Linear Algebra and Optimization Primer for Stats305a

John Duchi

January 9, 2025

Contents

1 Introduction

In this document, whose length grew longer than I planned as I wrote it, I review a few of the (reasonably) basic concepts in linear algebra that are important for work in applied statistics, machine learning, and optimization. This note is by no means comprehensive—there are entire books on the subject!—but it can hopefully provide a handy guide. For much more information, with great presentation, I recommend Horn and Johnson's *Matrix Analysis* [4]. In addition, I will touch on the basic concepts from optimization we will use; note that I will not treat these in any depth, and as these notes might be charitably called "quick and dirty," I provide no illustrations—the readers will have to give geometric pictures themselves.

2 Vectors, matrices, and basic decompositions

Much of the work one does with vectors and matrices involves solving (or approximately solving) the matrix equation

$$Ax = b$$

for a given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^n$. Key to solving this in practice—and to understand properties of solutions, including stability to, e.g., the noise from statistical processes—is to develop various decompositions of the matrix A to write it in simpler forms, and to understand how it acts as an operator on matrices. This is precisely what we do over the next several sections.

2.1 Vectors, independence, and norms

A vector $x \in \mathbb{R}^n$ is a collection of n real numbers; we *always* treat vectors as columns, as it is a crime against mathematics to do otherwise. We write

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

The *Euclidean norm* of a vector x, often termed its *length*, is

$$\|x\| = \sqrt{\sum_{i=1}^n x_i^2}.$$

In some contexts, one uses p-norms for the size of the vector x, that is, $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, in which case the Euclidean norm corresponds to the ℓ_2 norm $\|x\|_2 = \|x\|$, though we typically omit the subscript when clear from context. Throughout this note, we will not use arrows or bold or particular subscripts to distinguish vectors; the hope is that what a particular mathematical symbol or element is is clear from context.

Given k vectors $x_1, \ldots, x_k \in \mathbb{R}^n$, the vectors are *linearly independent*, or simply *independent*, if for all $\alpha \in \mathbb{R}^k$, we have

$$\sum_{i=1}^k \alpha_i x_i = 0 \quad \text{if and only if} \quad \alpha = 0.$$

That is, no vector can be written as a linear combination of the others. An evidently equivalent formulation to this is that for $\alpha, \beta \in \mathbb{R}^k$, we have $\sum_i \alpha_i x_i = \sum_i \beta_i x_i$ if and only if $\alpha_i = \beta_i$, as this is equivalent to $\sum_{i=1}^k (\alpha_i - \beta_i) x_i = 0$. Given a set of vectors $x_1, \ldots, x_k$, the *span* of the vectors is

$$\operatorname{span}\{x_i\} = \left\{ \sum_{i=1}^k \alpha_i x_i \mid \alpha \in \mathbb{R}^k \right\}.$$

This is a subspace—a collection of vectors closed under addition and scalar multiplication—of $\mathbb{R}^n$. Given any set S, its perpendicular subspace $S^\perp$ consists of those vectors orthogonal to all of S, i.e.,

$$S^\perp = \{v \mid v^T x = 0 \text{ for all } x \in S\}.$$

Note that no matter what the set S is, $S^\perp$ is *always* a subspace; in the case that S is a subspace as well, we have

$$S + S^\perp = \{v + w \mid v \in S, w \in S^\perp\} = \mathbb{R}^n = S \cup S^\perp,$$

and we can decompose $\mathbb{R}^n$ into the disjoint (excepting the zero vector) subspaces S and $S^\perp$.

A useful variant of these ideas concerns when exactly a collection of vectors can span $\mathbb{R}^n$. We avoid discussion of dimensionality, as we will not need the abstraction particularly, and focus on this case. The lemma states that we require at least n vectors to span $\mathbb{R}^n$.

Lemma 2.1. *Let $v_1, \ldots, v_m \in \mathbb{R}^n$ and $m < n$. Then there is a vector $v \notin \operatorname{span}\{v_1, \ldots, v_m\}$.*

So given a collection of vectors $\{v_1, \ldots, v_m\} \subset \mathbb{R}^n$, so long as $m < n$ we may always "add more" vectors to capture more of $\mathbb{R}^n$. The converse to this lemma—that a collection of n linearly independent vectors spans $\mathbb{R}^n$—will appear when we discuss matrix inverses in Proposition 1 to come.

Proof We prove the first statement by induction on the dimension n. For the base case $n = 1$, it is trivial: the span of the empty collection of scalars contains no non-zero scalar $v \neq 0$. Now, let us assume the result holds true for $n-1$, and let $v_1, \ldots, v_m \in \mathbb{R}^n$ for some $m < n$. Assume for the sake of contradiction that for any vector $v \in \mathbb{R}^n$, there exists some $x \in \mathbb{R}^m$ for which $v = \sum_{j=1}^m v_j x_j$.

Partition each vector into its first $n-1$ entries and last entry via

$$v = \begin{bmatrix} u \\ b \end{bmatrix}, \quad v_j = \begin{bmatrix} u_j \\ b_j \end{bmatrix},$$

where $b, b_j \in \mathbb{R}$ and $u, u_j \in \mathbb{R}^{n-1}$. Then because we may take $b \neq 0$, it is immediate that at least one of the $b_j \neq 0$; without loss of generality, assume that $b_1 \neq 0$. Observe then that $b_1 x_1 + b_2 x_2 + \cdots + b_m x_m = b$, so that $x_1 = (b - \sum_{j=2}^m b_j x_j)/b_1$. Incorporating this into the equality of the first $n-1$ components of the vectors, we have

$$\begin{aligned} u &= \sum_{j=1}^m u_j x_j = \frac{1}{b_1} u_1 \left(b - \sum_{j=2}^m b_j x_j \right) + \sum_{j=2}^m u_j x_j \\ &= \frac{b}{b_1} u_1 + \left(u_2 - \frac{b_2}{b_1} u_1 \right) x_2 + \left(u_3 - \frac{b_3}{b_1} u_1 \right) x_3 + \cdots + \left(u_m - \frac{b_m}{b_1} u_1 \right) x_m. \end{aligned}$$

That is, for the vectors $\widetilde{u}_j = u_j - \frac{b_j}{b_1} u_1$, we have $u - \frac{b}{b_1} u_1 = \sum_{j=2}^m \widetilde{u}_j x_j$. But of course, we may choose $u \in \mathbb{R}^{n-1}$ arbitrarily, and so in particular we have a collection $\{\widetilde{u}_2, \ldots, \widetilde{u}_m\}$ of $m-1$ vectors in $\mathbb{R}^{n-1}$ that spans $\mathbb{R}^{n-1}$. This contradicts the inductive hypothesis. □

