Methods and Theory of Modern Optimisation

AMSI ACE Network lectures for Optimisation 2024

Prof Andrew Eberhard[1] Dr Scott B. Lindstrom[2]

[1]Discipline of Mathematical Sciences, School of Science, RMIT University, PO Box 2476, Melbourne 3001, Australia

[2]Centre for Optimisation and Decision Science, Curtin University

Contents

List of Figures

Preface

Nonlinear and convex optimisation has gained wide use in data mining and machine learning in recent years. This newfound popularity has birthed renewed interest in the underlying mathematical foundations behind descent methods. The main tool of machine learning is stochastic gradient descent, which corresponds to a statistically unbiased approximation of a deterministic descent method, and is used widely in regression and classification problems. Consequently, a renewed interest in first order gradient-based convex optimisation has arisen out of the study of this and similar problems. The nonsmooth version arises in other Machine learning objectives, such as the problem of regularised risk minimisation with binary hinge loss. In signal processing, similar nonsmooth problems must be solved when recovering corrupted signals with sparse support (compressed sensing). Moreover, optimisation under uncertainty also draws its essential techniques from the same library; methods that are used to solve stochastic optimisation problems exploit decomposition methods to create parallel programming approaches based on constrained convex optimisation. A popular example is the so-called alternating direction method of multipliers (ADMM). Projection algorithms are special cases of these general methods, and also offer insights about the more general class.

In this course we will take a modern view of the study of these and related problems, and the algorithms used to solve them. We will introduce to the language of convex and nonsmooth optimisation, and show how this powerful mathematical machinery allows one to analyse optimisation problems and develop algorithms for solving them. We will study the application of ADMM to various problem sets, including stochastic optimisation and feasibility problems. The ability to apply these techniques within the hyperspace of symmetric matrices allows the same methods to be applicable to a wider class of problems, and this aspect will also be explored. Ultimately, we will look at the use of these techniques to some of the areas of application discussed above.

Different models abound for solving optimisation problems. Smooth models allow one to use derivatives, and approximations for them; this utility is often associated with faster convergence and stronger theory. Nonsmooth models often have lower dimensional structure, because they do not require the introduction of variables used in smoothing. The lower dimensional structure translates into improvements in computation time for the steps involved. A well known tradeoff exists between simpler gradient descent methods and more exotic methods that use second order information. In data science, first order methods are often preferred, because of their low computational cost per step.

Part I

Convexity and Variational Analysis

Chapter 1

Convexity

Convex analysis refers to an area an area of mathematics that can be thought of as a special part of real analysis. We concern ourselves with functions that (in the main) evaluate to a real value but impose restriction on the geometry that these functions may possess. This has the unexpected reward of allowing us to relax other assumptions usually associated with the theory of functions of several variables. These are:

1. Differentiability;
2. Finiteness.

This greatly increases the usefulness of the area and gives rise a wealth of mathematical techniques that have had wide application, especial in the areas of optimisation, control and other areas engineering mathematics. In making this step we arrive at the first example of a wider theory known as nonsmooth analysis.

The power of a theory is determined by the tool that it provided the analysis of new problems. In convex analysis we find a treasure chest of techniques and a mature calculus that facilitates its application and we can only be certain that more applications will arise in the future.

In the course we will attempt to:

1. Cover some of the fundamental theory that convex analysis is based on.
2. Show how these techniques can be applied to a number of optimisation problems. How we can formulate and reformulate problems using these techniques into potentially more tractable problems.
3. Study several popular optimisation algorithms and their convergence properties.

Convex analysis comes in two brands, finite dimensional and infinite dimensional, categorized by the dimensionality of the underlying linear space. We will only consider the case when the underlying space is a Euclidean space. The also allows us to consider the space of symmetric matrices, so touching on semi-definite programming.

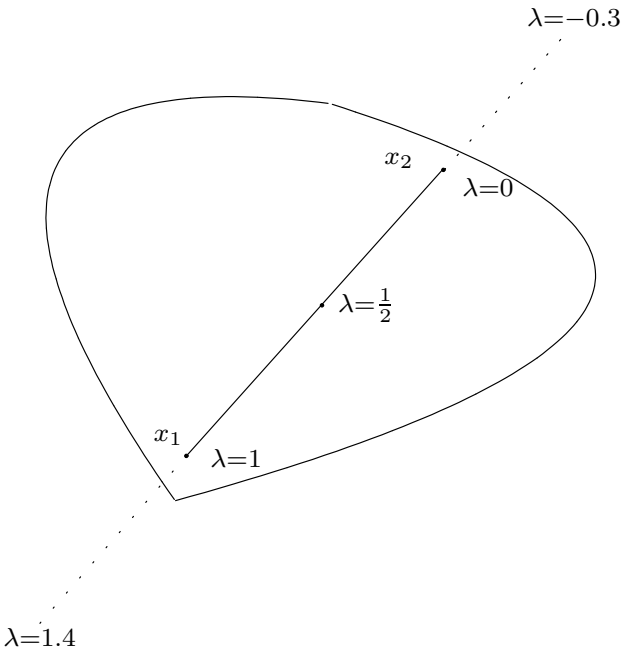
Convex optimisation has in recent years matured to a level where it has become a cornerstone of optimisation theory, a position that was long held only by linear programming (which is just convex optimisation over polyhedral functions). It is hoped that courses of this kind will alert more practitioners of applied mathematics to opportunities awaiting those who become proficient with the mathematical language of convex analysis.

1.1 What is Convexity?

Roughly speaking, convexity refers to a geometric property. A set in a vector space is convex if it contains all line segments connecting its points. lines that can be drawn between two points within this set. This corresponds to a rotund shape.

Definition 1.1.1 *Let $C \subseteq X$, where X is a vector space. A set C is convex if for all x_1, $x_2 \in C$ and $\lambda \in [0,1]$ we have*

$$\lambda x_1 + (1-\lambda)x_2 \in C.$$



Proposition 1.1.1 *The following are equivalent:*

(i) C is convex;

(ii) Any finite subset $\{x_i \mid i = 1, \ldots, m\} \subseteq C$ with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$ satisfies $\sum_i \lambda_i x_i \in C$

Consequently, it may be shown that if $C = \{x_1, \ldots, x_m\} \subseteq \mathbf{R}^n$ for $m > n$ then we only need to use sums from $i = 1, \ldots, m = n+1$ to represent a point $x \in C$ (this is called Carathéodory's theorem).

Proof. (ii) $\implies$ (i): This is obvious.
(i) $\implies$ (ii): This may be done by induction. Suppose $m = 3$ then we have

$$\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 = \underbrace{\left(\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3}\right)}_{y_1}\left(\underbrace{\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)}_{z_1} x_1 + \underbrace{\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)}_{z_2} x_2\right) + \underbrace{\left(\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}\right)}_{y_2} x_3.$$

Here $z_1 + z_2 = y_1 + y_2 = 1$, because of how $\lambda_1, \lambda_2, \lambda_3$ are defined. Because C is convex, the right hand side must therefore belong to C, and so $\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 \in C$. □

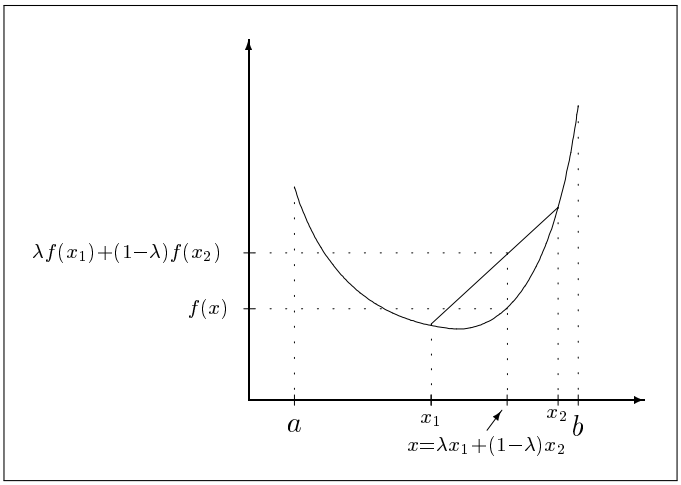


Figure 1.1: A strictly convex function has all chords above its graph.

Definition 1.1.2 *The convex hull of a subset $C \subseteq \mathbf{R}^n$ is the smallest convex set containing C and is denoted by $\operatorname{co} C$.*

Proposition 1.1.2 *If $C \subseteq \mathbf{R}^n$ then*

$$\operatorname{co} C = \left\{ \sum_i \lambda_i x_i \mid \sum_i \lambda_i = 1,\ \lambda_i \geq 0 \text{ and } x_i \in C \right\}. \tag{1.1}$$

Proof. Let $\hat{C}$ denote the right hand side (1.1). Clearly $\hat{C}$ is convex and $C \subseteq \hat{C}$ (just take $\lambda_1 = 1$ and $x_1 = x \in C$). As $\operatorname{co} C$ is the smallest convex set containing C and $\hat{C}$ is an example of such a set we have $\operatorname{co} C \subseteq \hat{C}$. Now suppose there exists $y \in \hat{C}$ with $y \notin \operatorname{co} C$. Then $y = \sum_i \lambda_i x_i$ with $x_i \in C \subseteq \operatorname{co} C$. Then by the convexity of $\operatorname{co} C$ we must have $y = \sum_i \lambda_i x_i \in \operatorname{co} C$, a contradiction. □

We may now define what is meant by a convex function on a real variable.

Definition 1.1.3 *A function is said to be convex on a convex set C, if for all x_1, $x_2 \in C$ and $\lambda \in [0,1]$ we have*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{1.2}$$

When the inequality is strict for all $\lambda \neq \{0,1\}$ and $x_1 \neq x_2$ we say the function is strictly convex.

This definition is the mathematical way of saying that all chords from points on the graph of f lie above the graph of f (see Figure 1.1).

Rules for extended real values

Note that if $f(x_1) = +\infty$ then the inequality (10.15) certainly holds. This leads us to allow the possibility that convex functions take the value $+\infty$. Let $\overline{\mathbf{R}}$ denote the extended real line i.e. $\overline{\mathbf{R}} := (-\infty, +\infty]$ or $\overline{\mathbf{R}} := \mathbf{R} \cup \{+\infty\}$. We may extend the usual arithmetic of the real line to the extended real line as follows:

- $a + \infty = +\infty + a = +\infty$ if $a \neq -\infty$
- $a - \infty = -\infty + a = -\infty$ if $a \neq +\infty$
- $a \times \pm\infty = \pm\infty \times a = \pm\infty$ if $a > 0$
- $a \times +\infty = +\infty \times a = -\infty$ if $a < 0$
- $a \times -\infty = -\infty \times a = +\infty$ if $a < 0$
- $a/\pm\infty = 0$ if $-\infty < a < +\infty$
- $\pm\infty/a = \pm\infty$ if $0 < a < +\infty$
- $+\infty/a = -\infty$ if $-\infty < a < 0$
- $-\infty/a = +\infty$ if $-\infty < a < 0$

Here, "$a + \infty$" means "$a + (+\infty)$" and "$a - \infty$" means both "$a - (+\infty)$" as well as "$a + (-\infty)$".

The expressions $\infty - \infty$, $0 \times \pm\infty$ and $\pm\infty/\pm\infty$ are left undefined (and are prohibited). These prohibited combinations will be avoided in all calculations. These rules are modeled on the laws for infinite limits. Note that $1/0$ is not defined as either $+\infty$ or $-\infty$, because although it is true that whenever $f(x) \to 0$ for a continuous function $f(x)$, we must have that $1/f(x)$ is eventually in every neighborhood of the set $\{-\infty, +\infty\}$, it is not true that $1/f(x)$ must converge to one of these points. An example is $f(x) = 1/(\sin(1/x))$.

In order to prevent the occurrence of prohibited products of infinities we usually assume our functions are *proper.* That is we assume f is not identically $+\infty$ everywhere and that f never takes the value $-\infty$ anywhere.

Properties of convex functions

We note the following: A subset $C \subseteq \mathbf{X} \times \overline{\mathbf{R}}$ is convex iff (x_1, α_1), $(x_2, \alpha_2) \in C$ implies

$$(\lambda x_1 + (1-\lambda)x_2, \lambda\alpha_1 + (1-\lambda)\alpha_2) \in C.$$

Remark 1.1.1 *Let $f : \mathrm{dom}(f) \to \mathbf{R}$ be convex and define*

$$\mathrm{epi}(f) = \{(x, \alpha) \mid f(x) \leq \alpha\}.$$

Then $\mathrm{epi}(f)$ *is convex. Take* (x_1, α_1), $(x_2, \alpha_2) \in \mathrm{epi}(f)$ *(and so* $f(x_1) \leq \alpha_1$ *and* $f(x_2) \leq \alpha_2$*) then*

$$\lambda(x_1, \alpha_2) + (1-\lambda)(x_2, \alpha_2) = (\lambda x_1 + (1-\lambda)x_2, \lambda\alpha_1 + (1-\lambda)\alpha_2),$$

and by the convexity of f we have

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \leq \lambda\alpha_1 + (1-\lambda)\alpha_2,$$

or

$$(\lambda x_1 + (1-\lambda)x_2, \lambda\alpha_1 + (1-\lambda)\alpha_2) \in \mathrm{epi}(f).$$

Example 1.1.1 *Given function* $f_1, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ *define*

$$f(x) = \max\{f_1(x), \ldots, f_m(x)\}.$$

Show f *is a convex function.*

Soln: *As each function is convex we have for any* $x_1, x_2 \in \mathbf{R}^n$ *and* $\lambda \in [0,1]$ *that*

$$f_i(\lambda x_1 + (1-\lambda)\, x_2) \leq \lambda f_i(x_1) + (1-\lambda)\, f_i(x_2), \text{ for all } i = 1, \ldots, m.$$

Then for each i *we have*

$$\begin{aligned}\lambda f(x_1) + (1-\lambda)\, f(x_2) &= \lambda \max\{f_1(x_1), \ldots, f_m(x_1)\} + (1-\lambda) \max\{f_1(x_2), \ldots, f_m(x_2)\} \\ &\geq \lambda f_i(x_1) + (1-\lambda)\, f_i(x_2) \geq f_i(\lambda x_1 + (1-\lambda)\, x_2).\end{aligned}$$

Thus $\lambda f(x_1) + (1-\lambda)\, f(x_2)$ *is an upper bound for all* i *and so*

$$\begin{aligned}\lambda f(x_1) + (1-\lambda)\, f(x_2) &\geq \max\{f_1(\lambda x_1 + (1-\lambda)\, x_2), \ldots, f_m(\lambda x_1 + (1-\lambda)\, x_2)\} \\ &= f(\lambda x_1 + (1-\lambda)\, x_2).\end{aligned}$$

Remark 1.1.2 *Another way of defining convexity of a function* $f : \mathbf{R}^n \to \overline{\mathbf{R}}$ *is to demand* $\operatorname{epi}(f)$ *to be a convex subset of* $\mathbf{X} \times \overline{\mathbf{R}}$.

The effective domain is defined as

$$\operatorname{dom}(f) := \{x \in \mathbf{X} \mid f(x) < +\infty\}.$$

Lemma 1.1.3 *If* $f, g : \mathbf{X} \to \overline{\mathbf{R}}$ *where* $\operatorname{dom}(f) \cap \operatorname{dom}(g) \neq \emptyset$ *and* f *and* g *are both convex then* $f + g$ *is also a proper convex function with* $\operatorname{dom}(f+g) = \operatorname{dom}(f) \cap \operatorname{dom}(g)$.

Proof. Take $x, y \in \mathbf{X}$ and $\lambda \in [0,1]$. Then

$$\begin{aligned}(f+g)(\lambda x + (1-\lambda)\, y) &= f(\lambda x + (1-\lambda)\, y) + g(\lambda x + (1-\lambda)\, y) \\ &\leq \lambda f(x) + (1-\lambda)\, f(y) + \lambda g(x) + (1-\lambda)\, g(y) \\ &= \lambda (f+g)(x) + (1-\lambda)(f+g)(y).\end{aligned}$$

□

Remark 1.1.3 *Note that in order for* $(f+g)$ *to be strictly convex we only require one of* f *or* g *to be strictly convex.*

How do we know when a function is convex or strictly convex? Derivatives can be of use.

Theorem 1.1.4 *Let* $f : [a,b] \to \mathbf{R}$ *be twice continuously differentiable. Then the following hold.*

(i) The following are equivalent statements:

(a) f *is convex on* $[a,b]$*;*
(b) $f(y) - f(x) \geq (y-x) f'(x)$ *for all* x, $y \in [a,b]$*;*

(c) $f''(x) \geq 0$ *for all* $x \in [a, b]$.

(ii) The function f *is strictly convex if and only if* $f(y) - f(x) > (y - x)f'(x)$ *for all* x, $y \in [a, b]$ *with* $x \neq y$.

(iii) When $f''(x) > 0$ *for all* $x \in [a, b]$ *then* f *is strictly convex.*

Proof. We will only prove the equivalence for the case of f convex. The case of f strictly convex follows along similar lines of argument.

(i)a$\Longrightarrow$(i)b: Consider $y > x$ and suppose f is convex. If f is convex, then for all $0 < \alpha < 1$ we have

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

implying

$$\begin{aligned} f'(x) &= \lim_{\alpha \to 0} \frac{f(x + \alpha(y - x)) - f(x)}{\alpha(y - x)} \leq \lim_{\alpha \to 0} \frac{(1 - \alpha)f(x) + \alpha f(y) - f(x)}{\alpha(y - x)} \\ &= \lim_{\alpha \to 0} \frac{\alpha(f(y) - f(x))}{\alpha(y - x)} = \frac{f(y) - f(x)}{(y - x)}, \\ \text{or} \qquad & f'(x)(y - x) \leq f(y) - f(x). \end{aligned} \tag{1.3}$$

Graphically this says that all tangent lines to the curve lie below the curve as depicted in Figure 1.2.

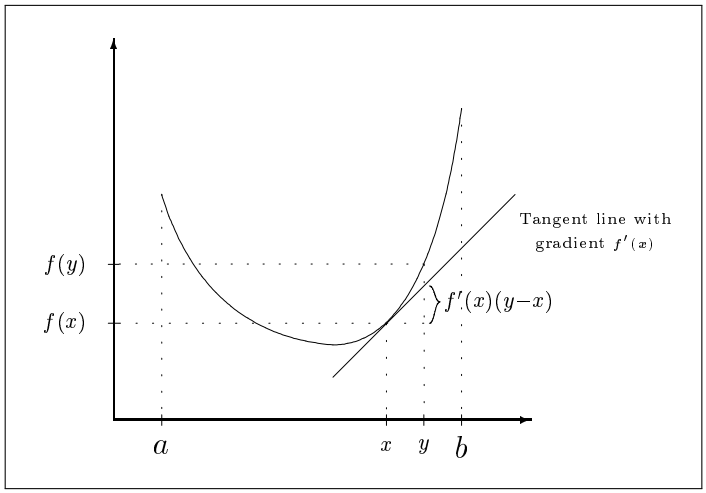


Figure 1.2: All the tangent lines lie below the curve

(i)b$\Longrightarrow$(i)c: Now suppose that $y = x + \delta$ then (1.3) becomes

$$f(x + \delta) - f(x) \geq \delta f'(x) \tag{1.4}$$

Using a Taylor expansion for $f(x + \delta)$ about $\delta = 0$ gives

$$f(x + \delta) - f(x) = \delta f'(x) + \frac{1}{2}\delta^2 f''(x) + o(\delta^2) \geq \delta f'(x),$$

(we have use the "little oh" notation again here to mean $\lim_{\delta \to 0} \frac{o(\delta^2)}{\delta^2} = 0$). This implies, after cancellation of $\delta f'(x)$, that

$$\frac{1}{2}\delta^2 f''(x) + o(\delta^2) \geq 0 \quad \text{or} \quad f''(x) \geq -2\frac{o(\delta^2)}{\delta^2} \to 0 \text{ as } \delta \to 0.$$

Thus $f''(x) \geq 0$ (recall the sufficiency condition for a local minimum).
(i)c$\Longrightarrow$(i)a: For convenience of notation, define

$$q(\alpha) = f(\alpha x + (1-\alpha)y) - \alpha f(x) - (1-\alpha)f(y).$$

Note that the definition of convexity, is equivalent to $q(\alpha) \leq 0$ for all $a \in [0,1]$. Now $q(0) = q(1) = 0$ and since f is differentiable there must exist a turning point. In fact since

$$q''(\alpha) = f''(\alpha x + (1-\alpha)y)(x-y)^2 \geq 0.$$

all turning points must be minima. Indeed on reflection we see that this means that there is only one minimum (see Figure 1.3). Thus $q(\alpha)$ can only decrease form zero and then

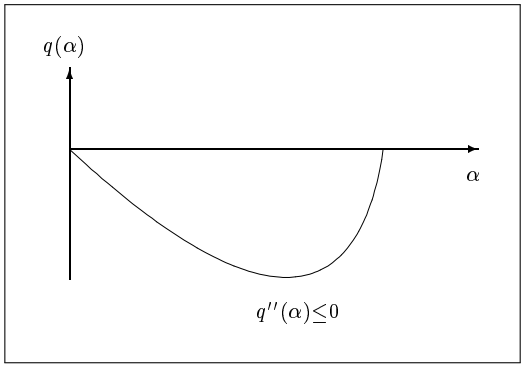


Figure 1.3: The graph of q lies below the α axis.

increase to zero implying $q(\alpha) \leq 0$, as desired. □

Remark 1.1.4 *The condition $f''(x) > 0$ for all $x \in [a,b]$ is not a necessary condition for strict convexity. Consider the example of $f(x) = x^3$ and $x \in [-1,1]$. Then $f''(0) = 0$ yet f is strictly convex on $[0,1]$ since, by polynomial division:*

$$\frac{y^3 - x^3}{y - x} = y^2 + xy + x^2. \tag{1.5}$$

Now when $x, y \in [0,1]$ and $y > x$ (and so $(y-x) > 0$) we have

$$y^2 + xy + x^2 > 3x^2.$$

Combining with (1.5), we obtain that (ii) holds. Alternatively, if $x > y$ (and so $y - x < 0$), then $y^2 + xy + x^2 < 3x^2$, which again combines with (1.5) to yield (ii).
Altogether, f is strictly convex on $[0,1]$. To obtain an example where $f''(0) = 0$ and f is strictly convex on $[-1,1]$ simply reflect the graph in the y axis i.e. define

$$f(x) = \begin{cases} x^3 & \text{for } x \geq 0 \\ -x^3 & \text{for } x < 0 \end{cases}.$$

Definition 1.1.4 *When $f : \mathbf{X} \to \mathbf{R}$ we may define a one dimensional function (see Figure 1.4)*

$$t \mapsto f(x + t(y - x)) = g_{x,y}(t).$$

Note that $g_{x,y}(0) = f(x)$ and $g_{x,y}(1) = f(y)$.

