13. Multivariate Differentiation

To describe optimization theory for multivariate functions we need a good theory of multivariate differentiation. A good description of the classical theory for undergraduates is given in Chapter 3 of Fleming [3]. A more extensive and thorough treatment, including important counterexamples is given in Chapter 3 of Ortega and Rheinboldt [4]. Only the essential definitions and results needed for this course will be described here.

Throughout the next few sections, $U$ is an open set in $\mathbb{R}^n$, $x^{(0)} \in U$ and $f : U \to \mathbb{R}$ is assumed to be a continuous function. A direction in $\mathbb{R}^n$ is a unit vector in $\mathbb{R}^n$. The set of all directions is the Euclidean unit sphere $S_1 := \{ x \in \mathbb{R}^n : \| x \|_2 = 1 \}$.

We say that $f$ has a derivative at $x^{(0)} \in U$ in the direction $h$, provided there is a real number $d$ such that

$$
\lim_{t \to 0^+} t^{-1} [f(x^{(0)} + th) - f(x^{(0)})] = d. \tag{13.1}
$$

When $h = e^{(j)}$, this derivative is called the $j$-th partial derivative of $f$ at $x^{(0)}$ and will be denoted $D_j f(x^{(0)})$. Note that, for our purposes here, derivatives are just "one-sided" derivatives. That is, they generalize $D_+$ and $D_-$ from the one-variable case.

$f$ is said to be Gateaux-differentiable at $x^{(0)}$ provided there is a vector $v \in \mathbb{R}^n$ such that

$$
\lim_{t \to 0^+} t^{-1} \left[ (f(x^{(0)} + th) - f(x^{(0)})) - v \cdot h \right] = 0 \quad \text{for all } h \in S_1. \tag{13.2}
$$

When this holds, $\nabla f(x^{(0)}) := v$ is called the gradient of $f$ at $x^{(0)}$. The $j$th component of $\nabla f(x^{(0)})$ is the $j$-th partial derivative $D_j f(x^{(0)})$. Note that if this holds, then the derivative may be taken in "all-directions" as $h$ is an arbitrary (direction)-vector in $S_1$.

Functions of the form $\varphi(t) := f(x^{(0)} + th)$ defined on an interval $(-\delta_1, \delta_2)$ which includes 0 will be used quite often. When $f$ is G-differentiable at $x^{(0)}$ then, $\varphi$ is differentiable at 0 for any $h \in S_1$ and from (13.2)

$$
\varphi'(0) = \nabla f(x^{(0)}) \cdot h. \tag{13.3}
$$

As usual the graph of the function $f$ is the set $G(f) := \{(x, f(x)) : x \in U\}$. When $f$ is G-differentiable at $x^{(0)}$, then the tangent hyperplane to this graph at $x^{(0)}$ is the graph of the function

$$
z = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle. \tag{13.4}
$$

The graph of $z = l(x) := f(x^{(0)}) + \langle a, x - x^{(0)} \rangle$ is said to be a support hyperplane for the (graph of the) function $f$ at $x^{(0)}$ provided $f(x) \geq l(x)$ for all $x \in U$. 

A vector \( \hat{x} \) in \( U \) is said to be a critical point of \( f \) provided \( f \) is G-differentiable at \( \hat{x} \) and \( \nabla f(\hat{x}) = 0 \in \mathbb{R}^n \).

The properties of G-derivatives include linearity and a product rule. However a function may be G-differentiable at a point \( x^{(0)} \) without being continuous there and the general chain rule does not hold. See page 61ff of [3] for discussions of this.

Let \( F : U \to \mathbb{R}^m \) be a continuous vector valued function on the open set \( U \). \( F \) is said to be \textit{G}-differentiable at \( x^{(0)} \) provided there is a \( m \times n \) matrix \( J \), such that

\[
\lim_{t \to 0^+} \| t^{-1}[F(x^{(0)} + th) - F(x^{(0)})] - Jh \| = 0 \quad \text{for all } h \in S_1.
\]

(13.5)

Here any norm on \( \mathbb{R}^m \) may be used. When this holds, we write \( DF(x^{(0)}) := J \) and this matrix is called the Jacobian of \( F \) at \( x^{(0)} \). Suppose

\[
F(x) = \begin{pmatrix}
F_1(x_1, x_2, \ldots, x_n) \\
F_2(x_1, x_2, \ldots, x_n) \\
\vdots \\
F_m(x_1, x_2, \ldots, x_n)
\end{pmatrix}
\]

(13.6)

When \( F \) is differentiable at \( x^{(0)} \), then each \( D_k F_j(x^{(0)}) \) is finite and

\[
DF(x^{(0)}) = \begin{pmatrix}
D_1 F_1(x^{(0)}) & D_2 F_1(x^{(0)}) & \cdots & D_n F_1(x^{(0)}) \\
D_1 F_2(x^{(0)}) & D_2 F_2(x^{(0)}) & \cdots & D_n F_2(x^{(0)}) \\
\vdots & \vdots & \ddots & \vdots \\
D_1 F_m(x^{(0)}) & D_2 F_m(x^{(0)}) & \cdots & D_n F_m(x^{(0)})
\end{pmatrix}.
\]

(13.7)

For the particular case when \( F(x) = \nabla f(x) = (D_1 f(x), \ldots, D_n f(x))^T \) this derivative matrix is called the Hessian of \( f \) and

\[
DF(x) = D^2 f(x) = \begin{pmatrix}
D_{11} f(x) & D_{21} f(x) & \cdots & D_{n1} f(x) \\
D_{12} f(x) & D_{22} f(x) & \cdots & D_{n2} f(x) \\
\vdots & \vdots & \ddots & \vdots \\
D_{1n} f(x) & D_{2n} f(x) & \cdots & D_{nn} f(x)
\end{pmatrix}
\]

(13.8)

is an \( n \times n \) matrix. The entry \( D_{jk} f(x) := D_j(D_k f)(x) \) is the \( j \)-th partial derivative of the function \( D_k f \).

In this case equation (13.5) becomes

\[
\lim_{t \to 0^+} \| t^{-1}[\nabla f(x^{(0)} + th) - \nabla f(x^{(0)})] - D^2 f(x^{(0)}) h \| = 0 \quad \text{for all } h \in S_1.
\]

(13.9)

Take inner products with \( h \) here, then

\[
t^{-1}[\nabla f(x^{(0)} + th) - \nabla f(x^{(0)})] \cdot h \rightarrow (D^2 f(x^{(0)}) h, h)
\]

as \( t \to 0 \) since (13.9) holds for both \( h, -h \).