2.2 Rank and invertibility of a matrix

As I mention above, a prototypical problem is to solve $Ax = b$ for x; this equation may not always have a solution (though for us it often will), and it is useful to understand when it may. We connect the existence of solutions to the independence properties of the rows and columns of A. To that end, the *range* of a matrix $A \in \mathbb{R}^{m \times n}$ or its *column space* is

$$\operatorname{range}(A) = \{Ax \mid x \in \mathbb{R}^n\} = \left\{ \sum_{i=1}^n a_i x_i \mid x_i \in \mathbb{R} \right\},$$

where $A = [a_1 \; a_2 \; \cdots \; a_n]$ has columns a_i. Notably, this is equivalent to the span of the columns of A. The *null space* of A is

$$\operatorname{null}(A) = \{y \in \mathbb{R}^n \mid Ay = 0\},$$

that is, those vectors that A maps to 0. The familiar *rank nullity* theorem states that the nullspace of A is the perpendicular subspace to $\operatorname{range}(A^T)$, that is,

$$\operatorname{null}(A) = \operatorname{range}(A^T)^\perp. \tag{1}$$

It is relatively easy to see this equality: we write

$$\begin{aligned}\text{range}(A^T)^\perp = \{v \mid v^T z = 0, \text{ all } z \in \text{range}(A^T)\} &= \{v \mid v^T A^T x = 0, \text{ all } x \in \mathbb{R}^m\} \\ &= \{v \mid Av = 0\}\end{aligned}$$

as $y^T x = 0$ for all x if and only if $y = 0$. By taking transposes, we similarly see that $\text{range}(A) = \text{null}(A^T)^\perp$. The *rank* $\text{rank}(A)$ of a matrix A is the dimension of its column space or range, equivalently, it is the number of independent columns of A. Sometimes, one considers the column rank (the number of linearly independent columns of A) and the row rank (the number of linearly independent rows of A) separately; these are always identical.

Finally, we come to inverses of matrices. We say a matrix $A \in \mathbb{R}^{m \times n}$ has a *left inverse* B if there exists $B \in \mathbb{R}^{n \times m}$ such that $BA = I$, and similarly, A has a right inverse if $AB = I$. Note that A can have a left inverse only if A is a tall matrix, while A can have a right inverse only if A is a wide matrix, just by dimension considerations. Notably, if A has both a left inverse B and right inverse C, then A must be square, and moreover, we have

$$BA = I = AC, \quad \text{or} \quad B = B(AC) = (BA)C = C,$$

so $C = B$ and we can write A^{-1} for the inverse of A. Given a square matrix $A \in \mathbb{R}^{n \times n}$, we have the following equivalent characterizations of invertibility. We derive several from one another.

Proposition 1. *For $A \in \mathbb{R}^{n \times n}$, the following are equivalent.*

(a) A has column rank n.

(b) A has row rank n.

(c) $\text{range}(A) = \mathbb{R}^n$ and $\text{null}(A^T) = \{0\}$.

(d) $\text{null}(A) = \{0\}$ and $\text{range}(A^T) = \mathbb{R}^n$.

(e) For any $y \in \mathbb{R}^n$, the matrix equation $Ax = y$ has a solution x.

(f) The matrix A is invertible.

Proof We use the notation

$$A = [a_1 \ \cdots \ a_n] = \begin{bmatrix} -\widetilde{a}_1^T - \\ \vdots \\ -\widetilde{a}_n^T - \end{bmatrix},$$

so that A has columns $a_i \in \mathbb{R}^n$ and rows $\widetilde{a}_i \in \mathbb{R}^n$. We prove several of the cases, leaving others as exercises for the reader. First, we show that (a) $\Rightarrow$ (d) $\Rightarrow$ (b).

- (a) implies (d): as A has n independent columns, we have $Ax = \sum_{i=1}^n a_i x_i = 0$ if and only $x = 0$, so that $\text{null}(A) = \{0\}$.
- (d) implies (b). If $\text{null}(A) = \{0\}$, then the rank-nullity decomposition (1) shows that $\text{range}(A^T) = \{0\}^\perp = \mathbb{R}^n$. Now, suppose for the sake of contradiction that the rows $\widetilde{a}_i$ are linearly dependent; without loss of generality, that there exists $y \in \mathbb{R}^{n-1}$ such that $\widetilde{a}_n = \sum_{i=1}^{n-1} \widetilde{a}_i y_i$. But then

  $$A^T z = \sum_{i=1}^{n-1} \widetilde{a}_i z_i + \left(\sum_{i=1}^{n-1} \widetilde{a}_i y_i \right) z_n \in \text{span}\{\widetilde{a}_1, \ldots, \widetilde{a}_{n-1}\},$$

  which is $n - 1$ dimensional and so cannot be $\mathbb{R}^n$ (by Lemma 2.1).

From a completely parallel (or really, perpendicular...) argument, we have (b) $\Rightarrow$ (c) $\Rightarrow$ (a), so that (a)–(d) are all equivalent.

We now show the equivalence of (e) and (f) to the others. If (e) holds, then by assumption $\text{range}(A) = \mathbb{R}^n$ (i.e. (a) holds), and so we have seen that $\text{range}(A^T) = \mathbb{R}^n$, and so additionally, for each $y \in \mathbb{R}^n$ there exists x such that $A^T x = y$. Variously taking $y = e_1, \ldots, e_n$, the n standard basis vectors, we find vectors $b_1, \ldots, b_n$ such that $Ab_i = e_i$ and $c_1, \ldots, c_n$ such that $A^T c_i = e_i$. Define the matrices B, C by $B = [b_1 \ \cdots \ b_n]$ and $C = [c_1 \ \cdots \ c_n]^T$, so that $AB = I$ and $A^T C^T = I$, i.e., $CA = I$. Then evidently A has a left and right inverse, which must be equivalent, and we have shown that (e) implies (f) and (a). That (f) implies (e) is immediate. Finally, to see that (a)–(d) imply (e), note that they are equivalent to $\{Ax \mid x \in \mathbb{R}^n\} = \text{range}(A) = \mathbb{R}^n$, that is, that for each $y \in \mathbb{R}^n$ there exists $x \in \mathbb{R}^n$ such that $Ax = y$. $\square$

2.3 QR decomposition

In many matrix problems, it is useful to decompose the matrix A into the product of simpler matrices; this can allow for easier or more efficient computation, and it can yield more numerical stability. With that in mind, we consider a few special structures here, showing how we can use them to compute quantities such as matrix inverses, or otherwise.

To begin we recall that a matrix $Q \in \mathbb{R}^{n \times n}$ is *orthogonal* if $Q^T Q = I_n$, the $n \times n$ identity matrix. By Proposition 1, this evidently implies that $QQ^T = I_n$, and so orthogonal matrices enjoy simple inverse properties. An equivalent way to write this is that the matrix $Q \in \mathbb{R}^{n \times n}$ has *orthonormal columns*, meaning that if $Q = [q_1 \ \cdots \ q_n]$, then

$$q_i^T q_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

When $Q \in \mathbb{C}^{n \times n}$, that is, it is complex, if Q^* denotes the Hermitian transpose of Q, we say Q is *orthonormal* or *unitary* if $Q^* Q = QQ^* = I$, though we shall not worry about complex matrices here. In some cases, we shall abuse language a bit and call a tall matrix $Q \in \mathbb{R}^{m \times n}$, where $m > n$, orthogonal: by this we mean that $Q = [q_1 \ \cdots \ q_n]$ has orthogonal columns, so that $Q^T Q = I_n$, though we can never have $QQ^T = I_m$ as $\text{rank}(QQ^T) = n < m$.