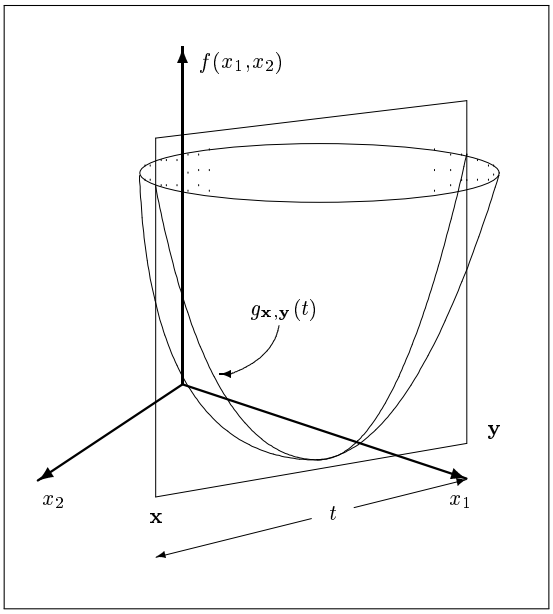


Figure 1.4: A vertical plane cuts f to form a convex function $g_{\mathbf{x},\mathbf{y}}$.

We now consider functions of several variables. In finite dimensions we consider $\mathbf{X}$ to be Euclidean space (see next section for details).

Lemma 1.1.5 *The function f is convex on a Euclidean space $\mathbf{X}$ if and only if $g_{x,y}$ is a convex function in $t \in [0,1] \subseteq \mathbf{R}$ for every $x, y \in \mathbf{X}$.*

Proof. Suppose f is convex then

$$\begin{aligned} g_{x,y}(\lambda t + (1-\lambda)t') &= f((x + (\lambda t + (1-\lambda)t')(y-x)) \\ &= f(\lambda(x + t(y-x)) + (1-\lambda)(x + t'(y-x))) \\ &\leq \lambda f(x + t(y-x)) + (1-\lambda) f(x + t'(y-x)) \\ &= \lambda g_{x,y}(t) + (1-\lambda) g_{x,y}(t') \end{aligned}$$

verifying $t \mapsto g_{x,y}(t)$ is convex. Now suppose $t \mapsto g_{x,y}(t)$ is convex. Then by definition

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= f(x + (1-\lambda)(y-x)) \\ &= g_{x,y}\left((1-\lambda)\right) \\ &= g_{x,y}\left(\lambda 0 + (1-\lambda)1\right) \\ &\leq \lambda g_{x,y}(0) + \ (1-\lambda) g_{x,y}(1) = \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

giving the definition of convexity of f. □

Remark 1.1.5 (Some vector calculus) *Recall the following vector calculus notation and facts:*

(i) For a function of more than one variable, we may write $f(\mathbf{x}) := f(x_1, \ldots, x_n)$, where it is understood that $\mathbf{x} := (x_1, \ldots, x_n)$.

(ii) $\frac{d}{dt} f(u_1(t), \ldots, u_n(t)) := g'(t)$ where $g : t \mapsto f(u_1(t), \ldots, u_n(t))$.

(iii) $\frac{\partial}{\partial u_1} f(u_1, \ldots, u_n) = h'(u_1)$ *where* $h := s \mapsto f(s, \ldots, u_n)$.

(iv) The chain rule for a function of more than 1 variable:

$$\frac{d}{dt} f(u_1(t), u_2(t), \ldots, u_n(t)) = \frac{\partial}{\partial u_1} f(u_1(t), u_2(t), \ldots, u_n(t)) \frac{du_1(t)}{dt} + \cdots + \frac{\partial}{\partial u_n} f(u_1(t), u_2(t), \ldots, u_n(t)) \frac{du_n(t)}{dt}.$$

(v) $D_u f(x_1, \ldots, x_n) \lim_{t \to 0} \frac{f(x+tu)-f(x)}{t} = \nabla f(x_1, \ldots, x_n) \cdot u$.

Lemma 1.1.5 tells us that some properties of convex functions in one dimension carry over to those in many dimensions. At times we will need to use the first and second derivative of the above function g with respect to the variable t. To calculate these we use the chain rule for partial derivatives. Recall that when we substitute $u_i(t)$ for the ith variable x_i in the function f we have (by the chain rule)

$$\frac{d}{dt} f(u_1(t), u_2(t), \ldots, u_n(t)) = \frac{\partial}{\partial u_1} f(u_1(t), u_2(t), \ldots, u_n(t)) \frac{du_1(t)}{dt} + \cdots + \frac{\partial}{\partial u_n} f(u_1(t), u_2(t), \ldots, u_n(t)) \frac{du_n(t)}{dt}.$$

We may wish to use one-dimensional functions like the one in Definition 1.1.4. In such a case, we have $u_i(t) = x_i + td_i$ for some $x_i, d_i \in E$ and so we have $\frac{du_i(t)}{dt} = d_i$ and so we may write

$$\begin{aligned} g'(t) = \frac{d}{dt} f(\mathbf{x} + t\mathbf{d}) &= \frac{\partial}{\partial x_1} f(\mathbf{x} + t\mathbf{d}) d_1 + \cdots + \frac{\partial}{\partial x_n} f(\mathbf{x} + t\mathbf{d}) d_n \\ &= \nabla_x f(\mathbf{x} + t\mathbf{d}) \cdot (d_1, \ldots, d_n) = \nabla_x f(\mathbf{x} + t\mathbf{d}) \cdot \mathbf{d} \end{aligned} \tag{1.6}$$

where $\nabla_x f(\mathbf{x} + t\mathbf{d}) := (\frac{\partial}{\partial x_1} f(\mathbf{x} + t\mathbf{d}), \ldots, \frac{\partial}{\partial x_n} f(\mathbf{x} + t\mathbf{d}))$ denotes the gradient vector of the function f. To calculate the second derivative of g we must apply the chain rule for partial derivatives to each component of (1.6). That is we need each of the derivatives

$$\begin{aligned} \frac{d}{dt} \frac{\partial}{\partial x_j} f(x_1 + td_1, x_2 + td_2, \ldots, x_n + td_n) &= \frac{d}{dt} f_{x_j}(x_1 + td_1, x_2 + td_2, \ldots, x_n + td_n) \\ \text{(chain rule)} \quad &= \frac{\partial}{\partial x_1} f_{x_j}(\mathbf{x} + t\mathbf{d}) \frac{d}{dt}(x_1 + td_1) + \cdots + \frac{\partial}{\partial x_n} f_{x_j}(\mathbf{x} + t\mathbf{d}) \frac{d}{dt}(x_n + td_n) \\ &= \frac{\partial}{\partial x_1} f_{x_j}(\mathbf{x} + t\mathbf{d}) d_1 + \cdots + \frac{\partial}{\partial x_n} f_{x_j}(\mathbf{x} + t\mathbf{d}) d_n. \end{aligned} \tag{1.7}$$

Altogether, we obtain

$$\begin{aligned} g''(t) = \frac{d^2}{dt^2} f(\mathbf{x} + t\mathbf{d}) &\overset{\text{(by (1.6))}}{=} \left(\frac{d}{dt} \frac{\partial}{\partial x_1} f(\mathbf{x} + t\mathbf{d}) \right) d_1 + \cdots + \left(\frac{d}{dt} \frac{\partial}{\partial x_n} f(\mathbf{x} + t\mathbf{d}) \right) d_n \\ &\overset{\text{(by (1.7))}}{=} \left(\frac{\partial}{\partial x_1} f_{x_1}(\mathbf{x} + t\mathbf{d}) d_1 + \cdots + \frac{\partial}{\partial x_n} f_{x_1}(\mathbf{x} + t\mathbf{d}) d_n \right) d_1 \\ &\quad + \left(\frac{\partial}{\partial x_1} f_{x_2}(\mathbf{x} + t\mathbf{d}) d_1 + \cdots + \frac{\partial}{\partial x_n} f_{x_2}(\mathbf{x} + t\mathbf{d}) d_n \right) d_2 \\ &\quad + \cdots \\ &\quad \vdots \end{aligned}$$

$$\cdots + \left(\frac{\partial}{\partial x_1} f_{x_n}(\mathbf{x}+t\mathbf{d})d_1 + \cdots + \frac{\partial}{\partial x_n} f_{x_n}(\mathbf{x}+t\mathbf{d})d_n \right) d_n$$

$$= \begin{pmatrix} d_1 & \dots & d_n \end{pmatrix} \begin{pmatrix} f_{x_1x_1}(\mathbf{x}+t\mathbf{d}) & f_{x_1x_2}(\mathbf{x}+t\mathbf{d}) & \dots & f_{x_1x_n}((\mathbf{x}+t\mathbf{d}) \\ f_{x_2x_1}(\mathbf{x}+t\mathbf{d}) & f_{x_2x_2}(\mathbf{x}+t\mathbf{d}) & \dots & f_{x_2x_n}(\mathbf{x}+t\mathbf{d}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_nx_1}(\mathbf{x}+t\mathbf{d}) & f_{x_nx_2}(\mathbf{x}+t\mathbf{d}) & \dots & f_{x_nx_n}(\mathbf{x}+t\mathbf{d}) \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix}$$

$$= \mathbf{d}^T \nabla_x^2 f(\mathbf{x}+t\mathbf{d})\mathbf{d}. \tag{1.8}$$

The matrix of second partial derivative $\nabla_x^2 f(\mathbf{x}+t\mathbf{d})$ is called the Hessian of the function f. Note that it is a square $n \times n$ matrix and is symmetric when

$$f_{x_ix_j}(\mathbf{x}+t\mathbf{d}) = f_{x_jx_i}(\mathbf{x}+t\mathbf{d})$$

which always occurs when f is twice continuously differentiable.
Now recall Taylor's expansion of $g(t)$ around $t = 0$. That is

$$\begin{aligned} f(\mathbf{x}+t\mathbf{d}) = g(t) &= g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + o(t^2) \\ &= f(\mathbf{x}) + \underbrace{\nabla_x f(\mathbf{x}+0\mathbf{d}) \cdot (t\mathbf{d})}_{\text{by (1.6)}} + \frac{1}{2}\underbrace{(t\mathbf{d})^T \nabla_x^2 f(\mathbf{x}+0\mathbf{d})(t\mathbf{d})}_{\text{by (1.8)}} + o(\|t\mathbf{d}\|^2) \\ &= f(\mathbf{x}) + \nabla_x f(\mathbf{x}) \cdot (t\mathbf{d}) + \frac{1}{2}(t\mathbf{d})^T \nabla_x^2 f(\mathbf{x})(t\mathbf{d}) + o(\|t\mathbf{d}\|^2). \end{aligned}$$

When $t = 1$ we obtain

$$f(\mathbf{x}+\mathbf{d}) = f(\mathbf{x}) + \nabla_x f(\mathbf{x}) \cdot \mathbf{d} + \frac{1}{2}\mathbf{d}^T \nabla_x^2 f(\mathbf{x})\mathbf{d} + o(\|\mathbf{d}\|^2). \tag{1.9}$$

Recall that the 'Oh' notation $o\left(\|\mathbf{d}\|^2\right)$ means

$$\lim_{\|\mathbf{d}\| \to 0} \frac{o\left(\|\mathbf{d}\|^2\right)}{\|\mathbf{d}\|^2} = 0$$

and we may absorb any constant into the little oh term i.e.

$$K \times o\left(\|\mathbf{d}\|^2\right) = o\left(\|\mathbf{d}\|^2\right).$$

Hessians and gradients us to extend the previous characterizations of convexity to functions of several variables. In what follows, we may write using the identity $\nabla^2 f(x)d = d^T \nabla^2 f(x)$, which holds because $\nabla^2 f(x)$ is symmetric.

Theorem 1.1.6 *Suppose $f : \mathbf{X} \to \mathbf{R}$ is twice differentiable on a Euclidean space $\mathbf{X}$. Then consider the following are statements are equivalent:*

(i) f is convex on $\mathbf{X}$.

(ii) $f(y) - f(x) \geq \langle \nabla f(x), (y-x) \rangle$ for all x, $y \in \mathbf{X}$ (called the subgradient inequality).

(iii) The Hessian $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbf{X}$ i.e. $\langle \nabla^2 f(x)(d), d \rangle \geq 0$ for all d and x.

Proof. (i) $\Longrightarrow$ (ii): Using Lemma 1.1.5, we know that the function f is convex on $\mathbf{X}$ if and only if $g_{x,y}$ is also convex in $\mathbf{R}$ for any $x, y \in \mathbf{X}$. Using this convexity, we have that

$$\begin{aligned} g_{x,y}(1) - g_{x,y}(0) &\overset{(Theorem\ 1.1.4(i)b)}{\geq} g'_{x,y}(0)(1-0) \\ \text{or } f(y) - f(x) &\geq \frac{d}{dt}\left(f(x+t(y-x))\right)|_{t=0} = \lim_{t\to 0}\frac{1}{t}\left(f(x+t(y-x)) - f(x)\right) \\ &= \langle \nabla f(x), (y-x)\rangle. \quad \text{(using Remark 1.1.5(v))} \end{aligned}$$

(ii) $\Longrightarrow$(i): Since (ii) holds at all $x, y \in \mathbf{X}$ we have

$$(\forall y \in \mathbf{X}) \quad f(y) = \sup\left\{f(x) + \langle \nabla f(x), (y-x)\rangle \mid x \in \mathbf{X}\right\}.$$

In particular, this identity holds if we replace y by $\lambda z + (1-\lambda)y$, and so:

$$\begin{aligned} f(\lambda z + (1-\lambda)y) &= \sup\left\{f(x) + \langle \nabla f(x), (\lambda z + (1-\lambda)y - x)\rangle \mid x \in \mathbf{X}\right\} \\ &= \sup\left\{f(x) + \langle \nabla f(x), \lambda(z-x) + (1-\lambda)(y-x)\rangle \mid x \in \mathbf{X}\right\} \\ &= \sup\{\lambda\left(f(x) + \langle \nabla f(x), z-x\rangle\right) \\ &\qquad + (1-\lambda)\left(f(x) + \langle \nabla f(x), y-x\rangle\right) \mid x \in \mathbf{X}\} \\ &\leq \lambda \sup\left\{f(x) + \langle \nabla f(x), z-x\rangle \mid x \in X\right\} \\ &\qquad + (1-\lambda)\sup\left\{f(x) + \langle \nabla f(x), y-x\rangle \mid x \in \mathbf{X}\right\} \\ &= \lambda f(z) + (1-\lambda) f(y). \end{aligned}$$

To show (ii) $\Longrightarrow$ (iii) build the Taylor series in several variables[1] for a function f:

$$f(y) - f(x) = \langle \nabla f(x), (y-x)\rangle + \frac{1}{2}\langle \nabla^2 f(x)(y-x), (y-x)\rangle + o\left(\|(y-x)\|^2\right).$$

Combining this identity with (ii), we obtain that

$$\langle \nabla f(x), (y-x)\rangle + \frac{1}{2}\langle \nabla^2 f(x)(y-x), (y-x)\rangle + o\left(\|(y-x)\|^2\right) \geq \langle \nabla f(x), (y-x)\rangle$$

for all $y \in \mathbf{X}$. Let $y = x + td$ with $d \in \mathbf{X}$ then

$$\langle \nabla^2 f(x)(d), d\rangle \geq 2\frac{o\left(t^2\|d\|^2\right)}{t^2} \to 0 \quad \text{as } t \to 0.$$

(i) $\Longrightarrow$ (iii): Since we have both (i) $\Longleftrightarrow$ (ii) and (ii) $\Longrightarrow$ (iii), this is true automatically. However, here is an alternative proof. Since f is convex, $g_{x,y}$ is convex by Lemma 1.1.5. $g_{x,y}$ then satisfies $g''_{x,y}(0) \geq 0$ by Theorem 1.1.4(i)c. Noting that it is true for all x, y, this is equivalent to

$$\begin{aligned} &\langle \nabla^2 f(x + t(y-x))(y-x), (y-x)\rangle|_{t=0} \geq 0 \text{ for all } x, y \\ &\qquad \text{or equivalently} \quad \langle \nabla^2 f(x)(d), d\rangle \geq 0 \quad \text{for all } x, d. \end{aligned}$$

Here the equivalence uses the fact that we may express any d as $d = y - x$ for some y. Hence $\nabla^2 f(\mathbf{x})$ is positive semi–definite.

(iii) $\Longrightarrow$ (i): Since (iii) holds, then $\langle \nabla^2 f(x + t(y-x))(y-x), (y-x)\rangle \geq 0$ for all y which is equivalent to $g''_{x,y}(t) \geq 0$ for all t. Hence for all $x, y \in \mathbf{X}$ and $t \in \mathbf{R}$ we have $g''_{x,y}(t) \geq 0$ and hence by Theorem 1.1.4(i)c and Lemma 1.1.5 that f is convex. □

[1] see, for example, `https://en.wikipedia.org/wiki/Taylor_series#Taylor_series_in_several_variables`

Remark 1.1.6 *We may develop similar results for strict convexity with the change that* $\langle \nabla^2 f(x)(d), d\rangle > 0$ *for all* d *and* x *only implies strict convexity (the reverse implication fails). The failure of the reversed implication may be seen by considering* $f : \mathbf{R} \to \mathbf{R}$ *defined by* $f(x) = x^4$ *(note that* $f''(0) = 0$*).*

Remark 1.1.7 *The following properties of matrices will come in handy.*

(i) If an n *by* n *matrix* Q *has* n *linearly independent eigenvectors* $v_1, \ldots, v_n$*, then the matrix* $P = [v_1, \ldots, v_n]$ *satisfies* $P^{-1}QP = \mathrm{diag}\lambda$ *where* $\lambda_1, \ldots, \lambda_n$ *are the eigenvalues that correspond respectively to the eigenvectors* $v_1, \ldots, v_n$*. If the eigenvectors are normalized and orthogonal, then* $P^{-1} = P^T$ *(See [23, 5C, Section 5.2])*

(ii) If Q *is symmetric, then its eigenvalues are real and its unit eigenvectors are orthogonal (although its eigenvalues need not be distinct; e.g. the identity matrix) [23, Section 5.6].*

(iii) (Spectral Theorem) A real symmetric matrix Q *can be written as* $Q = P(\mathrm{diag}\lambda)P^T$*, where* P *contains its orthonormal eigenvectors and* $P^T = P^{-1}$ *[23, Theorem 50, p.317]).*

(iv) For symmetric matrices X, Y *it holds that* $(XY)^T = YX$*.*

To see why (iv) holds, recall that for any matrices A, B*, we have that* $(AB)^T = B^T A^T$ *[23, Identity 1M]. Since* X, Y *are symmetric, they are self-transpose, and so we have* $(XY)^T = Y^T X^T = YX$*.*

In the following we will use the outer product of two vectors. Our motivation is to learn some tools that are useful for working with Hessians.

