Linear Algebra for Optimization

Robert Carlson
Department of Mathematics
University of Colorado at Colorado Springs
rcarlson@uccs.edu

0.1 Some Linear Algebra

0.1.1 Introduction

We have seen that the Hessian of a function

$$Hf = \left(\frac{\partial^2 f}{\partial x_m \partial x_n} \right)$$

is an important tool for recognizing when critical points of a function are local or global minimizers. The key property we wanted was positivity of the quadratic form associated to the Hessian,

$$X \bullet Hf(Z)X > 0, \quad X = [x_1, \dots, x_N] \neq [0, \dots, 0].$$

When the second partial derivatives are continuous, the mixed partial derivatives are equal,

$$\frac{\partial^2 f}{\partial x_m \partial x_n} = \frac{\partial^2 f}{\partial x_n \partial x_m},$$

and so the Hessian is a real symmetric matrix.

The positivity of the quadratic form of a real symmetric matrix $\mathbf{A}$ is trivial to verify if the matrix is diagonal. That is, if

$$X \bullet \mathbf{A}X = \sum_{n=1}^{N} a_n x_n^2,$$

then $Q_{\mathbf{A}} > 0$ if and only if each $a_n > 0$.

To generalize this observation, it helps to consider a linear change of variables that diagonalizes the quadratic form of a matrix. Here is an example. Let

$$f([x_1, x_2]) = 3x_1^2 - 2x_1 x_2 + 3x_2^2 = (x_1 + x_2)^2 + 2(x_1 - x_2)^2.$$

The Hessian is

$$Hf = \begin{pmatrix} 6 & -2 \\ -2 & 6 \end{pmatrix},$$

and

$$f = \frac{1}{2} X \bullet Hf X.$$

Make the linear change of variables

$$y_1 = x_1 + x_2, \quad y_2 = x_1 - x_2.$$

In terms of the new variables,

$$\frac{1}{2}X \bullet HfX = y_1^2 + 2y_2^2.$$

In this case the fact that Hf was positive definite was easy to spot. But if $\mathbf{A}$ is a larger matrix, some effective technique is required. To start moving in the right direction, it helps to notices that the matrix

$$\mathbf{A} = (1/2)Hf = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix},$$

has some special vectors called eigenvectors, satisfying $\mathbf{A}X = \lambda X$, with $\lambda \in \mathbb{R}$ and $X \neq 0$. In this example

$$\mathbf{A}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_1 = 2$$

and

$$\mathbf{A}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 4 \\ -4 \end{pmatrix} = 4\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \lambda_2 = 4.$$

Let

$$X_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Notice that

$$X_1 \bullet X_2 = 0,$$

and if

$$\mathbf{P} = (X_1, X_2) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

then

$$\mathbf{AP} = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 2 & -4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

In this case

$$\mathbf{P}^{-1} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix},$$

and

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

There is also some additional structure which appears if we normalize the eigenvectors. Dividing by $\sqrt{2}$ we would find

$$\mathbf{P} = (X_1/\|X_1\|, X_2/\|X_2\|) = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix},$$

with orthonormal columns. In this case $\mathbf{P^{-1}} = \mathbf{P}$, although this will not be typical. The normalization does not change the equation

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

0.1.2 The spectral theorem

In the rest of this section we aim to understand when such a change of variables is available (always for real symmetric matrices), how it is related to eigenvalues, and how these ideas can be used for optimization. Real symmetric matrices have very special structural features not shared by general $N \times N$ matrices. Many of these special features are consequences of a simple dot product identity.

Let's start with arbitrary $N \times N$ matrices with complex entries. As mentioned above, and eigenvalue for $\mathbf{A}$ is a number λ (possibly complex) such that

$$\mathbf{A}X = \lambda X$$

for some vector $X \neq 0$. The vector X is called an eigenvector. If X is an eigenvector with eigenvalue λ, so is cX for every $c \neq 0$. When we want to talk about all vectors X satisfying $\mathbf{A}X = \lambda X$ for some fixed λ, it is helpful to include $X = 0$. Then we can freely add the eigenvectors X, multiply by constants, and find that the *eigenspace* with eigenvalue λ is a vector subspace of $\mathbb{R}^N$.

