Notes on optimization

August 21, 2018

1 Optimization Problems

We consider the following problem:

\[
\text{Minimize } f(x) \text{ subject to } x \in S, \tag{1}
\]

where \( x = (x_1, x_2, \ldots, x_n) \) and \( S \subseteq \mathbb{R}^n \).

Example: \( f(x) = c^T x \) and \( S = \{ x \in \mathbb{R}^n : Ax = b, x \geq 0 \} \) – Linear Programming.

Local versus Global Optima: \( x^* \) is a global minimum if it is an actual minimizer in (1).

\( x^* \) is a local minimum if there exists \( \delta > 0 \) such that \( f(x^*) \leq f(x) \) for all \( x \in B(x^*) \cap S \), where \( B(x, \delta) = \{ y : |y - x| \leq \delta \} \) is the ball of radius \( \delta \), centred at \( x \).

See Diagram 1 at the end of these notes.

If \( S = \emptyset \) then we say that the problem is unconstrained, otherwise it is constrained.

2 Convex sets and functions

2.1 Convex Functions

A function \( f : \mathbb{R}^n \to \mathbb{R} \) is said to be convex if

\[
f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).
\]

See Diagram 2 at the end of these notes.

Examples of convex functions:
F1 A linear function \( f(x) = a^T x \) is convex.

F2 If \( n = 1 \) then \( f \) is convex iff
\[
f(y) \geq f(x) + f'(x)(y - x) \quad \text{for all } x, y. \tag{2}
\]

Proof. Suppose first that \( f \) is convex. Then for \( 0 < \lambda \leq 1 \),
\[
f(x + \lambda(y - x)) \leq (1 - \lambda)f(x) + \lambda f(y).
\]
Thus, putting \( h = \lambda(y - x) \) we have
\[
f(y) \geq f(x) + \frac{f((x + h) - f(x))}{h}(y - x).
\]
Taking the limit as \( \lambda \to 0 \) implies (2).

Now suppose that (2) holds. Choose \( x \neq y \) and \( 0 \leq \lambda \leq 1 \) and let \( z = \lambda x + (1 - \lambda)y. \) Then we have
\[
f(x) \geq f(z) + f'(z)(x - z) \quad \text{and} \quad f(y) \geq f(z) + f'(z)(y - z).
\]
Multiplying the first inequality by \( \lambda \) and the second by \( 1 - \lambda \) and adding proves that
\[
\lambda f(x) + (1 - \lambda)f(y) \geq f(z).
\]

F3 If \( n \geq 1 \) then \( f \) is convex iff
\[
f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \text{for all } x, y.
\]
Apply F2 to the function \( h(t) = f(tx + (1 - t)y) \).

F4 A \( n = 1 \) and \( f \) is twice differentiable then \( f \) is convex iff \( f''(z) \geq 0 \) for all \( z \in \mathbb{R} \).

Proof. Taylor’s theorem implies that
\[
f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(z)(y - x)^2 \quad \text{where } z \in [x, y].
\]
We now just apply (2).

F5 It follows from F3 that \( e^{ax} \) is convex for any \( a \in \mathbb{R} \).

F6 \( x^a \) is convex on \( \mathbb{R}_+ \) for \( a \geq 1 \) or \( a \leq 0 \). \( x^a \) is concave for \( 0 \leq a \leq 1 \).
Here \( f \) is concave iff \( -f \) is convex.

F7 Suppose that \( A \) is a symmetric \( n \times n \) positive semi-definite matrix. Then \( Q(x) = x^T A x \) is convex.

By positive semi-definite we mean that \( Q(x) \geq 0 \) for all \( x \in \mathbb{R}^n \).
We have
\[
Q(\lambda x + (1 - \lambda)y) - \lambda Q(x) - (1 - \lambda)Q(y)
= \lambda^2 Q(x) + (1 - \lambda)^2 Q(y) + 2\lambda(1 - \lambda)x^T A y - \lambda Q(x) - (1 - \lambda)Q(y)
= -\lambda(1 - \lambda)Q(y - x) \leq 0.
\]
F8 If \( n \geq 1 \) then \( f \) is convex iff \( \nabla^2 F = \begin{bmatrix} \frac{\partial^2 f}{\partial x_i \partial x_j} \end{bmatrix} \) is positive semi-definite for all \( x \).

