent of Δ, the error in the gradient is linear in Δ. One can see also that the error in the interpolation model $m(x)$ itself is quadratic in Δ.

Theorem 2.12. *Let Assumption* 2.1 *hold. The interpolation model satisfies, for all points y in $B(y^0;\Delta)$, an error bound of the form*

$$|f(y)-m(y)| \leq \kappa_{ef}\,\Delta^2, \tag{2.8}$$

where $\kappa_{ef}=\kappa_{eg}+\nu/2$ and κ_{eg} is given in Theorem 2.11.

Proof. We use the same arguments as in the proof of Theorem 2.11 to obtain

$$f(y)-f(y^0) \leq \nabla f(y^0)^\top(y-y^0)+\frac{\nu}{2}\|y-y^0\|^2.$$

From this we get

$$f(y)-f(y^0)-g^\top(y-y^0) \leq (\nabla f(y^0)-g)^\top(y-y^0)+\frac{\nu}{2}\|y-y^0\|^2.$$

The error bound (2.8) comes from combining this inequality with (2.7) and from noting that the constant term in the model can be written as $c=f(y^0)-g^\top y^0$. $\square$

Error bounds for linear regression

In the regression case we are considering a sample set $Y=\{y^0,y^1,\ldots,y^p\}$ with more than $n+1$ points contained in the ball $B(y^0;\Delta(Y))$ of radius

$$\Delta = \Delta(Y) = \max_{1\le i\le p}\|y^i-y^0\|.$$

We start by also rewriting the linear regression polynomial in the form

$$m(y) = c+g^\top y,$$

where $c=\alpha_0$ and $g_i=\alpha_i$, $i=1,\ldots,n$, are the components of the least-squares solution of the system (2.6).

Assumption 2.2. *We assume that $Y=\{y^0,y^1,\ldots,y^p\}\subset\mathbb{R}^n$, with $p>n$, is a poised set of sample points (in the linear regression sense) contained in the ball $B(y^0;\Delta(Y))$ of radius $\Delta=\Delta(Y)$.*

Further, we assume that the function f is continuously differentiable in an open domain Ω containing $B(y^0;\Delta)$ and ∇f is Lipschitz continuous in Ω with constant $\nu>0$.

The error bounds for the approximation are also derived in terms of the scaled matrix

$$\hat{L} = \frac{1}{\Delta}L = \frac{1}{\Delta}\left[y^1-y^0\cdots y^p-y^0\right]^\top,$$

which corresponds to a scaled sample set contained in a ball of radius 1 centered at y^0/Δ, i.e.,

$$\hat{Y} = \{y^0/\Delta, y^1/\Delta, \ldots, y^p/\Delta\} \subset B(y^0/\Delta; 1).$$

The proof of the bounds follows exactly the same steps as the proof for the linear interpolation case. For example, for the gradient approximation, once we are at the point in the proof where

$$\|L(\nabla f(y^0) - g)\| \leq (p^{\frac{1}{2}} \nu/2)\, \Delta^2,$$

or, equivalently,

$$\|\hat{L}(\nabla f(y^0) - g)\| \leq (p^{\frac{1}{2}} \nu/2)\, \Delta, \tag{2.9}$$

we "pass" $\hat{L}$ to the right-hand side by means of its left inverse $\hat{L}^\dagger$.[3] We can then state the bounds in the following format.

Theorem 2.13. *Let Assumption* 2.2 *hold. The gradient of the linear regression model satisfies, for all points y in $B(y^0;\Delta)$, an error bound of the form*

$$\|\nabla f(y) - \nabla m(y)\| \leq \kappa_{eg}\, \Delta,$$

where $\kappa_{eg} = \nu(1 + p^{\frac{1}{2}} \|\hat{L}^\dagger\|/2)$ and $\hat{L} = L/\Delta$.

The linear regression model satisfies, for all points y in $B(y^0;\Delta)$, an error bound of the form

$$|f(y) - m(y)| \leq \kappa_{ef}\, \Delta^2,$$

where $\kappa_{ef} = \kappa_{eg} + \nu/2$.

2.5 Other geometrical concepts

The notion of poisedness for linear interpolation is closely related to the concept of affine independence in convex analysis.

Affine independence

We will follow Rockafellar [200] to introduce affine independence as well as other basic concepts borrowed from convex analysis.