Example 1.1.2 *Consider the 'outer product' of two vectors. This differs from the inner product in the order in which we place the vectors. Instead of forming*

$$x^T x = \begin{matrix}(1,2)\end{matrix} \quad \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 1 + 4 = 5 \quad \text{a } 1 \times 1 \text{ matrix (or real number),}$$

we form

$$xx^T = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \quad \text{a } 2 \times 2 \text{ matrix.}$$

Similarly in three dimensions for $x^T = (1, 4, 2)$ *we have*

$$xx^T = \begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 4 & 2 \\ 4 & 16 & 8 \\ 2 & 8 & 4 \end{pmatrix}$$

a 3×3 *matrix. Note that this matrix is symmetric since* $\left(xx^T\right)^T = (x^T)^T x^T = xx^T$ *(which follows from the identity* $(AB)^T = B^T A^T$*).*

Let us also consider a arbitrary real symmetric matrix Q again. Recall that, by the finite-dimensional spectral theorem, a symmetric, real matrix can be diagonalized by an orthogonal matrix consisting of its eigenvectors. Thus there exists an orthogonal matrix that diagonalizes Q. For example take

$$Q = \begin{pmatrix} 9 & 1 & 2 \\ 1 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix}.$$

We can find the matrix P that contains the normalized eigenvectors of Q as columns and diagonalizes Q to a matrix with diagonal entries $\lambda_1 = 6$, $\lambda_2 = 4.6277$ and $\lambda_3 = 10.372$, the approximate eigenvalues of Q. The matrix P containing the corresponding (normalized) eigenvectors (in this order), is:

$$P = \begin{pmatrix} \frac{1}{\sqrt{6}} & -.33472 & .84928 \\ \frac{1}{\sqrt{6}} & .89907 & .15817 \\ \frac{-2}{\sqrt{6}} & .28218 & .50373 \end{pmatrix}.$$

Remember also that an orthogonal matrix P satisfies $P^T = P^{-1}$, and so we can write P^T in place of P^{-1}. By direct matrix multiplication

$$\begin{aligned} P^T Q P &= \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ -.33472 & .89907 & .28218 \\ .84928 & .15817 & .50373 \end{pmatrix} \begin{pmatrix} 9 & 1 & 2 \\ 1 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & -.33472 & .84928 \\ \frac{1}{\sqrt{6}} & .89907 & .15817 \\ \frac{-2}{\sqrt{6}} & .28218 & .50373 \end{pmatrix} \\ &= \begin{pmatrix} 6.0 & -2.4495 \times 10^{-5} & -2.4495 \times 10^{-5} \\ -2.4495 \times 10^{-5} & 4.6277 & 2.9037 \times 10^{-4} \\ -2.4492 \times 10^{-5} & 2.9037 \times 10^{-4} & 10.373 \end{pmatrix} \\ &\approx \begin{pmatrix} 6.0 & 0 & 0 \\ 0 & 4.6277 & 0 \\ 0 & 0 & 10.373 \end{pmatrix}. \end{aligned}$$

Thus, multiplying the left side of the above by P and the right side by P^T yields $P \operatorname{diag}[\lambda_1, \lambda_2, \lambda_3] P^T \simeq Q$. This prompts us to define (for real, positive definite, symmetric matrices)

$$\begin{aligned} Q^{\frac{1}{2}} &= P \operatorname{diag}\left[\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}\right] P^T \\ &= \begin{pmatrix} \frac{1}{\sqrt{6}} & -.33472 & .84928 \\ \frac{1}{\sqrt{6}} & .89907 & .15817 \\ \frac{-2}{\sqrt{6}} & .28218 & .50373 \end{pmatrix} \begin{pmatrix} \sqrt{6.0} & 0 & 0 \\ 0 & \sqrt{4.6277} & 0 \\ 0 & 0 & \sqrt{10.373} \end{pmatrix} \\ &\qquad \times \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ -.33472 & .89907 & .28218 \\ .84928 & .15817 & .50373 \end{pmatrix} \\ &= \begin{pmatrix} 2.9723 & .19351 & .35817 \\ .19351 & 2.2277 & -1.4125 \times 10^{-2} \\ .35817 & -1.4125 \times 10^{-2} & 2.6215 \end{pmatrix}. \end{aligned}$$

Again one can directly verify that

$$Q^{\frac{1}{2}} Q^{\frac{1}{2}} = \begin{pmatrix} 2.9723 & .19351 & .35817 \\ .19351 & 2.2277 & -1.4125 \times 10^{-2} \\ .35817 & -1.4125 \times 10^{-2} & 2.6215 \end{pmatrix}$$

$$\times \begin{pmatrix} 2.9723 & .19351 & .35817 \\ .19351 & 2.2277 & -1.4125 \times 10^{-2} \\ .35817 & -1.4125 \times 10^{-2} & 2.6215 \end{pmatrix}$$

$$= \begin{pmatrix} 9.0003 & 1.0012 & 2.0008 \\ 1.0012 & 5.0003 & 8.1453 \times 10^{-4} \\ 2.0008 & 8.1453 \times 10^{-4} & 7.0007 \end{pmatrix} \approx \begin{pmatrix} 9 & 1 & 2 \\ 1 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix} = Q.$$

The errors are due to numerical rounding in the calculation (which are only performed to about six decimal places).

Lemma 1.1.7 *The following hold:*

(i) $v^T(\operatorname{diag} z)v = \sum_{i=1}^n v_i^2 z_i$*; and*

(ii) $v^T(zz^T)v = (\sum_{i=1}^n v_i z_i)^2$.

(iii) If Q *is symmetric, then* $(Qx)^T = x^T Q$.

Proof. (i): Notice

$$(v_1, \ldots, v_n) \begin{pmatrix} z_1 & 0 & \ldots & 0 & 0 \\ 0 & z_2 & \ldots & 0 & 0 \\ \vdots & \ldots & \ddots & \ldots & \vdots \\ 0 & \vdots & \ldots & z_{n-1} & 0 \\ 0 & 0 & \ldots & 0 & z_n \end{pmatrix} \cdot v = \begin{pmatrix} v_1 z_1 \\ v_2 z_2 \\ \vdots \\ v_{n-1} z_{n-1} \\ v_n z_n \end{pmatrix} \cdot v = \sum_{i=1}^n v_i^2 z_i.$$

(ii): Notice

$$v^T(zz^T)v = v^T \begin{pmatrix} z_1 z_1 & z_1 z_2 & \ldots & z_1 z_{n-1} & z_1 z_n \\ z_2 z_1 & z_2 z_2 & \ldots & z_2 z_{n-1} & z_2 z_n \\ \vdots & \ldots & \ddots & \vdots & \\ z_{n-1} z_1 & \vdots & \ldots & z_{n-1} z_{n-1} & z_{n-1} z_n \\ z_n z_1 & z_n z_2 & \ldots & z_n z_{n-1} & z_n z_n \end{pmatrix} \cdot v$$

$$= \begin{pmatrix} v_1 z_1 z_1 & +v_2 z_1 z_2 & +\ldots & +v_{n-1} z_1 z_{n-1} & +v_n z_1 z_n \\ v_1 z_2 z_1 & +v_2 z_2 z_2 & +\ldots & +v_{n-1} z_2 z_{n-1} & +v_n z_2 z_n \\ \vdots & \ldots & \ddots & \vdots & \\ v_1 z_{n-1} z_1 & +\vdots & +\ldots & +v_{n-1} z_{n-1} z_{n-1} & +v_n z_{n-1} z_n \\ v_1 z_n z_1 & +v_2 z_n z_2 & +\ldots & +v_{n-1} z_n z_{n-1} & +v_n z_n z_n \end{pmatrix} \cdot v$$

$$= v_1 \left(v_1 z_1 z_1 + v_2 z_1 z_2 + \cdots + v_n z_1 z_n\right) + \cdots + v_n \left(v_1 z_n z_1 + v_2 z_n z_2 + \cdots + v_n z_n z_n\right)$$

$$= v_1 \left(z_1(v_1 z_1) + z_1(v_2 z_2) + \cdots + z_1(v_n z_n)\right) + \cdots + v_n \left(z_n(v_1 z_1) + z_n(v_2 z_2) + \cdots + z_n(v_n z_n)\right)$$

$$= v_1 z_1 \left(v_1 z_1 + v_2 z_2 + \cdots + v_n z_n\right) + \cdots + v_n z_n \left(v_1 z_1 + v_2 z_2 + \cdots + v_n z_n\right)$$

$$= (v_1 z_1 + \ldots v_n z_n)^2.$$

(iii): This is clear. □

Example 1.1.3 (Using symmetric matrices to show convexity) *Let $f : \mathbf{R}^n \to \mathbf{R}$ be defined by*

$$f(x) := \log\left(e^{x_1} + e^{x_2} + \cdots + e^{x_n}\right).$$

Show that f is convex.

Soln: *We show that the Hessian is positive definite. Note that*

$$\begin{aligned}
\frac{\partial f}{\partial x_i} &= \frac{1}{(e^{x_1} + e^{x_2} + \cdots + e^{x_n})} e^{x_i} \quad \textit{and so} \\
\frac{\partial^2 f}{\partial x_i x_j} &= -\frac{1}{(e^{x_1} + e^{x_2} + \cdots + e^{x_n})^2} e^{x_i} e^{x_j} \quad \textit{for } i \neq j \quad \textit{and otherwise} \\
\frac{\partial^2 f}{\partial x_i^2} &= -\frac{1}{(e^{x_1} + e^{x_2} + \cdots + e^{x_n})^2} e^{2x_i} + \frac{1}{(e^{x_1} + e^{x_2} + \cdots + e^{x_n})} e^{x_i}.
\end{aligned}$$

So

$$\nabla^2 f(x) = \frac{1}{(z_1 + z_2 + \cdots + z_n)^2} \left((z_1 + z_2 + \cdots + z_n)\operatorname{diag} z - zz^T\right)$$

where $z = (e^{x_1}, e^{x_2}, \ldots, e^{x_n})^T$. We note that

$$zz^T = \begin{pmatrix}
e^{2x_1} & e^{x_1}e^{x_2} & \ldots & e^{x_1}e^{x_{n-1}} & e^{x_1}e^{x_n} \\
e^{x_2}e^{x_1} & e^{2x_2} & \ldots & e^{x_2}e^{x_{n-1}} & e^{x_2}e^{x_n} \\
\vdots & \ldots & \ddots & \ldots & \vdots \\
e^{x_1}e^{x_{n-1}} & \vdots & \ldots & e^{2x_{n-1}} & e^{x_{n-1}}e^{x_n} \\
e^{x_1}e^{x_n} & e^{x_1}e^{x_{n-1}} & \ldots & e^{x_{n-1}}e^{x_n} & e^{2x_n}
\end{pmatrix} \quad \textit{and}$$

$$\operatorname{diag} z = \begin{pmatrix}
e^{x_1} & 0 & \ldots & 0 & 0 \\
0 & e^{x_2} & \ldots & 0 & 0 \\
\vdots & \ldots & \ddots & \ldots & \vdots \\
0 & \vdots & \ldots & e^{x_{n-1}} & 0 \\
0 & 0 & \ldots & 0 & e^{x_n}
\end{pmatrix}.$$

Then

$$\begin{aligned}
v^T \nabla^2 f(x) v &= \frac{1}{(z_1 + z_2 + \cdots + z_n)^2} \left(\left(\sum_i z_i\right) \overbrace{\left(\sum_i v_i^2 z_i\right)}^{v^T(\operatorname{diag} z)v} - \overbrace{\left(\sum_i v_i z_i\right)^2}^{v^T(zz^T)v} \right) \\
&= \frac{1}{(\sum_i z_i)^2} \left[\left\| \left(z_1^{\frac{1}{2}}, \ldots, z_m^{\frac{1}{2}}\right) \right\|^2 \left\| \left(v_1 z_1^{\frac{1}{2}}, \ldots, v_n z_m^{\frac{1}{2}}\right) \right\|^2 \right. \\
&\qquad \left. - \left(z_1^{\frac{1}{2}}, \ldots, z_m^{\frac{1}{2}}\right) \cdot \left(v_1 z_1^{\frac{1}{2}}, \ldots, v_n z_m^{\frac{1}{2}}\right) \right] \\
&\geq 0 \quad \textit{by the Cauchy Schwarz inequality}^2 \textit{i.e} \\
v \cdot w &\leq |v \cdot w| \leq \|v\| \, \|w\| .
\end{aligned}$$

[2] easily proven from $v \cdot w = |v||w| cos(\theta)$

An example of a function that take some infinite values is the indicator function of a convex set $C \subseteq \mathbf{X}$ defined as:

$$\delta_C(x) = \begin{cases} 0 & \text{if} \quad x \in C \\ +\infty & \text{otherwise} \end{cases}. \tag{1.10}$$

Clearly (1.10) δ_C is convex when C is convex. We may constrain a given function f to a set C by defining a new extended real valued convex function

$$g(x) := f(x) + \delta_C(x) = \begin{cases} f(x) & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

Definition 1.1.5 *A convex cone is a convex set C in $\mathbf{X}$ with the additional property that $\lambda x \in C$ whenever $x \in C$ and $\lambda \geq 0$.*

Given any convex set C we may generate the smallest convex cone containing C as

$$\operatorname{cone} C = \{\lambda x \mid x \in C \text{ and } \lambda \geq 0\}.$$

If C is not convex then using the characterization of the smallest convex set containing C in (1.1) one may see that the smallest convex cone containing C is given by

$$\operatorname{conv}\operatorname{cone} C = \left\{ \sum_i \lambda_i x_i \mid \lambda_i \geq 0 \text{ and } x_i \in C \right\}.$$

Example 1.1.4 *The simplest example of a convex cone is the positive ornathant in $\mathbf{R}^n$*

$$\mathbf{R}^n_+ := \{x := (x_1, \ldots, x_n)) \in \mathbf{R}^n \mid x_i \geq 0 \text{ for all } i = 1, \ldots, n\}.$$

(Check that this is indeed a convex cone).

Example 1.1.5 *A half-space in $\mathbf{X}$ is given by*

$$H_{y,\alpha} := \{x \in \mathbf{X} \mid f(x) \geq \alpha\}$$

where $f(x)$ is a linear mapping on the space $\mathbf{X}$. When $\mathbf{X} = \mathbf{R}^n$ then one could take $f(x) = x \cdot y$ (the usual dot product of vectors). Show $H_{y,\alpha}$ is a convex set.
Soln: *Take $x, z \in H_{y,\alpha}$ and $\lambda \in [0,1]$ then*

$$f(\lambda x + (1-\lambda) z) = \lambda f(x) + (1-\lambda) f(z) \geq \lambda\alpha + (1-\lambda)\alpha = \alpha$$

and so $\lambda x + (1-\lambda) z \in H_{y,\alpha}$. Having verified convexity we now show that $H_{y,0}$ is a cone. This follows immediately since $\lambda \geq 0$ and $x \in H_{y,0}$ implies

$$f(\lambda x) = \lambda f(x) \geq \lambda 0 = 0$$

and so $\lambda x \in H_{y,0}$ for all $\lambda \geq 0$.

Example 1.1.6 *Let*

$$P := \{p := (p_1, p_2, \ldots, p_n) \geq 0 \mid \sum_i p_i = 1\}$$

be the set of all probability measures on a discrete set $\{x_1, x_2, \ldots, x_n\}$. Define

$$S = \{p \in P \mid \alpha \leq \mathbf{E}_p f(x) \leq \beta\}$$

where $\mathbf{E}_p f(x) := \sum_i p_i f(x_i)$ and the function $f : \mathbf{R} \to \mathbf{R}$ is given. Show that S is a convex set.

Soln: Let p, $q \in S$ and $\lambda \in [0,1]$ then

$$r := \lambda p + (1-\lambda)\, q \geq 0 \quad \text{and} \quad \sum_i (\lambda p_i + (1-\lambda)\, q_i) = 1.$$

Also

$$\begin{aligned} \mathbf{E}_r f(x) &:= \sum_i r_i f(x_i) \\ &= \lambda \sum_i p_i f(x_i) + (1-\lambda) \sum_i q_i f(x_i) \\ &= \lambda \mathbf{E}_p f(x) + (1-\lambda)\, \mathbf{E}_q f(x) \end{aligned}$$

and so from $\alpha \leq \mathbf{E}_p f(x) \leq \beta$ and $\alpha \leq \mathbf{E}_q f(x) \leq \beta$ we may deduce

$$\alpha \leq \mathbf{E}_r f(x) \leq \beta.$$

Definition 1.1.6 *We denote the vector space of all symmetric $n \times n$ matrices by*

$$\mathcal{S}(n) := \left\{ X \in \mathbf{R}^{n \times n} \mid X = X^T \right\},$$

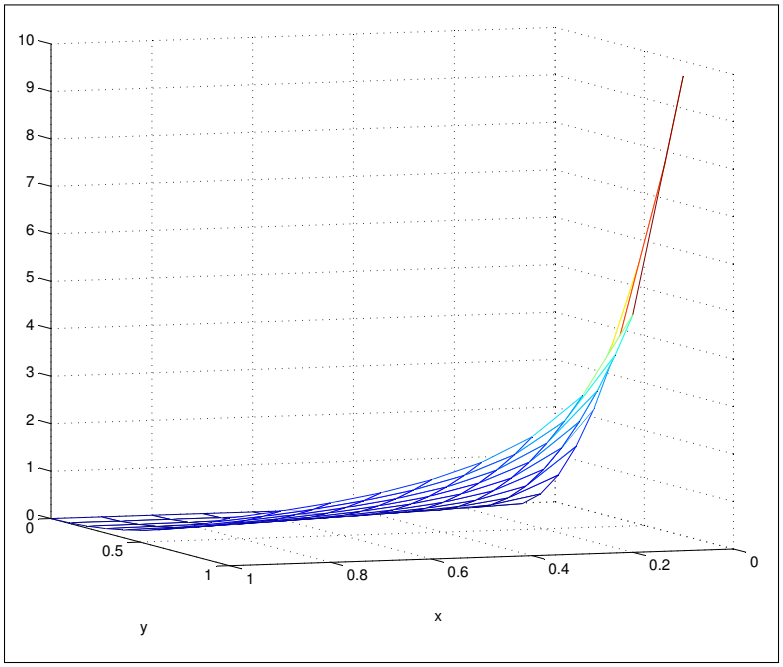
where the superscript T denotes the transpose operation. A natural cone in the space is the cone of positive semidefinite matrices given by

$$\mathcal{P}(n) := \left\{ X \in \mathcal{S}(n) \mid x^T X x \geq 0 \text{ for all } x \in \mathbf{R}^n \right\}.$$

If I fix an x and consider all matrices X that satisfy $x^T X x$, those matrices form a half-space. Thus $\mathcal{P}(n)$ is the intersection of infinitely many half-spaces. Even in two dimensions this set is complicated. For

$$X = \begin{pmatrix} x & y \\ y & z \end{pmatrix} \in \mathcal{P}(n)$$

we need $x \geq 0$, $z \geq 0$ and $\det X = xz - y^2 \geq 0$ or $xz \geq y^2$.



The boundary of the positive semidefinite cone $\mathcal{P}(2)$ ploted in $\mathbf{R}^3$.

Example 1.1.7 *It is clear from Lemma 1.1.7(ii) that any symmetric matrix of form zz^T will be positive semidefinite. However, not every symmetric matrix need be positive semidefinite. For example, take $x = (1, -1)$. We have from Lemma 1.1.7(ii) that*

$$zz^T = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

satisfies $v^T(zz^T)v = (v_1 - v_2)^2$, which can be easily verified. On the other hand, if we take the slightly modified matrix

$$M = \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix},$$

Then we have that $v^T M v = v_1^2 - 2v_1v_2 - v_2^2$. Choosing $v = (1, 1)$, M is clearly not positive semidefinite.

Lemma 1.1.8 (Positive semidefinite matrices) *We have the following properties of positive semidefinite matrices.*

(i) If Q is positive semidefinite, then its eigenvalues are nonnegative and real, and its (principal) square root $Q^{\frac{1}{2}} := P(\mathrm{diag}\sqrt{\lambda})P^T$ is also positive semidefinite;

(ii) If Q is positive semidefinite and $Q \preceq I$, then the eigenvectors $\lambda_1, \ldots, \lambda_n$ of Q satisfy $\lambda_1, \ldots, \lambda_n \in [0, 1]$.

Proof. (i): The eigenvalues are real by the spectral theorem (Remark 1.1.7(iii)). Suppose for a contradiction that Q possesses a negative eigenvalue λ_i. Then the corresponding eigenvector x_i satisfies $x_i^T Q x_i = x_i^T(\lambda_i x_i) = \lambda_i x_i^T x_i < 0$, which contradicts the requirement that Q be positive semidefinite.