When \( \varphi(t) := f(x^{(0)} + th) \) is substituted here, this and (13.3) yield that

\[
\varphi''(0) = \langle D^2 f(x^{(0)}) h, h \rangle.
\]

(13.10)
Another result from multivariate calculus is that when \( f \) is twice continuously differentiable on a neighborhood of \( x \), then \( D^2 f(x) \) will be a symmetric \( n \times n \) matrix and
\[
D_{jk} f(x) = D_{kj} f(x) \quad \text{for all } j, k \in I_n
\] (13.11)
See theorem 3.3 of Fleming [3] for a proof. Sections 3.5 and 3.6 of [3] also provide a different version of the following material.

**Exercises.**

**Exercise 13.1** Suppose \( r : \mathbb{R}^3 \to [0, \infty) \) is the Euclidean radius function \( r(x) := \|x\|_2 \). Let \( \psi : [0, \infty) \to \mathbb{R} \) be a continuous function that is \( C^2 \) on \((0, \infty)\) and let \( f(x) := \psi(r(x)) \).

(a) Evaluate \( \nabla r(x) \) and \( D^2 r(x) \) for \( x \in \mathbb{R}^3 \setminus \{0\} \).

(b) Find formulae for \( \nabla f(x) \), \( D^2 f(x) \) in terms of derivatives of \( \psi \) for \( x \in \mathbb{R}^3 \setminus \{0\} \).

(c) Prove that \( f(x) := r(x)^p \) is convex on \( \mathbb{R}^3 \) for \( 1 \leq p < \infty \) and evaluate this \( \nabla f(x) \).

**Exercise 13.2**  
(a) Suppose \( f(x) := \|x\|^p_0 \) with \( p \in (1, \infty) \). Find a formula for the gradient of this function. When is this function \( C^1 \) on \( \mathbb{R}^n \)?

(b) Suppose \( f(x) := \|x\|_1 \) with \( x \in \mathbb{R}^3 \). Describe the set of points where this function is not G-differentiable. Find an explicit expression for the gradient of this function at points where it is differentiable.

**Exercise 13.3** A real valued function \( f \) on \( \mathbb{R}^n \) is said to be homogeneous of degree \( p \) if \( f(cx) = |c|^p f(x) \) for all \( c \in \mathbb{R} \) and \( x \in \mathbb{R}^n \). Suppose \( f \) is also G-differentiable, prove Euler’s rule that
\[
\langle \nabla f(x), x \rangle = pf(x) \quad \text{for all } x \in \mathbb{R}^n.
\]

14. **Multivariate Minimization**

Just as for one dimensional optimization, there are necessary, and also sufficient, conditions for a point in \( U \) to be a local minimizer of \( f \) on \( U \). The necessary conditions may be expressed in terms or either the lower variation of \( f \) or, when \( f \) is differentiable, in terms of the derivative of \( f \).

The lower variation of \( f \) at \( x^{(0)} \) in the direction (ie a unit vector) \( h \) is
\[
\delta_t f(x^{(0)}, h) := \lim_{t \to 0^+} t^{-1} [f(x^{(0)} + th) - f(x^{(0)})].
\] (14.1)

This right hand side always exists and may be \( \pm \infty \). If \( f \) has a derivative \( d \) at \( x^{(0)} \) in the direction \( h \), then \( \delta_t f(x^{(0)}, h) = d \).

**Theorem 14.1.** (First Order Necessary conditions) Suppose \( f : U \to \mathbb{R} \) is continuous \( \hat{x} \) is a local minimizer of \( f \) on \( U \), then

(i) \( \delta_t f(\hat{x}, h) \geq 0 \) for all \( h \in S_1 \), and

(ii) if \( f \) is G-differentiable at \( \hat{x} \), then \( \nabla f(\hat{x}) = 0 \). (F)
Proof. (i) If \( \hat{x} \) is a local minimizer, then \( f(\hat{x} + th) \geq f(\hat{x}) \) for all \( h \in S_1 \) and \( t \) small enough. Thus
\[
t^{-1} [f(\hat{x} + th) - f(\hat{x})] \geq 0 \quad \text{for} \quad 0 < t < \delta(h)
\]
so
\[
\liminf_{t \to 0^+} t^{-1}[f(\hat{x} + th) - f(\hat{x})] \geq 0 \quad \text{or (i) holds.}
\]
(ii) If \( f \) is G-differentiable at \( \hat{x} \), then part (i) implies that \( \nabla f(\hat{x}) \cdot h \geq 0 \) for each \( h \in S_1 \). Take \( h = \pm e(j) \), then
\[
D_j f(\hat{x}) \geq 0 \quad \text{and} \quad D_j f(\hat{x}) \leq 0, \quad \text{so} \quad D_j F(\hat{x}) = 0
\]
for each \( j \in I_n \).

Note that (ii) here does not require that \( f \) be differentiable at any point except \( \hat{x} \).

In view of theorem 14.1, the only possible minimizers of a G-differentiable function \( f \) on an open set \( U \) occur at critical points of \( f \) on \( U \). For \( n \geq 2 \) the level sets of \( f \) near a critical point \( \hat{x} \) may be very complicated in general. A critical point \( \hat{x} \) of \( f \) on \( U \) may be either a local minimizer, a saddle point or a local maximizer of \( f \) on \( U \) - just as for univariate functions.

Also we say that a critical point \( \hat{x} \) of \( f \) is a degenerate critical point provided either
(i) \( D^2 f(\hat{x}) \) does not exist, or else
(ii) the matrix \( D^2 f(\hat{x}) \) is singular.

When \( D^2 f(\hat{x}) \) exists and is a non-singular matrix then \( \hat{x} \) is said to be a non-degenerate critical point of \( f \).

When \( f \) is twice differentiable at the critical point with a non-singular Jacobian, however, a number of results may be proved.

Theorem 14.2. (2nd order Necessary condition) Suppose \( \hat{x} \) is a local minimizer of \( f \) on \( U \) and \( f \) is \( C^1 \) on an open neighborhood of \( \hat{x} \) in \( U \). If \( D^2 f(\hat{x}) \) exists, then
\[
\langle D^2 f(\hat{x})h, h \rangle \geq 0 \quad \text{for all} \quad h \in \mathbb{R}^n.
\] (14.2)

Proof. Put \( \varphi(t) = f(\hat{x} + th) \) for some \( h \in S_1 \). \( \hat{x} \) is a local minimizer of \( f \) implies that 0 is a local minimizer of \( \varphi \). From theorem 7.2, if \( \varphi''(0) \) exists, then it is \( \geq 0 \). The analysis of the second G-derivative above yields that \( \varphi''(0) = \langle D^2 f(\hat{x})h, h \rangle \) so (14.2) follows.

Theorem 14.3. (Sufficient Condition) Suppose \( f \) is \( C^1 \) on a neighborhood of a critical point \( \hat{x} \), \( D^2 f(\hat{x}) \) exists and there is a \( c_1 > 0 \) such that
\[
\langle D^2 f(\hat{x})h, h \rangle \geq c_1 ||h||^2 \quad \text{for all} \quad h \in \mathbb{R}^n.
\] (14.3)

Then \( \hat{x} \) is an isolated, strict local minimizer of \( f \) on \( U \).