The second structure of interest is *triangular* matrices. We say that a matrix R with entries $r_{i,j}$ is *upper* or *right triangular* if $r_{i,j} = 0$ whenever $i > j$, that is, R has the form

$$R = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n-1} & r_{1,n} \\ 0 & r_{2,2} & \cdots & r_{2,n-1} & r_{2,n} \\ 0 & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & r_{n-1,n} \\ 0 & 0 & \cdots & 0 & r_{n,n} \end{bmatrix}.$$

It is evident that if we wish to solve the equation $Rx = b$ when R is upper triangular with all non-zero entries on its diagonal, we can do this in $O(n^2)$ time: indeed, we simply "walk up" the matrix R. First, if $x = [x_1 \ \cdots \ x_n]^T$, then we note that

$$r_{n,n} x_n = b_n \quad \text{or} \quad x_n = \frac{b_n}{r_{n,n}}.$$

We can then solve for x_{n-1}, where we recognize that $r_{n-1,n-1}x_{n-1} + r_{n-1,n}x_n = b_{n-1}$, or (substituting the known value of x_n) that $x_{n-1} = \frac{1}{r_{n-1,n-1}}(b_{n-1} - r_{n-1,n}b_n/r_{n,n})$. Continuing the obvious induction, we see that given $x_n, \ldots, x_{n-k+1}$, we can solve for x_{n-k} via the identity

$$x_{n-k} = \frac{1}{r_{n-k,n-k}}\Big(b_{n-k} - \sum_{i=0}^{k-1} r_{n-k,n-i}x_{n-i}\Big).$$

So long as no entries of $\mathrm{diag}(R)$ are zero, this evidently gives x.

If we wish to solve $Ax = b$ when A is square in more generality, then a natural idea is to find a way to transform A into an upper triangular matrix, and then solve the resulting equation. This is indeed feasible: we can use the *QR factorization* of the matrix A. By this, we mean we write $A = QR$, where Q is an orthogonal matrix and R is upper triangular. If we can do this, then inverting A and solving matrix equations is evidently easy: we have $Ax = b$ if and only if $QRx = b$, or (left multiplying by Q^T, so $Q^TQ = I$)

$$Rx = Q^Tb,$$

which we solve for x. This factorization always exists, and here we present a few variants of its existence.

Proposition 2. *Let $A \in \mathbb{R}^{n\times n}$ have rank n. Then A has a QR factorization $A = QR$ and the diagonal entries of R are all non-zero.*

Proof The idea of the proof is to use *Gram-Schmidt* orthonormalization, where we iteratively construct matrices Q_k and R_k, with $Q_k \in \mathbb{R}^{n\times k}$ and $R_k \in \mathbb{R}^{k\times k}$, such that $Q_k^TQ_k = I_k$ and $Q_kR_k = [a_1 \cdots a_k]$. That is, Q_kR_k reconstructs the first k columns of A, and the columns of Q_k form an orthonormal basis for the first k columns of A.[1]

We provide the proof inductively. We take $q_1 = a_1/\left\|a_1\right\|$ and $r_{1,1} = \left\|a_1\right\|$, which evidently satisfies $q_1r_{1,1} = a_1$. Now, for the inductive step, we assume that we have orthogonal matrix $Q_k = [q_1 \cdots q_k] \in \mathbb{R}^{n\times k}$ and an upper triangular $R_k \in \mathbb{R}^{k\times k}$ with entries $r_{i,j}$ such that $[a_1 \cdots a_k] = Q_kR_k$. The idea is to take a_{k+1}, project out all components of a_{k+1} belonging to $\mathrm{span}\{q_i\}_{i=1}^k$, and then treat the remaining vector as q_{k+1}. Recall that for a unit vector q, the projection of a onto $\mathrm{span}\{q\}$ is $(a^Tq)q$. To that end, write

$$v_{k+1} \coloneqq a_{k+1} - \sum_{i=1}^{k}(q_i^Ta_{k+1})q_i. \tag{2}$$

As a_{k+1} is independent of $a_1, \ldots, a_k$ by assumption, and $\mathrm{span}\{q_1, \ldots, q_k\} = \mathrm{span}\{a_1, \ldots, a_k\}$, we must have $v_{k+1} \neq 0$. We then define

$$q_{k+1} = \frac{1}{\left\|v_{k+1}\right\|}v_{k+1}.$$

By inspecting equation (2), we also see how to define $r_{i,k+1}$ for each i: as

$$a_{k+1} = \sum_{i=1}^{k}\underbrace{(q_i^Ta_{k+1})}_{=r_{i,k+1}}q_i + \underbrace{\left\|v_{k+1}\right\|}_{=r_{k+1,k+1}}q_{k+1},$$

[1] The procedure we describe works, abstractly, even in infinite dimensional space, though obviously we will not require this.

we can evidently set we set $r_{i,k+1} = q_i^T a_{k+1}$ for $i \le k$ and $r_{k+1,k+1} = \|v_{k+1}\| > 0$. Appending q_{k+1} and the r values as appropriate to construct $Q_{k+1} = [q_1 \ \cdots q_{k+1}]$ and

$$R_{k+1} = \begin{bmatrix} & & r_{1,k+1} \\ & R_k & r_{2,k+1} \\ & & \vdots \\ 0 & & r_{k+1,k+1} \end{bmatrix}$$

we have $[a_1 \ \cdots \ a_{k+1}] = Q_{k+1}R_{k+1}$. This completes the induction. □

As some minor extensions of this result, we can develop QR factorizations for tall matrices: if $A \in \mathbb{R}^{m\times n}$ with $m \ge n$, then we can similarly write $A = QR$ with $Q \in \mathbb{R}^{m\times n}$ orthogonal, so $Q^TQ = I_n$, and R upper triangular. If $A \in \mathbb{R}^{m\times n}$ is not full rank—that is, $\mathrm{rank}(A) < n$—then the QR factorization still exists, but R cannot be full rank. In this case, step (2) of the Gram-Schmidt process may fail: there is some k such that $a_{k+1} = \sum_{i=1}^k (a_{k+1}^T q_i) q_i$. Then $v_{k+1} = 0$, evidently, but this is fixable: we construct $r_{i,k+1}$ for $i \le k$ as before, but set $r_{k+1,k+1} = 0$, and choose q_{k+1} to be an arbitrary vector perpendicular to $\{q_1, \ldots, q_k\}$. We can continue this process to obtain $A = QR$, though R may have zeros on its diagonal.[2]

We have seen how for square matrices, the QR factorization immediately allows solving linear equations; it also allows getting strong best approximations to overdetermined equations. Indeed, suppose that $A \in \mathbb{R}^{m\times n}$, $m \ge n$ and $\mathrm{rank}(A) = n$, and we wish to find the x that is as close as possible to solving $Ax = b$—though there may be not consistent solution. We may write this as finding the x minimizing the quadratic form

$$\|Ax - b\|^2 = \|QRx - b\|^2 = (QRx - b)^T (QRx - b).$$

But then using that $Q^TQ = I$, we have the identity

$$(QRx-b)^T(QRx-b) = (Rx-Q^Tb)^T(Rx-Q^Tb)+b^Tb-b^TQQ^Tb = \left\|Rx - Q^Tb\right\|^2+b^T(I-QQ^T)b,$$

which is quite clearly minimized by $x = R^{-1}Q^Tb$, that is, the x solving $Rx = Q^Tb$. That is, even in the case that A is a tall matrix, the QR factorization gives us the "best" approximate solution to $Ax = b$ in a reasonably straightforward way.