Let P be the matrix of eigenvectors $x_1, \ldots, x_n$ that diagonalizes Q. Because $Q = Q^{\frac{1}{2}}Q^{\frac{1}{2}}$, it is straightforward to verify that Q and $Q^{\frac{1}{2}}$ have the same eigenvectors, and thereafter clear that the eigenvalues $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ of $Q^{\frac{1}{2}}$ are nonnegative. By the spectral theorem [23, Theorem 50, p.317],

$$\begin{aligned} Q = P(\mathrm{diag}\lambda)P^T &= \begin{bmatrix} | & & | \\ x_1 & \ldots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & | \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix} \\ &= \lambda_1 x_1 x_1^T + \lambda_2 x_2 x_2^T + \cdots + \lambda_n x_n x_n^T. \end{aligned}$$

From this, the spectral decomposition of Q, we see that Q is a sum of symmetric matrices $\lambda_i x_i x_i^T$. Similarly,

$$Q^{\frac{1}{2}} = P(\mathrm{diag}\sqrt{\lambda})P^T = \sqrt{\lambda_1} x_1 x_1^T + \sqrt{\lambda_2} x_2 x_2^T + \cdots + \sqrt{\lambda_n} x_n x_n^T,$$

where each of the $\sqrt{\lambda_i} x_i x_i^T$ is a symmetric matrix. Having shown that $Q^{\frac{1}{2}}$ is symmetric with nonnegative eigenvalues, we have that $Q^{\frac{1}{2}}$ is positive semidefinite.

(ii): It is straightforward to see that $I - Q$ has the same eigenvectors as Q and so is also diagonalized by P. More specifically, $I - Q = PIP^T - PDP^T = P(I - \mathrm{diag}\lambda)P^T$. Thus the matrix $I - \mathrm{diag}\lambda$ consists of the eigenvalues of $I - Q$. The definition of the inequality $Q \preceq I$ is that $I - Q$ must be positive semidefinite. Consequently, its eigenvalues are all nonnegative. This means that the members of $I - \mathrm{diag}\lambda$ are all nonnegative, and so each of the λ_i must be less than or equal to 1. □

1.1.1 Problem Set 1: Convexity of Functions

Problem 1.1.9 *Show that any norm is a convex function.*

Problem 1.1.10 *Use derivatives to show that the function*

$$f(x) = -\log(x)$$

is strictly convex in the positive reals. Use the definition of convexity to show that for any strictly positive reals $x^1, \ldots, x^m$ and nonnegative numbers $\lambda_1, \ldots, \lambda_m$ summing to one, we have the following (arithmetic-geometric mean) inequality holding:

$$\sum_i \lambda_i x^i \geq \Pi_i \left(x^i\right)^{\lambda_i}.$$

Problem 1.1.11 *Prove that the following function of $x \in \mathbf{R}^2$ is convex (i.e. $-f$ is concave)*

$$f(x) = \begin{cases} -(x_1 x_2 \ldots x_n)^{\frac{1}{n}} & \text{if } x \in \mathbf{R}^n_+ := \{(x_1, \ldots, x_2) \mid x_i \geq 0\} \\ +\infty & \text{otherwise} \end{cases}$$

[Hint: compute the Hessian; Lemma 1.1.7 will make it easier. Make use of the vector $(1, 1, \ldots, 1)$ and the Cauchy–Schwarz inequality.]

1.2 Convexity Preserving Operations

One can construct complicated convex sets from simpler ones by using various convexity preserving operations.

1.2.1 Intersection.

Lemma 1.2.1 *Convexity is preserved under arbitrary intersection. That is if $\{C_\alpha\}_{\alpha\in\Lambda}$ is a family of convex sets in a vector space $\mathbf{X}$ then so is*

$$C := \bigcap_{\alpha\in\Lambda} C_\alpha.$$

If all C_α are cones then so is C.

Proof. Let x_1, $x_2 \in C$ and $\lambda \in [0, 1]$ then x_1, $x_2 \in C$ for all $\alpha \in \Lambda$ and as C_α is convex we have $\lambda x_1 + (1-\lambda)x_2 \in C_\alpha$ for all $\alpha \in \Lambda$. But then

$$\lambda x_1 + (1-\lambda)x_2 \in \bigcap_{\alpha\in\Lambda} C_\alpha = C.$$

Now suppose all C_α are cones, $x \in C$ and $\lambda \geq 0$. As $x \in C_\alpha$ we have $\lambda x \in C_\alpha$ for all $\alpha \in \Lambda$ and so $\lambda x \in \bigcap_{\alpha\in\Lambda} C_\alpha = C$. □

Example 1.2.1 *The positive semidefinite cone $\mathcal{P}(n)$ in $\mathcal{S}(n)$ can be written as*

$$\mathcal{P}(n) = \bigcap_{x \in \mathbf{R}^n} \left\{ X \in \mathcal{S}(n) \mid x^T X x \geq 0 \right\}.$$

The simpler set $\left\{ X \in \mathcal{S}(n) \mid x^T X x \geq 0 \right\}$ is clearly convex since it is actually an half-space for each fixed x as $X \mapsto x^T X x$ is a linear mapping i.e. for all X_1, $X_2 \in \mathcal{S}(n)$ and α_1, $\alpha_2 \in \mathbf{R}$ we have

$$x^T(\alpha_1 X_1 + \alpha_2 X_2)x = \alpha_1 \left(x^T X_1 x \right) + \alpha_2 \left(x^T X_2 x \right).$$

As $\mathcal{P}(n)$ is an intersection of convex cones and so is also a convex cone.

This kind of argument may also be used to show that for an arbitrary convex set C we have

$$C = \bigcap \{ H \mid C \subseteq H, \text{ and } H \text{ is a half-space} \}. \tag{1.11}$$

Indeed the right hand side of (1.11) can be shown to be just $\operatorname{co} C$ even if C is not convex.

1.2.2 Cross Products

When Λ is finite index set $\Lambda = \{\alpha_1, \ldots, \alpha_k\}$ we denote

$$\bigotimes_{\alpha \in \Lambda} C_\alpha = C_{\alpha_1} \times C_{\alpha_2} \times \cdots \times C_{\alpha_k}.$$

When Λ is not finite, the cross product is more difficult to write. For this reason, we will prove the next result (Lemma 1.2.2) for 2-dimensions. The proof for a possibly infinite index set Λ is given below in Lemma 1.2.3, and is similar.

Lemma 1.2.2 (Λ finite) *Given two convex sets $\{C_i\}_{i \in \{1,2\}}$, the cross product*

$$C = \bigotimes_{\alpha \in \Lambda} C_i := \left\{ (c_i)_{i \in \{1,2\}} \mid c_i \in C_i \right\} = \{ (c_1, c_2) \mid c_1 \in C_1 \text{ and } c_2 \in C_2 \}.$$

is convex. If both C_i are cones then so is C.

Proof. If $(c_i^1)_{i \in \{1,2\}} = (c_1^1, c_2^1)$ and $(c_i^2)_{i \in \{1,2\}} = (c_1^2, c_2^2)$ are in C, then for any $\lambda \in [0,1]$ and each $i \in \{1,2\}$, we have

$$\lambda c_i^1 + (1-\lambda) c_i^2 \in C_i.$$

That is

$$(\lambda c_i^1 + (1-\lambda) c_i^2)_{i \in \{1,2\}} = (\lambda c_1^1 + (1-\lambda) c_1^2, \lambda c_2^1 + (1-\lambda) c_2^2) \in C_1 \times C_2 = \bigotimes_{i \in \{1,2\}} C_i = C.$$

The proof that both C_i are cones is similar. □

Lemma 1.2.3 (Λ general) *Given a family of convex sets $\{C_\alpha\}_{\alpha \in \Lambda}$ the cross product*

$$C = \bigotimes_{\alpha \in \Lambda} C_\alpha := \{ (c_\alpha)_{\alpha \in \Lambda} \mid c_\alpha \in C_\alpha \}$$

is convex. If all C_α are cones then so is C.

Proof. If $(c_\alpha^1)_{\alpha\in\Lambda}$ and $(c_\alpha^2)_{\alpha\in\Lambda}$ are in C then for any $\lambda \in [0,1]$ and each $\alpha \in \Lambda$ we have

$$\lambda c_\alpha^1 + (1-\lambda)c_\alpha^2 \in C_\alpha.$$

That is

$$\left(\lambda c_\alpha^1 + (1-\lambda)c_\alpha^2\right)_{\alpha\in\Lambda} \in \bigotimes_{\alpha\in\Lambda} C_\alpha = C.$$

If all C_α are cones $(c_\alpha)_{\alpha\in\Lambda} \in C$ and $\lambda \geq 0$ we have $\lambda(c_\alpha)_{\alpha\in\Lambda} = (\lambda c_\alpha)_{\alpha\in\Lambda} \in C$. □

1.2.3 Affine Images and Preimages

Let $h : \mathbf{X} \to \mathbf{Y}$ be a linear function i.e. for all α_1, $\alpha_2 \in \mathbf{R}$ and x_1, $x_2 \in \mathbf{X}$ we have

$$h(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 h(x_1) + \alpha_2 h(x_2).$$

We say $f : \mathbf{X} \to \mathbf{Y}$ is an affine mapping if it is a translation of a linear mapping i.e.

$$f(x) = h(x) + \alpha \quad \text{for some } \alpha \in \mathbf{Y}.$$

Then the image of a convex set C in $\mathbf{X}$ is a convex set $f(C)$ in $\mathbf{Y}$ where

$$f(C) := \{f(x) \mid x \in C\}.$$

Also the preimage of a convex set S in $\mathbf{Y}$ is a convex set $f^{-1}(S)$ in $\mathbf{X}$ where

$$f^{-1}(S) := \{x \in \mathbf{X} \mid f(x) \in S\}.$$

Exercise 1.2.1 *Prove that $f(C)$ and $f^{-1}(S)$ are convex for convex sets C in $\mathbf{X}$ and S in $\mathbf{Y}$.*

Proof. Consider first proving the convexity of $f(C)$. We take y_1, $y_2 \in f(C)$ and $\lambda \in [0,1]$. Then we note that $y_i = f(x_i)$ for some $x_i \in C$ so

$$\begin{aligned}\lambda y_1 + (1-\lambda) y_2 &= \lambda f(x_1) + (1-\lambda) f(x_2) \\ &= \lambda h(x_1) + (1-\lambda) h(x_2) + \lambda\alpha + (1-\lambda)\alpha \\ &= h(\lambda x_1 + (1-\lambda) x_2) \ + \alpha = f(\lambda x_1 + (1-\lambda) x_2) \in f(C).\end{aligned}$$

Now consider proving the convexity of $f^{-1}(C)$. Take x_1, $x_2 \in f^{-1}(C)$ and $\lambda \in [0,1]$. Then we have $f(x_i) \in C$ for each i and so

$$\begin{aligned}f(\lambda x_1 + (1-\lambda) x_2) &= h(\lambda x_1 + (1-\lambda) x_2) + \alpha \\ &= \lambda h(x_1) + (1-\lambda) h(x_2) + \lambda\alpha + (1-\lambda)\alpha \\ &= \lambda (h(x_1) + \alpha) + (1-\lambda) h(h(x_2) + \alpha) \\ &= \lambda f(x_i) + (1-\lambda) f(x_2) \in C \quad \text{(by the convexity of } C\text{)}.\end{aligned}$$

Thus it follows that

$$\lambda x_1 + (1-\lambda) x_2 \in f^{-1}(C).$$

□

Definition 1.2.1 *For $x, y \in \mathbb{R}^n$, we write $x \leq y$ if $y - x \in \mathbb{R}^n_+$.*

Example 1.2.2 *The following operations can all be shown to preserve convexity.*

- *scalar multiplication* $\alpha C = \{\alpha x \mid x \in C\}$*, for* $\alpha \in \mathbf{R}$

  Why? *This is the image of* C *under the affine mapping* $f(x) = \alpha x$*.*

- *translation* $C + y := \{x + y \mid x \in C\}$*, where* $y \in \mathbf{X}$

  Why? *This is the image of* C *under the affine transformation* $f(x) = x + y$*.*

- *the sum of two sets* S_1 *and* S_2 *in* $\mathbf{X}$ *given by*

$$S_1 + S_2 := \{x_1 + x_2 \mid x_1 \in S_1 \text{ and } x_2 \in S_2\}.$$

  Why? *This set may be viewed as the image of the convex set*

$$S_1 \times S_2 := \{(x_1, x_2) \mid x_1 \in S_1 \text{ and } x_2 \in S_2\}$$

  under the affine mapping

$$f(x_1, x_2) = x_1 + x_2.$$

- *The polyhedron*

$$C' := \{x \mid Ax \leq b,\ Cx = d\}$$

  where A *is an* $n \times m$ *matrix* $b \in \mathbf{R}^m$ *and* C *is an* $n \times l$ *matrix and* $d \in \mathbf{R}^l$*.*

  Why? *This may be shown to convex by noting that it is representable as* $f^{-1}(S)$ *where* $S = \mathbf{R}^m_+ \times \{0\}$ *is a convex set (the cross product of two convex sets) and*

$$f(x) := (b - Ax, d - Cx) = (b, d) - (A, C)\, x$$

  is an affine mapping. That is, $x \in f^{-1}\left(\mathbf{R}^m_+ \times \{0\}\right)$ *or* $f(x) \in \mathbf{R}^m_+ \times \{0\}$ *which means* $b - Ax \in \mathbf{R}^m_+$ *and* $d - Cx = 0$*.*

- *The solution set to the matrix linear inequality*

$$C' := \{x \in \mathbf{R}^n \mid x_1 A_1 + \cdots + x_n A_n \preceq B\}$$

  where $x_1 A_1 + \cdots + x_n A_n \preceq B$ *is interpreted as meaning* $B - (x_1 A_1 + \cdots + x_n A_n) \in \mathcal{P}(n)$*.*

  Why? *The set* C' *is representable as* $f^{-1}(\mathcal{P}(n))$ *under the affine function:*

$$f(x) : \mathbf{R}^n \to \mathcal{S}(n) : x \mapsto B - (x_1 A_1 + \cdots + x_n A_n).$$

  The function is clearly affine in x*. Moreover, notice that*

$$\begin{aligned} f^{-1}(\mathcal{P}(n)) = \{x | f(x) \in \mathcal{P}(n)\} &= \{x \mid f(x) \in \mathcal{P}(n)\} \\ &= \{x \mid 0 \preceq f(x)\} = C'. \end{aligned}$$

  Therefore C' *is convex, because it is the preimage under an affine function* f *of a convex set* $\mathcal{P}(n)$*.*

- *Suppose $C \subseteq \mathbf{X} \times \mathbf{Y} \times \mathbf{R}$ is a convex set then under the linear projection mapping $P(x,y,\beta) = (x,\beta)$ the set*

$$P(C) := \{(x,\beta) \mid \exists y \text{ such that } (x,y,\beta) \in C\}$$

is convex (put possibly not closed).

Why? *The projection P is linear because for any A, $B \in \mathbf{R}$ we have*

$$\begin{aligned} &P(A(x_1,y_1,\beta_1) + B(x_2,y_2,\beta_2)) \\ &= P(Ax_1 + Bx_2, Ay_1 + By_2, A\beta_1 + B\beta_2) \\ &= (Ax_1 + Bx_2, A\beta_1 + B\beta_2) = A(x_1,\beta_1) + B(x_2,\beta_2) \\ &= AP(x_1,y_1,\beta_1) + BP(x_2,y_2,\beta_2). \end{aligned}$$

It may not be closed as seen in the following example: let

$$C = \operatorname{epi}(\{(x,y) \mid y \geq 1/x\}) \quad \text{and} \quad P(x,y) = y.$$

Then

$$P(C) = (0,+\infty) \quad \text{is an open convex set.}$$

Example 1.2.3 *When $C = \operatorname{epi} f \subseteq \mathbf{X} \times \mathbf{Y} \times \mathbf{R}$ for a proper function $f : \mathbf{X} \times \mathbf{Y} \to \mathbf{R}$. Let*

$$\phi(x) := \inf\{f(x,y) \mid y \in \mathbf{Y}\}.$$

Then

$$\operatorname{epi}\phi = \{(x,\alpha) \mid (x,\beta) \in P(\operatorname{epi} f) \text{ for all } \beta > \alpha\},$$

where $P(x,y,\alpha) = (x,\alpha)$.

Proof. Let $E := P(\operatorname{epi} f)$ and

$$S := \{(x,\alpha) \mid (x,\beta) \in E \text{ for all } \beta > \alpha\}.$$

Next note that if $(x,y,\alpha) \in \operatorname{epi} f$ then $\alpha \geq f(x,y) \geq \phi(x)$ and so $P(x,y,\alpha) = (x,\alpha) \in \operatorname{epi}\phi$. Consequently when $\beta > \alpha$ we must have $(x,\beta) \in \operatorname{epi}\phi$ implying $S \subseteq \operatorname{epi}\phi$. Conversely let $(x,\alpha) \in \operatorname{epi}\phi$ then for any $\beta > \alpha \geq \phi(x) = \inf_{y\in Y} f(x,y)$ we have the existence of y such that $\beta \geq f(x,y)$. It follows that $(x,y,\beta) \in \operatorname{epi} f$ and so $(x,\beta) \in E$. □

When f is convex then $P(\operatorname{epi} f)$ is convex (as $\operatorname{epi} f$ is convex) and hence so is $\operatorname{epi}\phi$. Thus ϕ is a convex function but we don't know whether $\operatorname{epi}\phi$ is closed.

In Example 1.2.3, you can visualize the set $P(\text{epi}f)$ in the following way. Take epi$f$ plotted in 3-dimensions, where the domain of f is the xy-plane. Then project epi$f$ onto the xz-axis. As an example, let

$$f(x,y) = \begin{cases} e^y & \text{if } x > 0. \\ \infty & \text{otherwise} \end{cases} \quad \text{so that } P(\text{epi}f) = (0,\infty]\times(0,\infty] \text{ and } \text{epi}\phi = (0,\infty]\times[0,\infty].$$

1.2.4 Problem Set 2: Convexity Preserving Operations

Problem 1.2.4 *Consider the transformation $P(x,t) = \frac{x}{t}$ for $t > 0$. Show that $P(C)$ is convex when the set $C \subseteq X \times \mathbf{R}_{++}$ is convex.*
Hint: For any $\mu \in [0,1]$ and $s,t \geq 0$ with $s \neq t$, show that you can find θ such that

$$\mu = \frac{\theta t}{\theta t + (1-\theta)s} \in [0,1] \text{ and } \theta \in [0,1].$$

Problem 1.2.5 *The second-order cone is defined to be*

$$C := \left\{(x,t) \in \mathbf{R}^{n+1} \mid \|x\|_2^2 = x^T x \leq t^2\right\}. \tag{1.12}$$

1. *Show that the set given in (1.12) is a convex cone.*
2. *Hyperbolic cone is the set*

$$HC := \left\{x \mid x^T P x \leq \left(c^T x\right)^2,\ c^T x \geq 0\right\}$$

where $c \in \mathbf{R}^n$ and $P \in \mathcal{P}(n)$ as in Definition 1.1.6. Show that HC is convex.
[Hint: Consider the inverse image of the second order cone under the affine mapping $f(x) = \left(P^{\frac{1}{2}}x, c^T x\right)$, where P is the matrix such that $P^{\frac{1}{2}}P^{\frac{1}{2}} = P$. You may also find Lemma 1.1.8(i) and Lemma 1.1.7(iii) to be useful.]

Chapter 2

Inner Product Spaces

Some of the material of this section is included for completeness and should have been encountered in undergraduate courses. In the first part of this course we will consider the underlying "space" we are using is the Euclidean space. This is a finite dimensional space vector space $\mathbf{X}$ over the reals $\mathbf{R}$, equipped with an inner product $\langle \cdot, \cdot \rangle$. One example of this is the space $\mathbf{R}^n$ of real (column) n-vectors (with the standard inner product $\langle x, y \rangle = \sum_{i=1}^n x_i y_i = x^T y$). Any inner produce indices a "norm" (or a distance function); for example,

$$\langle x, x \rangle = \sum_{i=1}^n x_i x_i = x^T x = \sum_{i=1}^n x_i^2 = \|x\|^2 .$$

Given any inner product we may assign a norm by placing $\|x\| = \langle x, x \rangle^{\frac{1}{2}}$. By placing convex analysis in a more general inner product space we open up a number of possibilities for applications later on.

