MA427 – Mathematical Optimisation
Part III
Convex Programming[1]

Giacomo Zambelli
g.zambelli@lse.ac.uk

Department of Mathematics
London School of Economics and Political Science
2018/19

[1]Based in part on S. Boyd and L. Vandenberghe: Convex Optimization (Cambridge University Press, 2004).

Contents

Chapter 1

Convex optimization problems

1.1 Definitions and notations

For a vector $x \in \mathbb{R}^n$, we let $\|x\|$ denote the 2-norm, that is, $\|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$.

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, we denote by $\mathbf{dom}\, f$ the *domain* of f, that is, the set of points x in $\mathbb{R}^n$ for which $f(x)$ is defined. For example, the domain of the function $x \mapsto \log x$ is the set $\{x \in \mathbb{R} \,|\, x > 0\}$, whereas the domain of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $(x_1, x_2) \mapsto x_1/x_2$ is the set $\mathbf{dom}\, f = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_2 \neq 0\}$.

Graph of a function The *graph* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set $\{(x, f(x)) \in \mathbb{R}^{n+1} \,|\, x \in \mathbf{dom}\, f\}$, and the *epigraph* is the set of points in $\mathbb{R}^{n+1}$ that "lay above" the graph of f, that is, the set $\{(x, t) \in \mathbb{R}^{n+1} \,|\, x \in \mathbf{dom}\, f,\ f(x) \leq t\}$.

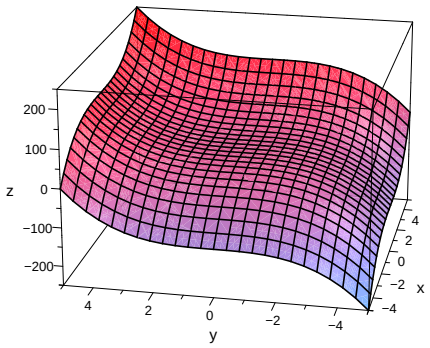


Figure 1.1: Graph of the function $f(x_1, x_2) = x_1^3 + x_2^3$. The epigraph of f is the region above the shaded surface.

Gradients The *gradient* of the function f at point $x \in \mathbf{dom}\, f$ is the vector

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

where $\frac{\partial f(x)}{\partial x_i}$ is the i-th partial derivative of f at point x, i.e., the derivative of f at point x taken with respect to the variable x_i. In particular, when $n = 1$ (i.e. f is a function of one variable), the gradient is simply the derivative of f.

Differentiable functions. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable* at a point x in the interior of $\mathbf{dom}\, f$ if

$$\lim_{z \in \mathbf{dom}\, f,\, z \to x} \frac{|f(z) - f(x) - \nabla f(x)^\top (z - x)|}{\|z - x\|} = 0.$$

Geometrically, differentiability at x means that, near x, f is well approximated by the affine function $h : \mathbf{dom}\,(f) \to \mathbb{R}$ defined by $h(z) = f(x) + \nabla f(x)^\top (z - x)$.

We say that a continuous function f is *differentiable* if $\mathbf{dom}\, f$ is an open set and f is differentiable at every point $x \in \mathbf{dom}\, f$.

For example, the function $f : \mathbb{R}^2 \to \mathbb{R}$, defined over $\mathbf{dom}\, f = \{x \in \mathbb{R}^2 \,|\, x_1, x_2 > 0\}$ by $f(x_1, x_2) = \log(x_1/x_2)$, is differentiable, and its gradient at any point $x \in \mathbf{dom}\, f$ is

$$\nabla f(x) := \begin{bmatrix} 1/x_1 \\ -1/x_2 \end{bmatrix}.$$

The function $f : x \mapsto |x|$ is not differentiable, because its derivative does not exist at $x = 0$.

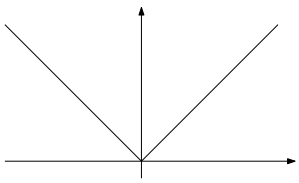


Figure 1.2: Graph of the function $x \mapsto |x|$.

Recall that, if f is differentiable at $\bar{x}$ in the interior of $\mathbf{dom}\, f$, then the gradient of f at $\bar{x}$ points in the direction of steepest ascent of the graph of f at point $(\bar{x}, f(\bar{x}))$. More formally, the hyperplane in $\mathbb{R}^{n+1}$ defined by

$$H = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \,|\, t = f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x})\},$$

is the hyperplane tangent to the graph of f at point $(\bar{x}, f(\bar{x}))$. The direction of steepest ascent for the tangent hyperplane H is the direction of the gradient $\nabla f(\bar{x})$, and the slope of H in the direction of $\nabla f(\bar{x})$ is the magnitude $\|\nabla f(x)\|$ of the gradient.

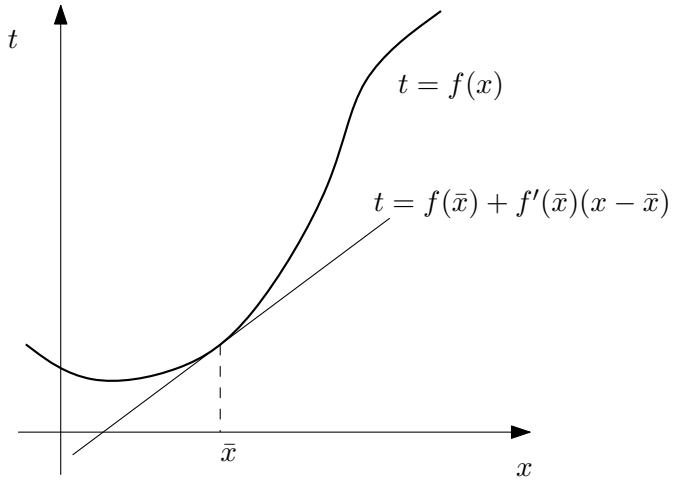


Figure 1.3: For functions of one variables, the gradient at point $\bar{x}$ is the slope of the tangent to the graph at point $(\bar{x}, f(\bar{x}))$.

Another geometric interpretation is that the gradient at point $\bar{x}$ is a vector orthogonal to the contour of f at point $\bar{x}$ – that is, the set $\{x \in \mathbb{R}^n \,|\, f(x) = f(\bar{x})\}$ – pointing in the direction of ascent of the function at $\bar{x}$ (see Figure 1.4).

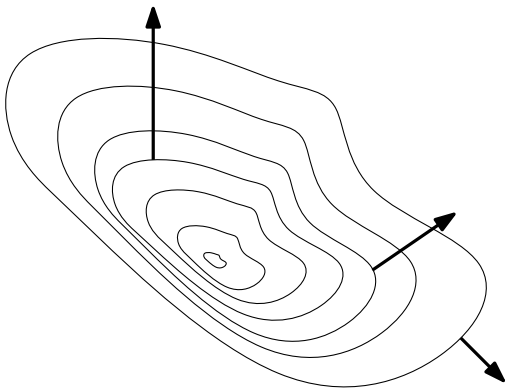


Figure 1.4: For functions of one variables, $\nabla f(\bar{x})$ is orthogonal to the contour at $\bar{x}$ and pointing in the direction of ascent.

Directional derivatives Given $f : \mathbb{R}^n \to \mathbb{R}$, a point $\bar{x} \in \mathbf{dom}\, f$ and a vector $p \in \mathbb{R}^n$, we can define the function of one variable $g : \mathbb{R} \to \mathbb{R}$, $t \mapsto f(\bar{x} + tp)$. The *directional derivative* of f at point $\bar{x}$ in the direction p is the derivative of g at $t = 0$, that is

$$\frac{\partial f}{\partial p}(\bar{x}) = g'(0).$$

The geometric meaning of the above is that the directional derivative $\partial f(\bar{x})/\partial p$ measures the rate of change of f at point $\bar{x}$ when moving in the direction of p.

If f is differentiable, then the following holds

$$\frac{\partial f}{\partial p}(\bar{x}) = \nabla f(\bar{x})^\top p. \tag{1.1}$$

Global and Local Optima Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and a set $X \subseteq \mathbf{dom}\, f$, we say that a point $x^* \in X$ is a *global minimum* for f in X if $f(x) \geq f(x^*)$ for all $x \in X$. We say that $x^* \in X$ is a *global maximum* for f in X if $f(x) \leq f(x^*)$ for all $x \in X$.

A point $x^* \in X$ is said a *local minimum* for f in X if there exists $\varepsilon > 0$ such that $f(x) \geq f(x^*)$ for all $x \in X$ such that $\|x - x^*\| \leq \varepsilon$. We say that $x^* \in X$ is a *local maximum* for f in X if there exists $\varepsilon > 0$ such that $f(x) \leq f(x^*)$ for all $x \in X$ such that $\|x - x^*\| \leq \varepsilon$.

When $X = \mathbf{dom}\, f$, we refer simply to global or local maxima or minima.

Consider, for example, the single variable function whose graph is represented in Figure 1.5. The point x' is a local minimum of f, because we can find a small interval around x' where no point has value smaller than x'. However, x' is not a global minimum because there are points with lower objective value.

We recall the following facts from calculus concerning local optima of unconstrained problems.

Theorem 1.1 (First-order necessary conditions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and let x^* be a point in* $\mathbf{dom}\, f$*. If x^* is a local maximum or a local minimum for f, then $\nabla f(x^*) = 0$.*

While the above condition is useful in finding a local optimum, it has two drawbacks. First, a point where the gradient is zero needs not be a local minimum or a local maximum (see Figure 1.6). Second, even if we were to determine that a certain point $x^* \in X$ is a local minimum, in general we would be unable to determine if it is a global minimum. In the next section we will see that, for an important class of functions, namely the convex functions, every local minimum is also a global minimum, and that, for differentiable convex functions, the first order necessary condition are also sufficient.

1.2 Convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if $\mathbf{dom}\, f$ is convex and, for every $x, y \in \mathbf{dom}\, f$, and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y). \tag{1.2}$$

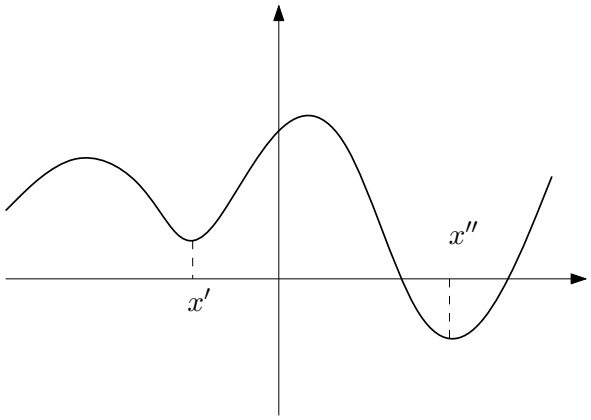


Figure 1.5: Point x' is a local minimum, but not a global minimum.