Proof. Choose \( \varphi \) as above, then (14.3) implies
\[
\varphi''(0) \geq c_1 ||h||^2 > 0
\]
Just as in the proof of theorem 7.3, when \( \hat{x} \) is a critical point of \( f \),
\[
\varphi(t) - \varphi(0) \geq c_1 t^2 / 4 \quad \text{since} \quad |\varphi'(t) - \varphi'(0)| \geq \frac{c_1}{2} |t| \quad \text{for} \quad 0 < |t| < \delta.
\]
Thus \( \varphi(t) - \varphi(0) = \int_0^t \varphi'(\tau)d\tau \geq \frac{c_1}{4}t^2 \) for \( t > 0 \).

Similarly when \( t < 0 \), so \( f(\hat{x} + th) - f(\hat{x}) \geq c_1t^2/4 \) for each direction \( h \in S_1 \) and \( 0 < |t| < \delta \). Thus \( \hat{x} \) is a strict local minimizer as claimed. Since \( \nabla f(\hat{x}) = 0 \) and \( \langle D^2 f(\hat{x})h, h \rangle \geq c_1||h||^2 \), then \( \hat{x} \) is an isolated critical point. \( \square \)

Comments 1. The necessary and sufficient conditions for local maximizers require the reverse inequalities in (14.2) or (14.3).

Exercises.

Exercise 14.1 Define a function \( G : \mathbb{R}^n \to \mathbb{R} \) by
\[
G(x) := \|x\|^4 - 2 \langle Ax, x \rangle
\]
where \( A \) is an \( n \times n \) symmetric matrix. Prove
(i) Prove that this function is bounded below and has minimizers on \( \mathbb{R}^n \).
(ii) Find the equation satisfied by the critical points of \( G \) on \( \mathbb{R}^n \).
(iii) What mathematical properties can you say about the critical points and/or minimizers of \( G \)? What can you say about the value of this problem?


Let \( C \) be a non-empty, open convex set in \( \mathbb{R}^n \) and \( f : C \to \mathbb{R} \) be \( G \)-differentiable at each point in \( C \). Then \( \nabla f : C \to \mathbb{R}^n \) is defined. The following results provide differential criteria for the function to be convex. Another treatment of this material is in section 3.6 of Fleming [3]. Chapter III, Section 3 of Berkowitz [1] has some more details and these results are extensions of the 1-dimensional results described in section 7.

Theorem 15.1. Suppose \( f, C \) as above, then \( f \) is convex on \( C \) if and only if
\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \text{for all } x, y \in C.
\]

Proof. Suppose \( f \) is convex, \( x, y \in C, x \neq y \) and \( x(t) := (1-t)x + ty \) with \( 0 < t < 1 \). Then
\[
f(x(t)) \leq (1-t)f(x) + tf(y)
\]
Rearranging this with \( 0 < t \leq 1 \) leads to
\[
t^{-1}[f(x(t)) - f(x)] \leq f(y) - f(x).
\]
Take limits as \( t \to 0^+ \), then
\[
\langle \nabla f(x), y - x \rangle \leq f(y) - f(x)
\]
so (15.1) holds.

Conversely when (15.1) holds, \( x, y \in C, x \neq y \), let \( z := (1-t)x + ty \) with \( 0 < t < 1 \). Substitute \( z \) for \( x \) in (15.1), then
\[
f(x) \geq f(z) + \langle \nabla f(z), x - z \rangle \quad \text{and } f(y) \geq f(z) + \langle \nabla f(z), y - z \rangle.
\]
Multiply first equation by \((1 - t)\), second by \(t\) and add to find that the convexity inequality (9.3) holds.

**Corollary 15.2.** Suppose \(z = l(x) = f(x^{(0)}) + \langle a, x - x^{(0)} \rangle\) is a support hyperplane for the graph of \(f\) at \(x^{(0)} \in C\). If \(f\) is convex on \(C\), then \(f(y) \geq l(y)\) for all \(y \in C\).

**Proof.**

The following only requires a simple modification of this proof.

**Corollary 15.3.** Suppose \(f, C\) as above, then \(f\) is strictly convex on \(C\) if and only if strict inequality holds in (15.1) when \(x \neq y\).

The next two theorems give the conditions that are usually used to check whether a particular differentiable function is convex on \(C\).

**Theorem 15.4.** Suppose \(f, C\) as above, then \(f\) is convex on \(C\) if and only if

\[
\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0 \quad \text{for all } x, y, \in C. \tag{15.2}
\]

It is strictly convex on \(C\) if strict inequality holds here when \(y \neq x\).

**Proof.** Suppose \(f\) is convex, \(x, y \in C, x \neq y\) and let \(\varphi(t) := f(x(t))\). Then \(\varphi\) is a convex differentiable function of \(t\) and \(\varphi'(t) = \langle \nabla f(x(t)), y - x \rangle\) from (13.3). From 1-d theory, \(\varphi'(1) \geq \varphi'(0)\) so (15.2) holds.

Conversely, when (15.2) holds, then \(\varphi'(t) \geq \varphi'(0)\) for all \(t > 0\). Then

\[
\varphi(t) = \varphi(0) + \int_0^t \varphi'(s) \, ds \geq \varphi(0) + \varphi'(0)t.
\]

That is

\[
f(x(t)) \geq f(x) + t \langle \nabla f(x), y - x \rangle.
\]

This implies that (15.1) holds, so \(f\) is convex on \(C\). The strictness part is similar.

When (15.2) holds, then \(\nabla f(x)\) is said to be a *monotone* mapping of \(C\) into \(\mathbb{R}^n\).

The basic results about 1-dimensional minimization and maximization of convex functions on a interval were summarized previously in theorem 9.1. Here the corresponding results for multivariate convex functions \(f\) on \(C\) will be given.

A general existence result is the following.

**Theorem 15.5.** Suppose \(C\) is a nonempty compact convex set in \(\mathbb{R}^n\) and \(f : C \to \mathbb{R}\) is lower semi-continuous and quasi-convex. Then

(i) the set of all minimizers of \(f\) on \(C\) is a nonempty closed convex subset of \(C\).

(ii) If \(f\) is strictly convex on \(C\), then this set consists of exactly one point.

**Proof.** The existence follows from Weierstrass’ theorem 5.1. The fact that each synoptic set is convex and closed means that the set \(S_c(f)\) is convex and closed so the set of all minimizers of \(f\) on \(C\) is this set with \(c = \alpha(f, C)\) so (i) holds. (ii) is the standard argument.
A very useful version of this is the following “unconstrained” version. Its proof just uses the same argument as in the 1 variable case.