2.1 Basics of Inner Product Spaces

We consider a function $\langle \cdot, \cdot \rangle : \mathbf{X} \times \mathbf{X} \to \mathbf{R}$ to be an inner product on the real vector space X if it satisfies the following properties

1. We have $\langle x, x \rangle \geq 0$ for all $x \in \mathbf{X}$ and $\langle x, x \rangle = 0$ implies $x = 0$.
2. For all α, $\beta \in \mathbf{R}$ and $x, y.z \in \mathbf{X}$ we have
$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$$
3. For all $x, y \in \mathbf{X}$ we have
$$\langle x, y \rangle = \langle y, x \rangle.$$

The last assumption signifies we are using the field of real numbers only (complex Hilbert spaces exist as well). A common calculation that occurs is the following:

$$\begin{aligned} \langle \alpha x + \beta y, \alpha x + \beta y \rangle &= \alpha \langle \alpha x + \beta y, x \rangle + \beta \langle \alpha x + \beta y, y \rangle \\ &= \alpha \left(\alpha \langle x, x \rangle + \beta \langle y, x \rangle \right) + \beta \left(\alpha \langle x, y \rangle + \beta \langle y, y \rangle \right) \\ &= \alpha^2 \langle x, x \rangle + \alpha\beta \langle y, x \rangle + \beta\alpha \langle x, y \rangle + \beta^2 \langle y, y \rangle \end{aligned}$$

$$\begin{aligned} &= \alpha^2\langle x,x\rangle + 2\alpha\beta\langle y,x\rangle + \beta^2\langle y,y\rangle \\ &= \alpha^2\|x\|^2 + 2\alpha\beta\langle y,x\rangle + \beta^2\|y\|^2 . \end{aligned}$$

An important consequence of these assumption is the Cauchy-Schwarz inequality.

Proposition 2.1.1 *Let* $\mathbf{X}$ *be a vector space endowed with the inner product* $\langle\cdot,\cdot\rangle$*. Then for all* x*,* $y \in \mathbf{X}$

$$|\langle x,y\rangle| \le \|x\|\,\|y\| .$$

Proof. For arbitrary $x, y \in \mathbf{X}$ we have

$$\begin{aligned} 0 &\le \langle x+\lambda|\langle x,y\rangle|y, x+\lambda|\langle x,y\rangle|y\rangle \\ &= \langle x,x\rangle + 2\lambda\,|\langle x,y\rangle|^2 + \lambda^2\,|\langle x,y\rangle|^2\,\langle y,y\rangle \end{aligned}$$

which is a polynomial in λ. Since this quadratic (in λ) is nonnegative, it cannot have two distinct real roots. Therefore, its discriminant must be nonpositive:

$$\left(2\,|\langle x,y\rangle|^2\right)^2 - 4\,|\langle x,y\rangle|^2\,\langle y,y\rangle\langle x,x\rangle \le 0$$

and hence (even if $\langle x,y\rangle = 0$) we have

$$|\langle x,y\rangle|^4 \le |\langle x,y\rangle|^2\,\|y\|^2\,\|x\|^2 \quad\Longrightarrow\quad |\langle x,y\rangle| \le \|x\|\,\|y\| .$$

□

A consequence of these properties is that the norm induced by an inner product satisfies the following properties:

1. For all $x \in \mathbf{X}$ we have $\|x\| \ge 0$ with $\|x\| = 0$ if and only of $x = 0$.
2. For all $x, y \in \mathbf{X}$ we have

$$\|\alpha x + \beta y\| \le |\alpha|\,\|x\| + |\beta|\,\|y\| .$$

Proof. Clearly $\|x\| \ge 0$ since $\|x\| = \langle x,x\rangle^{\frac{1}{2}}$ and $\langle x,x\rangle \ge 0$. Also $\|x\| = 0$ if and only if $\|x\|^2 = \langle x,x\rangle = 0$ implying $x = 0$ (by part 1 of the definition). Finally note that when $x = 0$ we have $\alpha x = 0$ for all $\alpha \in \mathbf{R}$ and so $\langle 0,0\rangle = \langle \alpha x, \alpha x\rangle = \alpha^2\langle x,x\rangle = \alpha^2\,\langle 0,0\rangle$. That is $(1-\alpha)\langle 0,0\rangle = 0$ implying $\langle 0,0\rangle = \|0\|^2 = 0$. Using the (bi-) linearity of inner product we have

$$\begin{aligned} \|\alpha x + \beta y\|^2 &= \langle \alpha x + \beta y, \alpha x + \beta y\rangle \\ &= \alpha^2\|x\|^2 + 2\alpha\beta\langle x,y\rangle + \beta^2\|y\|^2 \\ &\le \alpha^2\|x\|^2 + 2\,|\alpha|\,|\beta|\,\|x\|\,\|y\| + \beta^2\|y\|^2 \\ &= \left(|\alpha|\,\|x\| + |\beta|\,\|y\|\right)^2 \end{aligned}$$

and so

$$\|\alpha x + \beta y\| \le |\alpha|\,\|x\| + |\beta|\,\|y\|$$

□

Suppose $\{u_i\}_{i=1}^n$ are an orthonormal set in an inner product space (i.e. the vectors are orthogonal and of unit length). An identity that is often used is $u = \sum \langle u, u_i \rangle u_i$. To see this consider

$$w := \left(u - \sum_{i=1}^{n} \langle u, u_i \rangle u_i \right).$$

Note that for all j we have

$$\begin{aligned} \langle w, u_j \rangle &= \langle u, u_j \rangle - \sum_{i=1}^{n} \langle u, u_i \rangle \langle u_i, u_j \rangle = \langle u, u_j \rangle - \sum_{i=1}^{n} \langle u, u_i \rangle \delta_{ij} \\ &= \langle u, u_j \rangle - \langle u, u_j \rangle = 0, \end{aligned}$$

where the Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Now let $x = \sum_{i=1}^n \alpha_i u_i$ be an arbitrary vector in $\mathbf{R}^n$. Then

$$\langle w, x \rangle = \left\langle w, \sum_{i=1}^{n} \alpha_i u_i \right\rangle = \sum_{i=1}^{n} \alpha_i \langle w, u_i \rangle = 0$$

In particular, I can pick x to be w, and so I get

$$\|w\|^2 = \langle w, w \rangle = 0 \quad \Longrightarrow \quad w = 0 \quad \text{or} \quad u = \sum_{i=1}^{n} \langle u, u_i \rangle u_i.$$

From this identity follows

$$\begin{aligned} \|u\|^2 &= \langle \sum_{i=1}^{n} \langle u, u_i \rangle u_i, \sum_{j=1}^{n} \langle u, u_j \rangle u_j \rangle \\ &= \sum_{i,j=1}^{n} \langle u, u_i \rangle \langle u, u_j \rangle \langle u_i, u_j \rangle = \sum_{i,j=1}^{n} \langle u, u_i \rangle \langle u, u_j \rangle \delta_{ij} = \sum_{i=1}^{n} |\langle u, u_i \rangle|^2 . \end{aligned}$$

The space $\mathbf{R}^n$ is not the only example of inner product spaces.

Example 2.1.1 ***The symmetric matrices:*** *Denote by $\mathcal{S}(n)$ the $n \times n$ real, symmetric matrices and by $\mathcal{P}(n)$ the set of all positive semi–definite, real symmetric matrices i.e.*

$$\mathcal{P}(n) = \left\{ Q \in \mathcal{S}(n) \mid x^T Q x \geq 0 \text{ for all } x \in \mathbf{R}^n \right\}.$$

We will show later that all Q in $\mathcal{S}(n)$ are finitely decomposable and so $\mathcal{S}(n)$ is a finite dimensional space but its dimension is not n but actually $\frac{1}{2} n (n+1)$. To make it a linear space, define the trace of a matrix to be the sum of the diagonal elements i.e. $\operatorname{tr} Q = \sum_{j=1}^n q_{jj}$ for $Q = [q_{ij}] \in \mathcal{S}(n)$. We now show $\mathcal{S}(n)$ is a vector space with an inner product $\langle A, B \rangle = \operatorname{tr} B^T A (= \operatorname{tr} BA)$ via number of Lemmas and remarks. We define for $A, B \in \mathcal{S}(n)$ the sum $A + B = \{a_{ij} + b_{ij}\}$ (i.e. component wise sum). Multiplication by a scalar is similarly component wise i.e. $\alpha A = \{\alpha a_{ij}\}$. Clearly this defines a vector space (essentially isomorphic to a subspace of $\mathbf{R} \times \mathbf{R} = \mathbf{R}^{2n}$).

Lemma 2.1.2 *Show that* $\operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B$

Proof. *Let* $A = \{a_{ij}\}$ *and* $B = \{b_{ij}\}$ *then* $A + B = \{a_{ij} + b_{ij}\}$ *and so* $\operatorname{tr}(A+B) = \sum a_{ii} + b_{ii} = \sum a_{ii} + \sum b_{ii} = \operatorname{tr} A + \operatorname{tr} B$. □

Lemma 2.1.3 *For* $A \in \mathcal{S}(n)$ *we have* $\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i$ *where* λ_i *are the eigenvalues of* A.

Proof. We show that $\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i$ where λ_i are the eigenvalues of A. Let T be the matrix that diagonalizes A. The eigenvalues of A are the roots of the characteristic polynomial $p(\mu) = \det(A - \mu I)$. Recall the product rule of determinants [23, Chapter 4, rule 9]. Then the characteristic polynomial $p(\mu)$ satisfies

$$\begin{aligned} p(\mu) &= \det(A - \mu I) = \\ &= \overbrace{\det(I)}^{=1} \det(A - \mu I) \\ &= \det\left(T^{-1}T\right) \det(A - \mu I) \\ &= \det(T^{-1}) \det(A - \mu I) \det(T) \quad \text{(product rule of determinants)} \\ &= \det(T^{-1}(A - \mu I)T) \quad \text{(product rule of determinants again)} \\ &= \det\left(D - \mu I\right), \end{aligned}$$

where $D = T^{-1}AT$ is the diagonal matrix diagλ whose entries $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A (see Remark 1.1.7). Expanding

$$\begin{aligned} \det(A - \mu I) &= \begin{vmatrix} a_{1,1} - \mu & \ldots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \ldots & a_{n,n} - \mu \end{vmatrix} \\ &= (a_{1,1} - \mu)(\ldots)(a_{n,n} - \mu) + \ldots \\ &= (-\mu)^n + (-\mu)^{n-1} \underbrace{(a_{1,1} + \cdots + a_{n,n})}_{\operatorname{tr} A} + \ldots \end{aligned}$$

and collecting only powers in $(-\mu)^{n-1}$ we get the coefficient term $\sum a_{ii} = \operatorname{tr} A$. On the other hand, if we expand $\det(D - \mu I)$ as

$$\begin{aligned} \det(D - \mu I) &= \begin{vmatrix} \lambda_1 - \mu & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n - \mu \end{vmatrix} \\ &= (\lambda_1 - \mu)(\ldots)(\lambda_n - \mu) + \ldots \\ &= (-\mu)^n + (-\mu)^{n-1}(\lambda_1 + \cdots + \lambda_n) + \ldots \end{aligned}$$

then we get that the coefficient of term $(-\mu)^{n-1}$ is $\sum \lambda_i$. Because these are exactly equivalent ways of writing the polynomial $p(\mu)$, the coefficients must be the same, and so we know that $\operatorname{tr} A = \sum a_{ii} = \sum \lambda_i$. □

Lemma 2.1.4 *The form* $\langle A, B \rangle = \operatorname{tr} B^T A$ *(*$= \operatorname{tr} BA$*) is an inner product on the linear space* $\mathcal{S}(n)$.

Proof. The eigenvalues of A^2 are λ_i^2. Thus $\langle A, A\rangle = \|A\|^2 = \operatorname{tr} A^2 = \sum \lambda_i^2 = 0$ if and only if all $\lambda_i = 0$ implying $A = 0$. Clearly $\langle A, A\rangle = \operatorname{tr} A^2 = \sum_{j=1}^n \lambda_i^2 \geq 0$ and also

$$\begin{aligned}\langle \alpha A + \beta B, C\rangle &= \operatorname{tr} C(\alpha A + \beta B) \\ &= \operatorname{tr}(\alpha CA + \beta CB)) = \alpha \operatorname{tr} CA + \beta \operatorname{tr} CB = \alpha\langle A, C\rangle + \beta\langle B, C\rangle\end{aligned}$$

where we have used the fact that $\operatorname{tr} \alpha A = \sum \alpha a_{ii} = \alpha \sum a_{ii}$. Finally it is clear that $\operatorname{tr} Q = \operatorname{tr} Q^T$ for *any* Q, and so

$$\langle A, B\rangle = \operatorname{tr} \overbrace{BA}^{Q} = \operatorname{tr}(BA)^T = \operatorname{tr} A^T B^T = \operatorname{tr} AB = \langle B, A\rangle$$

where we have used the facts that as A and B are symmetric $A^T B^T = (BA)^T$ along with $A^T = A$ and $B^T = B$. □

An alternative way of writing this inner product is to use the vec operation to form vectors from the matrices and then place $\langle A, B\rangle = \langle \operatorname{vec} A, \operatorname{vec} B\rangle_2$. Here the vec operation stacks columns on top of each other e.g.

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,n} \end{pmatrix} \quad \Longleftrightarrow \quad \operatorname{vec} A = \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{n,1} \\ \vdots \\ a_{1,n} \\ \vdots \\ a_{n,n} \end{pmatrix}$$

and $\langle \operatorname{vec} A, \operatorname{vec} B\rangle_2$ denotes the usual "dot" product of vectors in $\mathbf{R}^{n^2}$. *Now* $AB = \{c_{ij}\} = \{\sum_{k=1}^n a_{ik} b_{kj}\}$. Taking the trace and using the fact that $b_{ij} = b_{ji}$ (due to symmetry),

$$\begin{aligned}\operatorname{tr}(AB) = \sum_{i=1}^n c_{ii} = \sum_{i=1}^n \sum_{k=1}^n a_{ik} b_{ki} &= \sum_{k=1}^n \sum_{i=1}^n a_{ik} b_{ik} \\ &= \operatorname{vec} A \cdot \operatorname{vec} B\end{aligned}$$

as required. In particular (where F in $\|\cdot\|_F$ stands for Frobenius),

$$\|A\|_F^2 = \operatorname{tr}(AA) = \operatorname{vec} A \cdot \operatorname{vec} A = \sum_{i,j=1}^n a_{ij}^2.$$

This provides an easy way to calculate this inner product and the associated norm $\|A\|_F$.

Here is a useful identity for the Frobenius inner product.

Lemma 2.1.5 *For $A \in \mathcal{S}(n)$ we have $\langle A, uu^T\rangle = \langle Au, u\rangle = u^T Au$ where uu^T is an $n \times n$ rank one matrix.*

Proof. Now

$$uu^T = \begin{pmatrix} u_1u_1 & u_1u_2 & \dots & u_1u_n \\ u_2u_1 & u_2u_2 & & u_2u_n \\ \vdots & & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \dots & u_nu_n \end{pmatrix}$$

and so using the last part

$$\langle A, uu^T \rangle = \operatorname{tr}(A(uu)^T) = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik} \underbrace{(uu^T)_{ki}}_{u_k u_i} = \sum_{i=1}^{n} \left(\sum_{k=1}^{n} a_{ik} u_k \right) u_i = u^T A u.$$

□

Remark 2.1.1 *The norm $\|A\|_2 = \max_{\|x\|=1} \|Ax\|_2$ is the norm 'induced by" the Euclidean norm $\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}$. Given any norm $\|\cdot\|_1$ on $\mathbf{R}^n$ and norm $\|\cdot\|_2$ on $\mathbf{R}^m$ along with $A : \mathbf{R}^n \to \mathbf{R}^m$ then the induced norm of A is*

$$\|A\|_{1,2} = \max \{ \|Ax\|_2 \mid \|x\|_1 = 1 \}.$$

Norms can exist without an inner product existing, so let's review the basic definition of a norm.

Definition 2.1.1 *A mapping $\|\cdot\| : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ where ($\mathbf{X}$ is a linear space over the real scalars) is a norm if and only if*

1. *$\|x + y\| \leq \|x\| + \|y\|$ for all x, $y \in X$,*
2. *$\|\alpha x\| = |\alpha| \, \|x\|$ for all $x \in X$ and $\alpha \in \mathbb{R}$,*
3. *$\|x\| \geq 0$ for all $x \in X$ and*
4. *$\|x\| = 0$ if and only if $x = 0$.*

Remark 2.1.2 *The norm is used to define strong convergence on the space. We say that*

1. *x_n converges $x_n \to x$ (strongly) if and only if $\|x_n - x\| \to 0$ as $n \to \infty$;*
2. *A set C is said to be (strongly) closed if and only if $\{x_n\} \subseteq C$ and $x_n \to x$ implies $x \in C$;*
3. *$B_\delta(\bar{x}) := \{y \in X \mid \|y - \bar{x}\| < \delta\}$ is an open ball of radius δ about y;*
4. *We say a point $x \in C$ is an interior point of C if and only if there exists $\delta > 0$ such that $B_\delta(x) \subseteq C$;*
5. *$\operatorname{int} C := \{x \in C \mid x$ is a interior point of $C\}$;*
6. *C is an open set if all points of C are interior points (i.e. $\operatorname{int} C = C$).*

2.1.1 Problem Set 3: Matrix norms and spectral radius

Problem 2.1.6 *Let $A \in \mathcal{S}(n)$. Define the Frobenius norm as $\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\operatorname{tra} A^2}$, $\rho(A) = \max\{|\lambda_i| \mid \{\lambda_i\}_{i=1}^n$ are the eigenvalues of $A\}$ (the "spectral radius" of A) and denote the induced 2-norm (aka Schatten ∞-norm, aka Spectral norm) by $\|A\|_2 = \max_{\|x\|=1} \|Ax\|_2$.*

1. *Show that*

$$\left(\max_{i=1,\ldots,n} \lambda_i\right) = \max_{\|u\|=1} \langle A, uu^T\rangle \leq \max_{\|u\|=1} \left|u^T A u\right| \leq \|A\|_2$$
$$= \rho(A) \leq \|A\|_F = \sqrt{\sum_{i=1}^n \lambda_i^2} \leq \sqrt{n}\rho(A) = \sqrt{n}\|A\|_2.$$

2. *Explain why $A = 0$ iff*

$$0 = \max_{\|u\|=1} \left|u^t A u\right| = \|A\|_F = \rho(A) = \|A\|_2 = 0.$$

[Hint: First consider $\langle A, u_i u_i^T\rangle$ where u_i is the ith unit eigenvector. Next consider representing an arbitrary vector u via $u = \sum\langle u, u_i\rangle u_i$ where $\{u_i\}_{i=1}^n$ are the orthogonal set of eigenvectors in $\langle A, uu^T\rangle$. Next use the identities $\|A\|_2^2 = \left(\max_{\|u\|=1} \|Au\|\right)^2 = \max_{\|u\|=1} \|Au\|^2 = \max_{\|u\|=1} (Au)^T Au = \max_{\|u\|=1}\langle A^2, uu^T\rangle$.]

2.2 Spectral Decomposition and Adjoint Operators

Spectral decomposition

Recall from Remark 1.1.7 that any real symmetric matrix has n real eigenvalues $\{\lambda_i\}_{i=1}^n$ and n orthonormal eigenvectors $\{v_i\}_{i=1}^n$. This allow us to represent any $Q \in \mathcal{S}(n)$ in the spectral form

$$Q = Q_{\text{spec}} := \sum_{i=1}^n \lambda_i v_i v_i^T. \tag{2.1}$$

Here $v_i v_i^T$ is the so called outer product (or $v_i v_i^T$ is a rank–1 matrix). We can prove this representation rather compactly. First note that $\left(xx^T\right) x = x\left(x^T x\right) = x\|x\|^2 = x$ (taking x to be of unit length). We will prove the decomposition (2.1) by showing that $Q - Q_{\text{spec}} = 0$. First notice

$$Q_{\text{spec}}\, v_j = \sum_{i=1}^n \lambda_i v_i \left(v_i^T v_j\right) = \sum_{i=1}^n \lambda_i v_i \langle v_i, v_j\rangle = \sum_{i=1}^n \lambda_i v_i \overbrace{\delta_{ij}}^{\text{Kronecker}} = \lambda_j v_j. \tag{2.2}$$

Thus we have that $Q_{\text{spec}}\, v_j = \lambda_j v_j$ for all $j = 1, \ldots, n$. As $\{v_j\}_{j=1}^n$ are orthonormal for any $x \in X$ we have some $\{\alpha_j\}_{j=1}^n$ such that $x = \sum_{j=1}^n \alpha_j v_j$ and hence

$$\left(Q_{\text{spec}} - Q\right) x = \sum_{j=1}^n \alpha_j Q_{\text{spec}} v_j - \sum_{i=1}^n \alpha_i Q v_i$$

$$= \sum_{j=1}^{n} \alpha_j \underbrace{\lambda_j v_j}_{(2.2)} - \sum_{j=1}^{n} \alpha_j \underbrace{\lambda_j v_j}_{(a)} = 0.$$

Here (a) is just the definition of v_j being an eigenvector with corresponding eigenvalue λ_j. In summary $(Q_{\text{spec}} - Q)\, x = 0$ for all $x \in X$, which shows the representation (2.1). Consequently

$$\max_{\|u\|=1} \left| x^T (Q_{\text{spec}} - Q)\, x \right| = 0 \quad \Longleftrightarrow \quad Q_{\text{spec}} - Q = 0 \quad \text{or} \quad Q = Q_{\text{spec}} = \sum_{i=1}^{n} \lambda_i v_i v_i^T.$$

Application to the square root matrix

The spectral decomposition provides an avenue to define the square root of a matrix $Q \in \mathcal{P}(n)$ e.g.

$$Q^{1/2} := \sum_{i=1}^{n} \sqrt{\lambda_i} v_i v_i^T. \tag{2.3}$$

We have

$$\begin{aligned} Q = Q^{1/2} Q^{1/2} &= \left(\sum_{i=1}^{n} \sqrt{\lambda_i} v_i v_i^T \right) \left(\sum_{j=1}^{n} \sqrt{\lambda_j} v_j v_j^T \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{\lambda_i} \sqrt{\lambda_j} v_i \left(\underbrace{v_i^T v_j}_{(a)} \right) v_j^T = \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{\lambda_i} \sqrt{\lambda_j} v_i v_j^T \delta_{ij} \\ &= \sum_{i=1}^{n} \lambda_i v_i v_i^T = Q. \end{aligned}$$

Here (a) is 1 if $i = j$ and 0 otherwise, and this is why we may simplify with the Kronecker delta in the following term.