2.4 Cholesky decomposition

We often treat quadratic forms with special care, that is, quantities such as $f(x) = x^TAx$ for a matrix A. In these, by noting that $x^TAx = x^TA^Tx = \frac{1}{2}x^T(A + A^T)x$, it is no loss of generality to assume that A is a symmetric matrix, so we shall do so. Of particular interest are *positive definite* matrices A, denoted $A \succ 0$, which are symmetric matrices satisfying

$$x^TAx > 0 \ \text{ whenever } x \ne 0.$$

A matrix is *positive semidefinite*, or PSD, denoted $A \succeq 0$, if $x^TAx \ge 0$ for all x. Note that by taking $x = e_1, \ldots, e_n$, the n standard basis vectors, we see that a positive definite matrix A always has $e_i^TAe_i = a_{ii} > 0$, or $\mathrm{diag}(A) \succ 0$.

For such positive definite matrices, an elegant decomposition known as the *Cholesky decomposition* guarantees that we can write A as the product of lower triangular matrices:

[2] There are many variants of this process, and factorizations that reveal the rank of A, but we shall not address these.

Proposition 3 (Cholesky factorization)**.** *Let $A \succ 0$. There exists a lower triangular $L \in \mathbb{R}^{n \times n}$ such that*

$$A = LL^T.$$

Proof The existence of L is fairly straightforward: we cheat a bit by looking ahead to the spectral theorem (Theorem 1) to note that we can write $A = V\Lambda V^T$ for an orthogonal matrix V and diagonal $\Lambda \succ 0$. Set $B = \Lambda^{1/2}V^T$, so that $A = B^TB$. Then B has QR factorization $B = QR$, and we write $A = R^TQ^TQR = R^TR$. As R is upper triangular, $L = R^T$ is lower triangular. □

Of course, an existence proof is a bit empty, and finding such a factorization is more computationally helpful. Conveniently, one can actually recover a reasonable (and numerically stable) Cholesky algorithm by writing down the matrix factorization directly. I leave deriving this as an exercise for the reader.

3 Spectral and singular value decompositions

Two additional decompositions are key to any use of matrices, both theoretical and applied: eigenvalue (or spectral) decompositions and the singular value decomposition. We shall not concern ourselves much with nonsymmetric matrices—their eigenvalues and eigenvectors arise less frequently in most statistical applications—though we develop other decompositions.

Recall that for a square matrix $A \in \mathbb{R}^{n \times n}$, a vector v is an *eigenvector* with *eigenvalue* λ is $Av = \lambda v$. Nonsingular matrices always have n eigenvalues—including multiplicities—a consequence of the characterization of eigenvalues as roots of the characteristic polynomial $p(\lambda) = \det(A - \lambda I)$. But we will not concern ourselves with these; instead, we focus on matrices whose eigenvalues and eigenvectors are (i) more easily visualized, (ii) avoid pathologies, and (iii) are most common in statistical, machine learning, and engineering applications.

3.1 The spectral theorem for symmetric matrices

We focus on *symmetric* matrices, that is, those for which $A = A^T$. In these cases, the eigenvectors and eigenvalues take the particularly nice form the spectral theorem guarantees:

Theorem 1 (Spectral theorem)**.** *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then A has n real eigenvector/eigenvalue pairs (λ_i, v_i), and the eigenvectors are orthogonal. Moreover, if $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of the eigenvalues and $V = [v_1 \ \cdots \ v_n] \in \mathbb{R}^{n \times n}$, then*

$$A = V\Lambda V^T. \tag{3}$$

For completeness, we provide a proof of the spectral theorem in Section 4.3, though we do so using *only* tools from optimization. This gives a variational perspective that can often be quite useful for matrix analysis and solving statistical (or other) problems, and has the advantage that it avoids the complex analysis usually required to develop eigenvalue decompositions.

A few side benefits of the formulation (3) are that it shows immediately how to invert A, assuming A is full rank: we simply have $A^{-1} = V\Lambda^{-1}V^T$, as $V\Lambda^{-1}V^TV\Lambda V^T = VIV^T = I$. Raising A to various powers is also trivial: we have $A^p = V\Lambda^pV^T$, which is easy to compute for Λ diagonal. As another consequence of the spectral theorem, we see that

Corollary 3.1. *A symmetric A is positive semidifinite if and only if its eigenvalues are nonnegative, and A is positive definite if and only if its eigenvalues are strictly positive.*

Proof To see this, note that for $A = V\Lambda V^T$, then A is positive definite if and only if $x^T\Lambda x = \sum_{j=1}^n x_j^2\lambda_j > 0$ whenever $x \neq 0$, that is, $\lambda_j > 0$ for each j. Similarly, A is semidefinite if and only if $\sum_{j=1}^n x_j^2\lambda_j \geq 0$ for all x, that is, $\lambda_j \geq 0$. $\square$

A few alternatives to the spectral theorem are available; the development of their proofs is similar to our proof of the spectral theorem itself, so we omit them. But, among other useful examples, is the following variational characterization of the first k eigenvectors of A:

Corollary 3.2. *Let $W_\star$ solve the following optimization problem (in W):*

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^k w_i^T A w_i \\ \text{subject to} & W^T W = I_k, \end{array}$$

where $W = [w_1 \ \cdots \ w_k] \in \mathbb{R}^{n\times k}$. Then $W_\star = [v_1 \ \cdots \ v_k]$, where v_i is the ith eigenvector of A.

The proof of this corollary is not completely immediate—typical proofs rely on a result known as Von Neumann's Trace Inequality or on majorization identities—but it provides alternative variational characterizations of eigenvectors. We shall see applications of this result and the next decompositions in principal component analysis and other multivariate analysis scenarios.

3.2 The singular value decomposition

The spectral theorem also allows us to show the existence of a perhaps even more powerful decomposition: the singular value decomposition, or SVD. The statement is as follows.

Theorem 2. *Let $A \in \mathbb{R}^{m\times n}$, where $m \geq n$. Then there exists an orthogonal $U \in \mathbb{R}^{m\times n}$, orthogonal $V \in \mathbb{R}^{n\times n}$, and nonnegative diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ such that*

$$A = U\Sigma V^T.$$

Proof For simplicity, we first assume that A has rank n, meaning that A^TA is positive definite. We use the spectral theorem to write $A^TA = VDV^T$, where $D \succ 0$, that is, $\mathrm{diag}(D)$ has all positive entries. By inspection, if we have $A = UD^{1/2}V^T$ then, inverting, we must take $U = AVD^{-1/2} \in \mathbb{R}^{m\times n}$. Note that with this choice, $U^TU = D^{-1/2}V^TA^TAVD^{-1/2} = D^{-1/2}DD^{-1/2} = I$, and so $A = UD^{1/2}V^T$ as desired.