The affine hull of a given set $S \subset \mathbb{R}^n$ is the smallest affine set containing S (meaning that it is the intersection of all affine sets containing S). The affine hull of a set is always uniquely defined and consists of all linear combinations of elements of S whose scalars sum up to one (see [200]).

Definition 2.14. *A set of $m+1$ points $Y = \{y^0, y^1, \ldots, y^m\}$ is said to be affinely independent if its affine hull* $\mathrm{aff}(y^0, y^1, \ldots, y^m)$ *has dimension m.*

The dimension of an affine set is the dimension of the linear subspace parallel to it. So, we cannot have an affinely independent set in $\mathbb{R}^n$ with more than $n+1$ points.

[3] $A^\dagger$ denotes the Moore–Penrose generalized inverse of a matrix A, which can be expressed by the singular value decomposition of A for any real or complex matrix A. In the current context, where L has full column rank, we have $\hat{L}^\dagger = (\hat{L}^\top \hat{L})^{-1} \hat{L}^\top$.

Given an affinely independent set of points $\{y^0, y^1, \ldots, y^m\}$, we have that

$$\mathrm{aff}(y^0, y^1, \ldots, y^m) \;=\; y^0 + \mathcal{L}(y^1 - y^0, \ldots, y^m - y^0),$$

where $\mathcal{L}(y^1 - y^0, \ldots, y^m - y^0)$ is the linear subspace of dimension m generated by the vectors $y^1 - y^0, \ldots, y^m - y^0$.

We associate with an affinely independent set of points $\{y^0, y^1, \ldots, y^m\}$ the matrix

$$\left[y^1 - y^0 \cdots y^m - y^0 \right] \in \mathbb{R}^{n \times m},$$

whose rank must be equal to m. This matrix is exactly the transpose of the matrix L that appeared when we linearly interpolated a function f on the set $Y = \{y^0, y^1, \ldots, y^n\}$, with $m = n$, choosing an appropriate basis for the space $\mathcal{P}_n^1$ of linear polynomials of degree 1 in $\mathbb{R}^n$.

Simplices

Similarly, the convex hull of a given set $S \subset \mathbb{R}^n$ is the smallest convex set containing S (meaning that it is the intersection of all convex sets containing S). The convex hull of a set is always uniquely defined and consists of all convex combinations of elements of S, i.e., of all linear combinations of elements of S whose scalars are nonnegative and sum up to one (see [200]).

Definition 2.15. *Given an affinely independent set of points $Y = \{y^0, y^1, \ldots, y^m\}$, its convex hull is called a simplex of dimension m.*

A simplex of dimension 0 is a point, of dimension 1 is a closed line segment, of dimension 2 is a triangle, and of dimension 3 is a tetrahedron.

The vertices of a simplex are the elements of Y. A simplex in $\mathbb{R}^n$ cannot have more than $n+1$ vertices. When there are $n+1$ vertices, its dimension is n. In this case,

$$\left[\; y^1 - y^0 \;\cdots\; y^n - y^0 \;\; -(y^1 - y^0) \;\cdots\; -(y^n - y^0) \;\right] \tag{2.10}$$

forms a (maximal) positive basis in $\mathbb{R}^n$. We illustrate this maximal positive basis in Figure 2.6.

The diameter of a simplex Y of vertices $y^0, y^1, \ldots, y^m$ is defined by

$$\mathrm{diam}(Y) = \max_{0 \le i < j \le m} \|y^i - y^j\|.$$

One way of approximating $\mathrm{diam}(Y)$ at y^0 is by computing the less expensive quantity

$$\Delta(Y) = \max_{1 \le i \le n} \|y^i - y^0\|.$$

Clearly, we can write $\Delta(Y) \le \mathrm{diam}(Y) \le 2\Delta(Y)$.

By the shape of a simplex we mean its equivalent class under similarity: the simplices of vertices Y and λY, $\lambda > 0$, share the same shape. The volume of a simplex of $n+1$ vertices $Y = \{y^0, y^1, \ldots, y^n\}$ is defined by