Remember that the gradient and Hessian of a convex function can be used to build first and second order approximations for the function. For finding such approximations, we will want to use the Taylor series applied to the eigenvalues of the spectral decomposition. We will see an example of this later on.

Norms and functions of matrices

Indeed we may use any Taylor series for a given function to generate a related function of the matrix. Consider any one of the following:

$$\begin{aligned} \ln(1+x) &= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \frac{1}{5}x^5 + O\left(x^6\right) \quad \text{if } |x| < 1, \\ \exp(x) &= 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + O\left(x^5\right) \quad \text{for all } x \text{ and} \\ (1+x)^m &= 1 + mx - \frac{1}{2}x^2 m(m-1) + \frac{1}{3!}x^3 m(m-1)(m-2) + O\left(x^4\right) \quad \text{if } |x| < 1 \end{aligned}$$

In functional analysis one finds the concept of bounded linear operator (matrix in our case). Suppose $A : \mathbf{R}^n \mapsto \mathbf{R}^m$ is a bounded operator i.e.

$$\|A\| := \sup_x \frac{\|Ax\|}{\|x\|} = \sup_x \|A\frac{x}{\|x\|}\| = \sup_{\|x\|=1} \|Ax\| \leq M < \infty. \tag{2.4}$$

Here $\|x\|$ correspond to the natural Euclidean norm but could be any other norm. The formula for $\|A\|$ is dependent on the vector norm used.

Consider now ABx. We note that sub-multiplicativity is immediate

$$\begin{aligned}\|AB\| :=& \sup_x \frac{\|ABx\|}{\|x\|} = \sup_x \frac{\|A(Bx)\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \\ &\leq \sup_x \frac{\|A(Bx)\|}{\|Bx\|} \sup_x \frac{\|Bx\|}{\|x\|} \\ &\leq \sup_y \frac{\|Ay\|}{\|y\|} \sup_x \frac{\|Bx\|}{\|x\|} = \|A\|\|B\|.\end{aligned}$$

By induction we have $\|A^k\| = \|A^{k-1}A\| \leq \|A^{k-1}\|\|A\| \leq \|A\|^k$. It is easily see that when may use Taylor expansion to define convergent series involving an operator A as long as $\|A\| < \rho$ where ρ is the radius of convergence of the series. Indeed when

$$\sum_n a_n x^n \quad \text{converges absolutely for } |x| < \rho \quad \text{we have}$$

$$\left\| \sum_n a_n A^n \right\| \leq \sum_n |a_n| \, \|A^n\| \leq \sum_n |a_n| \, \|A\|^n < +\infty \quad \text{if } \|A\| < \rho.$$

For example we may place

$$\exp(At) = I + At + \frac{1}{2}A^2t^2 + \frac{1}{3!}A^3t^3 + \frac{1}{4!}A^4t^4 + \dots$$

$$\text{or } (1+At)^m = I + mAt - \frac{1}{2}A^2m(m-1)t^2 + \frac{1}{3!}A^3m(m-1)(m-2)t^3 + \dots \quad \text{if } \|A\| < 1/t.$$

Remember from Remark 1.1.7 that when A is real symmetric and diagonalisable by P we have $D = \text{diag}\{\lambda_i\} = P^TAP$ for an orthogonal P. So we can write

$$\begin{aligned}P^T \exp(At) P &= I + P^TAPt + \frac{1}{2}\left(P^TAP\right)\left(P^TAP\right)t^2 + \frac{1}{3!}\left(P^TAP\right)\left(P^TAP\right)\left(P^TAP\right)t^3 + \dots \\ &= I + Dt + \frac{1}{2}D^2t^2 + \frac{1}{3!}D^3t^3 + \dots \\ &= \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} + \begin{pmatrix} \lambda_1 t & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n t \end{pmatrix} + \begin{pmatrix} \lambda_1^2\frac{t^2}{2!} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n^2\frac{t^2}{2!} \end{pmatrix} + \dots \\ &= \begin{pmatrix} 1+\lambda_1 t + \frac{(\lambda_1 t)^2}{2!} + \frac{(\lambda_1 t)^3}{3!}\dots & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1+\lambda_n t + \frac{(\lambda_n t)^2}{2!} + \frac{(\lambda_n t)^3}{3!}\dots \end{pmatrix} \\ &= \begin{pmatrix} \exp(\lambda_1 t) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \exp(\lambda_n t) \end{pmatrix}\end{aligned}$$

$$\text{so} \quad \exp(At) = P \operatorname{diag}\{\exp(\lambda_i t)\} P^T. \tag{2.5}$$

The positive semidefinite cone is an "order cone," by which we mean that it allows us to write down a partial ordering of matrices. Once I can write such inequalities, I can use them to express constraints. This is the idea behind positive semidefinite programming problems. As an example of why this is of interest, we can often take very hard (e.g. combinatorial) problems and relax them into semidefinite programming problems, where the relaxations are strong.

Corollary 2.2.1 *The space $\mathcal{S}(n)$ is a (Euclidean) inner product space that contains an "order cone" $\mathcal{P}(n)$ which induces a partial order $A \geq_{\mathcal{P}(n)} B$ if and only if $A - B \in \mathcal{P}(n)$.*

We say $S \in \mathcal{P}(n)$ is positive definite if $x^T S x > 0$ for all $x \in \mathbf{R}^n$. This clearly implies the eigenvalues of S are all strictly positive (and so nonzero) and hence $\det S = \lambda_1 \lambda_2 \ldots \lambda_n \neq 0$ implying S is invertible. It may be shown that S is positive definite if and only if $S \in \operatorname{int} \mathcal{P}(n)$.

Example 2.2.1 *Prove that for any $X \in \mathcal{S}(n)$ we have $\left(X^2\right)^{1/2} \succeq X$.*

[Hint: By definition $\left(X^2\right)^{1/2}$ and X are diagonalised by the same orthogonal matrix].
Soln: Let v_i be an eigenvector of X and λ_i its corresponding eigenvector. Then $X^2 v_i = \lambda_i^2 v_i$ implying the spectral decomposition of X^2 is given by

$$X^2 = \sum_i \lambda_i^2 v_i v_i^T.$$

The square root is then given by

$$\left(X^2\right)^{\frac{1}{2}} = \sum_i \left(\lambda_i^2\right)^{\frac{1}{2}} v_i v_i^T = \sum_i |\lambda_i| \, v_i v_i^T \succeq \sum_i \lambda_i v_i v_i^T = X.$$

Example 2.2.2 *Find two matrices X, $Y \in \mathcal{P}(2)$ such that $X \succeq Y$ but $X^2 \not\succeq Y^2$.*

Soln: Let

$$X := \begin{pmatrix} 5 & .01 \\ .01 & 2 \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$$

then

$$X - Y = \begin{pmatrix} 1 & -0.99 \\ -0.99 & 1 \end{pmatrix}$$

for which the product of the eigenvalues is $\det(X - Y) = 1 - (0.99)^2 > 0$ and the sum of the eigenvalues is $\operatorname{tr}(X - Y) = 2 > 0$. As the product of the two eigenvalues is positive, they must both have the same sign; since their sum is also positive, that sign is $+$. Thus $X - Y \succeq 0$. Next note that

$$X^2 - Y^2 = \begin{pmatrix} 5 & .01 \\ .01 & 2 \end{pmatrix}^2 - \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}^2 = \begin{pmatrix} 5.000\,1 & -9.93 \\ -9.93 & -0.999\,9 \end{pmatrix}$$

for which $\det\left(X^2 - Y^2\right) < 0$ and so one of the eigenvalues of $X^2 - Y^2$ is negative. Thus $X^2 \not\succeq Y^2$.

Definition 2.2.1 *Consider the inner product spaces X_1 and X_2 with inner products $\langle\cdot,\cdot\rangle_i$, $i=1,2$ and a mapping $A:\mathbf{X}_1\to\mathbf{X}_2$.*

1. *The the adjoint operator $A^*:\mathbf{X}_2\to\mathbf{X}_1$ defined by*

$$\langle Ax,y\rangle=\langle x,A^*y\rangle,\ \textit{for all } x\in X_1,\ y\in X_2.$$

2. *An operator Q is called orthogonal if*

$$\langle Qx_1,Qx_2\rangle=\langle x_1,x_2\rangle\quad \textit{for all } x_i,\ x_2\in X_1.$$

Example 2.2.3 *When Q is orthogonal we have $Q^*Q=I$. To see why, notice that*

$$(\forall x_1,x_2)\quad \langle x_1,(Q^*Qx_2)\rangle-\langle x_1,x_2\rangle=\langle Qx_1,Qx_2\rangle-\langle x_1,x_2\rangle=\langle x_1,x_2\rangle-\langle x_1,x_2\rangle=0.$$

Since it holds for any x_1, it holds for $x_1:=(Q^*Qx_2)-x_2$, and so

$$\begin{aligned}0&=\langle x_1,(Q^*Qx_2)\rangle-\langle x_1,x_2\rangle\quad \textit{(above equality)}\\&=\langle x_1,(Q^*Qx_2)-x_2\rangle\\&=\langle (Q^*Qx_2)-x_2,(Q^*Qx_2)-x_2\rangle=\|(Q^*Qx_2)-x_2\|^2,\end{aligned}$$

and so $Q^*Qx_2=x_2$. Since it is true for any x_2, we must have $Q^*Q=I$.
We also note that when $\mathbf{X}_1=\mathbf{R}^n$ and $\mathbf{X}_2=\mathbf{R}^m$ and $A:\mathbf{R}^n\to\mathbf{R}^m$ then $A^*=A^T$ since for all x and y we have

$$\langle Ax,y\rangle=y^TAx=\left(A^Ty\right)^Tx=\langle x,A^Ty\rangle.$$

Example 2.2.4 *For matrices $X,Y\in\mathcal{P}(n)$ show that $X\succeq Y$ implies $X^{\frac12}\succeq Y^{\frac12}$.*

[Hint: Try considering the relationship $\langle\left(X^{\frac12}+Y^{\frac12}\right)x,\left(X^{\frac12}-Y^{\frac12}\right)x\rangle=\langle(X-Y)\,x,x\rangle$ for eigenvectors x of $X^{\frac12}+Y^{\frac12}$.]
Soln: First we verify that

$$\begin{aligned}\left\langle\left(X^{\frac12}+Y^{\frac12}\right)x,\left(X^{\frac12}-Y^{\frac12}\right)x\right\rangle&=x^T\left(X^{\frac12}-Y^{\frac12}\right)\left(X^{\frac12}+Y^{\frac12}\right)x\\&=x^T\left(X-Y^{\frac12}X^{\frac12}+X^{\frac12}Y^{\frac12}-Y\right)x\\&=x^T(X-Y)\,x+x^T\left(X^{\frac12}Y^{\frac12}-Y^{\frac12}X^{\frac12}\right)x\\&=x^T(X-Y)\,x=\langle(X-Y)\,x,x\rangle\end{aligned}\tag{2.6}$$

where the last equality follows because

$$x^TX^{\frac12}Y^{\frac12}x=\langle x,X^{\frac12}Y^{\frac12}x\rangle\overset{(a)}{=}\langle(X^{\frac12}Y^{\frac12})^Tx,x\rangle\overset{(b)}{=}\langle Y^{\frac12}X^{\frac12}x,x\rangle=x^TY^{\frac12}X^{\frac12}x.$$

Here (a) holds because the transpose is the adjoint. To see why (b) holds, notice that because $X,Y\in\mathcal{P}(n)$, we have by Lemma 1.1.8 that $X^{\frac12},Y^{\frac12}\in\mathcal{P}(n)$, and so clearly they are symmetric, whereafter (b) holds by applying Remark 1.1.7(iv) for the matrices $Y^{\frac12}X^{\frac12}$. An exactly equivalent (but shorter) argument is to use the fact that a real number is equal to its transpose:

$$x^TX^{\frac12}Y^{\frac12}x=\left(x^TX^{\frac12}Y^{\frac12}x\right)^T=(Y^{\frac12}x)^T(x^TX^{\frac12})^T=x^TY^{\frac12}X^{\frac12}x.$$

Now $\mathcal{P}(n)$ is a cone and $X^{\frac{1}{2}}, Y^{\frac{1}{2}} \in \mathcal{P}(n)$, we also have that $X^{\frac{1}{2}} + Y^{\frac{1}{2}} \in \mathcal{P}(n)$. Therefore, let x be an eigenvector of $\left(X^{\frac{1}{2}} + Y^{\frac{1}{2}}\right)$ and it will have eigenvalue $\lambda \geq 0$.

Case: If $\lambda = 0$, then

$$(X^{\frac{1}{2}} + Y^{\frac{1}{2}})x = 0, \text{ and so } X^{\frac{1}{2}}x = -Y^{\frac{1}{2}}x \text{ and so } (X^{\frac{1}{2}} - Y^{\frac{1}{2}})x = 2X^{\frac{1}{2}}x,$$

This means that

$$\langle x, (X^{\frac{1}{2}} - Y^{\frac{1}{2}})x\rangle = 2\langle x, X^{\frac{1}{2}}x\rangle \geq 0 \quad (\text{since } X^{\frac{1}{2}} \in P(n)).$$

Case: If $\lambda > 0$, then

$$0 \leq \langle (X - Y)\, x, x\rangle \overset{(2.6)}{=} \langle \left(X^{\frac{1}{2}} + Y^{\frac{1}{2}}\right) x, \left(X^{\frac{1}{2}} - Y^{\frac{1}{2}}\right) x\rangle \overset{(*)}{=} \lambda \langle x, \left(X^{\frac{1}{2}} - Y^{\frac{1}{2}}\right) x\rangle.$$

Here the first inequality follows from $X \succeq Y$, while $(*)$ holds because we chose x to be an eigenvector with eigenvalue λ. Because $\lambda > 0$, we may divide it away to obtain

$$\langle x, \left(X^{\frac{1}{2}} - Y^{\frac{1}{2}}\right) x\rangle \geq 0.$$

We have shown that this inequality holds for all eigenvectors x of the real symmetric matrix $X^{\frac{1}{2}} + Y^{\frac{1}{2}}$. As these eigenvectors span $\mathbf{R}^n$ it follows that $\langle x, \left(X^{\frac{1}{2}} - Y^{\frac{1}{2}}\right) x\rangle \geq 0$ for all x and so $X^{\frac{1}{2}} \succeq Y^{\frac{1}{2}}$.

Example 2.2.5 *When $\mathbf{X}_1 = \mathcal{S}(n)$ and $\mathbf{X}_2 = \mathbf{R}^n$ and $S_1, \ldots, S_m \in \mathcal{S}(n)$ then the mapping $A : \mathcal{S}(n) \to \mathbf{R}^n$ (or operator) defined by*

$$X \mapsto (\langle X, S_1\rangle, \ldots, \langle X, S_n\rangle)$$

has an adjoint

$$A^* y = \sum_i y_i S_i.$$

This may be seen by considering $X \in \mathcal{S}(n)$ and $y \in \mathbf{R}^n$

$$\begin{aligned} \langle AX, y\rangle_{\mathbb{R}^n} &= \sum_i y_i \langle X, S_i\rangle \\ &= \sum_i y_i \operatorname{tr} XS_i = \operatorname{tr} \sum_i y_i X S_i = \operatorname{tr} X \left(\sum_i y_i S_i\right) \\ &= \langle X, \sum_i y_i S_i\rangle = \langle X, A^* y\rangle_{\mathcal{S}(n)}. \end{aligned} \tag{2.7}$$

This illustrates that what the adjoint of A looks like depends on the inner products on the spaces.

2.2.1 Problem Set 4: Norms and Eigenvalues

Problem 2.2.2 *Consider a matrix $A \in \mathcal{P}(n)$ with $I \succeq A$ and the iteration*

$$Y_0 = 0, \quad Y_{n+1} = \frac{1}{2}\left(A + Y_n^2\right) \quad n = 0, 2 \ldots$$

1. *Prove that $\{Y_n\}$ is nondecreasing (that is $Y_n \preceq Y_{n+1}$ for all n) and converges to the matrix*
$$I - (I - A)^{1/2}.$$
2. *Suggest how this procedure may be modified to calculate $A^{1/2}$.*
3. *(Optional) Try it out by coding it in MATLAB, MAPLE or your favourite programming language and hence find the square root of*
$$A = \begin{pmatrix} 15 & 20 & 18 & 23 \\ 20 & 38 & 24 & 40 \\ 18 & 24 & 25 & 28 \\ 23 & 40 & 28 & 51 \end{pmatrix}$$
and verify your answer by using a in-built function.

Problem 2.2.3 *To show that the Frobenius norm is sub- multiplicative i.e. $\|XS\|_F \leq \|X\|_F \|S\|_F$ perform the following steps:*

1. *Let $\hat{X}$ be the ith row of X and $\hat{S}$ be the jth row of S. Write down $(\operatorname{vec}\hat{X} \cdot \operatorname{vec}\hat{S})^2$ explicitly, and also write down $\|\operatorname{vec}\hat{X}\|_F^2 \|\operatorname{vec}\hat{S}\|_F^2$ explicitly. Then use the Cauchy–Schwarz inequality to show that*
$$\left(\operatorname{vec}\hat{X} \cdot \operatorname{vec}\hat{S}\right)^2 \leq \left\|\operatorname{vec}\hat{X}\right\|_F^2 \left\|\operatorname{vec}\hat{S}\right\|_F^2. \tag{2.8}$$
2. *Noting that we can write*
$$\|XS\|_F^2 = \sum_i \sum_j \left(\sum_k x_{ik} s_{kj}\right)^2,$$
you can use what you wrote down in step 1 to establish that $\|XS\|_F^2 \leq \|X\|_F^2 \|S\|_F^2$.

Chapter 3

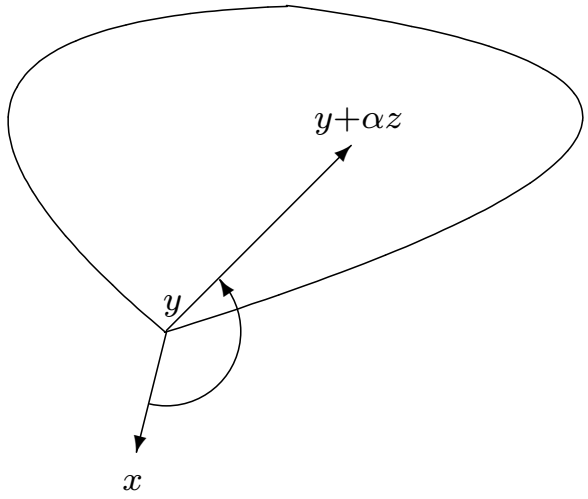
Supports, Separation and Subdifferentiability

3.1 Supports and Separation

Suppose C is a closed, convex set and $x \notin C$. The problem of finding the closest point in C to x is well defined for a convex set. This problem has many important application and may be formulated as:

$$\text{Find} \quad y \in \arg\min \{\|x - y\| \mid y \in C\}$$

where the argmin refers to the elements that are arguments of the minimized function value. When $\mathbf{X}$ is an inner product space the following classical analysis may be performed.



y is the closest point to x in the convex set C.