Apply F7 to the function \( h(t) = f(x + t d) \) for all \( x, d \in \mathbb{R}^n \).

**Operations on convex functions**

E1 If \( f, g \) are convex, then \( f + g \) is convex.

E2 If \( \lambda > 0 \) and \( f \) is convex, then \( \lambda f \) is convex.

E3 If \( f, g \) are convex then \( h = \max\{f, g\} \) is convex.

**Proof.**

\[
h(\lambda x + (1 - \lambda)y) = \max \{f(\lambda x + (1 - \lambda)y), g(\lambda x + (1 - \lambda)y)\}
\leq \max \{\lambda f(x) + (1 - \lambda)f(y), \lambda g(x) + (1 - \lambda)g(y)\}
\leq \lambda \max \{f(x), g(x)\} + (1 - \lambda) \max \{f(y), g(y)\}
= \lambda h(x) + (1 - \lambda)h(y).
\]

\hfill \Box

**Jensen’s Inequality**

If \( f \) is convex and \( a_i \in \mathbb{R}^n, \lambda_i \in \mathbb{R}_+, 1 \leq i \leq m \) and \( \lambda_1 + \lambda_2 + \cdots + \lambda_m = 1 \) then

\[
f \left( \sum_{i=1}^{m} \lambda_i a_i \right) \leq \sum_{i=1}^{m} f(\lambda_i a_i).
\]

The proof is by induction on \( m \). \( m = 2 \) is from the definition of convexity and then we use

\[
\sum_{i=1}^{m} \lambda_i a_i = \lambda_m a_m + (1 - \lambda_m) \sum_{i=1}^{m-1} \frac{\lambda_i}{1 - \lambda_m} a_i.
\]

**Application:** Arithmetic versus geometric mean.

Suppose that \( a_1, a_2, \ldots, a_m \in \mathbb{R}_+ \). Then

\[
\frac{a_1 + a_2 + \cdots + a_m}{m} \geq (a_1 a_2 \cdots a_m)^{1/m}.
\]

\(-\log(x)\) is a convex function for \( x \geq 0 \). So, applying (3),

\[
-\log \left( \sum_{i=1}^{m} \lambda_i a_i \right) \leq \sum_{i=1}^{m} -\log(\lambda_i a_i).
\]

Now let \( \lambda_i = 1/m \) for \( i = 1, 2, \ldots, m \).
2.2 Convex Sets

A set \( S \subseteq \mathbb{R}^n \) is said to be convex if \( x, y \in S \) then the line segment
\[
L(x, y) = \{ \lambda x + (1 - \lambda)y : 0 \leq \lambda \leq 1 \}.
\]

See Diagram 3 at the end of these notes.

Examples of convex sets:

C1 \( S = \{ x : a^T x = 1 \} \). \( x, y \in S \) implies that
\[
a^T(\lambda x + (1 - \lambda)y) = \lambda a^T x + (1 - \lambda)a^T y = \lambda + (1 - \lambda) = 1.
\]

C2 \( S = \{ x : a^T x \leq 1 \} \). Proof similar to C1.

C3 \( S = B(0, \delta) \): \( x, y \in S \) implies that
\[
|\lambda x + (1 - \lambda)y| \leq |\lambda x| + |(1 - \lambda)y| \leq \lambda \delta + (1 - \lambda)\delta = \delta.
\]

C4 If \( f \) is convex, then the level set \( \{ x : f(x) \leq 0 \} \) is convex.
\[
f(x), f(y) \leq 0 \text{ implies that } f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \leq 0.
\]

Operations on convex sets:

O1 \( S \) convex and \( \xi \in \mathbb{R}^n \) implies that \( x + S = \{ x + y : y \in S \} \) is convex.

O2 \( S, T \) convex implies that \( A = S \cap T \) is convex. \( x, y \in A \) implies that \( x, y \in S \) and so \( L = L(x, y) \subseteq S \). Similarly, \( L \subseteq T \) and so \( L \subseteq S \cap T \).

O3 Using induction we see that if \( S_i, 1 \leq i \leq k \) are convex then so is \( \bigcap_{i=1}^k S_i \).