$$\mathrm{vol}(Y) \;=\; \frac{|\det(L)|}{n!},$$

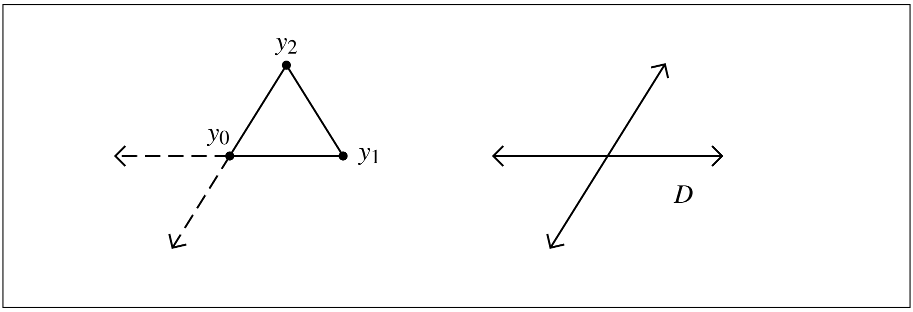


Figure 2.6. *How to compute a maximal positive basis from the vertices of a simplex.*

where

$$L = L(Y) = \left[y^1 - y^0 \cdots y^n - y^0 \right]^\top.$$

Since the vertices of a simplex form an affinely independent set, one clearly has that $\mathrm{vol}(Y) > 0$. It is also left as an exercise to see that the choice of the centering point in L is irrelevant for the definition of the volume of a simplex.

The volume of a simplex is not a good measure of the quality of its geometry since it is not scaling independent. To see this let

$$Y_t = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} t \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ t \end{bmatrix} \right\},$$

with $t > 0$. It is easy to see that $\mathrm{vol}(Y_t) \to 0$ when $t \to 0$. However, the angles between the vectors formed by the vertices are the same for all positive values of t (or putting it differently all these simplices have the same shape).

A measure of the quality of a simplex geometry must be independent of the scale of the simplex, given by either $\Delta(Y)$ or $\mathrm{diam}(Y)$. One such measure is given by

$$\| [L(Y)/\Delta(Y)]^\dagger \|,$$

which reduces to

$$\| [L(Y)/\Delta(Y)]^{-1} \|$$

for simplices of $n+1$ vertices. One alternative when there are $n+1$ vertices is to work with a normalized volume

$$\mathrm{von}(Y) = \mathrm{vol}\left(\frac{1}{\mathrm{diam}(Y)} Y \right) = \frac{|\det(L(Y))|}{n!\,\mathrm{diam}(Y)^n}.$$

Poisedness and positive spanning

Another natural question that arises is the relationship between the concepts of poised interpolation and regression sets and positive spanning sets or positive bases. To study this relationship, let us assume that we have a positive spanning set D formed by nonzero vectors. Recall that consequently the cosine measure $\mathrm{cm}(D)$ is necessarily positive. For the sake of simplicity in this discussion let us assume that the elements in D have norm one. Then

$$\mathrm{cm}(D) = \min_{\|v\|=1} \max_{d \in D} v^\top d \leq \min_{\|v\|=1} \max_{d \in D} |v^\top d| \leq \min_{\|v\|=1} \|D^\top v\|. \tag{2.11}$$

We immediately conclude that D has full row rank.

Thus, given a point y^0 and a positive spanning set D for which $\mathrm{cm}(D) > 0$, we know that the sample set $\{y^0\} \cup \{y^0 + d : d \in D\}$ is poised for linear regression. See Figure 2.7.

The contrary, however, is not true. Given a poised set $Y = \{y^0, y^1, \ldots, y^p\}$ for linear regression, the set of directions $\{y^1 - y^0, \ldots, y^p - y^0\}$ might not be a positive spanning set. It is trivial to construct a counterexample. For instance, let us take $n = 2$, $p = 3$, $y^0 = (0,0)$, and

$$\left[\begin{array}{ccc} y^1 & y^2 & y^3 \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & -1 \\ 0 & 1 & 0 \end{array}\right].$$

See also Figure 2.7.

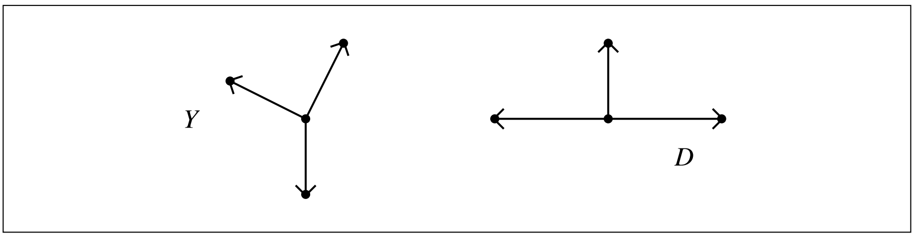


Figure 2.7. *For the positive spanning set on the left, the set Y marked by the bullets is poised for linear regression. However, given the poised set on the right, we see that the set of directions D marked by the arrows is not a positive spanning set.*