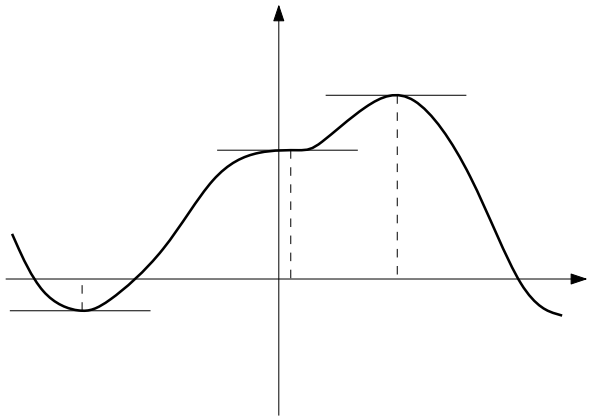


Figure 1.6: Three points with zero gradient. From left to right, the first point is a local minimum, the second a saddle point, and the third a local maximum.

Figure 1.7 provides a geometric interpretation of the above definition. Note the point $(\lambda x + (1 - \lambda)y, f(\lambda x + (1 - \lambda)y))$ in $\mathbb{R}^{n+1}$ is a point on the graph of f, therefore (1.2) says that the point $(\lambda x + (1 - \lambda)y, \lambda f(x) + (1 - \lambda)f(y))$ belongs to the epigraph of f. Observe that the set of all points $(\lambda x + (1 - \lambda)y, \lambda f(x) + (1 - \lambda)f(y))$ for $\lambda \in [0, 1]$ is the line segment joining $(x, f(x))$ to $(y, f(y))$, therefore (1.2) means that the epigraph of f contains the line segment joining any two points in the graph of f. This implies the following.

Proposition 1.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the epigraph of f is a convex set.*

The simplest example of convex functions are *affine functions*. The function $f : \mathbb{R}^n \to \mathbb{R}$ is *affine* if there exists $p \in \mathbb{R}^n$ and $r \in \mathbb{R}$ such that $f(x) = p^\top x + r$; note that linear functions are affine functions where $r = 0$.

It will be convenient to extend the definition of a convex function f to the points not in the domain, by considering $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, where $f(x) = +\infty$ for every $x \notin \mathbf{dom}\, f$. Note that this convention respects the definition of convexity given by (1.2). Indeed, if one among $x, y \in \mathbb{R}^n$ is not in $\mathbf{dom}\, f$, then the right-hand-side of (1.2) is $+\infty$, and the inequality is still verified. With this notation, it follows that $\mathbf{dom}\, f = \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$.

We remark, without proof, that convex functions are always continuous.

Theorem 1.3. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $\mathbf{dom}\, f$ is open, then f is continuous on $\mathbf{dom}\, f$.*

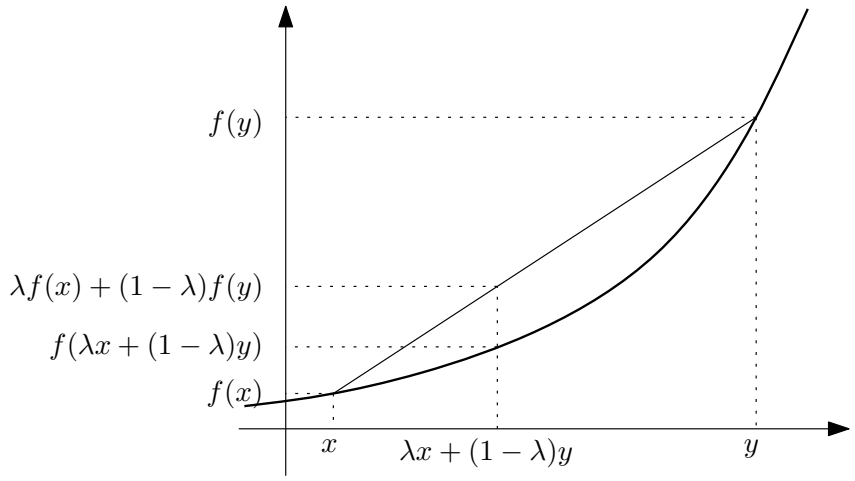


Figure 1.7: A function f is convex if the line segment joining any two points in the graph of f is contained in the epigraph of f.

1.2.1 Univariate convex functions

Let us recall the familiar case of univariate functions $f : \mathbb{R} \to \mathbb{R}$. A practical test involves the second derivatives.

Theorem 1.4. *Assume that $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable on* $\mathbf{dom}\, f$. *Then f is convex* if and only if *the second derivative f'' is nonnegative on* $\mathbf{dom}\, f$.

Using this criterion, we can easily verify the convexity of the following univariate functions.

- $f(x) = -\log x$, where $\mathbf{dom}\, f = \{x \in \mathbb{R} \,|\, x > 0\}$.
- $f(x) = e^x$, where $\mathbf{dom}\, f = \mathbb{R}$.
- $f(x) = 1/x$, where $\mathbf{dom}\, f = \{x \in \mathbb{R} \,|\, x > 0\}$. Observe that $f(x) = 1/x$ can be defined over $\mathbb{R} \setminus \{0\}$, but it is not convex over the negative reals.
- $f(x) = x \log x$, where $\mathbf{dom}\, f = \{x \in \mathbb{R} \,|\, x > 0\}$.

1.2.2 Simple constructions of convex functions

Let us show two simple operations that enable constructing convex functions from other convex functions. The first one is *nonnegative linear combination.*

Proposition 1.5. *If $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex and $\gamma_1, \ldots, \gamma_m \geq 0$, then $f = \gamma_1 f_1 + \cdots + \gamma_m f_m$ is convex.*

Proof. We need to show that for every $x, y \in \mathbf{dom}\, f$, and $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. This follows using the convexity of the f_i's:

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= \sum_{i=1}^m \gamma_i f_i(\lambda x + (1-\lambda)y) \leq \sum_{i=1}^m \gamma_i (\lambda f_i(x)) + (1-\lambda) f_i(y)) \\ &= \lambda \sum_{i=1}^m \gamma_i f_i(x) + (1-\lambda) \sum_{i=1}^m \gamma_i f_i(y) = \lambda f(x) + (1-\lambda) f(y). \end{aligned}$$

□

The second operation is taking *point-wise supremum.* This is illustrated in Figure 1.8.

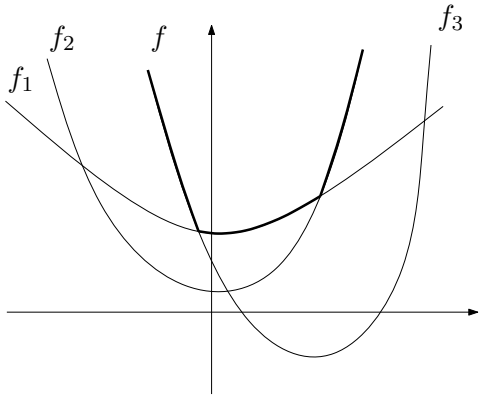


Figure 1.8: The graph of the point-wise maximum of the three convex functions in the picture is in boldface.

Proposition 1.6. *If $f_\alpha : \mathbb{R}^n \to \mathbb{R}$ $(\alpha \in \mathcal{A})$ is a family of convex functions indexed by the elements of a set $\mathcal{A}$ (possibly infinite), then the function f defined by*

$$f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

is convex.

Proof. Again, we need to show that for every $x, y \in \mathbf{dom}\, f$, and $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. This follows by showing that for any $\varepsilon > 0$, $f(\lambda x + (1-\lambda)y) - \varepsilon \leq \lambda f(x) + (1-\lambda) f(y)$.

Let us now select an arbitrary $\varepsilon > 0$. By the definition of supremum, there exists an $\alpha \in \mathcal{A}$ such that $f(\lambda x + (1-\lambda)y) - \varepsilon \leq f_\alpha(\lambda x + (1-\lambda)y)$. Then we use the convexity of f_α:

$$f(\lambda x + (1-\lambda)y) - \varepsilon \leq f_\alpha(\lambda x + (1-\lambda)y) \leq \lambda f_\alpha(x) + (1-\lambda) f_\alpha(y) \leq \lambda f(x) + (1-\lambda) f(y).$$

The last inequality follows by the definition of f.

We note that if $\mathcal{A}$ is a finite set, then ε is not needed. In this case, we always have and $\alpha \in \mathcal{A}$ such that $f(\lambda x + (1-\lambda)y) = f_\alpha(\lambda x + (1-\lambda)y)$. □

An alternative proof can be given via Proposition 1.2. The epigraph of f is the intersection of the epigraphs of f_α for all $\alpha \in \mathcal{A}$. These are all convex sets and the intersection of any number of convex sets is convex. Thus, the epigraph of f is also convex, and therefore, f is convex.

1.2.3 First order characterisation of convexity

Theorem 1.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then f is convex if and only if, for all $x, y \in \mathbf{dom}\, f$,*

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y).$$

Proof. "⇒" Assume f is convex. By (1.2), for every λ, $0 < \lambda \leq 1$, it follows

$$f(y) + \lambda(f(x) - f(y)) = \lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y) = f(y + \lambda(x-y)).$$

Subtracting $f(y)$ on both sides and dividing by λ on both sides, we get

$$f(x) - f(y) \geq \frac{f(y + \lambda(x-y)) - f(y)}{\lambda}$$

for $0 < \lambda \leq 1$. Taking the limit for $\lambda \to 0^+$,

$$f(x) - f(y) \geq \lim_{\lambda \to 0^+} \frac{f(y + \lambda(x-y)) - f(y)}{\lambda} = \frac{\partial f(y)}{\partial (x-y)} = \nabla f(y)^\top (x - y),$$

where $\partial f(y)/\partial(x-y)$ is the directional derivative at y along the vector $x-y$, and the last equation follows from (1.1).

"$\Leftarrow$" Assume $f(x) \geq f(y) + \nabla f(y)^\top (x-y)$ for all $x, y \in \mathbf{dom}\, f$. For all $y \in \mathbf{dom}\, f$, define the function $f_y : x \mapsto f(y) + \nabla f(y)^\top (x-y)$. Note that the function f_y is affine, therefore it is convex. Also, by assumption $f(x) \geq f_y(x)$ for all $x \in \mathbf{dom}\, f$, and by definition $f(x) = f_x(x)$. It follows that, for all $x \in \mathbf{dom}\, f$,

$$f(x) = \max_{y \in \mathbf{dom}\, f} f_y(x).$$

Thus $f(x)$ is the point-wise supremum of a family of convex functions, and is therefore convex by Proposition 1.6. □

For a geometric intuition of Theorem 1.7, recall that the hyperplane tangent to the graph of f at point $(y, f(y))$ is $H = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \,|\, t = f(y) + \nabla f(y)^\top (x-y)\}$, therefore the theorem states that a function is convex if and only if, for every y, the graph of the function lies above the hyperplane tangent to the graph of f at point $(y, f(y))$.

1.2.4 Minima of convex functions

The following is a fundamental property of convex functions.