O4 If \( S, T \) are convex sets and \( \alpha, \beta \in \mathbb{R} \) then \( \alpha S + \beta T = \{ \alpha x + \beta y \} \) is convex.
If \( z_i = \alpha x_i + \beta y_i \in T, i = 1, 2 \) then
\[
\lambda z_1 + (1 - \lambda)z_2 = \alpha(\lambda x_1 + (1 - \lambda)x_2) + \beta(\lambda y_1 + (1 - \lambda)y_2) \in T.
\]

It follows from C1, C2 and O3 that an affine subspace \( \{ x : Ax = b \} \) and a halfspace \( \{ x : Ax \leq b \} \) are convex for any matrix \( A \) any vector \( b \).

We now prove something that implies the importance of the above notions. Most optimization algorithms can only find local minima. We do however have the following theorem:

**Theorem 2.1.** Let \( f, S \) both be convex in \( I \). Then if \( x^* \) is a local minimum, it also a global minimum.
Proof.
See Diagram 4 at the end of these notes.

Let \( \delta \) be such that \( x^* \) minimises \( f \) in \( B(x^*, \delta) \cap S \) and suppose that \( x \in S \setminus B(x^*, \delta) \). Let \( z = \lambda x^* + (1 - \lambda)y \) be the point on \( L(x^*, y) \) at distance \( \delta \) from \( x^* \). Note that \( x \in S \) by convexity of \( S \). Then by the convexity of \( f \) we have

\[
 f(x^*) \leq f(x) \leq \lambda f(x^*) + (1 - \lambda)f(y)
\]

and this implies that \( f(x^*) \leq f(x) \).

The following shows the relationship between convex sets and functions.

Lemma 2.2. let \( f_1, f_2, \ldots, f_m \) be convex functions on \( \mathbb{R}^n \). Let \( b \in \mathbb{R}^m \) and let

\[
 S = \{ x \in \mathbb{R}^n : f_i(x) \leq b_i, i = 1, 2, \ldots, m \}.
\]

Then \( S \) is convex.

Proof. It follows from O3 that we can consider the case \( m = 1 \) only and drop the subscript. Suppose now that \( x, y \in S \) i.e. \( f(x), f(y) \leq b \). Then for \( 0 \leq \lambda \leq 1 \)

\[
 f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \leq \lambda b + (1 - \lambda)b = b.
\]

So, \( \lambda x + (1 - \lambda)y \in S \).

3 Algorithms

3.1 Line search – \( n = 1 \)

Here we consider the simpler problem of minimising a convex (more generally unimodal) function \( f : \mathbb{R} \to \mathbb{R} \).

See Diagram 5 at the end of these notes.

We assume that we are given \( a_0, a_1 \) such that \( a_0 \leq x^* \leq a_1 \) where \( x^* \) minimises \( f \). This is not a significant assumption. We can start with \( a_0 = 0 \) and then consider the sequences \( \zeta_i = f(2^i) \), \( \xi_i = f(-2^i) \) until we find \( \zeta_{i-1} \leq \min \{ \zeta_0, \zeta_i \} \) (resp. \( \xi_{i-1} \leq \min \{ \xi_0, \xi_i \} \)). Then we know that \( x^* \in [\zeta_0, \zeta_i] \) (resp. \( x^* \in [\xi_0, \xi_i] \)).

Assume then that we have an interval \( [a_0, a_1] \) of uncertainty for \( x^* \). Furthermore, we will have evaluated \( f \) at two points in this interval, two points inside the interval at \( a_2 = a_0 + \alpha(a_1 - a_0) \) and \( a_3 = a_0 + \alpha(a_1 - a_0) \) respectively. We will determine \( \alpha \) shortly. And at each iteration we make one new function evaluation and decrease the interval of uncertainty by a factor \( \alpha \). There are two possibilities:
(i) \( f(a_2) \leq f(a_3) \). This implies that \( x^* \in [a_0, a_3] \). So, we evaluate \( f(a_0 + \alpha^2(a_3 - a_0)) \) and make the changes \( a_i \rightarrow a'_i \):

\[
a'_0 \leftrightarrow a_0, \ a'_1 \leftrightarrow a_3, \ a'_2 \leftrightarrow a_0 + \alpha^2(a_3 - a_0), \ a'_3 \leftrightarrow a_2.
\]

(ii) \( f(a_2) > f(a_3) \). This implies that \( x^* \in [a_2, a_1] \). So, we evaluate \( f(a_0+) \) and make the changes \( a_i \rightarrow a'_i \):

\[
a'_0 \leftrightarrow a_2, \ a'_1 \leftrightarrow a_1, \ a'_2 \leftrightarrow a_3, \ a'_3 \leftrightarrow a_2 + \alpha^2(a_1 - a_0).
\]

In case (i) we see that \( a'_1 - a'_0 = a_3 - a_0 = \alpha(a_1 - a_0) \) and so the interval has shrunk by the required amount. Next we see that \( a'_2 - a'_0 = \alpha^2(a_3 - a_0) = \alpha^2(a'_1 - a_0) \). Furthermore, \( a'_3 - a'_0 = a_2 - a_0 = \alpha^2(a_1 - a_0) = \alpha(a'_1 - a'_0) \).