**Theorem 15.6.** Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) is proper, lower semi-continuous, quasi-convex and weakly coercive. Then \( \alpha(f) := \inf_{x \in \mathbb{R}^n} f(x) \) is finite and the set of minimizers of \( f \) on \( \mathbb{R}^n \) is non-empty, bounded, closed and convex.

The following theorem says that, when \( f \) is differentiable and convex on an open convex set \( C \) then the only critical points of \( f \) on \( C \) are the minimizers of \( f \) on \( C \). Alternatively being a critical point of \( f \) is a necessary and sufficient condition to be a minimizer - on an open set - when \( f \) is convex and differentiable everywhere.

**Theorem 15.7.** Suppose \( C \) is a non-empty open convex set in \( \mathbb{R}^n \) and \( f : C \to \mathbb{R} \) is convex and G-differentiable on \( C \). A vector \( \hat{x} \in C \) minimizes \( f \) on \( C \) if and only if \( \hat{x} \) is a critical point of \( f \).

**Proof.** When \( \hat{x} \) is a critical point of \( f \), then (15.1) implies that \( f(y) \geq f(\hat{x}) \) for all \( y \in C \). Thus \( \hat{x} \) minimizes \( f \) on \( C \).

Conversely, if \( \hat{x} \) minimizes \( f \) on \( C \), then \( f(\hat{x} + td) \geq f(\hat{x}) \) for all \( t > 0, d \in S_1 \). Thus \( \langle \nabla f(\hat{x}), d \rangle \geq 0 \) for all \( d \in S_1 \). This implies \( \hat{x} \) is a critical point of \( f \).

A number of important mathematical problems can be written as convex optimization problems. The important questions about each of them usually include

(i) does the optimization problem have a finite value and a minimizer?
(ii) what equations do the (local) minimizers satisfy? and
(iii) how can we find the minimizers and/or critical points?

To answer (i) we usually show that \( f \) is continuous (or l.s.c.) and coercive - or that some synoptic set is non-empty, closed and bounded. To answer (ii) just find the G-derivatives while (iii) often involves developing algorithms for finding the minimizers.

**Exercises.**

Exercise 15.1 Suppose \( S := \{a^{(1)}, \ldots, a^{(m)}\} \) is a finite set of distinct points in space and \( F : \mathbb{R}^3 \to [0, \infty) \) is defined by

\[
F(x) := \sum_{j=1}^{m} c_j \|x - a^{(j)}\|^2_2
\]

with each \( c_j > 0 \).

(a) Show that \( F \) is convex and coercive and that there is a unique minimizer of this function.
(b) What equations hold at the minimizer of this problem?
(c) Suppose \( S = \{0, e^{(1)}, e^{(2)}, e^{(3)}\} \) and each \( c_j = 1 \). Find the solution of this problem.

Exercise 15.2 Suppose \( \psi : [0, \infty)^n \to \mathbb{R}_+ \) is a continuous, convex function with \( \psi(0) = 0 \) and \( \psi(x) > 0 \) for \( x \neq 0 \).
(a) Show that there are constants $c_2 \geq c_1 > 0$ such that

\[ \psi(x) \geq c_1 \|x\|_1 \quad \text{and} \quad \psi(x) \geq c_2 \|x\|_\infty \quad \text{when} \quad \|x\|_\infty \geq 1. \]  

(15.3)

and that $\psi$ is weakly coercive on $[0, \infty)^n$.

(b) Let $A : \mathbb{R}^n \to [0, \infty)^n$ be the map defined by $A(x) := (|x_1|, |x_2|, \ldots, |x_n|)$. Define $p : \mathbb{R}^n \to [0, \infty)$ by

\[ p(x) := \inf \{ s > 0 : \psi(A(x)/s) \leq 1 \} \]

Show that $p$ is a norm on $\mathbb{R}^n$.

(c) Take $n = 2$ and $\psi(x_1, x_2) := e^{x_1} + e^{x_2} - 2$. Find the equation satisfied by the set $B$ of vectors $x \in \mathbb{R}^2$ that obey $p(x) \leq 1$ with $p$ as in (b). Is $B$ a subset of the unit ball with respect to the $\infty$-norm? Is the unit ball with respect to the $\infty$-norm a subset of $B$?

16. Matrices, Norms and Quadratic Forms.

Let $A := (a_{jk})$ be an $m \times n$ real matrix. Its transpose is the $n \times m$ real matrix with $A^T := (a_{kj})$. That is the rows and columns of $A$ are interchanged. The set of all $m \times n$ real matrices will be denoted by $M_{mn}$ and is a real vector space. For $p \in [1, \infty]$, the $p$-norm of a matrix $A$ is defined by

\[ \|A\|_p := \sup_{\|x\|_p = 1} \|Ax\|_p. \]  

(16.1)

It is a good exercise (see Ex 16.1 below) to prove that this is a norm on $M_{mn}$. In general there is no explicit formula for this norm in terms of the entries in $A$ - even when $p = 2$. It is a number defined as the maximum of a convex function on the (non-convex) compact set $S_{1p}$ in $\mathbb{R}^n$. This can be changed to a convex domain by modifying the constraint set to the convex hull, namely $B_{1p} := \{ x : \|x\| \leq 1 \}$. (Why are the maximum values the same?)

When $m = n$, the set of all $n \times n$ real matrices will be denoted $M_n$ and is a (non-commutative) algebra. An $n \times n$ matrix $A$ is said to be symmetric if $A^T = A$, it is skew-symmetric if $A^T = -A$. The quadratic form associated with a square matrix $A$ is the function $q : \mathbb{R}^n \to \mathbb{R}$ defined by

\[ q(x) := \langle Ax, x \rangle := \sum_{j,k=1}^n a_{jk} x_j x_k. \]  

(16.2)

It is straightforward to evaluate the partial derivatives of $q$ and find that

\[ \nabla q(x) = (A + A^T)x \quad \text{and also} \quad D^2 q(x) \equiv A + A^T \]  

(16.3)

with $A^T$ being the transpose matrix of $A$. When $A$ is a skew-symmetric matrix then $\nabla q(x) \equiv 0$ on $\mathbb{R}^n$ so $q(x) \equiv 0$ on $\mathbb{R}^n$.

The symmetric part of $A$ is the matrix $A_S := (A + A^T)/2$, so (16.3) becomes

\[ \nabla q(x) = 2A_S x, \quad D^2 q(x) \equiv 2A_S. \]
Observe that the quadratic forms associated with $A$ and $A_S$ are identical. That is $q(x) \equiv q_S(x)$ where $q_S(x) := \langle A_s, x, x \rangle$. So, without loss of generality, $A$ may be assumed to be real symmetric.

A real symmetric $n \times n$ matrix $A$ is defined to be positive semi-definite (p.s.d.) if $q(x) \geq 0$ for all $x \in \mathbb{R}^n$. It is positive definite (p.d.) if $q(x) > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.