Theorem 1.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, and let $X \subseteq \mathbf{dom}\, f$ be a convex set. Then every local minimum for f in X is a global minimum for f in X.*

Proof. Let x^* be a local minimum for f in X. By definition, there exists $\varepsilon > 0$ such that $f(x^*) \leq f(x)$ for all $x \in X$ such that $\|x - x^*\| < \varepsilon$. Suppose by contradiction that x^* is not a global minimum. Then there exists a point $y \in X$ such that $f(y) < f(x^*)$. By convexity of X, the line segment S joining x^* and y is contained in X. Consider a point $\bar{x} \in S \setminus \{x^*\}$ such that $\|x - x^*\| < \varepsilon$. Since $\bar{x} \in S$, $\bar{x} = \lambda x^* + (1-\lambda) y$ for some λ such that $0 \leq \lambda < 1$. By convexity of f,

$$\begin{aligned} f(\bar{x}) &= f(\lambda x^* + (1-\lambda) y) \\ &\leq \lambda f(x^*) + (1-\lambda) f(y) = f(x^*) + (1-\lambda)(f(y) - f(x^*)) < f(x^*), \end{aligned}$$

a contradiction. □

Furthermore, the first order necessary conditions given in Theorem 1.1 are also sufficient for convex functions.

Theorem 1.9 (First order conditions for unconstrained convex minimization)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function. A point $x^* \in \mathbf{dom}\, f$ is a global minimum of f if and only if $\nabla f(x^*) = 0$.*

Proof. We have already established in Theorem 1.1 that $\nabla f(x^*) = 0$ if x^* is a global minimum of f. Conversely, assume $\nabla f(x^*) = 0$. By Theorem 1.7, for every $x \in \mathbf{dom}\, f$,

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) = f(x^*).$$

It follows that $f(x^*)$ is a global minimum. □

The above theorem holds for "unconstrained" convex minimization problems, i.e., problems where we want to find the global minimum over the entire domain. The next theorem gives necessary and sufficient conditions for the case where we want to find the minimizer in a given convex set X.

Theorem 1.10 (First order conditions for constrained convex minimization)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function, and let $X \subseteq \mathbf{dom}\, f$ be a convex set. A point $x^* \in X$ is a global minimum of f over X if and only if*

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in X.$$

Proof. We first prove the "if" direction. That is, assume $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all $x \in X$. It follows by Theorem 1.7, for every $x \in \mathbf{dom}\, f$,

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) \geq f(x^*),$$

which implies that x^* is a global minimum for f over X.

For the "only if" direction, assume x^* is a global minimum over X. Given any $x \in X$, note that the point $\lambda x + (1-\lambda)x^*$ is in X for every λ such that $0 < \lambda \leq 1$, because X is convex. Since x^* is a global minimum over X, and noting that $\lambda x + (1-\lambda)x^* = x^* + \lambda(x - x^*)$ it follows that

$$f(x^*) \leq f(x^* + \lambda(x - x^*))$$

for every λ such that $0 < \lambda \leq 1$. Subtracting $f(x^* + \lambda(x - x^*))$ and dividing by λ on both sides, we have

$$0 \leq \frac{f(x^* + \lambda(x - x^*)) - f(x^*)}{\lambda}.$$

Taking the limit for λ that goes to 0^+, the above inequality yields

$$0 \leq \lim_{\lambda \to 0^+} \frac{f(x^* + \lambda(x - x^*)) - f(x^*)}{\lambda} = \nabla f(x^*)^\top (x - x^*),$$

which shows that $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all $x \in X$. □

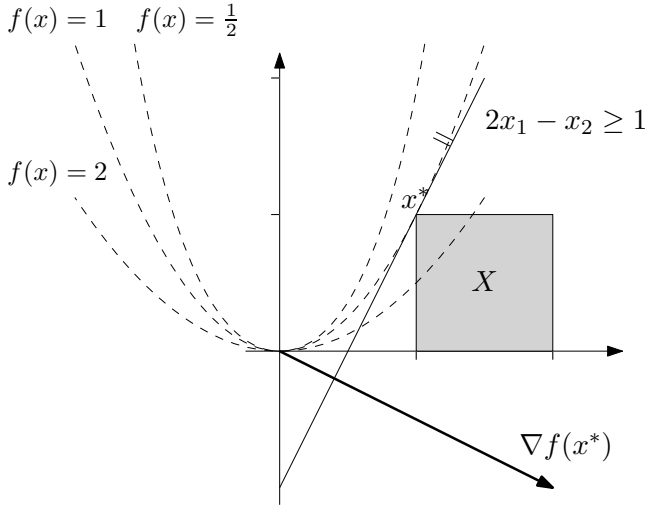
To interpret the previous theorem geometrically, let us consider the two following cases

- If x^* is in the interior of X, the previous theorem implies that x^* is a global minimum if and only if $\nabla f(x^*) = 0$. Indeed, since x^* is in the interior, then for $\varepsilon > 0$ sufficiently small the point $x = x^* - \varepsilon \nabla f(x^*)$ is also X, thus $\nabla f(x^*)^\top (x - x^*) = -\varepsilon \nabla f(x^*)^\top \nabla f(x^*) = -\varepsilon \|\nabla f(x^*)\|^2 \leq 0$ and so the only way for inequality $\nabla f(x^*)^\top (x - x^*) \geq 0$ to hold is that $\nabla f(x^*) = 0$.
- If x^* is a point on the boundary of X and $\nabla f(x^*) \neq 0$, then the previous theorem states that x^* is a minimizer in X if and only if X is contained in the half-space $\{x \,|\, \nabla f(x^*)^\top x \geq \nabla f(x^*)^\top x^*\}$. This means that, all directions pointing towards X starting from x^* on the boundary are directions of ascent.

Example 1.11. Let us consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as $f(x) = \frac{x_1^2}{x_2}$ over $\mathbf{dom}\, f = \{x \in \mathbb{R}^2 \,|\, x_2 > 0\}$. (We will show later on, in Example 1.17, that it is convex.) Let $X = \{x \in \mathbb{R}^2 \,|\, 1 \leq x_1 \leq 2,\ 0 \leq x_2 \leq 1\}$. We will show that $x^* = (1,1)^\top$ minimizes f in X. The gradient of f is

$$\nabla f(x) = \begin{pmatrix} 2x_1/x_2 \\ -x_1^2/x_2^2 \end{pmatrix},$$

thus $\nabla f(x^*) = (2,-1)^\top$. According to Theorem 1.10, x^* is a minimizer if and only if $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all $x \in X$. Noting that $\nabla f(x^*)^\top x^* = 1$, we need to verify that X is contained in the half-space $\{x \in \mathbb{R}^2 \,|\, (2,-1)x \geq 1\}$. This is indeed the case, as shown in the figure below.



Concave functions A function f is *concave* if $-f$ is convex. Observe that Theorems 1.8 and 1.10 remain true if we replace "convex" with "concave" and "minimum" with "maximum". Also, we will consider every concave function f as having values in $\mathbb{R} \cup \{-\infty\}$, where $f(x) = -\infty$ for all $x \notin \mathbf{dom}\, f$.

1.2.5 Second order characterisation of convexity

We now present a characterisation of twice differentiable convex functions. As an immediate use of this characterisation, we will be able to decide whether a twice differentiable function is convex.

Taylor expansion of univariate functions Let us first focus on univariate functions, and recall the second order Taylor-expansion.

Theorem 1.12. *Assume that $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable on $\mathbf{dom}\, f$. Then for every $x, y \in \mathbf{dom}\, f$, we can write*

$$f(x) = f(y) + f'(y)(x-y) + \frac{1}{2} f''(\bar{x})(x-y)^2$$

for some $\bar{x} \in [x, y]$.

The Taylor expansion shows that for a small $\varepsilon > 0$, the function f in the interval $[y-\varepsilon, y+\varepsilon]$ can be well approximated by the linear function $f(y)+f'(y)(x-y)$. Recall that according to Theorem 1.4, f'' is nonnegative for convex functions. Therefore, the Taylor expansion gives $f(x) \geq f(y)+f'(y)(x-y)$, as in Theorem 1.7.

Hessians The Taylor expansion can be generalised for multivariate functions. First, we need to define the notion of second derivative for multivariate functions. The *Hessian* of f at point $x \in \mathbf{dom}\, f$ is the $n \times n$ matrix

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

that is, the (i, j) entry of $\nabla^2 f(x)$ is the second partial derivative of f at x with respect to the variables x_i and x_j. Note that $\nabla^2 f(x)$ is symmetric since $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$.

The function f is said to be *twice differentiable* if $\mathbf{dom}\, f$ is an open set and the Hessian of f exists at every point in $\mathbf{dom}\, f$. Note that if $n = 1$, then $\nabla^2 f$ is simply the second derivative of f. Assume that $f(x) = \sum_{i=1}^n f_i(x_i)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is a univariate function in x_i for each $i = 1, 2, \ldots, n$. Then the Hessian is a diagonal matrix where the ith entry is $f_i''(x_i)$.

Taylor expansion of multivariate functions We now give the extension of Theorem 1.12 to multivariate functions.

Theorem 1.13. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable on $\mathbf{dom}\, f$. Then for every $x, y \in \mathbf{dom}\, f$, we can write*

$$f(x) = f(y) + \nabla f(y)^\top (x-y) + \frac{1}{2}(x-y)^\top \nabla^2 f(\bar{x})(x-y)$$

for some $\bar{x} \in [x, y]$.

Here, $[x, y] = \{\lambda x + (1-\lambda)y \,|\, 0 \leq \lambda \leq 1\}$ denotes the line segment between the vectors x and y. Theorem 1.7 now states that for convex functions, the third term on the right hand side must be nonnegative. This property will be guaranteed for postive semidefinite matrices, as defined next.

Positive definite and semidefinite matrices. Let $A \in \mathbb{R}^{n\times n}$ be a *symmetric* matrix, that is, $a_{ij} = a_{ji}$ for every i, j. We say that A is *positive definite* if, for all $x \in \mathbb{R}^n \setminus \{0\}$, $x^\top Ax > 0$. We say that A is *positive semidefinite* if, for all $x \in \mathbb{R}^n$, $x^\top Ax \geq 0$.

Further, we say that A is *negative definite* if $-A$ is positive definite, that is, if, for all $x \in \mathbb{R}^n \setminus \{0\}$, $x^\top Ax < 0$. Similarly, A is *negative semidefinite*, if, for all $x \in \mathbb{R}^n$, $x^\top Ax \leq 0$.

Clearly, the identity matrix is positive definite. Also, a diagonal matrix with all positive entries is positive definite, and a diagonal matrix with all nonnegative entries is positive semidefinite.

A well-known example of positive semidefinite matrices is the covariance matrix of a random vector. Indeed, recall that, if $p \in \mathbb{R}^n$ is a random vector with mean $\bar{p}$, then its covariance matrix is the matrix Σ whose (i, j) entry is $\Sigma_{ij} = \mathbb{E}[(p_i - \bar{p}_i)(p_j - \bar{p}_j)]$. One can easily compute that, for all $x \in \mathbb{R}^n$ $\mathrm{Var}(p^\top x) = x^\top \Sigma x$. Since the variance of random variable is always nonnegative, it follows that $x^\top \Sigma x \geq 0$ for all $x \in \mathbb{R}^n$.

A matrix can be neither positive nor negative semidefinite. Such matrices are called *indefinite*. For example, let $A = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix}$. Then, for $x = (1, 0)$, $x^\top Ax = 1$, and for $y = (0, 1)$, $y^\top Ay = -2$. We can recognise positive (semi)definite matrices based on their eigenvalues.