In case (ii) we see that \( a'_1 - a'_0 = a_1 - a_2 = a_1 - (a_0 + \alpha^2(a_1 - a_0)) = (1 - \alpha^2)(a_1 - a_0) \). So, shrink by \( \alpha \) in this case we choose \( \alpha \) to satisfy \( 1 - \alpha^2 = \alpha \). This gives us

\[
\alpha = \frac{1 + \sqrt{5}}{2} \text{ – the golden ratio.}
\]

Next we see that \( a'_2 - a'_0 = a_3 - a_2 = (\alpha - \alpha^2)(a_1 - a_0) = \frac{\alpha - \alpha^2}{\alpha}(a'_1 - a'_0) = (1 - \alpha)(a'_1 - a'_0) = \alpha^2(a'_1 - a'_0) \). Finally, we have \( a'_3 - a'_0 = a_2 + \alpha^2(a_1 - a_0) - a_2 = \alpha^2(a_1 - a_0) = \alpha(a'_1 - a'_0) \).

Thus to achieve an accuracy within \( \delta \) of \( x^* \) we need to take \( t \) steps, where \( \alpha' D \leq \delta \) where \( D \) is our initial uncertainty.

### 3.2 Gradient Descent

See Diagram 6 at the end of these notes.

Here we consider the unconstrained problem. At a point \( x \in \mathbb{R}^n \), if we move a small distance \( h \) in direction \( d \) then we have

\[
f(x + h d/|d|) = f(x) + h(\nabla f)^T \frac{d}{|d|} + O(h^2) \geq f(x) - h |\nabla f| + O(h^2).
\]

Thus, at least infinitesimally, the best direction is \(-\nabla f\). So, for us, the steepest algorithm will follow a sequence of points \( x_0, x_1, \ldots, x_k, \ldots \), where

\[
x_{k+1} = x_k - \varepsilon_k \nabla f(x_k).
\]

Then we have

\[
|x_{k+1} - x^*|^2 = |x_k - x^*|^2 - 2\alpha_k \nabla f(x_k)^T(x_k - x^*) + \alpha_k^2 |\nabla f(x_k)|^2
\leq |x_k - x^*|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 |\nabla f(x_k)|^2.
\]

The inequality comes from F3.
Applying (4) repeatedly we get

$$|x_k - x^*|^2 \leq |x_0 - x^*|^2 - 2 \sum_{i=1}^{k} \alpha_i (f(x_i) - f(x^*)) + \sum_{i=1}^{K} \alpha_i^2 |\nabla f(x_k)|^2.$$  \hspace{1cm} (5)

Putting $R = |x_0 - x^*|$, we see from (5) that

$$2 \sum_{i=1}^{k} \alpha_i (f(x_i) - f(x^*)) \leq R^2 + \sum_{i=1}^{K} \alpha_i^2 |\nabla f(x_k)|^2.$$  \hspace{1cm} (6)

On the other hand,

$$\sum_{i=1}^{k} \alpha_i (f(x_i) - f(x^*)) \geq \left( \sum_{i=1}^{k} \alpha_i \right) \min \{ f(x_k) - f(x^*) : i \in [k] \} = \left( \sum_{i=1}^{k} \alpha_i \right) (f(x_{min}) - f(x^*)),$$  \hspace{1cm} (7)

where $f(x_{min}) = \min \{ f(x_i) : i \in [k] \}$.

Combining (6) and (7) we get

$$f(x_{min}) - f(x^*) \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i},$$

where $G = \max \{|\nabla f(x_i)| : i \in [k]\}$.