The following theorem says that the convexity of $q$ is determined by the positivity, or otherwise, of $q$ on $\mathbb{R}^n$.

**Theorem 16.1.** Let $A$ be a real symmetric $n \times n$ matrix. The function $q$ defined by (16.2) is convex if and only if $A$ is positive semi-definite. $q$ is strictly convex if and only if $A$ is a positive definite matrix.

**Proof.** From theorem 15.1 $q$ will be convex on $\mathbb{R}^n$ if and only if

$$q(y) \geq q(x) + 2 \langle Ax, y - x \rangle$$

for all $x, y, \in \mathbb{R}^n$.

Substitute for $q$, then this holds if and only if

$$\langle A(y - x), y - x \rangle \geq 0$$

for all $x, y, \in \mathbb{R}^n$.

Write $z = y - x$, then this becomes $\langle Az, z \rangle \geq 0$ for all $z \in \mathbb{R}^n$ - or the matrix $A$ is p.s.d.

From the last part of theorem 15.4, the function $q$ will be strictly convex on $\mathbb{R}^n$ provided $\langle Az, z \rangle > 0$ for all $z \in \mathbb{R}^n \setminus \{0\}$. □

Consider the bilinear form $B : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$[x, y]_A := \langle Ax, y \rangle.$$ (16.4)

If $A$ is positive definite then this bilinear form will actually be an inner product on $\mathbb{R}^n$ and the following holds.

**Corollary 16.2.** Suppose $A$ be a positive definite symmetric $n \times n$ matrix. Then (16.4) defines an inner product on $\mathbb{R}^n$ and there are constants $0 < c_1 \leq c_2$ such that

$$c_1 \|x\|^2_2 \leq [x, x]_A = q(x) \leq c_2 \|x\|^2_2$$

for all $x \in \mathbb{R}^n$. (16.5)

In particular $q$ is strictly convex and coercive on $\mathbb{R}^n$.

**Proof.** To show that (16.4) defines an inner product we need only verify that $[x, x]_A \geq 0$ and equality here holds only when $x = 0$. This is just the requirement that $A$ be p.d. The quadratic form $q$ is a continuous function on $\mathbb{R}^n$. Consider the problem of minimizing, or maximizing, $q$ on the unit sphere $S_1$. These minimum and maximum values exist and are finite from Weierstrass’ theorem as $S_1$ is compact. Let them be $c_1, c_2$ respectively. One sees that $c_1 > 0$ as otherwise the matrix $A$ would not be p.d.

(16.5) holds for $x = 0$. Suppose $x \neq 0$ and define $e := x/\|x\|$. Then $c_1 \leq [e, e]_A \leq c_2$ from the definitions of $c_1, c_2$. Multiply both sides by $\|x\|^2$ and (16.5) follows. □

Essentially this says that when $A$ is a positive definite symmetric matrix, this inner product (16.4) and the corresponding norm are "equivalent" to the Euclidean inner product and norm.
Suppose now that \( f : C \to \mathbb{R} \) is twice continuously differentiable \((C^2-\) on \( C \) and define \( \varphi(t) := f(x(t)) \) - as in the proof of theorem 15.4. Then
\[
\varphi'(t) = \langle \nabla f(x(t)), y - x \rangle \quad \text{and} \quad \varphi''(t) = \langle D^2 f(x(t))(y - x), (y - x) \rangle.
\]
From the 1-d Taylor’s theorem for \( \varphi \), one has
\[
f(y) = f(x) + \langle \nabla f(x), y - x \rangle + (1/2)\langle D^2 f(x(\tau))(y - x), (y - x) \rangle
\]
for some \( \tau \in (0, 1) \).

This leads to the following second derivative criterion for convexity of a function. It is theorem 3.3 of Berkowitz chapter III - and a proof is given there.

**Theorem 16.3.** Suppose \( f, C \) as above with \( f \) of class \( C^2- \) on \( C \). Then \( f \) is convex on \( C \) if and only if \( D^2 f(x) \) is p.s.d on \( C \). If \( D^2 f(x) \) is p.d. on \( C \), then \( f \) is strictly convex on \( C \).

### 16. Exercises.

Exercise 16.1: Given \( p \in [1, \infty] \), prove that the function \( f_p : M_{mn} \to [0, \infty) \) defined by \( f_p(A) := \|A\|_p \) is both a norm and a convex function on \( M_{mn} \).

Exercise 16.2: (a) Show that if \( A \in M_{mn} \), then
\[
\|A\|_1 = \sup_{\|x\|_1 \leq 1} \sup_{\|y\|_1 \leq 1} \langle Ax, y \rangle.
\]
(b) Find an explicit formula for the 1-norm of the matrix \( A \).
(c) Find conditions on the entries in \( A \) for \( \|A^T\|_1 = \|A\|_1 \).

Exercise 16.3: (a) Suppose \( A \in M_{mn} \) show that
\[
\|A\|_2 = \sup_{\|x\|_2 \leq 1} \sup_{\|y\|_2 \leq 1} \langle Ax, y \rangle, \quad \text{and}
\]
\[
\|A^T\|_2 = \|A\|_2.
\]
(b) \( \|A^T\|_2 = \|A\|_2 \).

Exercise 16.4: Suppose \( A, B \) are \( n \times n \) matrices and \( p \in [1, \infty] \), prove that
\[
\|AB\|_p \leq \|A\|_p \|B\|_p.
\]

Exercise 16.5: Let \( A \) be an \( n \times n \) skew-symmetric matrix. Show that the quadratic form \( q(x) := \langle Ax, x \rangle \) is identically zero.

### 17. Energy Principles for Linear Equations

This section describes an optimization problem that yields existence, uniqueness and solvability criteria for linear equations involving a symmetric \( n \times n \) matrices. Suppose we want to find solutions of the system
\[
Ax = b \quad \text{(LEq)}
\]
where $A$ is an $n \times n$ symmetric matrix. Define $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{E}(x) := \langle Ax, x \rangle - 2 \langle b, x \rangle = \sum_{j,k=1}^n a_{jk}x_jx_k - 2 \sum_{j=1}^n b_jx_j \quad (17.1)$$

The energy principle for (LEq) is the problem of finding the minimizers of $\mathcal{E}$ on $\mathbb{R}^n$ and

$$\alpha(\mathcal{E}) := \inf_{x \in \mathbb{R}^n} \mathcal{E}(x).$$

This function is continuous and has gradient $\nabla \mathcal{E}(x) = 2(Ax - b)$, so $\hat{x}$ is a critical point of $\mathcal{E}$ if and only if it is a solution of (LEq). The following result justifies the study of this optimization problem.

**Lemma 17.1.** Assume $A$ is a positive semi-definite symmetric matrix and $\mathcal{E}$ is defined by (17.1). A vector $\hat{x} \in \mathbb{R}^n$ minimizes $\mathcal{E}$ on $\mathbb{R}^n$ if and only if $\hat{x}$ is a solution of (LEq).