Theorem 1.14. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then, A is positive definite if and only if all its eigenvalues are positive. A is positive semidefinite if and only if all its eigenvalues are nonnegative.*

Proof. We only prove the statements for positive definite matrices; the proofs easily extend to the positive semidefinite case.

"only if" direction: Assume A has a negative eigenvalue $\lambda \leq 0$. For the corresponding eigenvector $v \in \mathbb{R}^n$, $Av = \lambda v$. Then, $v^\top Av = v^\top(\lambda v) = -\lambda\|v\|^2 \leq 0$.

"if" direction: Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A (with multiplicities). It is well-known that a symmetric matrix A can be orthogonally diagonalised. That is, we can write

$$A = P^\top DP,$$

where P is an orthogonal matrix, and D is a diagonal matrix with the diagonal entries being the eigenvalues: $D_{ii} = \lambda_i$. Consider now any vector $x \in \mathbb{R}^n$, and let $y = Px$. Then,

$$x^\top Ax = x^\top P^\top DPx = y^\top Dy = \sum_{i=1}^{n} \lambda_i y_i^2.$$

Since P is nonsingular, $y = 0$ if and only if $x = 0$. If $\lambda_i > 0$ for all i, then it follows that $x^\top Ax > 0$ whenever $x \neq 0$. □

Second order characterisation of local extrema Recall that for a twice differentiable univariate function $f : \mathbb{R} \to \mathbb{R}$, if x^* is a local minimum (maximum), then $f'(x^*) = 0$, and $f''(x^*) \geq 0$ ($f''(x^*) \leq 0$). This statement naturally extends to multivariate functions. In Theorem 1.1, we have seen that for a differentiable function f, $\nabla f(x^*) = 0$ must hold for every local minimum and local maximum x^*. For the Hessians, the following holds.

Theorem 1.15. *Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable on $\mathbf{dom}\, f$. If $x^* \in \mathbf{dom}\, f$ is a local minimum, then $\nabla^2 f(x^*)$ is positive semidefinite. If $x^* \in \mathbf{dom}\, f$ is a local maximum, then $\nabla^2 f(x^*)$ is negative semidefinite.*

Second order characterisation of convexity If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then the Hessian reveals if the function is convex.

Theorem 1.16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable.*

(i) If $\nabla^2 f(z)$ is positive semidefinite for every $z \in \mathbf{dom}\, S$, then f is convex.

(ii) Assume that f is convex, and $\nabla^2 f$ is continuous on $\mathbf{dom}\, f$*. Then, $\nabla^2 f(z)$ is positive semidefinite for every $z \in \mathbf{dom}\, f$.*

We only prove part *(i)*. This is immediate from Theorems 1.7 and 1.13. Let us select any $x, y \in \mathbf{dom}\, f$, and consider the Taylor expansion. By assumption, $\nabla^2 f(\bar{x})$ is positive semidefinite, and therefore

$$f(x) - f(y) - \nabla f(y)^\top (x - y) = \frac{1}{2}(x - y)^\top \nabla^2 f(\bar{x})(x - y) \geq 0.$$

Example 1.17. Consider the function as in Example 1.11, that is, $f(x) = \dfrac{x_1^2}{x_2}$, where $\mathbf{dom}\, f = \{x \in \mathbb{R}^2 \,|\, x_2 > 0\}$. Then,

$$\nabla^2 f(x) = \begin{pmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{pmatrix} = \frac{2}{x_2^3} \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix}.$$

Pick any vector $v \in \mathbb{R}^2$. Then,

$$v^\top \nabla^2 f(x) v = \frac{2}{x_2^3} \cdot (v_1, v_2)^\top \begin{pmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{pmatrix} (v_1, v_2) = \frac{2}{x_2^3} \cdot (v_1 x_2 - v_2 x_1)^2 \geq 0,$$

using that $x_2 > 0$. We have thus shown that $f(x)$ is convex.

Quadratic functions By a *quadratic function* $f : \mathbb{R}^n \to \mathbb{R}$ we mean a degree two polynomial function. For example, $f(x_1, x_2) = -x_1^2 + 3x_1 x_2 + 2x_2^2 - 5x_1 + 6x_2 + 3$. We can write every quadratic function in the form

$$f(x) = x^\top Q x + p^\top x + r,$$

where $Q \in \mathbb{R}^{n \times n}$ is a symmetric matrix, $p \in \mathbb{R}^n$, $r \in \mathbb{R}$. For the particular example above, the representation is $Q = \begin{pmatrix} -1 & 1.5 \\ 1.5 & 3 \end{pmatrix}$, $p = (-5, 6)$, $r = 3$.

Theorem 1.18. *Let $Q \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $p \in \mathbb{R}^n$, and $r \in \mathbb{R}$. The quadratic function $f(x) = x^\top Q x + p^\top x + r$ is convex if and only if Q is positive semidefinite.*

Proof. The function is twice continuously differentiable, and it is easy to see that $\nabla f^2(x) = 2Q$. The claim follows using Theorem 1.16. □

In the special case $n = 1$, we obtain the well-know fact that $f(x) = ax^2 + bx + c$ is convex if and only if $a \geq 0$.

1.3 Convex Optimization Problems

Recall that a general optimization problem is of the form

$$\begin{array}{ll} \min & f_0(x) \\ & f_i(x) \leq 0 \quad i = 1, \ldots, m \\ & h_i(x) = 0 \quad i = 1, \ldots, k \end{array} \tag{1.3}$$

The *domain* of problem (1.3) is the set of points for which the objective function and the constraints functions are defined, that is

$$\mathcal{D} = \left(\bigcap_{i=0}^{m} \mathbf{dom}\, f_i \right) \bigcap \left(\bigcap_{i=1}^{k} \mathbf{dom}\, h_i \right).$$

The *feasible region* is the set X of all points in $\mathcal{D}$ satisfying the constraints.

We say that the above problem is a *convex optimization problem* if $f_0, \ldots, f_m$ are convex functions, and $h_1, \ldots, h_k$ are affine functions; that is, there exist $a_1, \ldots, a_k \in \mathbb{R}^m$ and $b_1, \ldots, b_k \in \mathbb{R}$ such that $h_i(x) = a_i^\top x - b_i$ $(i = 1, \ldots, k)$. The equality constraints can therefore be expressed as $a_i^\top x = b_i$.

Note that the requirement that $h_1, \ldots, h_k$ are affine is needed in order for the definition to be consistent. Indeed, if we replaced each equality constraint $h_i(x) = 0$ with the two inequality constraints $h_i(x) \leq 0$, $-h_i(x) \leq 0$, then a convex problem should satisfy that both h_i and $-h_i$ are convex, that is, h_i needs to be both concave and convex. The only functions that are both concave and convex are the affine ones, therefore we need to require that h_i are affine.

Note that, since $f_1, \ldots, f_m$ are convex functions, the sets $\{x \mid f_i(x) \leq 0\}$ are convex sets (this is easy to show). The sets $\{x \mid a_i^\top x = b_i\}$ are hyperplanes, and therefore convex.

These facts and Theorem 1.8 imply the following important facts.

Remark 1.19. *If problem* (1.3) *is convex, then*

1. *The feasible region X is convex, because it is the intersection of convex sets.*
2. *Every local optimum for f_0 in X is also a global optimum.*

Example 1.20. The Markowitz portfolio optimization problem (seen already in MA423):

$$\begin{gathered} \min\ x^\top \Sigma x \\ \text{s.t.} \\ \bar{p}^\top x \geq r_{\min} \\ \textstyle\sum_{i=1}^n x_i = B \\ x \geq 0 \end{gathered}$$

where $\bar{p}$ is the vector of means of the (random) vector of returns p, and Σ is the covariance matrix of p.

Note that all constraints are defined by affine functions. The objective function $x^\top \Sigma x$ is a quadratic convex function, because Σ is positive semidefinite. To verify this, we need to show that $x^\top \Sigma x \geq 0$ for all $x \in \mathbb{R}^n$. Recall that $x^\top \Sigma x \geq 0$ is the variance of the return of portfolio x, i.e. $x^\top \Sigma x = \mathbb{E}[(p^\top x - \bar{p}^\top x)^2]$, so it is non-negative because the variance of a random variable is always non-negative.

Therefore Markowitz's model is a convex optimization problem.

Concave maximization Convex optimization problems have been defined as minimization problems. However, if in a maximization problem of the form

$$\begin{array}{ll} \max & f_0(x) \\ & f_i(x) \leq 0 \quad i = 1, \ldots, m \\ & h_i(x) = 0 \quad i = 1, \ldots, k \end{array}$$

the objective function f_0 is concave, while $f_1, \ldots, f_m$ are convex and $h_1, \ldots, h_k$ affine, we will also say that the problem is a convex optimization problem. This is justified by the fact that the equivalent minimization problem obtained by replacing "$\max f_0(x)$" with "$\min -f_0(x)$" is a convex optimization problem, because $-f_0$ is convex.

Chapter 2

Lagrangian Duality and the Karush-Kuhn-Tucker conditions

2.1 The Lagrangian Dual

Lagrangian The *Lagrangian* of problem (1.3) is the function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k :\to \mathbb{R}$ defined by

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{k} \nu_i h_i(x), \tag{2.1}$$

where $\mathbf{dom}\, L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^k$. Here λ and ν are vectors of variables in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Variable λ_i is the *Lagrange multiplier* of constraint $f_i(x) \leq 0$, whereas ν_i is the Lagrange multiplier of constraint $h_i(x) = 0$.

Lemma 2.1. *For every feasible point $\bar{x}$ for (1.3) and every $(\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^k$, $\lambda \geq 0$, we have that*

$$L(\bar{x}, \lambda, \nu) \leq f_0(\bar{x}) \tag{2.2}$$

Proof. Since $\bar{x}$ is feasible, it follows that $f_i(\bar{x}) \leq 0$ and $h_i(x) = 0$. Therefore

$$L(\bar{x}, \lambda, \nu) = f_0(\bar{x}) + \sum_{i=1}^{m} \lambda_i f_i(\bar{x}) + \sum_{i=1}^{k} \nu_i h_i(\bar{x}) \leq f_0(\bar{x}),$$

because $\lambda_i \geq 0$ and thus $\lambda_i f_i(x) \leq 0$ for $i = 1, \ldots, m$. □

Lagrange dual function We define the *Lagrange dual function* $g : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$ as

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu). \tag{2.3}$$

Recall that $X \subseteq \mathcal{D}$ is the feasible region of the problem. If we denote by p^* the optimal value of problem (1.3), that is $p^* = \inf_{x \in X} f_0(x)$, it follows from (2.2) that, for all $(\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^k$, $\lambda \geq 0$,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq \inf_{x \in X} L(x, \lambda, \nu) \leq \inf_{x \in X} f_0(x) = p^*. \tag{2.4}$$

That is, for every choice of $(\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^k$, $\lambda \geq 0$, the value $g(\lambda, \nu)$ is a lower-bound on the optimal value p^*.

Next we point out an interesting property of the Lagrange dual function g.