So, if we choose $\alpha_k$ so that $\sum_{i=1}^{\infty} \alpha_i = \infty$ and $\sum_{i=1}^{\infty} \alpha_i^2 = O(1)$ then

$$|f(x_{min}) - f(x^*)| \rightarrow 0 \text{ as } k \rightarrow \infty.$$

As an example, we could let $\alpha_i = 1/i$.

4 Separating Hyperplane

See Diagram 7 at the end of these notes.

**Theorem 4.1.** Let $C$ be a convex set in $\mathbb{R}^n$ and suppose $x \notin C$. Then there exists $0 \neq a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that (i) $a^T x \geq b$ and (ii) $C \subseteq \{ y \in \mathbb{R}^n : a^T y \leq b \}$.

**Proof.**

**Case 1:** $C$ is closed.

Let $z$ be the closest point in $C$ to $x$. Let $a = x - z \neq 0$ and $b = (x - z)^T z$. Then

$$a^T x - b = (x - z)^T x - (x - z)^T z = |x - z|^2 > 0.$$  

This verifies (i). Suppose (ii) fails and there exists $y \in C$ such that $a^T y > b$. Let $w \in C$ be the closest point to $x$ on the line segment $L(y, z) \subseteq C$. The triangle formed by $x, w, z$ has a right angle at $w$ and an acute angle at $z$. This implies that $|x - w| < |x - z|$, a contradiction.
Case 2: \( x \notin \bar{C} \).
We observe that \( \bar{C} \supseteq C \) and is convex (exercise). We can thus apply Case 1, with \( \bar{C} \) replacing \( C \).

Case 3: \( x \in \bar{C} \setminus C \). Every ball \( B(x, \delta) \) contains a point of \( \mathbb{R}^n \setminus \bar{C} \) that is distinct from \( x \). Choose a sequence \( x_n, n \geq 1 \) that tends to \( x \). For each \( x_n \), let \( a_n, b_n = a_n^T z_n \) define a hyperplane that separates \( x_n \) from \( \bar{C} \), as in Case 2. We can assume that \( |a_n| = 1 \) (scaling) and that \( b_n \) is in some bounded set and so there must be a convergent subsequence of \( (a_n, b_n), n \geq 1 \) that converges to \( (a, b), |a| = 1 \). Assume that we re-label so that this subsequence is \( (a_n), n \geq 1 \). Then for \( y \in \bar{C} \) we have \( a_n^T y \leq b_n \) for all \( n \). Taking limits we see that \( a^T y \leq b \). Furthermore, for \( y \notin \bar{C} \) we see that for large enough \( n \), \( a_n^T y > b_n \). taking limits we see that \( a^T y \leq b \).

Corollary 4.2. Suppose that \( S, T \subseteq \mathbb{R}^n \) are convex and that \( S \cap T = \emptyset \). Then there exists \( a, b \) such that \( a^T x \leq b \) for all \( x \in S \) and \( a^T x \geq b \) for all \( x \in T \).

Proof. Let \( W = S + (-1)T \). Then \( 0 \notin W \) and applying Theorem 4.1 we see that there exists \( a \) such that \( a^T z \leq 0 \) for all \( z \in W \). Now put

\[
b = \frac{1}{2} \left( \sup_{x \in S} a^T x + \inf_{x \in T} a^T x \right).
\]

Corollary 4.3 (Farkas Lemma). For an \( m \times n \) matrix and \( b \in \mathbb{R}^m \), exactly one of the following holds:

(i) There exists \( x \in \mathbb{R}^n \) such that \( x \geq 0, Ax = b \).

(ii) There exists \( u \in \mathbb{R}^m \) such that \( u^T A \geq 0 \) and \( u^T b < 0 \).

Proof. We cannot have both (i), (ii) holding. For then we have

\[
0 \leq u^T Ax = u^T b < 0.
\]

Suppose then that (i) fails to hold. Let \( S = \{ y : y = Ax \text{ for some } x \geq 0 \} \). Then \( b \notin S \) and since \( S \) is closed there exists \( \alpha, \beta \) such that (a) \( \alpha^T b \leq \beta \) and (b) \( \alpha^T Ax \geq \beta \) for all \( x \geq 0 \). This implies that \( \alpha^T (b - Ax) \leq 0 \) for all \( x \geq 0 \). This then implies that \( u = \alpha \) satisfies (ii).