**Proof.** When $A$ is p.s.d then the functional $\mathcal{E}$ will be convex and continuously differentiable on $\mathbb{R}^n$ with $\nabla \mathcal{E}(x) = 2(Ax - b)$. Then theorem 15.7 yields that $\hat{x}$ minimizes $\mathcal{E}$ if and only if (LEq) holds. \qed

It is reasonable to ask what happens when $A$ is not positive definite? In the case when $A$ is not p.s.d, there is a $d \in \mathbb{R}^n$ such that $\langle Ad, d \rangle < 0$. Then $\mathcal{E}(td) \to -\infty$ as $t \to \infty$. Thus the value of the problem is $-\infty$ and there is no minimizer of $\mathcal{E}$ The function $\mathcal{E}$ is non-convex as the quadratic part is non-convex from theorem 16.1.

When $A$ is a positive definite symmetric matrix, the results we have obtained lead to the following existence-uniqueness theorem for the linear equation (LEq).

**Theorem 17.2.** Let $A$ be a positive definite symmetric matrix, then there is a unique minimizer $\hat{x}$ of $\mathcal{E}$ on $\mathbb{R}^n$. This $\hat{x}$ is the unique solution of (LEq) and the matrix $A$ has an inverse $A^{-1}$ that is positive definite and symmetric with

$$\|A^{-1}b\|_2 \leq c_1^{-1}\|b\|_2 \quad \text{and} \quad \alpha(\mathcal{E}) = \langle A^{-1}b, b \rangle. \quad (17.2)$$

**Proof.** When $A$ is positive definite then Corollary 16.2 says that there is a $c_1 > 0$ such that

$$\langle Ax, x \rangle \geq c_1 \|x\|_2^2 \quad \text{for all} \quad x \in \mathbb{R}^n.$$  

Thus

$$\mathcal{E}(x) \geq c_1 \|x\|_2^2 - \|b\| \|x\|_2 \quad \text{for all} \quad x \in \mathbb{R}^n$$

so $\mathcal{E}$ is coercive on $\mathbb{R}^n$. From corollary 5.4, there is an $\hat{x} \in \mathbb{R}^n$ that minimizes $\mathcal{E}$ on $\mathbb{R}^n$. When $A$ is p.d., then $\mathcal{E}$ is strictly convex from theorem 16.1, and theorem 15.5 implies that this minimizer is unique. The previous lemma 17.1 yields that $\hat{x}$ will be the unique solution of (LEq).

Choose $b = e^{(j)}$, and denote the unique minimizer of the associated $\mathcal{E}$ to be $\xi^{(j)}$ for $j = 1, 2, \ldots n$ respectively. Let $A^{-1}$ be the matrix whose column vectors are these $\xi^{(j)}$. It is straightforward to verify that the solution of (LEq) will be $\hat{x} = A^{-1}b$ with this specification of $A^{-1}$. Suppose (LEq) holds and $A\hat{x} = d$, then

$$\langle A^{-1}b, d \rangle = \langle \hat{x}, d \rangle = \langle \hat{x}, A\hat{x} \rangle = \langle A\hat{x}, \hat{x} \rangle$$
using the definitions and the fact that $A$ is symmetric. The last inner product equals $\langle b, A^{-1}d \rangle$ so $A^{-1}$ is symmetric.

Take inner products of (LEq) with $\hat{x}$, then $\langle A\hat{x}, \hat{x} \rangle = \langle b, \hat{x} \rangle$. Use the first inequality in (16.5) and Cauchy’s inequality to obtain

$$c_1 \|x\|^2 \leq \|b\| \|\hat{x}\|$$

This yields the first inequality in (17.2). Also observe that

$$\mathcal{E}(\hat{x}) = -\langle b, \hat{x} \rangle = -\langle A^{-1}b, b \rangle$$

so the value of the problem is as claimed. When $b \neq 0$, then this value must be negative as if the value is zero then 0 would be a minimizing vector. Hence $A^{-1}$ is positive definite. □

It is worth noting that this is a proof that (LEq) has a solution for any $b \in \mathbb{R}^n$ based on optimization theory and analysis. No non-trivial results from linear algebra have been used - just formulae for derivatives and properties of convexity.

When $A$ is p.s.d but not p.d., then $\mathcal{E}$ will be convex and the a more careful analysis is needed. Let $q(x) := \langle Ax, x \rangle$ be the quadratic form defined as in (16.2) and consider the problem of minimizing $q$ on $\mathbb{R}^n$. The following lemma holds.

**Lemma 17.3.** Suppose $A$ is an $n \times n$ real symmetric matrix that is p.s.d. A vector $\hat{x}$ minimizes $q$ on $\mathbb{R}^n$ if and only if it obeys $Ax = 0$. The set of all such minimizers is a subspace of $\mathbb{R}^n$.

**Proof.** Since $A$ is p.s.d. then $\alpha(q) = 0$ and it is attained. From theorem 16.1, $q$ is convex so theorem 15.7 says that the minimizers of $q$ on $\mathbb{R}^n$ are precisely the solutions of $Ax = 0$. The set of all solutions of this equation is a closed subspace $N(A)$ of $\mathbb{R}^n$. □

This theorem implies that $A$ is positive definite if and only if it is p.s.d. and 0 is not an eigenvalue of $A$. The subspace $N(A)$ is called the null space of $A$. Suppose $A$ is p.s.d and $\dim N(A) = m \geq 1$. Then $N(A)$ has an orthogonal complement $W$ in $\mathbb{R}^n$ which will be a subspace of dimension $n - m$. Each vector $x \in \mathbb{R}^n$ will have a unique decomposition

$$x := y + z \quad \text{where} \ y \in N(A) \ \text{and} \ z \in W \quad (17.3)$$

Substitute this in the definition of $\mathcal{E}$, then

$$\mathcal{E}(x) = \mathcal{E}(z) - 2 \langle b, y \rangle \quad (17.4)$$

This formulation leads to both non-existence and existence results for (LEq). First a solvability (or non-existence) result for the linear equation.

**Theorem 17.4.** Suppose $A$ is a p.s.d. symmetric matrix with $\dim N(A) = m \geq 1$. If there is a $y \in N(A)$ such that $\langle b, y \rangle \neq 0$, then there is no minimizer of $\mathcal{E}$ on $\mathbb{R}^n$ and no solution of (LEq).