If A is rank deficient—so $r = \mathrm{rank}(A) < n$—then the argument is more tedious, as we must argue we can ignore the rank deficient "parts" of A. We still have $A^TA = VDV^T$, except that $D = \mathrm{diag}(d_1, \ldots, d_r, 0, \ldots, 0) \succeq 0$, though D has a square root. Let $D_r = \mathrm{diag}(d_1, \ldots, d_r) \in \mathbb{R}^{r\times r}$ be the positive diagonal of D, and let $V_r = [v_1 \ \cdots \ v_r] \in \mathbb{R}^{n\times r}$ be the first r columns of V and V_{n-r} the last $n-r$ columns. Define $U_r = AV_rD_r^{-1/2} \in \mathbb{R}^{m\times r}$, noting that $U_r^TU_r = I_r$. We claim that $AV_rV_r^T = A$: indeed, as $V_{n-r}^TA^TAV_{n-r} = 0$ we have $\|AV_{n-r}\|_{\mathrm{Fr}} = 0$, and noting the identity

$$I = VV^T = \sum_{i=1}^n v_iv_i^T = V_rV_r^T + V_{n-r}V_{n-r}^T,$$

we thus obtain

$$A = \underbrace{AV_{n-r}V_{n-r}^T}_{=0} + AV_rV_r^T = AV_rV_r^T.$$

Then we see that the choice $U_r = AV_rD_r^{-1/2}$ satisfies $U_rD_r^{1/2}V_r^T = AV_rV_r^T = A$. To expand this into U, V as in the theorem statement, note that we may complete U by setting $U = [U_r \ U_{n-r}]$ for any U_{n-r} such that $U^TU = I$—which is possible—and setting

$\Sigma = \operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_r}, 0, \ldots 0) \in \mathbb{R}^{n \times n}$. Then $A = U\Sigma V^T$ as desired. $\square$

A few slight refinements and corollaries of the result are possible. First, if A is rank deficient, meaning $\operatorname{rank}(A) = r$, we can develop a "reduced" SVD:

Corollary 3.3. *If $A \in \mathbb{R}^{m \times n}$, $m \le n$, has $\operatorname{rank}(A) = r < n$, then we may write $A = U_r \Sigma_r V_r^T$, where $U_r \in \mathbb{R}^{m \times r}$, $V_r \in \mathbb{R}^{n \times r}$ satisfy $V_r^T V_r = U_r^T U_r = I_r$, and $\Sigma_r \in \mathbb{R}^{r \times r}$ is positive definite and diagonal.*

When A is a wide matrix instead of a tall matrix, i.e., $A \in \mathbb{R}^{m \times n}$ with $n \ge m$, we have a similar decomposition: we write $A = U\Sigma V^T$ with $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times m}$, where $UU^T = U^T U = I_m$ and $V^T V = I_m$. It is easy to see this by considering A^T.

3.3 Heuristic connections to statistical problems and data matrices as operators

Let us make a few heuristic applications to statistical problems here; these are perhaps more for intuition than anything completely rigorous at this point, but may serve to ground some of the mathematics we have been doing. Consider the linear model

$$Y = X\beta + \varepsilon, \quad X \in \mathbb{R}^{n \times d} \text{ and } \varepsilon \sim \mathsf{N}(0, \sigma^2 I).$$

Let us consider, at a high level, questions of recovering $\beta \in \mathbb{R}^d$ from the noisy observations $Y \in \mathbb{R}^n$ and data matrix $X \in \mathbb{R}^{n \times d}$, where the number of observations $n \ge d$. First, let us consider a special case that X is rank deficient, so that $\mathcal{N} := \operatorname{null}(X) \neq \{0\}$ contains more than just the zero vector. Roughly, the nullspace $\mathcal{N}$ gives directions in which we can infer *nothing* about β. Making this somewhat less heuristic, $\Delta \in \mathcal{N}$ implies $X\Delta = 0$. Now consider trying to distinguish two possible vectors β_0, β_1, where $\beta_0 \in \mathbb{R}^d$ and $\beta_1 = \beta_0 + \Delta$ for some $\Delta \in \mathcal{N}$. Then clearly $X\beta_0 = X(\beta_0 + \Delta) = X\beta_1$, so that X the observations contain *no* information about directions in $\mathcal{N}$, that is, we can never distinguish β_0 from β_1: the distributions of Y under β_0 and β_1 are identical.

As being low rank is a fairly brittle property—arbitrarily small perturbations of the matrix X make it full rank (i.e. rank d)—it is interesting to consider X being "nearly" low rank. With this in mind, consider the singular value decomposition of X,

$$X = U\Sigma V^T, \quad \Sigma = \operatorname{diag}(s_1, \ldots, s_d),$$

where $U \in \mathbb{R}^{n \times d}$, $V \in \mathbb{R}^{d \times d}$, and $U^T U = V^T V = I_d$. Consider the action of X on a vector β: first, we project β into the basis represented by V via $\alpha = V^T \beta$. Then we scale the resulting vector $\alpha \in \mathbb{R}^d$ by the diagonal matrix Σ, and reproject the result into the span of U (noting that $\operatorname{range}(U) = \operatorname{range}(X)$). For those singular values s_j that are near zero, X is evidently quite insensitive to changes of β in the directions v_j. As a somewhat more precise description of this, suppose that $s_d = \epsilon > 0$, but ϵ is quite small—say, 10^{-3}. Then if for $t \in \mathbb{R}$ we consider two vectors,

$$\beta_0 \text{ and } \beta_t := \beta_0 + t v_d,$$

so that β_t is perturbed in the direction of the smallest right singular vector v_d, then we see (or can measure) only very little about changes in β_t: we have

$$X\beta_t = X\beta_0 + tXv_d = X\beta_0 + t\epsilon u_d,$$

where $u_d \in \mathbb{R}^n$ is the dth left singular vector. In particular, to achieve even a constant factor change in the measurements $Y = X\beta + \varepsilon$, we must modify β_0 by a number of units $t \approx 1/\epsilon$ in the v_d direction. Thus, nearly singular data matrices—those with some small singular values—often make inferences about parameters β extremely challenging. Thinking contrapositively, small changes in the output Y may correspond to massive changes in the input β, as we must make large modifications in the v_d direction to change $X(\beta_0 + tv_d)$.

In sum, the singular value decomposition tells us essentially everything about a matrix: it is rank deficient if and only if some of the singular values are zero, and being nearly rank-deficient (having small singular values) can equally make statistical estimation and inference hard. (This of course is heuristic now, but we can make this substantially more precise.)

4 Optimization

Optimization problems involve minimizing or maximizing an objective subject to various constraints, and we typically write them as

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_i(x) = 0, \; i = 1, \ldots, m \\ & g_i(x) \le 0, \; i = 1, \ldots, p. \end{array}$$

Here, f is the *objective function*, x is the *optimization variable*, and h_i and g_i are constraint functions. We shall use relatively little optimization technology in this class, though a few ideas will be important and will appear quite frequently.[3]

The most important optimization problems we come against in this class will be *least squares* problems. Consider the linear model as motivation: we assume

$$Y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n \times d}$, and given (X, Y) wish to (approximately) find β. Then a natural optimization problem is to minimize the squared error, that is, to solve

$$\text{minimize} \; \frac{1}{2} \|X\beta - Y\|^2$$

in β. In some situations that we cover, we will use more sohpisticated losses, in effort, e.g., to gain robustness or to regularize problems. Letting X have rows $x_1^T, \ldots, x_n^T$, i.e., $X = [x_1 \; \cdots \; x_n]^T$ we can then consider a more general formulation and ask to solve

$$\text{minimize} \; \sum_{i=1}^n \ell(x_i^T \beta - y_i), \tag{4}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is some loss measuring the fidelity of the linear prediction $x_i^T \beta$ to y_i. Of course, $\ell(t) = \frac{1}{2}t^2$ recovers the typical least-squares problem, but other examples include the ℓ_1 loss, $\ell(t) = |t|$, or smoothed ℓ_1-type losses, including the Huber loss

$$\ell(t) = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \le 1 \\ |t| - \frac{1}{2} & \text{if } |t| \ge 1 \end{cases}$$

or $\ell(t) = \log(1 + e^t) + \log(1 + e^{-t})$. We will see numerous examples of *M estimation* problems of the form (4) (M for maximization/minimization). We at least will touch on the tools necessary to recognize when such problems are (easily) solvable.