4.1 Convex Hulls

See Diagram 8 at the end of these notes.
Given a set $S \subseteq \mathbb{R}^n$, we let

$$\text{conv}(S) = \left\{ \sum_{i \in I} \lambda_i x_i : (i) \ |I| < \infty, \ (ii) \ \sum_{i \in I} \lambda_i = 1, \ (iii) \ \lambda_i > 0, \ i \in I, \ (iv) \ x_i \in S, \ i \in I \right\}.$$  

Clearly $S \subseteq \text{conv}(S)$, since we can take $|I| = 1$.

**Lemma 4.4.** $\text{conv}(S)$ is a convex set.

**Proof.** Let $x = \sum_{i \in I} \lambda_i x_i$, $y = \sum_{j \in J} \mu_j y_j \in \text{conv}(S)$. Let $K = I \cup J$ and put $\lambda_i = 0, i \in J \setminus I$ and $\mu_j = 0, j \in I \setminus J$. Then for $0 \leq \alpha \leq 1$ we see that

$$\alpha x + (1 - \alpha)y = \sum_{i \in K} (\alpha \lambda_i + (1 - \alpha)\mu_i) x_i$$

and

$$\sum_{i \in K} (\alpha \lambda_i + (1 - \alpha)\mu_i) = 1$$

implying that $\alpha x + (1 - \alpha)y \in \text{conv}(S)$ i.e. $\text{conv}(S)$ is convex. $\square$

**Lemma 4.5.** If $S$ is convex, then $S = \text{conv}(S)$.

**Proof.** Exercise. $\square$

**Corollary 4.6.** $\text{conv}(\text{conv}(S)) = \text{conv}(S)$ for all $S \subseteq \mathbb{R}^n$.

**Proof.** Exercise. $\square$

### 4.1.1 Extreme Points

A point $x$ of a convex set $S$ is said to be an extreme point if there do not exist $y, z \in S$ such that $x \in L(y, z)$. We let $\text{ext}(S)$ denote the set of extreme points of $S$.

**EX1** If $n = 1$ and $S = [a, b]$ then $\text{ext}(S) = \{a, b\}$.

**EX2** If $S = B(0, 1)$ then $\text{ext}(S) = \{x : |x| = 1\}$.

**EX3** If $S = \{x : Ax = b\}$ is the set of solutions to a set of linear equations, then $\text{ext}(S) = \emptyset$.

**Theorem 4.7.** Let $S$ be a closed, bounded convex set. Then $S = \text{conv}(\text{ext}(S))$.

**Proof.** We prove this by induction on the dimension $n$. For $n = 1$ the result is trivial, since then $S$ must be an interval $[a, b]$.

Inductively assume the result for dimensions less than $n$. Clearly, $S \supseteq T = \text{conv}(\text{ext}(S))$ and suppose there exists $x \in S \setminus T$. Let $z$ be the closest point of $T$ to $x$ and let $H = \{y : a^T y = b\}$ be the hyperplane defined in Theorem 4.1. Let $b^* = \max \{a^T y : y \in S\}$. We have $b^* < \infty$ since $S$ is bounded. Let $H^* = \{y : a^T y = b^*\}$ and let $S^* = S \cap H^*$.
We observe that if $w$ is a vertex of $S^*$ then it is also a vertex of $S$. For if $w = \lambda w_1 + (1 - \lambda)w_2$, $w_1, w_2 \in S, 0 < \lambda < 1$ then we have

$$b^* = a^T w = \lambda a^T w_1 + (1 - \lambda) a^T w_2 \leq \lambda b^* + (1 - \lambda) b^* = b^*.$$ 

This implies that $a^T w_1 = a^T w_2 = b^*$ and so $w_1, w_2 \in S^*$, contradiction.

Now consider the point $w$ on the half-line from $z$ through $x$ that lies in $S^*$ i.e

$$w = z + \frac{b^* - b}{a^T x - b}(x - z).$$

Now by induction, we can write $w = \sum_{i=1}^{k} \lambda_i w_i$ where $w_1, w_2, \ldots, w_k$ are extreme points of $S^*$ and hence of $S$. Also, $x = \mu w + (1 - \mu)z$ for some $0 < \mu \leq 1$ and so $x \in ext(S)$. □

The following is sometimes useful.

**Lemma 4.8.** Suppose that $S$ is a closed bounded convex set and that $f$ is a convex function. The $f$ achieves its maximum at an extreme point.