**Proof.** If there is such a $y \in N(A)$, then $\mathcal{E}(ty) = -t\langle b, y \rangle$ for all real $t$. Thus $\alpha(\mathcal{E}) = -\infty$ and there is no minimizer of $\mathcal{E}$ on $\mathbb{R}^n$. The energy principle lemma 17.1 then implies there is no solution of (LEq). □
The existence result is the following

**Theorem 17.5. (Solvability Condition)** Suppose $A$ is a p.s.d. symmetric matrix with $\dim N(A) = m \geq 1$. If $\langle b, y \rangle = 0$ for all $y \in N(A)$ then

(i) $\alpha(\mathcal{E})$ is finite and there are minimizers $\bar{x}$ of $\mathcal{E}$ on $\mathbb{R}^n$, and

(ii) the set of all solutions of (LEq) is $\{ \bar{x} + y : y \in N(A) \}$, where $\bar{x}$ is a minimizer of $\mathcal{E}$.

**Proof.** Under these assumptions, (17.4) shows that minimizing $\mathcal{E}$ on $\mathbb{R}^n$ is equivalent to minimizing $\mathcal{E}$ on $W$. Given a nonzero $z \in W$, $q(z)$ is non-zero, so $q$ is positive definite on $W$. Since $q$ is a quadratic form, then Corollary 16.2 shows that $q$, and thus $\mathcal{E}$, is coercive on $W$. Thus $\mathcal{E}$ is bounded below and attains its minimum value on $W$ from theorem 17.2. Thus (i) holds while (ii) holds by linearity.

The last two results constitute a special case (namely when $A$ is p.s.d and symmetric) of the Fredholm alternative for solving (LEq). They say that, for this case, (LEq) has a solution if and only if

$$\langle b, u \rangle = 0, \quad \text{for all } u \in N(A). \quad (17.5)$$

The preceding results may be combined to prove the following theorem which says that whenever an energy function $\mathcal{E}$ is bounded below, then it has minimizers - a result that is quite special to this class of problems.

**Theorem 17.6. (Existence)** Suppose $A$ is an $n \times n$ symmetric matrix and $\mathcal{E}$ is defined by (17.1). If $\mathcal{E}$ is convex and bounded below on $\mathbb{R}^n$, then there is at least one minimizer of $\mathcal{E}$ on $\mathbb{R}^n$. If $\mathcal{E}$ is strictly convex this minimizer is unique.

**Proof.** When $\mathcal{E}$ is convex then $q(x) = \langle Ax, x \rangle$ is convex as it only differs from $\mathcal{E}$ by a linear function. Theorem 16.1 implies that $A$ is p.s.d. When $A$ is positive definite this theorem follows from theorem 17.2. When $A$ is only p.s.d., the assumption that $\mathcal{E}$ is bounded below implies $\langle b, y \rangle = 0$, for all $y \in N(A)$ from theorem 17.4. Then theorem 17.5 implies that there are minimizers of $\mathcal{E}$ on $\mathbb{R}^n$. \[ \square \]

18. **Least Squares Optimization and Linear Equations.**

You might think that equations involving the class of symmetric p.s.d square matrices are quite a special class of equations. Here we shall show that any system of the form (LEq) with $A$ an $m \times n$ matrix may be formulated as such problems. Suppose $A$ is an $m \times n$ matrix, $b \in \mathbb{R}^m$ and we are interested in solving (LEq). When $m < n$, the system (LEq) is said to be an *underdetermined* system of linear equations. When $m > n$, it is an *overdetermined* linear system.

Define the *least squares* function $\mathcal{F} : \mathbb{R}^n \to [0, \infty)$ by

$$\mathcal{F}(x) := ||Ax - b||_2^2 \quad (18.1)$$

The problem of minimizing $\mathcal{F}$ on $\mathbb{R}^n$ is called the *least squares* optimization problem for (LEq).
Obviously a solution \( \hat{x} \) of (LEq) is a minimizer of \( \mathcal{F} \) on \( \mathbb{R}^n \) and then \( \mathcal{F}(\hat{x}) = 0 \). In general the minimizers of \( \mathcal{F} \) may not be solutions of (LEq) - but they can provide important information about the equation. Least squares methods are almost always used for studying overdetermined systems with \( m \geq n \) and are very important in numerical and statistical applications. Much of the fundamental work on least squares methods was done by C.F. Gauss in connection with his work in geodesy and astronomy. Apparently it was the primary topic that he lectured on when he was asked to give university lectures at Goettingen.

From the definition,
\[
\mathcal{F}(x) = \langle Ax - b, Ax - b \rangle = \langle A^T Ax, x \rangle - 2 \langle Ax, b \rangle + ||b||_2^2 \quad (18.2)
\]
So
\[
\mathcal{F}(x) = \mathcal{E}_2(x) + ||b||_2^2
\]
where \( \mathcal{E}_2 \) is an energy function with \( A^T A, A^T b \) replacing \( A \) and \( b \) for the function \( \mathcal{E} \) used in the previous section. Thus this \( \mathcal{F} \) is an example of a convex energy function \( \mathcal{E} \) that is bounded below.

Some basic results about this optimization problem may be summarized as follows

**Theorem 18.1.** Assume \( A \) is an \( m \times n \) matrix, \( b \in \mathbb{R}^m \) and \( \mathcal{F} \) is defined by (18.1), then there are minimizers of \( \mathcal{F} \) on \( \mathbb{R}^n \). A vector \( \tilde{x} \) minimizes \( \mathcal{F} \) on \( \mathbb{R}^n \) if and only if it satisfies the equation
\[
A^T A x = A^T b. \tag{LSeq}
\]
A minimizer of \( \mathcal{F} \) on \( \mathbb{R}^n \) will be a solution of (LEq) if \( N(\mathcal{A}^T) = \{0\} \).

*Proof.* From (18.2), \( \mathcal{E}_2 \) is a convex energy function associated with the p.s.d. matrix \( A^T A \). Since \( \mathcal{F} \) is positive, \( \mathcal{E} \) is bounded below on \( \mathbb{R}^n \) by \(-||b||_2^2\). Thus theorem 17.6 implies there are minimizers of \( \mathcal{F} \) on \( \mathbb{R}^n \).

The derivatives of \( \mathcal{F} \) are given by
\[
\nabla \mathcal{F}(x) = 2 A^T (Ax - b) \quad \text{and} \quad D^2 \mathcal{F}(x) = 2 A^T A. \quad (18.3)
\]
Since \( \mathcal{F} \) is convex, a vector \( \tilde{x} \) minimizes \( \mathcal{F} \) if and only if (LSeq) holds. This equation may be written as \( A^T (Ax - b) = 0 \), so the last sentence holds. \( \square \)

A vector \( x_{LS} \in \mathbb{R}^n \) is a said to be a least squares solution of (LEq) provided it satisfies (LSeq). The theorem says that such vectors are exactly the minimizers of \( \mathcal{F} \) on \( \mathbb{R}^n \). When the minimal value is positive, there is no solution of (LEq). Sometimes least squares solutions are called generalized solutions of (LEq).