The distance form x to $y + \alpha z$ is larger than that to y. Thus $\|x - (y + \alpha z)\| \geq \|x - y\|$ and so

$$\|x - y\|^2 \leq \|x - (y + \alpha z)\|^2 = \|x - y\|^2 - \alpha\langle z, x - y\rangle + \alpha^2 \|z\|^2$$

implying

$$\alpha \left\|z^2\right\| \geq \langle z, x - y\rangle.$$

As this is true for all sufficiently small $\alpha \geq 0$ we obtain

$$0 \geq \langle z, x - y \rangle. \tag{3.1}$$

That is the angle between the vector $x - y$ and αz is obtuse. Placing $w := x - y$ we find that for any arbitrary point $h := y + \alpha z \in C$ we have

$$\begin{aligned} \langle w, h \rangle &= \langle x - y, y + \alpha z \rangle \\ &= \alpha \langle x - y, z \rangle + \langle x - y, y \rangle \\ &\leq \langle x - y, y \rangle = \langle w, y \rangle. \end{aligned}$$

Put this another way we have

$$y \in \arg\max \{ \langle w, h \rangle \mid h \in C \}.$$

That is y is a maximizing element of the following problem

$$\sup \{ \langle w, h \rangle \mid h \in C \}.$$

This prompts us to define the support function to C to be

$$v \mapsto S(C, v) := \sup \{ \langle v, h \rangle \mid h \in C \} \tag{3.2}$$

for which $S(C, w) = \langle w, y \rangle$ if we take $w = x - y$ (with y the closest point to x in C). In passing we note that for any v

$$S(C, v) - S(C, w) \geq \overbrace{\langle y, v \rangle}^{\leq S(C,v)} - \overbrace{\langle y, w \rangle}^{= S(C,w)} = \langle y, v - w \rangle. \tag{3.3}$$

Finally, when $x \notin C$ and C is closed we have $\|x - y\|^2 > 0$ and so

$$0 < \|x - y\|^2 = \langle x - y, x - y \rangle = \langle x - y, w \rangle = \langle x, w \rangle - \langle y, w \rangle$$

or

$$\langle x, w \rangle > \langle y, w \rangle = S(C, w) \geq \langle h, w \rangle \quad \text{for all } h \in C. \tag{3.4}$$

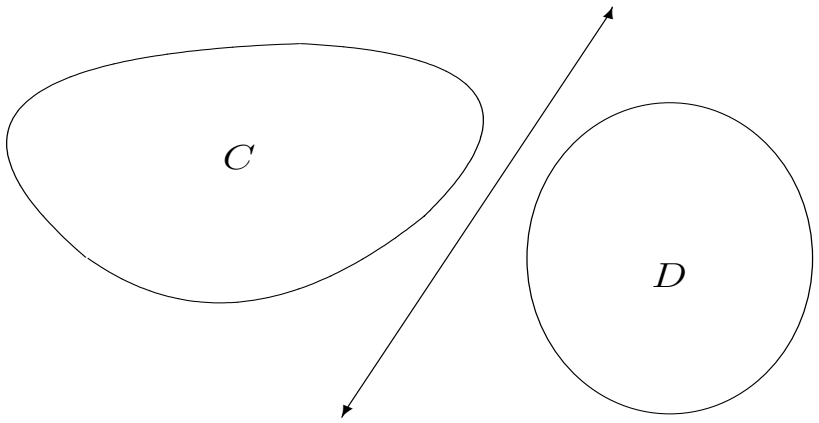
When such a w can be found we say that we have separated x from C. It can be shown that for two convex sets C and D with $\operatorname{int} D \neq \emptyset$ and

$$C \cap \operatorname{int} D = \emptyset$$

then there exists w such that

$$\langle c, w \rangle \geq \langle d, w \rangle \quad \text{for all } c \in C \text{ and } d \in D.$$

A similar result [6, Hahn–Banach Theorem 1.6] even holds in infinite dimensions (in real Banach spaces) and is usually referred to as strong separation of convex sets.



The two sets C and D are separated by a hyperplane.

The inner product is very important in the study of convexity since it can be used to generate a half-space via the linear mapping $x \mapsto \langle x, y\rangle$ i.e.

$$\begin{aligned} H &:= \{x \in \mathbf{X} \,|\langle x, y\rangle \geq \alpha\} \quad \text{and} \\ H' &:= \{x \in \mathbf{X} \,|\langle x, y\rangle \leq \alpha\} \quad \text{are a half space for each } y. \end{aligned}$$

We denote the smallest closed convex set containing C by $\overline{\text{co}}C$. Half spaces are always closed and convex. We have noted earlier in example 1.1.5 that half spaces are convex. To see that they are closed take a sequence $x_n \to x$ then when $x_n \in H'$ (say) for all n we have

$$\langle x, y\rangle = \langle x - x_n, y\rangle + \langle x_n, y\rangle \leq \overbrace{\|y\|\,\|x - x_n\|}^{\geq \langle x - x_n, y\rangle \text{ (C-S)}} + \overbrace{\alpha}^{\geq \langle x_n, y\rangle} \to_{n\to\infty} \alpha.$$

Consequently $x \in H'$ establishing closedness.

Lemma 3.1.1 *Suppose $C \subseteq \mathbf{X}$, an inner product space, then*

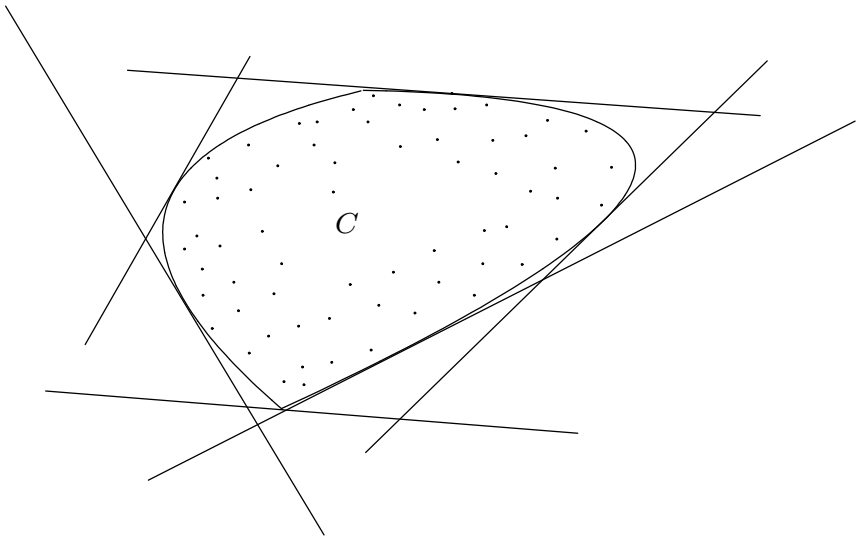
$$\overline{\text{co}}C = \bigcap \{H \mid C \subseteq H, \text{ and } H \text{ is a half-space}\}.$$

Proof. Let $\hat{C} := \bigcap \{H \mid C \subseteq H$, and H is a half-space$\}$. Clearly $C \subseteq \hat{C}$ and $\hat{C}$ is closed and convex. Thus $\overline{\text{co}}C \subseteq \hat{C}$ as $\overline{\text{co}}C$ is the smallest convex set containing C. Now suppose $\overline{\text{co}}C$ is a proper subset of $\hat{C}$. Then there exists $x \notin \overline{\text{co}}C$ with $x \in \hat{C}$. Then, using the half spaced defined by $w = x - y$, where y is the closest point to x in $\overline{\text{co}}C$, we find on using (3.4) that

$$\overline{\text{co}}C \subseteq H := \{z \in \mathbf{X} \,|\langle z, w\rangle \leq \alpha := \langle y, w\rangle\} \text{ but } x \notin H,$$

where $x \notin H$ is true because $\langle x, w\rangle > \alpha$. However, this cannot be true, as we have assumed $x \in \hat{C}$. Thus no such x can exist and $\overline{\text{co}}C \supseteq \hat{C}$. □

We say that y is supported by the hyperplane $H := \{x \in \mathbf{X} \mid \langle x, w\rangle = S(C, w)\}$ when $y \in C \cap H$.



A convex set is uniquely defined by its supporting hyperplanes.

A better way to write this relation is to use the support functional. For any $y \in \mathbf{X}$ we need to find an α so that $C \subseteq \{x \in \mathbf{X} \,|\langle x, y\rangle \leq \alpha\}$. The smallest such value would be

$$\alpha = S(C, y) := \sup_{x' \in C} \langle x', y\rangle \geq \langle x, y\rangle \quad \text{for all } x \in C.$$

Corollary 3.1.2 *Suppose $C \subseteq \mathbf{X}$, an inner product space, then*

$$\overline{\text{co}}C = \{x \in \mathbf{X} \,|\langle x, y\rangle \leq S(C, y)\text{, for all } y \in \mathbf{X}\}.$$

When we only have a finite number of points in C then $\overline{\text{co}}C$ is a polyhedral set. A convex polyhedral set is simply the intersection of a finite number of half spaces and so can be represented as:

$$\begin{aligned} P &:= \{x \in \mathbf{X} \,|\langle x, y_i\rangle \leq \alpha_i,\, i = 1, \ldots, m\} \\ &= \{x \in \mathbf{X} \,|Ax \leq b\} \end{aligned}$$

where A contains y_i as rows and b contains α_i as coordinates.

Example 3.1.1 *(Integer programming and Chvatal–Gomory cuts). Given a set*

$$C = \{x \mid Ax \leq b,\, x \in \mathbf{Z}^n\}$$

where $\mathbf{Z}$ denotes the integral points $\mathbf{Z} = \{\ldots, -1, 0, 1, 2, 3, \ldots\}$. How can we find $\overline{\text{co}}C$? This can be done (in principle) by the addition of valid inequalities (i.e. those that separate points not in $\text{co}\,C$ from the set $\overline{\text{co}}C$). Consider for example let C be defined by the inequalities

$$\begin{aligned} 7x_1 - 2x_2 &\leq 14 \\ x_2 &\leq 3 \\ 2x_1 - 2x_2 &\leq 3 \\ x \geq 0,\, x &\in \mathbf{Z}^2. \end{aligned}$$

Try first adding up the constraints with the weights $\left(\frac{2}{7}, \frac{37}{63}, 0\right)$ to obtain a valid constraint

$$\left(\frac{2}{7}\right)(7x_1 - 2x_2 - 14) + \left(\frac{37}{63}\right)(0x_1 + x_2 - 3) + 0(2x_1 - 2x_2 - 3) \leq 0,$$

which simplifies to

$$2x_1 + \frac{1}{63}x_2 \leq \frac{121}{21}.$$

Next note that the we may reduce the left hand side to the nearest integer value and obtain a weaker constraint (because of constraint $x \geq 0$). This gives

$$2x_1 + 0x_2 \leq \frac{121}{21} = 5 + \frac{16}{21}.$$

As the left hand side should take an integral value we may reduce the right hand side to an integral (integer) value without removing any integral points. Thus we obtain

$$2x_1 \leq 5$$

We could even continue this process by writing

$$x_1 \leq \frac{5}{2} \quad \Longrightarrow \quad x_1 \leq 2.$$

Then the convex set B defined by

$$7x_1 - 2x_2 \leq 14$$

$$\begin{aligned} x_2 &\leq 3 \\ 2x_1 - 2x_2 &\leq 3 \\ x_1 &\leq 2 \\ x &\geq 0 \end{aligned}$$

satisfies

$$B \supseteq \overline{\text{co}} C.$$

This is known as a Gomory cut. We have cut off a part of the original region defined by the original set of inequalities. It is a surprising but true fact that all valid inequalities can be generated finitely in this manor to obtain the convex hull of the allowable integer solutions. That is a finite (but possibly large) set of inequalities generated in the way defines $\overline{\text{co}} C$.

Theorems of the Alternative - Farkas Lemma

Farkas Lemma is about containment of a polyhedral set in a half space but can be formulated as an alternative. More precisely two logical propositions P and Q are said to form an alternative iff one and only one is true:

$$P \quad \Longleftrightarrow \quad \text{not } Q \quad [\text{or} \quad Q \quad \Longleftrightarrow \quad \text{not } P]$$

Farkas Lemma is really just about containment of convex cones.

Lemma 3.1.3 *Let* b, $y_1, \ldots, y_m$ *be given data in* $\mathbf{R}^n$. *Then the convex cone*

$$C := \text{co cone}\{y_1, \ldots, y_m\} = \left\{ \sum_{i=1}^{m} \lambda_i y_i \mid \lambda_i \geq 0,\ i = 1, \ldots, m \right\}$$

is closed.

Proof. Case 1: Consider the case when $\{y_1, \ldots, y_m\}$ are linearly independent. Then, when $y^k = \sum_{i=1}^m \lambda_i^k y_i \to y$, we know that y can expressed as $y = \sum_{i=1}^m \lambda_i y_i$ and $y^k - y = \sum_{i=1}^m \left(\lambda_i^k - \lambda_i\right) y_i \to 0$ by linear independence $\lambda_i^k - \lambda_i \to 0$. Consequently $\lambda_i \geq 0$ and $y \in C$.

Case 2: Now we deal with the case when $\{y_1, \ldots, y_m\}$ are not linearly independent. Suppose I take a linearly independent subset $\left\{y_{\alpha_1^1}, \ldots, y_{\alpha_l^1}\right\}$ and build a new cone $C_{\alpha^1} := \text{co cone}\left\{y_{\alpha_1^1}, \ldots, y_{\alpha_l^1}\right\}$. Then C_{α^1} is closed by our argument in case 1. Also, clearly $C_{\alpha^1} \subset C$. There are only finitely many (K many) ways to take a linearly independent combination of the y_i, and so the union of all such cones $\hat{C} := \cup_{k=1..K} C_{\alpha^k}$ is a closed cone. We will show that $\hat{C} = C$. Since $C_{\alpha^k} \subset C$ for all k, we clearly have $\hat{C} \subset C$. Therefore, we need only to show $\hat{C} \supset C$.

Therefore, take any $x \in C$, and we will show that it lives in one of the C_{α^k}. As the $\{y_1, \ldots, y_m\}$ are not linearly independent, we know that $\sum_{i=1}^m \beta_i y_i = 0$ has a nonzero solution, and we may suppose there is a j with $\beta_j < 0$ (take $-\beta$ if not, as $\sum_{i=1}^m (-\beta_i) y_i = 0$). Then note that if $x \in C$ then there exists $\lambda_i \geq 0$ and a $t^*(x) > 0$ such that

$$x = \sum_{i=1}^{m} \lambda_i y_i = \sum_{i=1}^{m} \left(\lambda_i + t^*(x)\, \beta_i\right) y_i \quad \text{since } t^*(x) \sum_{i=1}^{m} \beta_i y_i = 0$$

$$= \sum_{i \neq j(x)} \lambda_i' y_i \quad \text{where } \lambda_i' := \lambda_i + t^*(x)\,\beta_i \geq 0 \text{ where } j(x) \text{ is defined below.}$$

When $\beta_i \geq 0$ it is clear that $\lambda_i' \geq 0$ and if $\beta_i < 0$ in order that $\lambda_i' \geq 0$ we require

$$t^*(x) \leq \frac{-\lambda_i}{\beta_i} \quad \text{for all } i \text{ for which } \beta_i < 0.$$

If we choose $j(x) \in \arg\min_j \left\{ \frac{-\lambda j}{\beta_j} \mid \beta_j < 0 \right\}$ and $t^*(x) := \frac{-\lambda_{j(x)}}{\beta_{j(x)}}$ then $\lambda_i' \geq 0$ for all i and $\lambda_{j(x)}' = 0$. In effect we have removed one element $y_{j(x)}$ from this sum representing x. If the resultant collection $\{y_i\}_{i \neq j(x)}$ is still linearly dependent we can repeat this process and remove another y_i from the collection. Eventually we arrive at a linearly independent set of the y_i that allows us to represent x with a nonnegative combination λ. Thus x is a member of one of the C_{α^k}. □

Theorem 3.1.4 *Let b, $y_1, \ldots, y_m$ be given data in $\mathbf{R}^n$. Then the set*

$$\{x \in \mathbf{R}^n \mid \langle y_i, x\rangle \leq 0 \quad \text{for } i = 1, \ldots, m\} \subset \{x \in \mathbf{R}^n \mid \langle b, x\rangle \leq 0\} \tag{3.5}$$

if and only if

$$b \in C := \operatorname{co}\operatorname{cone}\{y_1, \ldots, y_m\}. \tag{3.6}$$

Proof. (3.6) $\Longrightarrow$ (3.5): Let (3.6) hold. Then take any

$$z \in \{x \in \mathbf{R}^n \mid \langle y_i, x\rangle \leq 0 \quad \text{for } i = 1, \ldots, m\},$$

and it holds that

$$\langle b, z\rangle = \langle \overbrace{\sum_{i=1}^{m} \lambda_i y_i}^{=b}, z\rangle = \sum_{i=1}^{m} \lambda_i \langle y_i, z\rangle \leq 0.$$

Thus we have (3.5).

(3.5) $\Longrightarrow$ (3.6): We prove this by contrapositive argument. Let $b \notin C$, so that (3.6) does not hold. Then we may separate b from this *closed* convex cone. Thus there is a d such that $\langle d, y\rangle \leq \alpha$ for all $y \in C$ and with $\langle b, d\rangle > \alpha$ where $\alpha := S(C, d)$. Note now that as $0 \in C$ we have $\alpha \geq 0$. To see that $\alpha = 0$, suppose to the contrary that some $y \in C$ attains $\langle y, d\rangle > 0$. Then since $\lambda y \in C$ for any $\lambda \geq 0$, we can take λ to be large enough that $\langle \lambda y, d\rangle > \langle b, d\rangle$, which contradicts the separation. This shows that $\alpha = 0$. Altogether, we have shown that $\langle b, d\rangle > 0$ and $d \notin \{x \in \mathbf{R}^n \mid \langle b, x\rangle \leq 0\}$ but $\langle d, y_i\rangle \leq 0$ for all i or $d \in \{x \in \mathbf{R}^n \mid \langle y_i, x\rangle \leq 0 \text{ for } i = 1, \ldots, m\}$, and so d is a counterexample to (3.5). □

Another way to look at the containment $b \in \operatorname{co}\operatorname{cone}\{y_1, \ldots, y_m\}$ is to say that the system of equations

$$b = \sum_{i=1}^{m} \lambda_i y_i, \quad \lambda_i \geq 0,\ i = 1, \ldots, m \tag{3.7}$$

has a solution.

Lemma 3.1.5 ***(Farkas lemma)*** *Let b, $y_1, \ldots, y_m$ be given data in $\mathbf{R}^n$. Then exactly one of the following statements are true:*

(P) The system of equations (3.7) has a solution;

(Q) The system of inequalities:

$$\langle b, x\rangle > 0, \quad \langle y_i, x\rangle \leq 0 \text{ for } i = 1, \ldots, m \tag{3.8}$$

has a solution.

Proof. Observe that if (3.7) has a solution then $b \in \operatorname{co}\operatorname{cone}\{y_1, \ldots, y_m\}$ and so, by the Theorem 3.1.4, we know that (3.8) has no solution. On the other hand, when $b \notin \operatorname{co}\operatorname{cone}\{y_1, \ldots, y_m\}$ then Theorem 3.1.4 establishes that (3.8) has a solution. □

It is useful to observe how these observations relate to the notion of consistency of linear equations.

Let $A : \mathbf{X} \to \mathbf{Y}$ be a linear mapping. Recall the *kernel* or *nullspace* of A is $\ker A = \{x \in \mathbf{X} | Ax = 0\}$ which is a subspace of the space $\mathbf{X}$ and the *range* of A is $\operatorname{Range} A = \{Ax | x \in \mathbf{X}\}$, a subspace of $\mathbf{Y}$. As before we denote the annihilator of a set $S \subseteq X$ by $S^\perp := \{x^* \in X^* : \langle x^*, y\rangle = 0, \forall y \in S\} \subseteq X^*$ and the annihilator of a set $D \subseteq \mathbf{X}^*$ by ${}^\perp D := \{x \in X : \langle x, y^*\rangle = 0, \forall y^* \in D\}$. People often informally say "perp" instead of annihilator. Recall that adjoint of A is a linear mapping $A^* : \mathbf{Y}^* \to \mathbf{X}^*$ where $\mathbf{Y}^*$ is the space with the bilinear form $\langle \cdot, \cdot\rangle_Y$ and $\mathbf{X}^*$ is the space with the bilinear form $\langle \cdot, \cdot\rangle_X$.