Lemma 2.2. *The Lagrange dual function g is concave.*

Proof. Indeed, for every $x \in \mathcal{D}$, the function $\theta_x : \mathbb{R}^m \times \mathbb{R}^k :\to \mathbb{R}$ defined by $\theta_x(\lambda, \nu) = L(x, \lambda, \nu)$ is affine, and therefore concave. By definition,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} \theta_x(\lambda, \nu),$$

therefore g is the pointwise infimum of the family of concave functions $\{\theta_x \,|\, x \in \mathcal{D}\}$, and is therefore concave. This follows since Proposition 1.6 is applicable to $-g$, showing that a function defined as the pointwise supremum of convex functions is convex. Consequently, g is concave. □

Lagrangian dual problem As we have seen, for every $\lambda \geq 0$ and $\nu \in \mathbb{R}^k$, the value $g(\lambda, \nu)$ provides a lower-bound to the optimal value of (1.3). The *Lagrangian dual* is the problem of finding the best such lower bound. That is,

$$\begin{array}{ll} \max & g(\lambda, \nu) \\ \text{s.t.} & \lambda \geq 0 \\ & (\lambda, \nu) \in \mathbf{dom}\, g \end{array} \tag{2.5}$$

By Lemma 2.2, the above problem is a convex optimization problem (because the constraints are linear and the objective is to maximize a concave function). Note that this is the case even when the primal problem (1.3) is not convex!

The following is an immediate consequence of (2.4).

Theorem 2.3 (Week Lagrangian Duality)**.** *Let p^* be the optimal value of the primal problem* (1.3)*, and let d^* be the optimal value of the dual problem* (2.5)*. Then*

$$d^* \leq p^*.$$

We call $p^* - d^*$ the *duality gap* of problem (1.3). As we have seen, strong duality always holds for LP problems, but does not hold in general for convex problems (see Example 2.6).

Example 2.4. Consider the problem

$$\begin{array}{ll} \min & x^2 + 1 \\ & (x-2)(x-4) \leq 0 \end{array}$$

Note that it is a convex optimization problem, since both functions $x \mapsto x^2+1$ and $x \mapsto (x-2)(x-4)$ are convex quadratic functions. Since both functions are defined over all of $\mathbb{R}$, the domain of the problem is $\mathcal{D} = \mathbb{R}$.

The feasible region is the interval $[2, 4]$, and the minimum is attained at $x^* = 2$, with optimal objective value $p^* = 5$.

The Lagrangian of the above problem is the function of two variables

$$\begin{aligned} L(x, \lambda) &= x^2 + 1 + \lambda(x-2)(x-4) \\ &= (1+\lambda)x^2 - 6\lambda x + 8\lambda + 1. \end{aligned}$$

To compute the Lagrangian dual function, we need to compute, for all $\lambda \in \mathbb{R}$,

$$g(\lambda) = \inf_{x \in \mathbb{R}} L(x, \lambda) = \inf_{x \in \mathbb{R}} (1+\lambda)x^2 - 6\lambda x + 8\lambda + 1.$$

Observe that for $\lambda \leq -1$ the above infimum is $-\infty$. For $\lambda > 1$, the function $(1+\lambda)x^2 - 6\lambda x + 8\lambda + 1$ is a convex quadratic function, therefore its global minima in $\mathbb{R}$ are the points with zero derivative. We compute the derivative of $L(x, \lambda)$ with respect to x and set it to zero.

$$\frac{\partial L(x, \lambda)}{\partial x} = 2(1+\lambda)x - 6\lambda = 0.$$

The zero of the above equation is the point

$$\bar{x} = \frac{3\lambda}{1+\lambda},$$

thus, for all $\lambda > -1$,

$$\begin{aligned} g(\lambda) &= L(\bar{x}, \lambda) = (1+\lambda)\left(\frac{3\lambda}{1+\lambda}\right)^2 - 6\lambda\frac{3\lambda}{1+\lambda} + 8\lambda + 1 \\ &= \frac{-\lambda^2 + 9\lambda + 1}{1+\lambda}, \end{aligned}$$

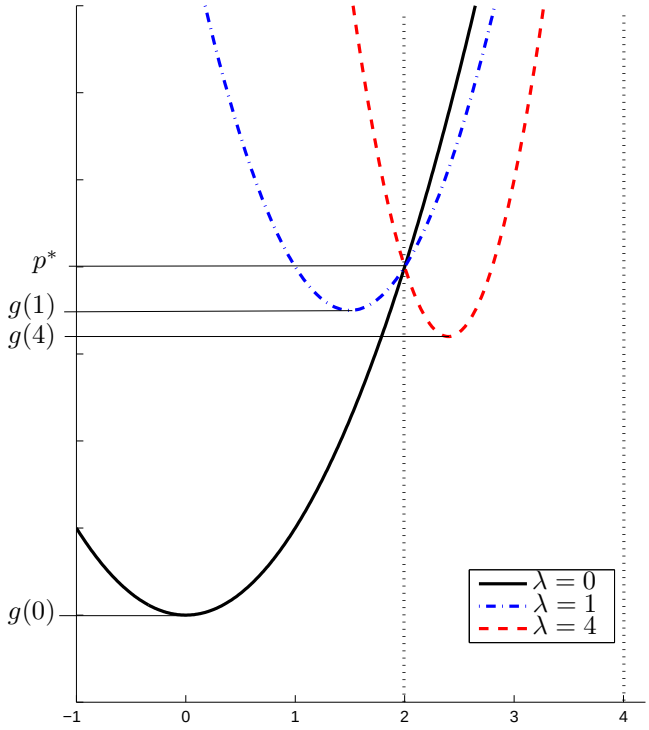


Figure 2.1: Graph of the function $L(x, \lambda)$ for $\lambda = 0, 1, 4$, and corresponding values of $g(\lambda)$. Note that $g(\lambda) \leq p^*$

and $\mathbf{dom}\,(g) = \{\lambda \in \mathbb{R} \,|\, \lambda > -1\}$. The Lagrangian dual is therefore

$$\begin{array}{ll} \max & \dfrac{-\lambda^2 + 9\lambda + 1}{1 + \lambda} \\ \text{s.t.} & \lambda \geq 0 \end{array}$$

To solve the above, we compute the derivative of g and set it to zero

$$\begin{aligned} g'(\lambda) &= \frac{(-2\lambda + 9)(1 + \lambda) - (-\lambda^2 + 9\lambda + 1)}{(1 + \lambda)^2} \\ &= -\frac{\lambda^2 + 2\lambda - 8}{(1 + \lambda)^2} = 0 \end{aligned}$$

The only non-negative solution to the above equation is $\lambda^* = 2$, which is therefore the dual optimal solution. The optimal value of the dual is therefore $d^* = g(2) = 5$. Note that in this case $p^* = d^*$, thus strong duality holds.

Example 2.5. (Dual of an LP problem.) Consider an LP problem of the form

$$\begin{array}{ll} \min & c^\top x \\ \text{s.t.} & Ax \geq b \end{array}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Rewriting the constraints in the form

$$b - Ax \leq 0,$$

the Lagrangian of the above problem is

$$L(x, \lambda) = c^\top x + \lambda^\top (b - Ax) = b^\top \lambda + (c^\top - \lambda^\top A)x,$$

where λ is a vector of m variables. Observe that $L(x, \lambda)$ is an affine function in x, therefore it is either a constant function or it not lower bounded. Note that $L(x, \lambda)$ is constant if and only if $A^\top \lambda - c = 0$, therefore the lagrangian dual function is

$$g(\lambda) = \inf_{x \in \mathbb{R}^n} L(x, \lambda) = \begin{cases} b^\top \lambda & \text{if} A^\top \lambda = c \\ -\infty & \text{otherwise.} \end{cases}$$

It follows that the Lagrangian dual function is given by $g(\lambda) = b^\top \lambda$, and it is defined over $\mathbf{dom}\, g = \{\lambda \in \mathbb{R}^m \mid A^\top \lambda = c\}$. The Lagrangian dual of the LP function is therefore

$$\begin{array}{ll} \max & b^\top \lambda \\ \text{s.t.} & A^\top \lambda = c \\ & \lambda \geq 0. \end{array}$$

This is the usual LP dual.

Example 2.6. (Convex problem with strict duality gap) Consider the problem

$$\begin{array}{rll} p^* = \min & e^{-x_1} \\ \text{s.t.} & \dfrac{x_1^2}{x_2} \leq 0 \end{array}$$

defined over $\mathcal{D} = \{x \in \mathbb{R}^2 \mid x_2 > 0\}$. The above problem is convex, as one can verify. Observe that the feasible region of the problem is the set $X = \{x \in \mathbb{R}^2 \mid x_1 = 0,\, x_2 > 0\}$. In particular, every feasible solution has objective value 1, therefore $p^* = 1$.

Let us now compute the lagrangian dual. The lagrangian is $L(x, \lambda) = e^{-x_1} + \lambda \frac{x_1^2}{x_2}$. We need to compute $g(\lambda) := \inf_{x \in \mathcal{D}} L(x, \lambda)$ for $\lambda \geq 0$. Observe that $L(x, \lambda) \geq 0$ for all $x \in \mathcal{D}$ and all $\lambda \geq 0$, therefore $g(\lambda) \geq 0$. On the other hand, there are $x \in \mathcal{D}$ for which $L(x, \lambda)$ takes arbitrarily small value (it suffices to consider any sequence of points in $\mathcal{D}$ for which $x_1 \to +\infty$ and $\frac{x_1^2}{x_2} \to 0$). If follows that $g(\lambda) = 0$ for all $\lambda \geq 0$, and therefore $d^* = 0$. Thus the duality gap is $p^* - d^* = 1 - 0 = 1$.

Despite cases in which convex optimization problems have positive duality gaps, such as the one in Example 2.6, strong duality holds for convex optimization problems under fairly general conditions.

Definition 2.7 (Slater Conditions)**.** *We say that a convex optimization problem* (1.3) *satisfies the Slater conditions if there exists a feasible solution $\bar{x}$ in the interior of $\mathcal{D}$ such that $f_i(\bar{x}) < 0$ for $i = 1, \ldots, m$.*

Theorem 2.8 (Strong duality under Slater conditions)**.** *Strong duality holds for every convex optimization problem satisfying Slater condition.*

We do not present the proof of the above theorem here. Note that the problem in Example 2.6 does not satisfy Slater's conditions, because every feasible solution satisfies the only constraint of the problem to equality.

2.2 Karush-Kuhn-Tucker conditions

Consider a general problem (not necessarily convex) of the form (1.3), and assume that the functions $f_0, f_1, \ldots, f_m, h_1, \ldots, h_k$ are differentiable.