Suppose \( N(A) = \{0\} \). Then the symmetric matrix \( A^T A \) is non-singular and positive definite, as \( ||Ax||_2 > 0 \) for non-zero \( x \), so theorem 17.2 applies to \( \mathcal{E}_2 \). For each \( b \in \mathbb{R}^m \), there is a unique minimizer \( x_{LS} \) of \( \mathcal{F} \) on \( \mathbb{R}^n \) and then \( x_{LS} \) is the unique solution of (LSq).

If \( \dim N(A) \geq 1 \), and \( x_{LS} \) is a minimizer of \( \mathcal{F} \) on \( \mathbb{R}^n \), then \( x_{LS} + y \) is again a minimizer for any \( y \in N(A) \) as \( \mathcal{F}(x_{LS} + y) = \mathcal{F}(x_{LS}) \). Thus there is an affine subspace of minimizers of \( \mathcal{F} \) on \( \mathbb{R}^n \) - or the set of all minimizers is
\[
\mathcal{M}(\mathcal{F}) = \{x_{LS}\} + N(A) = \{x_{LS} + y : y \in N(A)\}.
\]
When \( \dim N(A^T) \geq 1 \), then a least squares solution of (LSq) need not be a solution of (LEq). For a general \( m \times n \) matrix \( A \), the fundamental theorem of linear algebra is that

\[
\mathbb{R}^m = N(A^T) \oplus R(A) \quad \text{and} \quad \mathbb{R}^n = N(A) \oplus R(A^T).
\]

See Strang [7], chapter 2, section 4. Moreover the ranks of \( A \) and \( A^T \) are the same, or

\[
\dim R(A) = \dim R(A^T) = r. \quad \text{If} \quad m > n, \text{then}
\]

\[
\dim N(A^T) = m - r > \dim N(A) = n - r.
\]

Thus, for overdetermined systems, the last part of theorem 18.1 shows that the minimizers of \( \mathcal{F} \) need not be solutions of (LEq).

18.1. **Preconditioned Least Squares Problems.**

The previous optimization problem could be regarded as a special case of the following problem for the solutions of (LEq). Again this analysis is primarily used when \( m \geq n \).

Let \( A \) be a \( m \times n \) matrix, \( b \in \mathbb{R}^n \) and \( M \) be a p.d. \( m \times m \) symmetric matrix. Define the **preconditioned least squares** function \( \mathcal{F}_M : \mathbb{R}^n \to [0, \infty) \) by

\[
\mathcal{F}_M(x) := \langle M(Ax - b), Ax - b \rangle.
\]

(18.4)

The **preconditioned least squares optimization problem** is to minimize \( \mathcal{F} \) on \( \mathbb{R}^n \). When \( M := I \), this reduces to the least squares problem described above.

Note that \( \mathcal{F}_M(x) \geq 0 \) so \( \alpha(\mathcal{F}_M) \geq 0 \) and

\[
\mathcal{F}_M(x) = \langle A^TMAx, x \rangle - 2 \langle MAx, b \rangle + \langle Mb, b \rangle.
\]

(18.5)

Thus

\[
\mathcal{F}_M(x) = \mathcal{E}(x) + \langle Mb, b \rangle.
\]

Here \( \mathcal{E} \) as in the energy method but with \( A^TMA, A^TMb \) replacing \( A, b \).

If \( m = n \) and \( A \) is p.d symmetric, then so is \( A^{-1} \), and choosing \( M = A^{-1} \) yields the energy function of the previous section. Thus functions \( \mathcal{F} \) of this form include both energy and least squares functions as special cases. \( M \) is called a **preconditioner** for (LEq).

This function has derivatives

\[
\nabla \mathcal{F}_M(x) = 2 A^TMAx - b \quad \text{and} \quad D^2 \mathcal{F}_M(x) = 2 A^TMA.
\]

(18.6)

and \( \mathcal{F}_M \) is convex on \( \mathbb{R}^n \) from the analysis of quadratic forms in section 16.

The introduction of the matrix \( M \) does not change the existence and uniqueness results but may significantly change the convergence properties of algorithms for finding the minimizers. The proof of Theorem 18.1 is easily modified to prove the following result about this optimization problem.

**Theorem 18.2.** Assume \( A \) is an \( m \times n \) matrix, \( M \) is a p.d. \( m \times m \) symmetric matrix, \( b \in \mathbb{R}^m \) and \( \mathcal{F}_M \) is defined by (18.4). There are minimizers of \( \mathcal{F}_M \) on \( \mathbb{R}^n \). A vector \( \hat{x} \) minimizes \( \mathcal{F}_M \) on \( \mathbb{R}^n \) if and only if it satisfies the equation

\[
A^TMAx = A^TMb.
\]

(18.7)

If \( N(A^T) = \{0\} \), then any minimizer of \( \mathcal{F}_M \) on \( \mathbb{R}^n \) will be a solution of (LEq)
Exercises.

Exercise 18.1  Let $A$ be the $3 \times 3$ matrix

$$A := \begin{pmatrix} 1 & 0 & 1 \\ 0 & a_1 & 0 \\ 1 & 0 & a_2 \end{pmatrix} \quad (18.8)$$

Evaluate the associated quadratic form on $\mathbb{R}^3$ and find conditions on $a_1, a_2$ for $A$ to be positive semi-definite. When is it positive definite? When $A$ is p.s.d under what conditions on $b \in \mathbb{R}^3$ are there solutions of $Ax = b$? Find all solutions of the equation in this case.

Exercise 18.2  Let $A$ be an $m \times n$ matrix with rank $A = m$. Let $V := N(A)$ be the null space of $A$. What is the dimension of $V$? In Euclidean geometry, the projection of a vector $b \in \mathbb{R}^n$ onto $V$ is the point $\hat{x} \in V$ that is closest to $b$. That is $\hat{x} := P_V b$ minimizes

$$d(x) := \| x - b \|_2^2 \quad \text{subject to} \quad Ax = 0.$$

Find the equations obeyed by $P_V b$. What is the distance of $b$ from $V$?

Exercise 18.3  Let $S := \{a^{(j)} : 1 \leq j \leq m\}$ be a finite set of points in $\mathbb{R}^3$ with $m \geq 4$. Consider the problem of finding a plane in space that provides the ”best approximation” to this data. Suppose this plane $\Sigma$ has the equation

$$\langle c, x \rangle = \sum_{k=1}^{3} c_k x_k = \gamma.$$

Find the expression for the Euclidean distance $d_j$ of the point $a^{(j)}$ from a point in this plane. This distance is a function of $c, \gamma$. Define the function $F : \mathbb{R}^4 \rightarrow \mathbb{R}$ by

$$F(c, \gamma) := \sum_{j=1}^{m} d_j^2 + \| c \|_2^2 - 1.$$

This is a quadratic function of $c, \gamma$. The above analysis shows that there are minimizers of this function. Find the linear equations satisfied by the minimizers. How many equations are there? When does your matrix have a non-trivial null space?