[3] All students should take ee364a.

4.1 Convexity

The major watershed between solvable and (broadly) unsolvable problems is optimization is the division between *convex* optimization problems and non-convex optimization. In short: convex optimization problems are solvable, and solving them is a technology; we can solve them and it does not matter a wink how we have done it. For non-convex optimization, well, as Tolstoy writes in *Anna Karenina*, "Happy families are all alike; every unhappy family is unhappy in its own way," each non-convex optimization problem is non-convex for its own special reason. What, then, is a convex optimization problem?

The starting points are convex sets and functions. A set $C \subset \mathbb{R}^n$ is convex if

$$x \in C, y \in C \;\; \text{imply} \;\; \lambda x + (1-\lambda)y \in C$$

for all $\lambda \in [0,1]$, that is, C contains all lines between its points. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{5}$$

for all $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^n$. These are "bowl-shaped" functions. Convex problems are then those for which the objective f is a convex function, the constraints h_i are affine, and the constraints g_i are convex. You should convince yourself that this means that the set of *feasible points*, those x for which $h_i(x) = 0$ and $g_i(x) \le 0$ for each of the constraint functions, necessarily form a convex set. We will not concern ourselves with the constraints in problems (at least for now), except to reiterate that they make (essentially) no difference in our ability to computationally efficiently solve the problems.

Convex functions are convenient from a computational perspective because they remain convex under a number of compositions and other operations. In fact, some large-scale optimization toolboxes are based on careful reformulation of various composition rules, allowing the solution of numerous optimization problems [3]. For us, a few simple rules suffice:

- if f is convex, then for any matrix A and vector b, $h(x) = f(Ax + b)$ is convex.
- if f_0, f_1 are convex, then $f = f_0 + f_1$ is convex.

These are both immediate.

Among many reasons that we like convex functions are that their minimizers have enjoy a few key properties: if f is a convex function, then its minimizers form a convex set, and local optimality necessarily implies global optimality. To see these is relatively straightforward by a picture (again, drawing such a picture is a good exercise for the reader), but for completeness we include an argument in Appendix A.1, as it relies on some analysis.

Lemma 4.1. *Let f be convex. Then any local minimizer of f is a global minimizer, and if the collection of minimizers of f form a convex set.*

Note that a convex function f may have no minimizers—consider $f(x) = e^x$ or $f(x) = -\log x$—but when they exist, they form a convex set.

Another key property of convex f more or less follows the preceding lemma, and connects to the idea of local information providing sufficient global information for convex f: if f is differentiable, then $\nabla f(x) = 0$ if and only if x is a global minimizer of f. Indeed, we have the following proposition; a proof by picture again should suffice, though we provide a formal argument in Appendix A.2.

Proposition 4. *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then f is convex if and only if for each $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$,*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x). \tag{6}$$

If f is convex, then x minimizes f if and only if $\nabla f(x) = 0$. If additionally f is twice continuously differentiable, then f is convex if and only if its Hessian is positive semidefinite.

Using Proposition 4, we can provide a few simple examples that are essential for the class. As one immediate consequence, if we have a convex minimization problem and we can solve $\nabla f(x) = 0$, then immediately we recover minimizers of f. Let us apply this to the least-squares problem: we wish to solve

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2} \left\| X\beta - Y \right\|_2^2 .$$

Note that the function $h(z) = \frac{1}{2} \left\| z \right\|_2^2$ is convex, as $\nabla^2 h(z) = I \succ 0$, and so the least-squares objective is as well (it is the composition of a convex function with the affine mapping $\beta \mapsto X\beta - Y$). Consequently, by taking derivatives of $f(\beta) = \frac{1}{2} \left\| X\beta - Y \right\|_2^2$, we have

$$\nabla f(\beta) = X^T (X\beta - Y),$$

and β minimizes the least-squares objective if and only if

$$X^T X \beta = X^T Y. \tag{7}$$

These are the *normal equations*, and they allow us to recover substantial structure about the properites of estimates $\widehat{\beta}$ solving least-squares. Note that if $X \in \mathbb{R}^{n \times d}$ is rank d, then $X^T X$ is invertible and we uniquely obtain $\widehat{\beta} = (X^T X)^{-1} X^T Y$. We develop the properties of such estimators extensively throughout the class.

4.2 Lagrange Multipliers

With or without convexity, it is still of interest to solve constrained optimization problems—indeed, in Section 3 we claimed that the eigenvectors and eigenvalues of a symmetric matrix were recoverable via solving a constrained optimization problem. With that in mind, we review and recover a few of the basic results in optimization. While I certainly do not expect expertise—and this material is not essential for the course—I find it useful to provide alternative proofs and approaches to some of the basic results in linear algebra and optimization.

Throughout this general treatment, I will assume $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are $\mathcal{C}^1$, that is, continuously differentiable, so that the gradient $\nabla f(x)$ of f and the derivative matrix of $h(x) = (h_1(x), \ldots, h_m(x))$ given by

$$Dh(x) = \begin{bmatrix} \nabla h_1(x)^T \\ \vdots \\ \nabla h_m(x)^T \end{bmatrix}$$

are both continuous. That is, we have that as $\Delta \in \mathbb{R}^n$ tends to zero, $f(x + \Delta) = f(x) + \nabla f(x)^T \Delta + o(\left\| \Delta \right\|)$ and similarly $h(x + \Delta) = h(x) + Dh(x)\Delta + o(\left\| \Delta \right\|)$.

We consider two problems: the unconstrained problem

$$\underset{x}{\text{minimize}} \ f(x) \tag{8}$$

and the constrained problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0. \end{array} \tag{9}$$

While there is a deep optimality theory for *convex* optimization problems—see, for example, the book [2] and the course ee364a—we will consider only the general forms above. We recall that a point $x_\star$ is a *local minimum* of f if there is a neighborhood of $x_\star$ such that $f(x_\star) \le f(x)$ for all x in the neighborhood, that is, if there exists a positive radius $b > 0$ such that $f(x_\star) \le f(x)$ for x such that $\|x - x_\star\| \le b$. We have the following immediate result:

Proposition 5. *If $x_\star$ is a local minimizer of f in the unconstrained problem (8), then $\nabla f(x_\star) = 0$.*

Proof For x near $x_\star$, we have $f(x) = f(x_\star) + \nabla f(x_\star)^T(x - x_\star) + o(\|x - x_\star\|)$. Fix an $x \in \mathbb{R}^n$ and let $t > 0$. Then for small enough t, we have

$$0 \le \frac{f(x_\star + t(x - x_\star)) - f(x_\star)}{t} \xrightarrow[t \searrow 0]{} \nabla f(x_\star)^T(x - x_\star)$$

by definition of the derivative. As x is arbitrary, we see that $\nabla f(x_\star)^T y = 0$ for all $y \in \mathbb{R}^n$, or $\nabla f(x_\star) = 0$. □

The more advanced result, known as the Lagrange Multiplier theorem, allows us to develop stationary conditions for the constrained problem (9). The intuition is the following: we place a penalty λ_i on each of the constraints $h_i(x) = 0$. Then we consider the alternative problem, or *Lagrangian*,

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i h_i(x).$$

Intuitively, by adjusting the penalties on λ_i to force payment whenever $h_i(x) \neq 0$ (i.e., increasing λ_i when $h_i(x) > 0$ and decreasing it when $h_i(x) < 0$), we find that local minima of problem (9) are stationary for the Lagrangian. A geometric picture may help as well here (but for now, I don't have one).