Suppose that we know that strong duality holds, and that both the primal and the dual problem admit an optimal solution, say x^* for the primal and (λ^*, ν^*) for the dual problem. It follows that

$$\begin{array}{rcl} f_0(x^*) & = & g(\lambda^*, \nu^*) \\ & = & \inf\limits_{x \in \mathcal{D}} L(x, \lambda^*, \mu^*) \\ & \leq^* & L(x^*, \lambda^*, \mu^*) \\ & = & f_0(x^*) + \sum\limits_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum\limits_{i=1}^{k} \nu_i^* h_i(x^*) \\ & \leq^{**} & f_0(x^*). \end{array}$$

This shows that equality must holds throughout in the above chain of inequalities. For inequality (*), this means that

$$\inf_{x\in\mathcal{D}} L(x,\lambda^*,\mu^*) = L(x^*,\lambda^*,\mu^*).$$

That is, x^* must be a global minimum for $L(x,\lambda^*,\mu^*)$. In particular, since $f_0, f_1, \ldots, f_m$ are differentiable, it follows that $L(x,\lambda^*,\mu^*)$ is differentiable. By Theorem 1.1, it follows that the gradient of $L(x,\lambda^*,\mu^*)$ computed at x^* must be the zero vector. That is

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{k} \nu_i^* \nabla h_i(x^*) = 0. \tag{2.6}$$

For inequality (**) to be satisfied at equality (observing that $h_i(x^*) = 0$, because x^* is feasible, and that $\lambda_i^* f_i(x^*) \le 0$ because x^* is feasible and $\lambda^* \ge 0$) it must be that

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m. \tag{2.7}$$

Equations (2.7) are called *complementary slackness conditions*. They state that, if an optimal primal solution satisfies the ith constraint as strict inequality, then the corresponding dual variable λ_i should be zero in an optimal dual solution.

We summarize the above discussion in the following statement.

Lemma 2.9. *Let $f_0, f_1, \ldots, f_m$, $h_1, \ldots, h_k : \mathbb{R}^n \to \mathbb{R}$ be differentiable functions. Assume that strong duality holds for the optimization problem* (1.3)*. If* (1.3) *admits an optimum x^* and its dual admits an optimum (λ^*, ν^*), then these must satisfy the following conditions*

$$\begin{aligned} f_i(x^*) &\le 0 \quad (i = 1, \ldots, m) \\ h_i(x^*) &= 0 \quad (i = 1, \ldots, k) \\ \lambda_i^* &\ge 0 \quad (i = 1, \ldots, m) \\ \lambda_i^* f_i(x^*) &= 0 \quad (i = 1, \ldots, m) \\ \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{k} \nu_i^* \nabla h_i(x^*) &= 0. \end{aligned} \tag{2.8}$$

The above are known as Karush-Kuhn-Tucker (KKT) conditions. Observe that the above result only says that the KKT conditions are necessarily satisfied by every pair of optimal primal/dual solutions (x^*, λ^*, ν^*) if strong duality holds. However, the next example illustrates that it is not true in general that for every solution (x^*, λ^*, ν^*) to the KKT system the point x^* is a primal optimum.

Example 2.10. Consider the (non-convex) optimization problem $\min\{x^3 \,|\, x^2 \le 1\}$. Clearly, the only optimal solution is $x = -1$, with value -1. The Lagrangian is $L(x,\lambda) = x^3 + \lambda(x^2 - 1)$, thus the KKT conditions are

$$\begin{aligned} x^2 - 1 &\le 0 \\ \lambda &\ge 0 \\ \lambda(x^2 - 1) &= 0 \\ 3x^2 + 2\lambda x &= 0. \end{aligned}$$

The point $(x^*, \lambda^*) = (0, 0)$ is a solution for the KKT conditions, but the point $x^* = 0$ is not a primal optimum (it has value 0, whereas the optimal value is -1).

For convex optimization problems, however, the KKT conditions are also sufficient, as shown in the following theorem.

Theorem 2.11. *Let $f_0, f_1, \ldots, f_m$ be convex differentiable functions and $h_1, \ldots, h_k$ be affine functions. If the KKT conditions for problem* (1.3) *have a solution (x^*, λ^*, ν^*), then x^* is optimal for problem* (1.3)*, (λ^*, ν^*) is optimal for its dual, and strong duality holds.*

Proof. Assume that (x^*, λ^*, ν^*) is a solution to the KKT conditions (2.8). In particular, x^* is feasible for (1.3) and $\lambda^* \geq 0$. Thus it suffices to show that $f_0(x^*) \leq g(\lambda^*, \nu^*)$.

We only need to show that $f_0(x^*) = g(\lambda^*, \nu^*)$. We have

$$\begin{aligned} f_0(x^*) &= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{k} \nu_i^* h_i(x^*) \\ &= \inf_{x \in \mathcal{D}} f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{k} \nu_i^* h_i(x) \\ &= g(\lambda^*, \nu^*). \end{aligned}$$

The first equality follows from the fact that $h_i(x^*) = 0$ $(i = 1, \ldots, k)$ and $\lambda^* f_i(x^*) = 0$ $(i = 1, \ldots, m)$. The second equality follows from Theorem 1.10, since the function $L(x, \lambda^*, \nu^*)$ is convex (because $f_1, \ldots, f_m$, $h_1, \ldots, h_k$ are convex), and its gradient at x^* is zero (from condition (2.6)). □

Example 2.12. Consider the problem

$$\begin{array}{ll} \min & \frac{1}{2} x^\top P x + q^\top x + r \\ s.t. & -1 \leq x_i \leq 1 \qquad i = 1, 2, 3 \end{array}$$

where

$$P = \begin{pmatrix} 13 & 12 & -2 \\ 12 & 17 & 6 \\ -2 & 6 & 12 \end{pmatrix}, \quad q = \begin{pmatrix} -22 \\ -29/2 \\ 13 \end{pmatrix}, \quad r = 1.$$

We will show that the point $x^* = (1, 1/2, -1)$ is a global optimum.

The above is a convex optimization problem, because the constraint functions are affine, while the objective function is convex quadratic (indeed, matrix P is positive semidefinite). To show that x^* is optimal, we will find a dual solution that satisfies the KKT conditions together with x^*.

The constraints of the problems can be written as $-x_i - 1 \leq 0$ and $x_i - 1 \leq 0$, $i = 1, 2, 3$. We assign lagrange multipliers λ_i^0 to constraint $-x_i - 1 \leq 0$, and λ_i^0 to constraint $x_i - 1 \leq 0$, $i = 1, 2, 3$. We denote by λ^0, λ^1 the corresponding vectors. The Lagrangian is

$$L(x, \lambda^0, \lambda^1) = \frac{1}{2} x^\top P x + q^\top x + r + \sum_{i=1}^{3} \lambda_i^0 (-x_i - 1) + \sum_{i=1}^{3} \lambda_i^1 (x_i - 1).$$

Recall that the gradient of the objective function f at any given point x is

$$\nabla f(x) = Px + q.$$

Therefore, the KKT conditions are

$$\begin{aligned} -x_i - 1 &\leq 0 \quad i = 1, 2, 3 \\ x_i - 1 &\leq 0 \quad i = 1, 2, 3 \\ \lambda^0, \lambda^1 &\geq 0 \\ \lambda_i^0 (-x_i - 1) &= 0 \quad i = 1, 2, 3 \\ \lambda_i^1 (x_i - 1) &= 0 \quad i = 1, 2, 3 \\ Px + q - \lambda^0 + \lambda^1 &= 0 \end{aligned}$$

Clearly $-1 \leq x_i^* \leq 1$. Furthermore, from the complementary slackness conditions we get

$$\begin{aligned} x_1^* > -1 &\Rightarrow \lambda_1^0 = 0 \\ -1 < x_2^* < 1 &\Rightarrow \lambda_2^0, \lambda_2^1 = 0 \\ x_3^* < 1 &\Rightarrow \lambda_3^1 = 0 \end{aligned}$$

Finally, substituting x^* into the last KKT equation and setting Lagrangian multiplier to zero as above, we obtain

$$Px^* + q - \lambda^0 + \lambda^1 = \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \lambda_3^0 \end{pmatrix} + \begin{pmatrix} \lambda_1^1 \\ 0 \\ 0 \end{pmatrix} = 0.$$

The only solution to the above system is $\lambda_3^0 = 2 \geq 0$, $\lambda_1^1 = 1 \geq 0$. It follows that x^* is an optimal primal solution. An optimal dual solution is defined by $\lambda^0 = (0, 0, 2)^\top$ and $\lambda^1 = (1, 0, 0)^\top$.

Example 2.13. (KKT conditions for an LP problem.) Consider again an LP problem of the form

$$\begin{array}{ll} \min & c^\top x \\ \text{s.t.} & Ax \geq b \end{array}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. As before, the Lagrangian of the above problem is

$$L(x, \lambda) = c^\top x + \lambda^\top (b - Ax) = b^\top \lambda + (c^\top - \lambda^\top A)x.$$

Thus

$$\nabla L(x, \lambda) = c - A^\top \lambda.$$

It follows that the KKT conditions are

$$\begin{aligned} Ax &\geq b \\ \lambda &\geq 0 \\ (b - a_i^\top x)\lambda_i &= 0 \qquad i = 1, \ldots, m \\ A^\top \lambda &= c \end{aligned}$$

Thus the KKT conditions, when specialized to Linear Programming, enforce that x is a primal feasible solution, λ is a dual feasible solution, and that x and λ are in complementary slackness. These are the optimality conditions that we have already seen for Linear Programming.

Chapter 3

Gradient descent

In this chapter, we focus on simple algorithmic approaches to solving convex programs. We start with the simplest case, *unconstrained minimisation*, that is, for a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we want to solve

$$\min_{x \in \mathbf{dom}\, f} f(x).$$

Descent methods A *descent method* uses the gradient to construct a sequence of points that converge to the optimum. We start from an *initial point* $x^{(0)} \in \mathbf{dom}\, f$, and find further points $x^{(1)}, x^{(2)}, \ldots$ such that the function value decreases at each step: $f(x^{(k)}) > f(x^{(k+1)})$.

The general scheme of such an algorithm is that given the point $x^{(k)}$, we select a *search direction* $\Delta x^{(k)}$ to move, and a *step size* $t_k > 0$. We choose the next point as

$$x^{(k+1)} = x^{(k)} + t_k \Delta x^{(k)}. \tag{3.1}$$

There are various methods for choosing the search direction and the step size. Let us consider the simplest such method, *gradient descent*.

3.1 The gradient descent method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Recall from Theorem 1.9 that x^* is a global minimum point if and only if $\nabla f(x^*) = 0$.

Recall from the Taylor expansion that the gradient $\nabla f(x)$ provides a linear approximation of f around x. In particular, for a vector $x \in \mathbf{dom}\, f$, direction Δx, and step size $t > 0$,

$$f(x + t\Delta x) \approx f(x) + t\nabla f(x)^\top \Delta x \tag{3.2}$$

Since we wish to decrease the function value, we need to select a direction t such that $\nabla f(x)^\top \Delta x < 0$. A natural choice is $\Delta x = -\nabla f(x)$, the direction opposite to the gradient.

Hence, in the update formula (3.1), we use $\Delta x^{(k)} = -\nabla f(x^{(k)})$. We now discuss possible ways of determining the step size t_k.

Exact line search A natural choice is to find the minimiser of f in the search direction. That is, we select

$$t_k := \mathrm{argmin}_{t \geq 0} f(x^{(k)} - t\nabla f(x^{(k)})) \tag{3.3}$$

Solving this problem amounts to minimising a univariate convex function $g(t) = f(x^{(k)} - t\nabla f(x^{(k)}))$ over $t \geq 0$. In some cases this can be done very efficiently, but it could be computationally expensive for other functions.