**Proof.** Suppose the maximum occurs at $x = \lambda_1 x_1 + \cdots + \lambda_k x_k$ where $0 \leq \lambda_1, \ldots, \lambda_k \leq 1$ and $\lambda_1 + \cdots + \lambda_k = 1$ and $x_1, \ldots, x_k \in ext(S)$. Then by Jensen’s inequality we have $f(x) \leq \lambda_1 f(x_1) + \cdots + \lambda_k f(x_k) \leq \max \{f(x_i) : 1 \leq i \leq k\}$. □

This explains why the solutions to linear programs occur at extreme points.

## 5 Lagrangean Duality

See Diagram 9 at the end of these notes.

Here we consider the **primal problem**

Minimize $f(x)$ subject to $g_i(x) \leq 0, i = 1, 2, \ldots, m,$ 

where $f, g_1, g_2, \ldots, g_m$ are convex functions on $\mathbb{R}^n$.

The Lagrangean

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

The **dual problem** is

Maximize $\phi(\lambda)$ subject to $\lambda \geq 0$ where $\phi(\lambda) = \min_{x \in \mathbb{R}^n} L(x, \lambda)$. 

(9)

We note that $\phi$ is a concave function. It is the minimum of a collection of convex (actually linear) functions of $\lambda$ – see E3.
D1 : Linear programming. Let \( f(x) = c^T x \) and \( g_i(x) = -a_i^T x + b_i \) for \( i = 1, 2, \ldots, m \). Then
\[
L(x, \lambda) = (c^T - \lambda^T A)x + b^T \lambda \quad \text{where} \quad A \text{ has rows} \quad a_1, \ldots, a_m.
\]

D2

Weak Duality: If \( \lambda \) is feasible for (9) and \( x \) is feasible for (8) then \( f(x) \geq \phi(\lambda) \).

\[
\phi(\lambda) \leq L(x, \lambda) \leq f(x) \quad \text{since} \quad \lambda_i \geq 0, g_i(x) \leq 0, \quad i = 1, 2, \ldots, m. \quad (10)
\]

Now note that \( \phi(\lambda) = -\infty \), unless \( c^T = \lambda^T A \), since \( x \) is unconstrained in the definition of \( \phi \). And if \( c^T = \lambda^T A \) then \( \phi(\lambda) = b^T \lambda \). So, the dual problem is to

Maximize \( b^T \lambda \) subject to \( c^T = \lambda^T A \) and \( \lambda \geq 0 \), i.e. the LP dual.

Strong Duality: We give a sufficient condition Slater’s Constraint Condition for tightness in (10).

Theorem 5.1. Suppose that there exists a point \( x^* \) such that \( g_i(x^*) < 0, i = 1, 2, \ldots, m \). Then
\[
\max_{\lambda \geq 0} \phi(\lambda) = \min_{x: g_i(x) \leq 0, i \in [m]} f(x).
\]

Proof. Let
\[
A = \{u, t) : \exists x \in \mathbb{R}^n, g_i(x) \leq u_i, i = 1, 2, \ldots, m \text{ and } f(x) \leq t\}\,
\]
\[
B = \{(0, s) \in \mathbb{R}^{m+1} : s < f^*\}\,
\]
where \( f^* = \min_{x: g_i(x) \leq 0, i \in [m]} f(x) \).

Now \( A \cap B = \emptyset \) and so from Corollary 4.2 there exists \( \lambda, \gamma, b \) such that \( (\lambda, \gamma) \neq 0 \) and
\[
b \leq \min \\{ \lambda^T u + \gamma t : (u, t) \in A\}\,
\]
\[
b \geq \max \\{ \lambda^T u + \gamma t : (u, t) \in B\}. \quad (12)
\]

We deduce from (11) that \( \lambda \geq 0 \) and \( \bar{g} \geq 0 \). If \( \gamma < 0 \) or \( \lambda_i < 0 \) for some \( i \) then the minimum in (11) is \( -\infty \). We deduce from (12) that \( \gamma t < b \) for all \( t < f^* \) and so \( \gamma f^* \leq b \). And from (11) that
\[
\gamma f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \geq b \geq \gamma f^* \quad \text{for all} \quad x \in \mathbb{R}^n. \quad (13)
\]

If \( \gamma > 0 \) then we can divide (13) by \( \gamma \) and see that \( L(x, \lambda) \geq f^* \), and together with weak duality, we see that \( L(x, \lambda) = f^* \).