Rewrite the eigenvalue equation as

$$[\lambda I - \mathbf{A}]X = 0.$$

Such a number λ and vector $X \neq 0$ exist if and only if there is a λ such that the matrix $\lambda I - \mathbf{A}$ is not invertible. The test for noninvertibiity is

$$p(\lambda) = \det(\lambda I - \mathbf{A}) = 0.$$

The polynomial $p(\lambda)$ has degree N and leading coefficient 1. Every nonconstant polynomial has at least one complex root, so

Theorem 0.1.1. *Every $N \times N$ matrix of (real or complex) numbers has at least one (possibly complex) eigenvalue λ, and a corresponding nonzero eigenvector X.*

Since eigenvalues and eigenvectors can be complex, it helps to extend the dot product to complex vectors, taking

$$X \bullet Y = \sum_{n=1}^{N} x_n \overline{y_n}.$$

Notice that

$$Y \bullet X = \overline{X \bullet Y},$$

and if $c \in \mathbb{C}$ then

$$X \bullet cY = \overline{c} X \bullet Y.$$

Real symmetric matrices satisfy a simple but very important identity.

Lemma 0.1.2. *Suppose* $\mathbf{A} = (a_{mn})$ *is an* $N \times N$ *real symmetric matrix. Then for any* $X, Y \in \mathbb{C}^N$,

$$\mathbf{A}X \bullet Y = X \bullet \mathbf{A}Y.$$

Proof. With the help of the symmetry $a_{mn} = a_{nm}$, and the fact that a_{mn} is real, so $a_{mn} = \overline{a_{mn}}$, a computation gives

$$\mathbf{A}X \bullet Y = \sum_{m=1}^{N} \overline{y_m} (\sum_{n=1}^{N} a_{mn} x_n) = \sum_{m=1}^{N} \sum_{n=1}^{N} \overline{y_m} a_{mn} x_n = \sum_{n=1}^{N} \sum_{m=1}^{N} \overline{y_m} a_{mn} x_n$$

$$= \sum_{n=1}^{N} x_n \sum_{m=1}^{N} \overline{y_m a_{mn}} = \sum_{n=1}^{N} x_n \sum_{m=1}^{N} \overline{y_m a_{nm}} = X \bullet \mathbf{A}Y.$$

□

Theorem 0.1.3. *Suppose* $\mathbf{A}$ *is an* $N \times N$ *real symmetric matrix. Every eigenvalue* λ *of* $\mathbf{A}$ *is real. If* λ_1, λ_2 *are distinct eigenvalues with eigenvectors* X_1, X_2 *respectively, then*

$$X_1 \bullet X_2 = 0.$$

Proof. Suppose $\mathbf{A}X = \lambda X$, $X \neq 0$. Then

$$\lambda X \bullet X = \mathbf{A}X \bullet X = X \bullet \mathbf{A}X = X \bullet \lambda X = \overline{\lambda} X \bullet X.$$

Since $X \bullet X > 0$, we conclude that $\lambda = \overline{\lambda}$, so $\lambda \in \mathbb{R}$.

For the second part, using that $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$(\lambda_1 - \lambda_2) X_1 \bullet X_2 = AX_1 \bullet X_2 - X_1 \bullet AX_2 = AX_1 \bullet X_2 - AX_1 \bullet X_2 = 0.$$

Since $\lambda_1 \neq \lambda_2$,

$$X_1 \bullet X_2 = 0.$$

□

With a bit more work we can also show that

Theorem 0.1.4. *Suppose* $\mathbf{A}$ *is an* $N \times N$ *real symmetric matrix. Then there are* N *orthonormal eigenvectors* $X_1, \ldots, X_N$ *for* $\mathbf{A}$*. The vectors* $X_1, \ldots, X_N$ *are a basis for* $\mathbb{R}^N$*.*

The orthonormal basis of eigenvectors can be used to diagonalize $\mathbf{A}$. Let $\mathbf{P}$ be the $N \times N$ matrix whose columns are the vectors $X_1, \ldots, X_N$. Then

$$\mathbf{AP} = \mathbf{A}(X_1, \ldots, X_N) = (\lambda_1 X_1, \ldots, \lambda_N X_N) = (X_1, \ldots, X_N)\mathrm{diag}(\lambda_1, \ldots, \lambda_N),$$

where $\mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix with λ_n as the n-th diagonal entry.

Since the columns of $\mathbf{P}$ are orthonormal, the inverse of $\mathbf{P}$ is the matrix whose rows are $X_1, \ldots, X_N$. This is the transpose of $\mathbf{P}$, so $\mathbf{P}^{-1} = \mathbf{P}^T$. Real $N \times N$ matrices such that $\mathbf{P}^{-1} = \mathbf{P}^T$ are said to be orthogonal. (Unitary in the complex case.)

Theorem 0.1.5. *Suppose* $\mathbf{U}$ *is a real* $N \times N$ *matrix. The following are equivalent.*
(i) $\mathbf{U}^T = \mathbf{U}^{-1}$.
(ii) The columns of $\mathbf{U}$ *are orthonormal.*
(ii) The rows of $\mathbf{U}$ *are orthonormal.*

Proof. If $\mathbf{U}^T = \mathbf{U}^{-1}$, then $\mathbf{U}^T\mathbf{U} = I = \mathbf{U}\mathbf{U}^T$. The equation $\mathbf{U}^T\mathbf{U} = I$ shows that the columns U_n of $\mathbf{U}$ satisfy $U_m \bullet U_n = \delta_{mn}$. The equation $\mathbf{U}\mathbf{U}^T = I$ shows the same for the rows. Thus (i) implies (ii) and (iii).