Constant step size A common choice is using a constant step size $t_k = \mu$ throughout. However, this step size must be selected carefully. If μ is chosen too small, the method may converge but may take too many iterations. If μ is chosen too large, then we might not even obtain a descent method: $f(x^{(k+1)}) > f(x^{(k)})$ could be possible!

Backtracking line search A fast alternative to exact line search that is able to calibrate the step size is backtracking line search. We select two parameters: $0 < \alpha < \frac{1}{2}$, and $0 < \beta < 1$. Instead of finding the optimal step size, we wish to obtain t such that

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leq f(x^{(k)}) - \alpha t \|\nabla f(x^{(k)})\|^2. \tag{3.4}$$

From (3.2), using that $\Delta x^{(k)} = -\nabla f(x^{(k)})$, we see that this must be true for small enough $t > 0$. We wish to approximately identify the largest value of t where this holds.

This can be done by starting form a (relatively large value) $t = 1$, and as long as (3.4) is not satisfied, we decrease t by a factor β. This can be formally described as follows.

1. Set $t := 1$.
2. *While* $f(x^{(k)} - t\nabla f(x^{(k)})) > f(x^{(k)}) - \alpha t \|\nabla f(x)\|^2$, *update* $t := \beta t$.

Setting the value of β gives a trade-off between speed and accuracy. For larger values of β (that is, close to 1), we might need several calibrating iterations. On the other hand, a smaller value (e.g. $\beta = 0.3$) provides fast convergence, but the resulting t might be a factor $1/\beta$ worse than the best choice for (3.4). The value of α is typically chosen between 0.01 and 0.3.

Example 3.1. Assume we want to minimise the function $f(x_1, x_2) = 4x_1^2 - 4x_1x_2 + 2x_2^2 + 2x_1$ over $\mathbb{R}^2$. This function is convex, since it can be written as $f(x_1, x_2) = (2x_1 - x_2)^2 + x_2^2 + 2x_1$. The gradient is

$$\nabla f(x) = \begin{pmatrix} 8x_1 - 4x_2 + 2 \\ -4x_1 + 4x_2 \end{pmatrix}$$

We see that the optimal solution is $x^* = (-0.5, -0.5)$ with $f(x^*) = -0.5$. Let us start the gradient descent method from the point $x^{(0)} = (0, 0)$ with $f(x^{(0)}) = 0$. Then, $\nabla f(x^{(0)}) = \binom{2}{0}$. For exact line search, we need to minimise

$$g(t) = f(-2t, 0).$$

We set the derivate to 0 using (1.1):

$$\begin{aligned} 0 = g'(t) &= \nabla f(-2t, 0)^\top (-\nabla f(0, 0)) \\ &= -(-16t + 2, 8t)^\top \begin{pmatrix} 2 \\ 0 \end{pmatrix} \\ &= 32t - 4. \end{aligned}$$

Therefore, $t_0 = 0.125$. The next iterate is

$$x^{(1)} = x^{(0)} - t_0 \nabla f(x^{(0)}) = (-0.125 \cdot 2, 0) = (-0.25, 0),$$

with $f(x^{(1)}) = -0.25$.

Let us now recompute the first iteration using backtracking line search instead. Let us set $\alpha = 0.1$, and $\beta = 0.4$. We have $\alpha \|\nabla f(x^{(0)})\|^2 = 0.4$. The iterations of backtracking line search are as follows (with rounded values).

t	$f(-2t, 0)$	$-0.4t$
1	12	-0.4
0.4	0.96	-0.16
0.16	-0.23	-0.064

Thus, we stop with $t = 0.16$, and set $x^{(1)} = (-0.32, 0)$ with $f(x^{(1)}) \approx -0.23$. This is only slightly worse than the value obtained by optimal line search.

Approximate solutions and stopping criteria In most cases of nonlinear optimisation, it is impossible to find the exact optimal solution. (For instance, the optimal solution could be irrational, or even non-algebraic.) In most cases, the goal is to find a good enough *approximate solution*. Let $p^* = f(x^*)$ denote the optimum value. We say that $x \in \mathbf{dom}\, f$ is an ε-approximate solution, if

$$f(x) \geq p^* - \varepsilon.$$

We can set our error tolerance $\varepsilon > 0$, and aim for an ε-approximate solution. A difficulty is that, given an iterate $x^{(k)}$ of the descent method, we may not be able to decide whether it is already ε-approximate.

A usual *stopping criterion* is that $\|\nabla f(x^{(k)})\| < \delta$ for some threshold $\delta > 0$, which is easy to check. Ideally, we would like to give a bound $\delta = \delta(\varepsilon)$ such that $\|\nabla f(x^{(k)})\| < \delta$ implies that $x^{(k)}$ is ε-approximate. We will provide such a bound in Section 3.2.1 under certain assumptions on the function f.

3.2 Conditioning the function

We now introduce sufficient conditions on the function f under which we can provide convergence guarantees for gradient descent. First, we restrict the domain of the function to the *sublevel set* defined by the initial point $x^{(0)}$:

$$S = \{x \in \mathbf{dom}\, f \mid f(x) \leq f(x^{(0)})\}. \tag{3.5}$$

Every descent method will produce points $x^{(1)}, x^{(2)}, \ldots \in S$, and therefore it suffices to provide conditions on S.

Ordering of positive semidefinite matrices For positive semidefinite (PSD) matrices $P, Q \in \mathbb{R}^{n\times n}$, we say that P is PSD-smaller than Q, denoted by $P \preceq Q$, if $Q - P$ is also PSD matrix. Equivalently, this means that for any vector $v \in \mathbb{R}^n$, $v^\top P v \leq v^\top Q v$.

3.2.1 Strong convexity

Let $m > 0$. We say that the function f is *strongly convex* on S with parameter m, if

$$\nabla^2 f(x) \succeq m I_n \quad \text{for every } x \in S.$$

Here, I_n denotes the n-dimensional identity matrix. Equivalently, this means that

$$v^\top \nabla^2 f(x) v \geq m\|v\|^2 \quad \text{for every } x \in S, v \in \mathbb{R}^n.$$

A further equivalent definition is that all eigenvalues of $\nabla^2 f(x)$ are $\geq m$.

Let us recall the Taylor-expansion from Theorem 1.13. If f is is strongly convex on S with parameter m, then for any $x, y \in S$, we get

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{m}{2}\|x - y\|^2. \tag{3.6}$$

Note that an affine function $f(x) = c^\top x + d$ is *not* strongly convex: we have $\nabla f(y) = c$ for every $y \in \mathbb{R}^n$, and therefore $f(x) = f(y) + \nabla f(y)^\top (x - y)$ holds for any $x, y \in \mathbb{R}^n$. Hence, (3.6) can be fulfilled only for $m = 0$. In fact, unlike affine functions, strongly convex functions have unique optimal solutions.

Proposition 3.2. *If f is strongly convex on S, then there exists a unique global minimum point x^*.*

Proof. Assume that x^* be a global minimum and $x \in \mathbf{dom}\, f$ be an arbitrary point. We must have $x^* \in S$, since $f(x^*) \leq f(x^{(0)})$. By optimality, $\nabla f(x^*) = 0$. If $x \notin S$, then $f(x) > f(x^{(0)}) \geq f(x^*)$. Thus, $x \in S$. Using (3.6) for x and $y = x^*$, we obtain

$$f(x) \geq f(x^*) + \frac{m}{2}\|x - x^*\|^2 > f(x^*).$$

This completes the proof. ∎

Convex quadratic functions An important example is convex quadratic functions. Let $f(x) = x^\top Qx + p^\top x + r$; recall from Theorem 1.18 that f is convex if and only if Q is positive semidefinite. Then, $\nabla f(x) = 2Qx + p$ the Hessian is $\nabla^2 f(x) = 2Q$. In this case, the second order Taylor approximation is the function itself (that is, $\bar{x} = x$):

$$\begin{aligned} f(x) &= f(y) + \nabla f(y)^\top (x-y) + \frac{1}{2}(x-y)^\top \nabla^2 f(y)(x-y) \\ &= f(y) + (2y^\top Q + p^\top)(x-y) + (x-y)^\top Q(x-y). \end{aligned}$$

Let $\lambda_1 \geq 0$ be the smallest eigenvalue of Q. Then, $(x-y)^\top Q(x-y) \geq \lambda_1 \|x-y\|^2$. Thus, if Q is positive definite, that is, if $\lambda_1 > 0$, then $f(x)$ is strongly convex with $m = 2\lambda_1$. If Q is positive semidefinite but not positive definite, then $\lambda_1 = 0$ and therefore the function is not strongly convex.

Next, we show that if $p^* = f(x^*)$ is the minimum value of $f(x)$, then a small gradient $\|\nabla f(x)\|$ indicates that $f(x)$ is approximately optimal. This justifies the stopping criterion based on the length of the gradient.

Lemma 3.3. *For $x \in S$, we have*

$$f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|^2.$$

In particular, if $\|\nabla f(x)\| \leq (2m\varepsilon)^{1/2}$, then $f(x) - p^* \leq \varepsilon$.

Proof. From (3.6) (by swapping x and y), we see that for any $x, y \in S$, we have

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|^2. \tag{3.7}$$

Let us now fix x, and consider the function $g(y) = \nabla f(x)^\top (y-x) + \frac{m}{2}\|y-x\|^2$. This is a convex quadratic function in y. The minimum is taken where the gradient is 0, which is at

$$0 = \nabla g(y) = \nabla f(x) + m(y-x).$$

Consequently, the minimiser of $g(y)$ is $\tilde{y} = x - \frac{1}{m}\nabla f(x)$, that is, for any $y \in \mathbb{R}$, $g(y) \geq g(\tilde{y}) = -\frac{1}{2m}\|\nabla f(x)\|^2$. From (3.7), we obtain

$$f(y) \geq f(x) + g(y) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2.$$

This holds true for any $y \in S$, in particular, for $y = x^*$, in which case we obtain the desired

$$p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2. \tag{3.8}$$

□

3.2.2 Lipschitz smooth functions

For some $M > 0$, we say that the function f is M-Lipschitz smooth, or simply, M-smooth on S if

$$\nabla^2 f(x) \preceq M I_n.$$

This is equivalent ot saying that all eigenvalues of $\nabla^2 f(x)$ are $\leq M$. Analogously to (3.6), we obtain

$$f(x) \leq f(y) + \nabla f(y)^\top (x-y) + \frac{M}{2}\|x-y\|^2. \tag{3.9}$$

From this property, we can show the converse of Lemma 3.3: if the gradient is large, then the current point is far from the optimum value p^*:

Lemma 3.4. *For $x \in S$,*

$$f(x) - p^* \geq \frac{1}{2M}\|\nabla f(x)\|^2.$$

The proof is similar to that of Lemma 3.3.