If \( \gamma = 0 \) then substituting \( x^* \) into (13) we see that \( \sum_{i=1}^{m} \lambda_i g_i(x^*) \geq 0 \) which then implies that \( \lambda = 0 \), contradiction. \( \square \)
6 Conditions for a minimum: First Order Condition

6.1 Unconstrained problem

We discuss necessary conditons for \( a \) to be a (local) minimum. (We are not assuming that \( f \) is convex.) We will assume that our functions are differentiable. Then Taylor’s Theorem

\[
f(a + h) = f(a) + (\nabla f(a))^T h + o(|h|)
\]

implies that

\[
\nabla f(a) = 0 \tag{14}
\]

is a necessary condition for \( a \) to be a local minimum. Otherwise,

\[
f(a - t\nabla f(a)) \leq f(a) - t|\nabla f(a)|^2/2
\]

for small \( t > 0 \).

Of course (14) is not sufficient in general, \( a \) could be a local maximum. Generally speaking, one has to look at second order conditions to distinguish between local minima and local maxima.

However, \textbf{Lemma 6.1.} \textit{If \( f \) is convex then (14) is also a sufficient condition.}

\textit{Proof.} This follows directly from F3. \( \square \)

6.2 Constrained problem

We will consider Problem \( (8) \), but we will not assume convexity, only differentiability. The condition corresponding to (14) is the Karush-Kuhn-Tucker or KKT condition. Assume that \( f, g_1, g_2, \ldots, g_m \) are differentiable. Then (subject to some regularity conditions, a necessary condition for \( a \) to be a local minimum (or maximum) to Problem \( (8) \) is that there exists \( \lambda \) such that

\[
\nabla f(a) + \sum_{i=1}^{m} \lambda_i \nabla g_i(a) = 0. \tag{15}
\]

\[
\lambda_i \geq 0, \quad 1 \leq i \leq m. \tag{16}
\]

\[
\lambda_i g_i(a) = 0, \quad 1 \leq i \leq m. \quad \text{Complementary Slackness} \tag{17}
\]

The second condition says that only active constraints \( (g_i(a) = 0) \) are involved in the first condition.

One deals with \( g_i(x) \geq 0 \) via \(-g_i(x) \leq 0 \) (and \( \lambda_i \leq 0 \)) and \( g_i(x) = 0 \) by \( g_i(x) \geq 0 \) and \(-g_i(x) \leq 0 \) (and \( \lambda_i \) not constrained to be non-negative or non-positive).

In the convex case, we will see that (15), (16) and (17) are sufficient for a global minimum.
6.2.1 Heuristic Justification of KKT conditions

See Diagram 10 at the end of these notes.

Suppose that \( \mathbf{a} \) is a local minimum and assume w.l.o.g. that \( g_i(\mathbf{a}) = 0 \) for \( i = 1, 2, \ldots, m \). Then (heuristically) Taylor’s theorem implies that if (i) \( \mathbf{h}^T \nabla g_i(\mathbf{a}) \leq 0, i = 1, 2, \ldots, m \) then (ii) we should have \( \mathbf{h}^T \nabla f(\mathbf{a}) \geq 0 \). (The heuristic argument is that (i) holds then we should have (iii) \( \mathbf{a} + \mathbf{h} \) feasible for small \( \mathbf{h} \) and then we should have (ii) since we are at a local minimum. You need a regularity condition to ensure that (ii) implies (iii).)

Applying Corollary 4.3 we see that the KKT conditions hold. We let \( \mathbf{A} \) have columns \( \nabla g_i(\mathbf{a}), i = 1, 2, \ldots, m \). Then the KKT conditions are \( \mathbf{A} \lambda = -\nabla f(\mathbf{a}) \).

For much more on this subject see [Convex Optimization] by Boyd and Vandenbergh.
Diagram 1

Local Minimum

Global Minimum
Diagram 2

Convex Function

Non-Convex Function
Diagram 3

Convex Set

Non-Convex Set
Diagram 4

$B(x^*_S)$
Diagram 6

\[ \varphi(x) = c \geq c_2 \]

\[ \varphi(y) = c_2 \]

\( x^* \)
Diagram 7
Diagram 8

Extreme Point
Diagram 9

\[ \Delta u + t = \phi(t) \]

\[ \Delta u + t = \text{constant} \]