In any case, we have the following proposition:

Proposition 6 (Lagrange multipliers)**.** *Let $x_\star$ be a local minimizer in the constrained problem (9), and assume that the m vectors $\nabla h_1(x_\star), \ldots, \nabla h_m(x_\star)$ are linearly independent. Then there exists a* Lagrange multiplier $\lambda \in \mathbb{R}^m$ *such that*

$$\nabla f(x_\star) + \sum_{i=1}^m \lambda_i \nabla h_i(x_\star) = 0.$$

Notably, if $h : \mathbb{R}^n \to \mathbb{R}$ is a single function, then the *constraint qualification* that the gradients $\nabla h_i(x_\star)$ be linearly independent is trivially satisfied. We provide a proof here following Bertsekas [1]. Naturally, there are a number of proofs, many involving implicit function theorems, but the following result makes clean use of analysis and only first-order derivative definitions.

Proof By definition of $x_\star$ as a local minimizer in problem (9), there exists a compact set $S = \{x \in \mathbb{R}^n \mid \|x - x_\star\| \le b\}$, where $b > 0$, such that $f(x) \ge f(x_\star)$ for all $x \in S$ with

$h(x) = 0$. We consider optimization over this set. Let $\alpha > 0$ be otherwise arbitrary, and for $k \in \mathbb{N}$ consider the sequence of objectives

$$f_k(x) \coloneqq f(x) + \frac{k}{2} \|h(x)\|^2 + \frac{\alpha}{2} \|x - x_\star\|^2 ,$$

and let x_k minimize $f_k(x)$ over S. As S is compact and f_k is continuous, such a minimizer certainly exists, and moreover, as $h(x_\star) = 0$ and $x_\star - x_\star = 0$, we have

$$f_k(x_k) = f(x_k) + \frac{k}{2} \|h(x_k)\|^2 + \frac{\alpha}{2} \|x_k - x_\star\|^2 \le f_k(x_\star) = f(x_\star). \tag{10}$$

In particular, as $f(x)$ is bounded on S (it is continuous and S is compact), we obtain that for some finite $C < \infty$,

$$\frac{k}{2} \|h(x_k)\|^2 + \frac{\alpha}{2} \|x_k - x^\star\|^2 \le C,$$

and so $\|h(x_k)\| \to 0$. Considering any limit point x_∞ of x_k, taking $k \to \infty$ in inequality (10) thus implies

$$f(x_\infty) + \frac{\alpha}{2} \|x_\infty - x_\star\|^2 \le f(x_\star),$$

so $x_\infty = x_\star$. In particular, $x_k \to x_\star$ as $\alpha > 0$.

With this in mind, we now use Proposition 5: we see that $x_k \in \operatorname{int} S$ eventually, that is, the constraint S is eventually immaterial, and so x_k is a local minimizer for the unconstrained problem

$$\text{minimize} \;\; f_k(x) = f(x) + \frac{k}{2} \|h(x_k)\|^2 + \frac{\alpha}{2} \|x - x_\star\|^2 .$$

Taking derivatives, we obtain

$$0 = \nabla f(x_k) + k Dh(x_k)^T h(x_k) + \alpha(x_k - x_\star), \tag{11}$$

where we recall the derivative matrix $Dh(x)^T = [\nabla h_1(x) \; \cdots \; \nabla h_m(x)] \in \mathbb{R}^{n \times m}$. As $Dh(x_\star) \in \mathbb{R}^{m \times n}$ is assumed to be rank m, we have $Dh(x_k)$ also rank m for all large enough k by the continuity of $\nabla h_i(x)$. Using the shorthand $G_k = Dh(x_k) \in \mathbb{R}^{m \times n}$, $G_k G_k^T$ is a symmetric positive definite matrix, and multiplying this into the preceding display yields

$$0 = G_k \nabla f(x_k) + k G_k G_k^T h(x_k) + \alpha G_k (x_k - x_\star).$$

Equivalently, inverting yields

$$k h(x_k) = -(G_k G_k^T)^{-1} G_k \nabla f(x_k) - \alpha (G_k G_k^T)^{-1} (x_k - x_\star).$$

Of course, as $x_k \to x_\star$, continuity implies that $G_k \to G_\star \coloneqq Dh(x_\star)$ and $\nabla f(x_k) \to \nabla f(x_\star)$, so taking $k \to \infty$ yields

$$\lim_{k \to \infty} k h(x_k) = -(G_\star G_\star^T)^{-1} G_\star \nabla f(x_\star).$$

Let $\lambda = -(G_\star G_\star^T)^{-1} G_\star \nabla f(x_\star) \in \mathbb{R}^m$. Then substituting in equality (11) and taking $k \to \infty$ we obtain

$$0 = \nabla f(x_\star) + G_\star^T \lambda = \nabla f(x_\star) + \sum_{i=1}^m \lambda_i \nabla h_i(x_\star),$$

as desired. □

4.3 Applications to the spectral theorem

Finally, we turn to a delightful application of the Lagrange multiplier theorems: proving the spectral theorem for symmetric matrices. In most linear algebra classes, one develops the existence of eigenvectors and eigenvalues by using characteristic polynomials and determinants, objects that—to my eyes—have little to do with the actual structure of matrices and their actions as operators. To that end, here we use Lagrange multiplier results to show that eigenvectors and eigenvalues exist for symmetric matrices *without* any complex analysis (or even needing polynomials to have roots), instead, purely as an optimization concept. Indeed, we revisit the spectral theorem:

To prove the theorem, we begin by showing that there exists a single eigenvector/eigenvalue pair, where the eigenvector solves

$$\begin{array}{ll} \text{maximize}_x & x^T A x \\ \text{subject to} & \|x\| = 1 \end{array}$$

when A is symmetric.