3.2.3 The condition number

Assume now that the function is m-strongly convex and M-smooth at the same time. Then, we have that for every $x \in S$,

$$mI_n \preceq \nabla^2 f(x) \preceq MI_n.$$

Thus, the eigenvalues of $\nabla^2 f(x)$ fall in the range $[m, M]$. We call the ratio

$$\kappa = \frac{M}{m}$$

the *condition number* of f. We will see that the convergence of gradient descent can be bounded in terms of this condition number.

3.3 Convergence analysis of the gradient descent method

Let us apply gradient descent to a convex function f, starting with an initial point $x^{(0)} \in \mathbf{dom}\, f$. Assume that on the sublevel set S, the function is both m-strongly convex and M-smooth for $0 < m \leq M$; thus, it has a condition number $\kappa = M/m$.

Let us apply gradient descent with *exact line search*. Let $x := x^{(k)}$ be the current iterate, and $x^+ := x^{(k+1)}$ be the next iterate. We have $x^+ = x - t_k \nabla f(x)$. Recall that t_k is the minimiser of $g(t) = f(x - t\nabla f(x))$ over $t \geq 0$. Using M-smoothness, we can estimate

$$\begin{aligned} g(t) = f(x - t\nabla f(x)) &\leq f(x) + \nabla f(x)(-t\nabla f(x)) + \frac{M}{2}\| - t\nabla f(x)\|^2 \\ &= f(x) + \left(\frac{Mt^2}{2} - t\right)\|\nabla f(x)\|^2. \end{aligned} \tag{3.10}$$

The right hand side is a quadratic function in t, which is minimised at $t = 1/M$. Substituting this value provides an upper bound on the minimum of $g(t)$, which is taken at t_k:

$$g(t_k) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2. \tag{3.11}$$

Using that $f(x^+) = g(t_k)$, and subtracting the optimum value p^* from both sides, we see that

$$f(x^+) - p^* \leq f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|^2.$$

Let us now combine this with the bound in Lemma 3.3, giving $\|\nabla f(x)\|^2 \geq 2m(f(x) - p^*)$:

$$f(x^+) - p^* \leq \left(1 - \frac{m}{M}\right) \cdot (f(x) - p^*)$$

Hence, the distance from optimality decreases by a factor $1 - m/M = 1 - 1/\kappa$ in every iteration. By a recursive application of this argument, we see that for every $k \geq 1$,

$$f(x^{(k)}) - p^* \leq \left(1 - \frac{1}{\kappa}\right)^k \cdot \left(f(x^{(0)}) - p^*\right)$$

This shows that $f(x^{(k)})$ converges to p^* as $k \to \infty$.

Thus, we have proved the following theorem.

Theorem 3.5. *Given an m-strongly convex and M-smooth function, gradient descent with exact line search obtains a solution $f(x^{(k)}) - p^* \leq \varepsilon$ within*

$$k \leq \frac{\log \frac{f(x^{(0)}) - p^*}{\varepsilon}}{\log \frac{\kappa}{\kappa - 1}}$$

iterations.

If $\kappa = M/m$ is relatively close to 1, then this provides a rapid convergence speed. On the other hand, if κ is large compared to 1, then

$$\log \frac{\kappa}{\kappa - 1} = \log \left(1 + \frac{1}{\kappa - 1}\right) \approx \frac{1}{\kappa},$$

and therefore, the bound on the number of iterations grows linearly with κ: we obtain approximately

$$\kappa \log \frac{f(x^{(0)}) - p^*}{\varepsilon}.$$

3.3.1 Analysis for backtracking line search

Let us now consider gradient descent using backtracking line search. Recall that this method finds a step size $t_k = t$ satisfying (3.4). We now show that this must be satisfied for every $0 \le t \le \frac{1}{M}$.

Lemma 3.6. *If f is M-smooth, then*

$$f(x - t\nabla f(x)) \le f(x) - \alpha t \|\nabla f(x)\|^2.$$

holds for every $\alpha \le \frac{1}{2}$ and $0 \le t \le \frac{1}{M}$.

Proof. Again, let us denote $x = x^{(k)}$. Note that for any value $0 \le t \le 1/M$, we have

$$\frac{Mt^2}{2} - t \le -\frac{t}{2}.$$

From M-smoothness, we have (3.10). Therefore, we obtain

$$\begin{aligned} f(x - t\nabla f(x)) &\le f(x) + \left(\frac{Mt^2}{2} - t\right) \|\nabla f(x)\|^2 \\ &\le f(x) - \frac{t}{2}\|\nabla f(x)\|^2 \\ &\le f(x) - \alpha \|\nabla f(x)\|^2. \end{aligned}$$

The last estimation simply used $\alpha \le 1/2$. □

Using this statement, we see that the backtracking line search terminates either with $t = 1$ (i.e. if the initial value already satisfies (3.4)), or with $t \ge \beta/M$. Hence,

$$t \ge \min\left\{1, \frac{\beta}{M}\right\}.$$

Therefore,

$$f(x^{(k+1)}) \le f(x^{(k)}) - \min\left\{\alpha, \frac{\alpha\beta}{M}\right\} \|\nabla f(x^{(k)})\|^2.$$

We can use this bound instead of (3.11), and by the same argument as in the previous proof, we can derive

$$f(x^{(k)}) - p^* \le \left(1 - \frac{1}{\kappa'}\right)^k \cdot \left(f(x^{(0)}) - p^*\right),$$

where

$$\kappa' = \max\left\{\frac{1}{2\alpha m}, \frac{M}{2\alpha\beta m}\right\}.$$

Thus, we see that the convergence guarantee for backtracking line search is almost as good as the one for exact line search.

3.3.2 Analysis for M-smooth functions

What if we our function is not m-strongly convex, or it is but the value of m is very small? Then, the above analyses give rather weak bounds. However, we can still bound the convergence rate using solely M-smoothness.

In fact, we can obtain convergence bounds even for a constant step size μ, as long as $\mu \leq 1/M$. This follows by Lemma 3.6: for step size $t = \mu \leq 1/M$, we obtain

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{\mu}{2}\|\nabla f(x^{(k)})\|^2.$$

Using this property, one can show the following convergence bound. We do not include the proof.

Theorem 3.7. *If f is M-smooth, then gradient descent with constant step size $\mu \leq 1/M$ finds a solution $x^{(k)}$ with $f(x_k) - p^* \leq \varepsilon$ for*

$$k \leq \frac{C}{\mu\varepsilon}\|x^{(0)} - x^*\|^2$$

for some constant $C > 0$.

We get the best bound for $\mu = 1/M$, namely,

$$k \leq \frac{CM}{\varepsilon}\|x^{(0)} - x^*\|^2.$$

The same convergence guarantee can be given for exact or backtracking line search (with better constants).

Note that this guarantee is typically weaker than the one in Theorem 3.5. There, the dependence on ε was $\log(1/\varepsilon)$, in contrast to the $1/\varepsilon$ dependence here. Further, this bound also depends on the distance $\|x^{(0)} - x^*\|^2$, which may be very large.

3.4 Constrained optimisation: the Frank-Wolfe algorithm

Let us now consider the setting when we are given a convex set $X \subseteq \mathbf{dom}\, f$, and we wish to minimise f over the feasible region X, that is,

$$\begin{aligned} &\min f(x) \\ &\text{s. t. } x \in X. \end{aligned} \tag{3.12}$$

Gradient descent is not directly applicable, as the sequence of iterates could leave the feasible region. There are multiple different approaches to solve this problem. Here, we discuss only one of them, the *Frank-Wolfe algorithm*, also called the *conditional gradient method*. For this method, we need to assume that X is a *compact convex* set.

The Frank-Wolfe-algorithm is a descent method using feasible points. That is, given $x^{(k)} \in X$, we are looking for an update of the form

$$x^{(k+1)} = x^{(k)} + t_k \Delta x^{(k)}$$

such that $x^{(k+1)} \in X$ and $f(x^{(k+1)}) < f(x^{(k)})$. Here, $\Delta x^{(k)}$ is the search direction. In order to decrease the function value, we need to move in a *decreasing direction*, that is,

$$\nabla f(x^{(k)})^\top \Delta x^{(k)} < 0.$$

Recall from Theorem 1.10 that $x^* \in X$ is a global minimum of f over X if and only if

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in X.$$

Thus, if $x^{(k)}$ is not optimal, then there is a vector $s \in X$ such that

$$\nabla f(x^{(k)})^\top (s - x^{(k)}) < 0.$$

Direction finding subproblem The Frank-Wolfe method finds a search direction $\Delta x^{(k)} = s - x^{(k)}$ via solving the following optimisation problem.

$$\begin{aligned} &\min \nabla f(x^{(k)})^\top y \\ &\text{s. t. } y \in X. \end{aligned} \tag{3.13}$$

By the assumption that X is compact, this problem admits an optimal solution. In case when $x^{(k)}$ is an optimal solution to this problem, we conclude that $x^{(k)}$ is an optimal solution to the original problem, as the optimality conditions are satisfied for $x^* = x^{(k)}$. Otherwise, for the optimal solution $s \in X$, we have $\nabla f(x^{(k)})^\top (s - x^{(k)}) < 0$. Thus, we use $s - x^{(k)}$ as the search direction.

This problem (3.13) is itself a constrained optimisation problem. However, the objective function is *linear*, which makes the problem typically simpler than the original problem (3.12). If the feasible region X is given by linear constraints, the direction finding subproblem amounts to solving a linear program.

Finding the step size Using $\Delta x^{(k)} = s - x^{(k)}$ as the search direction, we compute the next iterate as

$$x^{(k+1)} = x^{(k)} + t_k(s - x^{(k)}).$$

Proposition 3.8. *For any value $0 \le t_k \le 1$, we have $x^{(k+1)} \in X$.*

Proof. We can rewrite the update formula as $x^{(k+1)} = (1 - t_k)x^{(k)} + t_k s$. Thus, $x^{(k+1)}$ is on the line segment between $x^{(k)}$ and s. The convexity of X implies that $x^{(k+1)} \in X$. □

Just as for gradient descent, there are various methods for choosing the step size. One option is *exact line search*, which finds

$$t_k := \operatorname{argmin}_{0 \le t \le 1} f((1 - t)x^{(k)} + ts).$$

A common choice is using simply the decreasing sequence $t_k = 2/(k + 1)$.

Convergence rate The search direction $s - x^{(k)}$ also provides information on the optimum value $p^* = f(x^*)$. Indeed, using convexity, we see that

$$\begin{aligned} p^* = f(x^*) &\ge f(x^{(k)}) + \nabla f(x^{(k)})^\top (x^* - x) \\ &\ge f(x^{(k)}) + \min_{y \in X} \nabla f(x^{(k)})^\top (y - x) \\ &= f(x^{(k)}) + \nabla f(x^{(k)})^\top (s - x). \end{aligned}$$

Hence, every iteration provides a lower bound on p^*. Using this bound, one can derive the following convergence guarantee. We omit the proof.

Theorem 3.9. *If f is M-smooth, then the Frank-Wolfe method with exact line search or with step size $t_k = 2/(k + 1)$ finds a solution $x^{(k)}$ with $f(x_k) - p^* \le \varepsilon$ for*

$$k \le \frac{CM}{\varepsilon} \|x^{(0)} - x^*\|^2$$

for some constant $C > 0$.