From the Courant–Fischer-type inequalities for singular values (see [131]), we conclude from (2.11) that

$$\mathrm{cm}(D) \leq \min_{\|v\|=1} \|D^\top v\| = \sigma_{min}(D^\top),$$

where $\sigma_{min}(D^\top)$ represents the smallest singular value of $D^\top$. Hence,

$$\|(D^\top)^\dagger\| = \frac{1}{\sigma_{min}(D^\top)} \leq \frac{1}{\mathrm{cm}(D)},$$

which shows that if the cosine measure of D is sufficiently away from zero, then the set $\{y^0\} \cup \{y^0 + d : d \in D\}$ is sufficiently "well poised."

2.6 Simplex gradients

Given a set $Y = \{y^0, y^1, \ldots, y^n\}$ with $n+1$ sample points and poised for linear interpolation, the simplex gradient at y^0 is defined in the optimization literature (see Kelley [141]) by

$$\nabla_s f(y^0) = L^{-1} \delta f(Y),$$

where

$$L = \left[y^1 - y^0 \cdots y^n - y^0 \right]^\top$$

and

$$\delta f(Y) = \begin{bmatrix} f(y^1) - f(y^0) \\ \vdots \\ f(y^n) - f(y^0) \end{bmatrix}.$$

However, it requires little effort to see that the simplex gradient is nothing else than the gradient of the linear interpolation model $m(x) = c + g^\top x$:

$$\nabla_s f(y^0) = g.$$

When the number of sample points exceeds $n+1$, simplex gradients are defined in a regression sense as the least-squares solution of $L \nabla_s f(y^0) = \delta f(Y)$, where

$$L = \left[y^1 - y^0 \cdots \cdots y^p - y^0 \right]^\top$$

and

$$\delta f(Y) = \begin{bmatrix} f(y^1) - f(y^0) \\ \vdots \\ \vdots \\ f(y^p) - f(y^0) \end{bmatrix}.$$

Again, one points out that a simplex gradient defined in this way is the gradient g of the linear regression model $m(x) = c + g^\top x$. We note that simplex gradients when $p > n$ are also referred to as stencil gradients (see the papers by C. T. Kelley). The set $\{y^1, \ldots, y^p\}$ is a stencil centered at y^0. For instance, the stencil could take the form $\{y^0 \pm h e_i,\ i = 1, \ldots, n\}$, where h is the stencil radius, and e_i, $i = 1, \ldots, n$, are the coordinate vectors.

It is then obvious that under the assumptions stated for linear interpolation and linear regression the simplex gradient satisfies an error bound of the form

$$\|\nabla f(y^0) - \nabla_s f(y^0)\| \leq \kappa_{eg} \, \Delta,$$

where $\kappa_{eg} = p^{\frac{1}{2}} \nu \|\hat{L}^\dagger\|/2$ and $\hat{L} = L/\Delta$. In the case $p = n$, one has $\|\hat{L}^\dagger\| = \|\hat{L}^{-1}\|$.

2.7 Exercises

1. Prove that a set of nonzero vectors forms a positive basis for $\mathbb{R}^n$ if and only if their positive combinations span $\mathbb{R}^n$ and no proper subset exhibits the same property.

2. Show that $[I \ -I]$ is a (maximal) positive basis for $\mathbb{R}^n$ with cosine measure $1/\sqrt{n}$.

3. Prove Corollary 2.5.

4. Show that the value of a in (2.1) must be equal to $-1/n$.

5. Show that the cosine measure of a positive spanning set is necessarily positive.

6. In addition to the previous exercise, prove the reverse implication; that is, if the cosine measure of a set is zero, then the set cannot span $\mathbb{R}^n$ positively.

7. Show that the cosine measure of a minimal positive basis with uniform angles in $\mathbb{R}^n$ is $1/n$.

8. Prove, for linear interpolation and regression, that if Y is poised for some basis, then it is poised for any other basis in $\mathcal{P}_n^1$. Show that the definition of $m(x)$ in linear interpolation and regression is also independent of the basis chosen.

9. Show that (2.10) forms a (maximal) positive basis.

10. From (2.9), conclude the proof of Theorem 2.13.

11. Show that $\Delta(Y) \leq \mathrm{diam}(Y) \leq 2\Delta(Y)$ and $\mathrm{vol}(Y) > 0$ for any simplex of vertices Y.