Lemma 4.2. *Let $A \in \mathbb{R}^{n \times n}$ be symmetric and v maximize $x^T A x$ subject to $\|x\| = 1$. Then $Av = \lambda v$ for some $\lambda \in \mathbb{R}$.*

Proof We can equivalently write this as maximization of $x^T A x$ subject to $\|x\|_2^2 = 1$, in which case Proposition 6 immediately implies that there exists some $\lambda \in \mathbb{R}$ such that

$$Av - \lambda v = 0.$$

Rearrange. □

So now we have shown the existence of a first eigenvalue/eigenvector pair: there is (λ_1, v_1) such that $v_1^T A v_1 = \lambda_1 \geq x^T A x$ for all $\|x\| = 1$. We now perform an induction by finding the maximizer of the quadratic $x^T A x$ subject to being orthogonal to all previously found eigenvectors: assume that we have $k-1$ eigenvectors in decreasing order of eigenvalues, i.e., $Av_1 = \lambda_1 v_1, \ldots, Av_{k-1} = \lambda_{k-1} v_{k-1}$, and $v_i^T v_j = 1$ if $i = j$ and 0 otherwise. Consider the constrained problem

$$\begin{array}{ll} \text{maximize}_x & x^T A x \\ \text{subject to} & \|x\| = 1, \quad v_i^T x = 0, \; i = 1, \ldots, k-1. \end{array}$$

Let v_k be a maximizer above. Then evidently the constraints have independent gradients, as the gradients are $v_1, \ldots, v_{k-1}$ and $\nabla \|x\|^2 = 2x$, which is v_k at optimum, so Proposition 6 implies there exist $\beta_1, \ldots, \beta_{k-1}$ and λ_k (choosing signs as convenient) such that

$$Av_k + \sum_{i=1}^{k-1} \beta_i v_i - \lambda_k v_k = 0.$$

Taking the inner product of this equality with v_i for each $i < k$ yields

$$0 = v_i^T \left(Av_k + \sum_{i=1}^{k-1} \beta_i v_i + \lambda_k v_k \right) = \lambda_i v_i^T v_k + \beta_i v_i^T v_i - \lambda_k v_k^T v_i = \beta_i,$$

as $v_i^T v_k = 0$ by construction. In particular, $\beta_i = 0$, and so returning to the previous display, we have shown

$$Av_k - \lambda_k v_k = 0.$$

Summarizing, we have the following proposition, which is similar to the spectral theorem:

Proposition 7. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then there exist n real eigenvalue/eigenvector pairs (λ_i, v_i) with $\lambda_1 \geq \cdots \geq \lambda_n$ and such that v_k solves*

$$\begin{array}{ll} \text{maximize}_x & x^T A x \\ \text{subject to} & \|x\| = 1, \quad v_i^T x = 0, \; i = 1, \ldots, k-1. \end{array}$$

Finally, we use Proposition 7 to prove the spectral theorem.

Proof of Theorem 1 We first show that $v_i^T v_j = 0$ if $i \neq j$. Indeed, we have $v_i^T A v_j = \lambda_j v_i^T v_j$, and by construction in Proposition 7, these are orthogonal. In particular, the matrix $V = [v_1 \; \cdots \; v_n] \in \mathbb{R}^{n\times n}$ is orthogonal, with $V^T V = I = V V^T$. Letting $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ be the diagonal matrix of eigenvalues, we see that because $Av_i = \lambda_i v_i$, $AV = V\Lambda$ and so

$$A = AVV^T = V\Lambda V^T$$

as desired. □

A Technical proofs

A.1 Proof of Lemma 4.1

First we show that the minimizers of f, assuming they exist, form a convex set. Suppose that x_0, x_1 are both minimizers of f, so that

$$f(x_0) = f(x_1) = \min_x f(x)$$

(where we write min because the minimum is attained by assumption). Then certainly $f(\lambda x_0 + (1-\lambda)x_1) \leq \lambda f(x_0) + (1-\lambda) f(x_1) = \min_x f(x)$ whenever $\lambda \in [0,1]$, that is, $\lambda x_0 + (1-\lambda)x_1$ also minimizes f.

Second, we argue that local minimizers are global minimizers. Suppose x_0 is a local minimizer of f, so that $f(x_0) \leq f(x)$ for all x in a neighborhood of x_0; say this neighborhood is of radius $\epsilon > 0$. Let $x_1 \in \mathbb{R}^n$ be arbitrary. Then for $t \in [0,1]$, we can write

$$x_t = (1-t)x_0 + t x_1,$$

noting that $f(x_t) \leq (1-t)f(x_0) + t f(x_1)$. Now, take t small enough that $\|x_t - x_0\| \leq \epsilon$; the choice $t = \epsilon / \|x_1 - x_0\|$ suffices. Then we have by assumption that $f(x_0) \leq f(x_t) \leq (1-t)f(x_0) + t f(x_1)$, and rewriting, $t f(x_1) \geq t f(x_0)$, or $f(x_1) \geq f(x_0)$. In particular, x_0 must be a global minimizer.

A.2 Proof of Proposition 4

First, assume f is convex. Then for all $t \in (0,1)$, we have

$$f((1-t)x + ty) = f(x + t(y-x)) \leq (1-t)f(x) + t f(y),$$

and rearranging by subtracting $f(x)$ from each side and dividing by $t > 0$ yields

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

Taking $t \downarrow 0$ on the left hand side and noting that $\lim_{t \downarrow 0} \frac{f(x+t(y-x))-f(x)}{t} = \nabla f(x)^T(y-x)$ gives inequality (6). Conversely, assume that inequality (6) holds; we show that f is convex. Indeed, we have

$$f(x) \geq f((1-t)x+ty)+\nabla f((1-t)x+ty)^T(x-((1-t)x-ty)) = f((1-t)x+ty)+t\nabla f((1-t)x+ty)^T(x-y)$$

for all $t \in [0,1]$, and similarly

$$f(y) \geq f((1-t)x+ty) + (1-t)\nabla f((1-t)x+ty)^T(y-x).$$

Multiplying the first equation by $(1-t)$ and the second by t, then adding gives

$$(1-t)f(x) + tf(y) \geq f((1-t)x+ty),$$

valid for all $t \in [0,1]$, so that f is convex.

The argument for the second claim is immediate from the first-order inequality (6). First, suppose that $\nabla f(x) = 0$; then $f(y) \geq f(x) + 0 = f(x)$ for all y, and x is a global minimizer of f. Conversely, if $f(y) \geq f(x)$ for all y, then using the definition of derivatives we have $0 \leq \frac{f(x+tv)-f(x)}{t}$ for all $t > 0$ and vectors v. Taking the limit as $t \downarrow 0$ yields

$$0 \leq \lim_{t \downarrow 0} \frac{f(x+tv)-f(x)}{t} = \nabla f(x)^T v.$$

As v was arbitrary, we have $\nabla f(x) = 0$.

Finally, we turn to the claim about the Hessian. Assume that f is convex. Fix any $x, y \in \mathbb{R}^n$, and note that

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) \quad \text{and} \quad f(x) \geq f(y) + \nabla f(y)^T(x-y),$$

so that adding these yields

$$(\nabla f(y) - \nabla f(x))^T(y-x) \geq 0$$

for all x, y. Fix an arbitrary vector v. Then by setting $y = x + tv$ and taking $t > 0$ to zero, we find that

$$0 \leq \frac{1}{t^2}\left[(\nabla f(x+tv) - \nabla f(x))^T(tv)\right] = \frac{1}{t}\left[(\nabla f(x+tv) - \nabla f(x))^T v\right] \to v^T \nabla^2 f(x) v.$$

As x, v are arbitrary, we have $\nabla^2 f(x) \succeq 0$. For the converse, assume that $\nabla^2 f(x) \succeq 0$ for all x. Then using Taylor's remainder theorem that

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}\underbrace{(y-x)^T \nabla^2 f(\widetilde{x})(y-x)}_{\geq 0}$$

for some $\widetilde{x} = tx + (1-t)y$, where $t \in [0,1]$, we have $f(y) \geq f(x) + \nabla f(x)^T(y-x)$.

References

[1] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx/, 2011.

[4] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.