If (ii) holds, then $\mathbf{U}^T\mathbf{U} = I$. If (iii) holds, then $\mathbf{U}\mathbf{U}^T = I$.

□

Notice that the product of two orthogonal matrices is orthogonal and the identity is orthogonal. That is

$$(\mathbf{U}_1\mathbf{U}_2)^T = \mathbf{U}_2^T\mathbf{U_1}^T = \mathbf{U}_2^{-1}\mathbf{U_1}^{-1} = (\mathbf{U}_1\mathbf{U}_2)^{-1}.$$

The $N \times N$ orthogonal matrices thus form a group under matrix multiplication.

By expanding in the basis of eigenvectors we obtain the following result.

Theorem 0.1.6. *Suppose* $\mathbf{A}$ *is an* $N \times N$ *real symmetric matrix.*

(a) $\mathbf{A}$ *is positive (negative) definite if and only if all eigenvalues are strictly positive (negative).*

(b) $\mathbf{A}$ *is positive (negative) semidefinite if and only if all eigenvalues are* ≥ 0 *(*≤ 0*).*

Proof. Recall that by definition $\mathbf{A}$ is positive definite if

$$Q_{\mathbf{A}}(Y) = Y \bullet \mathbf{A}Y > 0, \quad Y \neq 0.$$

Expand Y in the orthonormal basis of eigenvectors of $\mathbf{A}$,

$$Y = c_1 X_1 + \cdots + c_N X_N.$$

If $Y \neq 0$ then at least one value c_n is not zero. Use the orthonormality to compute

$$Y \bullet \mathbf{A}Y = [\sum_{n=1}^{N} c_n X_n] \bullet \mathbf{A}[\sum_{n=1}^{N} c_n X_n] = [\sum_{n=1}^{N} c_n X_n] \bullet [\sum_{n=1}^{N} \lambda_n c_n X_n] = \sum_{n=1}^{N} \lambda_n c_n^2.$$

If all λ_n are strictly positive (resp. ≥ 0) then $Q_{\mathbf{A}}(Y) > 0$ (resp. ≥ 0). If some λ_n is less than 0, then $Q_{\mathbf{A}}(X_n) < 0$.

$\square$

0.1.3 An Optimization Problem for Real Symmetric Matrices

Some optimization problems are directly related to the structure of positive semidefinite real symmetric matrices. The most common of these problems is to find a vector X of norm 1 which maximizes the quadratic form

$$X \bullet \mathbf{A}X.$$

It is easy to see that this vector must be an eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$. Before solving this problem, we mention an example where it arises.

The first application area is called principal component analysis in statistics. You are given a collection of K vectors, $X(k) = (x_1(k), \ldots, x_N(k))$ which might be thought of as an ellipsoidal cloud of data. You may believe that there are far fewer than N significant measurements, so you wish to find a single vector (or collection of J vectors) whose multiples (linear combinations) best model the data. The principal component analysis goes as follows. Form the matrix

$$\Sigma = (\sigma_{ij}) = (E([x_i - \mu_i][x_j - \mu_j])$$

where μ_j is the sample mean of the measurements of the $j-th$ component,

$$\mu_j = \sum_{k=1}^{K} x_j(k)/K$$

and the sample covariances are

$$E([x_i - \mu_i][x_j - \mu_j]) = \sum_{k=1}^{K} [x_i(k) - \mu_i][x_j(k) - \mu_j]/K.$$

To get a feel for what is happening, suppose the μ_j are all 0, and most of the variation of the data is in the first coordinate. Then $E([x_i - \mu_i][x_j - \mu_j]) \simeq 0$ unless $i = j = 1$. If $i = j = 1$ we get a large value. The covariance matrix looks like

$$\Sigma = \begin{pmatrix} v & \epsilon \\ \epsilon & \epsilon \end{pmatrix},$$

where v is large and ϵ respresent small entries. We expect the eigenvalues to be approximately $v, 0, \ldots, 0$, where v is the variance in the first component. The eigenvector should be close to the first standard basis vector $[1, 0, \ldots, 0]$.

Going back to the general case, the matrix Σ is real, symmetric, and positive semidefinite. The desired vector(s) is the normalized eigenvector for the largest eigenvalue (assuming the eigenspace is one dimensional). (The eigenvectors corresponding to the J largest eigenvalues, counting multiplicity are used when J vectors are desired.)