Proposition 3.1.6 *Let $A : \mathbf{X} \to \mathbf{Y}$ be a linear mapping. Then $\ker A = {}^\perp(\operatorname{Range} A^*)$ and $(\ker A)^\perp = \operatorname{Range} A^*$.*

Proof. We first show that ${}^\perp(\operatorname{Range} A^*) \supseteq \ker A$. Let $x^* \in (\operatorname{Range} A^*)$. Then $x^* = A^* y^*$ for some y^*. Let $x \in \ker A$ (i.e. $Ax = 0$). Then $\langle x, x^*\rangle = \langle x, A^* y^*\rangle = \langle Ax, y^*\rangle = 0$, as $Ax = 0$. Hence $x \perp x^*$ and so ${}^\perp(\operatorname{Range} A^*) \supseteq \ker A$. Now we must show that $\ker A \supseteq {}^\perp(\operatorname{Range} A^*)$. Let $y \in {}^\perp(\operatorname{Range} A^*)$ so, $\forall x^* \in \mathbf{X}^*$, $0 = \langle y, A^* x^*\rangle = \langle Ay, x^*\rangle$. As this is true $\forall x^* \in \mathbf{X}^*$ it follows that $Ay = 0$. Hence $y \in \ker A$ which gives the result. □

Exercise 3.1.1 *Prove the second assertion $(\ker A)^\perp = \operatorname{Range} A^*$ yourself. Note that in $\mathbf{X} = \mathbf{R}^n$ then $A^* = A^T$ and so $\left(\ker A^T\right)^\perp = \operatorname{Range} A$.*

Exercise 3.1.2 *Show that:*

1. *For any set C, $(C^\perp)^\perp \supset C$.*
2. *If $b \in C^\perp$ then $\{b\}^\perp \supset C$.*

A system of equations $Ax = b$ is consistent iff

$$b \in \operatorname{Range} A = \left(\ker A^T\right)^\perp$$
$$\text{or} \quad \{b\}^\perp \supseteq \ker A^T \quad \text{or} \quad \left\{x \in \mathbf{R}^n \mid A^T x = 0\right\} \subseteq \left\{x \in \mathbf{R}^n \mid b^T x = 0\right\}.$$

Stating this in terms of systems of equations we have (using $A = [y_1, \ldots, y_m]$)

$$b \in \operatorname{span}\{y_1, \ldots, y_m\}$$
$$\text{iff} \quad b^T x = 0 \quad \text{whenever} \quad \langle y_i, x\rangle = 0,\ i = 1, \ldots, m.$$

3.1.1 Problem Set 5: Farkas Lemma and Fritz John condition

Problem 3.1.7 *Let $A : x \mapsto \langle a, x\rangle$ be a linear operator. Find $S(\ker A, \cdot)$.*
Hint: any y can be written as $y = y_a + y_{a^\perp}$ where $y_a =\in \operatorname{span}\{a\}$ and $y_{a^\perp} \in \{a\}^\perp$. Consider what we can say about $S(\ker A, y)$ when $y_{a^\perp} = 0$. Then consider what happens when $y_{a^\perp} \neq 0$.

Problem 3.1.8 *Consider the space $\mathbf{R}^n \times \mathbf{R}$ and an element $b := (\mathbf{0}, 1)$ and a convex cone*

$$\operatorname{cone}\operatorname{co}\{(y_i, 1) \mid i = 1, \ldots, m\}.$$

Apply Farkas Lemma to show that exactly one of the following systems has a solution

$$\text{find } \lambda \text{ such that } \sum_{i=1}^m \lambda_i y_i = 0,\ \sum_{i=1}^m \lambda_i = 1,\ \lambda_i \geq 0,\ i = 1, \ldots, m \quad \text{or}$$
$$\text{find } x \text{ such that } \langle y_i, x\rangle < 0, \text{ for } i = 1, \ldots, m.$$

This is Gordon's Theorem.

Problem 3.1.9 *Consider the optimisation problem*

$$\begin{aligned} &\inf \quad f(x) \\ &\text{subject to} \\ &\qquad g_i(x) \leq 0,\ i = 1, \ldots, m. \end{aligned}$$

Suppose all f, g_i, $i = 1, \ldots, m$ are differentiable functions. We say that $\bar{x}$ is stationary for this problem when

$$\max\{\langle \nabla f(\bar{x}), d\rangle, \langle \nabla g_i(\bar{x}), d\rangle \mid i \in I(\bar{x})\} \geq 0,$$

where $I(\bar{x}) = \{i \mid g_i(\bar{x}) = 0\}$. Use Gordon's Theorem to show that $\bar{x}$ is stationary iff there exist positive multipliers λ_0, λ_i, $i = 1, \ldots, m$, not all zero such that

$$\lambda_0 \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0.$$

This is known as the Fritz John optimality condition.

3.2 Finite Differentiable Convex Functions on a Euclidean Space

We will discuss the special case of differential convex functions on a Euclidean space in this section. As a Euclidean space is finite dimensional the usual calculus rules apply in this situation. The directional derivative of a function $f : \mathbf{X} \to \overline{\mathbf{R}}$ is given by

$$f'(\bar{x}, d) = \lim_{t \downarrow 0} \frac{1}{t}\left(f(\bar{x} + td) - f(\bar{x})\right) \quad \text{for all } d \in X.$$

A handy way of viewing the directional derivative is the following. Let $g_{\bar{x},d}(t) := f(\bar{x}+td) : \mathbf{R} \to \mathbf{R}$ then

$$\frac{d}{dt} g_{\bar{x},d}(t)|_{t=0} = g'_{\bar{x},d}(0) = \lim_{t \downarrow 0} \frac{1}{t} \left(f(\bar{x} + td) - f(\bar{x}) \right) = f'(\bar{x}, d).$$

We say a function $f : \mathbf{X} \to \mathbf{R}$ is (Gâteaux) differentiable if $d \mapsto f'(\bar{x}, d)$ is actually linear (i.e. $f'(\bar{x}, d) = \langle a, d \rangle$ for some $a \in \mathbf{X}'$). Then we say f is (Gâteaux) differentiable at $\bar{x}$ with (Gâteaux) derivative $\nabla f(\bar{x}) = a$. Higher derivatives may be defined in an iterative fashion. The second derivative is a bilinear form $(d_1, d_2) \mapsto f''(\bar{x}, d_1, d_2)$ etc.

A vector valued function $\nabla f(x) : \mathbf{X} \to \mathbf{X}'$ is the Fréchet derivative of a function $f : \mathbf{X} \to \mathbf{R}$ if

$$\lim_{\delta x \to 0} \frac{|f(x + \delta x) - f(x) - \langle \nabla f(x), \delta x \rangle|}{\|\delta x\|} = 0.$$

It is well known that if f is Gâteaux differentiable everywhere and the resultant gradient mapping $x \mapsto \nabla f(x)$ is continuous then f is Fréchet differentiable with Fréchet derivative $\nabla f(x)$ and $g'_{\bar{x},d}(0) = \langle \nabla f(\bar{x}), d \rangle$.

In the main we are not particularly interested in the bilinear form $(d_1, d_2) \mapsto f''(\bar{x}, d_1, d_2)$ but rather the quadratic form $d \mapsto f''(\bar{x}, d, d)$ which appears in the second order expansion

$$f(\bar{x}) + \langle \nabla f(\bar{x}), d \rangle + \frac{1}{2} f''(\bar{x}, d, d) + o\left(\|d\|^2 \right).$$

When $x \mapsto \nabla f(x)$ is continuous we say f is continuously differentiable and write $f \in C^1$ and when $f''(\bar{x}, d, d)$ is defined by a linear operator (i.e. $f''(\bar{x}, d, d) = \langle \nabla^2 f(\bar{x}) d, d \rangle$) we say $f \in C^2$. In this case we may relate $f''(\bar{x}, d, d)$ to the function $g_{\bar{x},d}(t)$ by noting that

$$g''_{\bar{x},d}(0) = \frac{d^2}{dt^2} g_{\bar{x},d}(t)|_{t=0} = f''(\bar{x}, d, d).$$

Next, let's study the "barrier function" for the cone of positive definite matrices. Remember that the determinant gives you the product of the eigenvalues, which must be strictly positive for a positive definite matrix. Thus, the logarithm of the determinant of a positive definite matrix is defined.

Example 3.2.1 *Consider the function $f(X) = -\ln \det(X) : \operatorname{int} \mathcal{P}(n) \to \mathbf{R}$. We claim that when X is positive definite, $\nabla f(X) = -X^{-1}$ and the Hessian is the linear operator on $\mathcal{S}(n)$ given by $Q \mapsto X^{-1}(Q) X^{-1}$.*

Soln: Let $\Delta X \in \mathcal{S}(n)$ and X be positive definite. As X is positive definite, it possesses a positive definite inverse[1]. Since the inverse is positive definite, it must have a positive definite square root, so we may write $X^{-1/2} := \sqrt{X^{-1}}$. Denote the eigenvalues of $(\Delta X) X^{-1}$ by $\lambda_1, \ldots, \lambda_n$ and the eigenvectors $v_1, \ldots, v_n$. Note that the vectors $X^{-1/2} v_i$ are the eigenvectors of $X^{-1/2} (\Delta X) X^{-1/2}$ with corresponding eigenvalues λ_i, since

$$\begin{aligned} \left[X^{-1/2} (\Delta X) X^{-1/2} \right] \left(X^{-1/2} v_i \right) &= X^{-1/2} \left(\Delta X X^{-1} \right) v_i \\ &= X^{-1/2} (\lambda_i v_i) \\ &= \lambda_i \left(X^{-1/2} v_i \right). \end{aligned}$$

[1] See, for example, `https://mathworld.wolfram.com/PositiveDefiniteMatrix.html`

Now it also holds that $X^{-1/2}(\Delta X)X^{-1/2}$ is symmetric (because if two matrices A, B are symmetric then $(ABA)^T = A^T(AB)^T = A^TB^TA^T = ABA$)[2]. Altogether, $X^{-1/2}(\Delta X)X^{-1/2}$ is symmetric and its eigenvalues are real. Therefore, if we look at the Frobenius-inducing inner product, we have:

$$\langle X^{-1}, \Delta X\rangle_F \overset{\text{Lemma 2.1.4}}{=} \operatorname{tr}(\Delta X)^T X^{-1} \overset{\text{(symmetry)}}{=} \operatorname{tr}(\Delta X)X^{-1} \overset{\text{Lemma 2.1.3}}{=} \sum_i \lambda_i. \quad (3.9)$$

Then as $-\ln\det(X) = \ln\left(\frac{1}{\det(X)}\right) = \ln\det\left(X^{-1}\right)$ (since $1 = \det I = \det\left(X^{-1}X\right) = \det X^{-1}\det X$ and so $\det X^{-1} = (\det X)^{-1}$) it follows that:

$$\begin{aligned} f(X+\Delta X) - f(X) &= -\ln\det(X+\Delta X) - (-\ln\det(X)) \quad \text{(by definition)} \\ &= -\left[\ln\det(X+\Delta X) + \ln\det\left(X^{-1}\right)\right] \quad \text{(by above equalities)} \\ &= -\ln(\det(X+\Delta X)(\det X^{-1})) \quad \text{(log sum property)} \\ &= -\ln\det\left(I + (\Delta X)X^{-1}\right) \quad \text{(determinants product rule)} \\ &= -\ln\Pi_i(1+\lambda_i) \quad \text{(product of eigenvalues rule)} \\ &= -\sum_i \ln(1+\lambda_i). \quad \text{(log sum property)} \end{aligned} \quad (3.10)$$

The sub-multiplicitivity of the Frobenius norm (Problem 2.2.3) gives

$$\begin{aligned} \left\|(\Delta X)X^{-1}\right\|_F / \left\|X^{-1}\right\|_F \le \|\Delta X\|_F \le \left\|(\Delta X)X^{-1}\right\|_F \|X\|_F, \quad (3.11) \\ \text{since } \|\Delta X\|_F = \left\|(\Delta X)X^{-1}X\right\|_F \le \left\|(\Delta X)X^{-1}\right\|_F \|X\|_F \\ \text{and} \quad \left\|(\Delta X)X^{-1}\right\|_F \le \left\|X^{-1}\right\|_F \|\Delta X\|_F. \end{aligned}$$

Notice that (3.11) is a sandwich, so:

$$\|(\Delta X)X^{-1}\|_F \to 0 \quad \iff \quad \|\Delta X\|_F \to 0.$$

Now recall that $\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\operatorname{tra} A^2}$. Then note that $\left\|(\Delta X)X^{-1}\right\|_F^2 = \operatorname{tr}\left((\Delta X)X^{-1}\right)^2 = \sum\lambda_i^2 = \|(\lambda_1,\ldots,\lambda_n)\|^2$. Thus our sandwich relationship becomes

$$\|\Delta X\|_F \to 0 \quad \iff \quad \|(\lambda_1,\ldots,\lambda_n)\| \to 0. \quad (3.12)$$

Now we are ready to show $\nabla f(X) = -X^{-1}$. Just notice:

$$\begin{aligned} &\lim_{\|\Delta X\|_F\to 0} \frac{\left|f(X+\Delta X) - f(X) - \langle -X^{-1}, \Delta X\rangle\right|}{\|\Delta X\|_F} \quad \text{(Frechét derivative form)} \\ &= \overbrace{\lim_{\|\lambda\|\to 0}}^{(3.12)} \frac{|\sum_i \overbrace{-\ln(1+\lambda_i)}^{(3.10)} + \overbrace{\lambda_i}^{(3.9)}|}{\|\nabla X\|_F} \\ &\le \lim_{\|\lambda\|\to 0} \frac{\sum_i |-\ln(1+\lambda_i) + \lambda_i|}{\|\nabla X\|_F} \\ &\overset{(a)}{\le} \lim_{\|\lambda\|\to 0} \frac{\sum_i |\ln(1+\lambda_i) - \lambda_i|}{(1/\sqrt{n})(\sum_i |\lambda_i|)/\|X^{-1}\|} \end{aligned}$$

[2]Note that, in general, the product of two symmetric matrices AB will not be symmetric unless $AB = BA$.

$$= \lim_{\|\lambda\| \to 0} \sqrt{n} \left\| X^{-1} \right\| \frac{\sum_i |\ln(1+\lambda_i) - \lambda_i|}{\sum_i |\lambda_i|} \overset{(b)}{=} 0.$$

To obtain (a), notice that

$$\|\nabla X\|_F \overset{(3.11)}{\geq} \left\| (\Delta X) X^{-1} \right\|_F / \left\| X^{-1} \right\|_F \overset{\text{Lemma 2.1.4}}{=} \sqrt{\left(\sum_i \lambda_i^2 \right)} / \left\| X^{-1} \right\|_F \overset{\text{C-S}}{\geq} (1/\sqrt{n})(\sum_i |\lambda_i|) / \left\| X^{-1} \right\|_F .$$

To obtain (b), simply build the Taylor series expansion:

$$\ln(1+\lambda_i) - \lambda_i = -\lambda_i^2 \left(\frac{1}{2} + \dots \right).$$

Altogether, this shows our first claim: that $\nabla f(X) = -X^{-1}$. In practice, the complicated derivation we have used is generally foregone in favor of using a Taylor series (such as we did in (2.5)).

We will show such an approach for computing the Hessian. I'll start with the term $-\nabla f(X + \Delta X) = (X + \nabla X)^{-1}$. Notice that:

$$(X + \Delta X)^{-1} \overset{\text{(rewriting)}}{=} \left[(I + (\Delta X) X^{-1}) X \right]^{-1} \overset{\text{(rewriting)}}{=} X^{-1} \underbrace{\left(I + (\Delta X) X^{-1} \right)^{-1}}_{(a)}$$

$$= X^{-1} \underbrace{\sum_{k=0}^{\infty} (-1)^k \left[(\Delta X) X^{-1} \right]^k}_{(a)}.$$

Here we are setting $y = (\nabla X) X^{-1}$ so that the term (a) corresponds to $1/(1+y)$, for which the geometric expansion was given by Carl Neumann as $1 - y + y^2 - y^3 \dots$.

I can subtract the first two terms from the series from both sides, obtaining:

$$(X + \Delta X)^{-1} - X^{-1} + X^{-1} (\nabla X) X^{-1} = X^{-1} \sum_{k=2}^{\infty} (-1)^k \left[(\Delta X) X^{-1} \right]^k$$

Multiplying by -1 yields:

$$-(X + \Delta X)^{-1} + X^{-1} - X^{-1} (\nabla X) X^{-1} = -X^{-1} \sum_{k=2}^{\infty} (-1)^k \left[(\Delta X) X^{-1} \right]^k$$

Because I know ∇f, I may now make the substitutions $-\nabla f(X) = X^{-1}$ and $\nabla f(X + \Delta X) = -(X + \Delta X)^{-1}$, whereupon the above equation becomes:

$$\nabla f(X + \Delta X) - \nabla f(X) - X^{-1} (\Delta X) X^{-1} = -X^{-1} \sum_{k=2}^{\infty} (-1)^k \left[(\Delta X) X^{-1} \right]^k . \quad (3.13)$$

When we look at the form of the Frechét derivative (below), and look at the above equation, we are led to suspect that $\nabla^2 f(X)(\Delta X) = X^{-1} (\Delta X) X^{-1}$. Let's prove it.

$$\limsup_{\|\Delta X\|_F \to 0} \frac{\| \nabla f(X + \Delta X) - \nabla f(X) - \overbrace{X^{-1} (\Delta X) X^{-1}}^{\text{(claim) } = \nabla^2 f(X)(\Delta X)} \|_F}{\|\Delta X\|_F} \quad \text{(Frechét derivative form)}$$

$$
\begin{aligned}
&= \limsup_{\|\Delta X\|_F \to 0} \frac{\left\| -X^{-1} \sum_{k=2}^{\infty} (-1)^k \left[(\Delta X)\, X^{-1} \right]^k \right\|_F}{\|\Delta X\|_F} \quad \text{by (3.13)} \\
&= \limsup_{\|\Delta X\|_F \to 0} \frac{\left\| -X^{-1}((\Delta X)X^{-1})^2 \sum_{k=0}^{\infty} (-1)^k \left[(\Delta X)\, X^{-1} \right]^k \right\|_F}{\|\Delta X\|_F} \quad \text{(pulling out common factors)} \\
&\leq \limsup_{\|\Delta X\|_F \to 0} \frac{\|\Delta X\|_F^2 \left\|X^{-1}\right\|_F^3 \sum_{k=0}^{\infty} \left(\|\Delta X\|_F \left\|X^{-1}\right\|_F \right)^k}{\|\Delta X\|_F} = 0 \quad \text{(sub-multiplicativity).}
\end{aligned}
$$

This shows what we claimed about the Hessian. Note that this does induce a bilinear form in that

$$
\begin{aligned}
(A, B) \mapsto \langle \nabla^2 f(X)(A), B \rangle &= \langle X^{-1} A X^{-1}, B \rangle \\
&= \operatorname{tr} B X^{-1} A X^{-1} \quad \text{(Lemma 2.1.4)} \\
&= \langle A X^{-1}, X^{-1} B \rangle \quad \text{is linear separately in } A \text{ and } B.
\end{aligned}
$$

Consequently we may continue to use the notation $\langle \nabla^2 f(X)(Q), Q \rangle$ for the quadratic form in the second order term in the Taylor expansion.

Remark 3.2.1 *Remember that from time to time we will use the "little Oh" notation. By $o(t)$ we mean any quantity with the property that*

$$\lim_{t \to 0} \frac{o(t)}{t} = 0.$$

This could be a term like $o(t) = K \|d\| t^{1+\delta}$ for $\delta > 0$ etc. These may be used to shorten expression for Taylor expansion of smooth functions f e.g.

$$
\begin{aligned}
f(x + td) &= f(x) + t\langle \nabla f(x), d \rangle + o(t \|d\|) \quad \text{or} \\
f(x + td) &= f(x) + t\langle \nabla f(x), d \rangle + \frac{1}{2} t^2 \langle \nabla^2 f(x)(d), d \rangle + o\left(t^2 \|d\|^2 \right).
\end{aligned}
$$

Chapter 4

Convex Functions and Sets

Convex analysis refers to an area an area of mathematics that can be thought of as a special part of both real analysis and nonsmooth analysis. We concern ourselves with functions that (in the main) evaluate to a real value but impose restriction on the geometry that these functions may possess. We are reward with a wealth of mathematical relations that are much richer than for extended real valued functions. We recall the definitions of a convex function.

Definition 4.0.1 *A function $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is called convex iff for all x_1, $x_2 \in \operatorname{dom} f := \{x \mid f(x) < +\infty\}$ and $\lambda \in [0,1]$ we have*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

A convex function is proper iff $f(x) > -\infty$ for all $x \in \operatorname{dom} f$ and $\operatorname{dom} f \neq \emptyset$.

Definition 4.0.2 *A convex function $f : \operatorname{dom}(f) \subseteq \mathbf{R}^n \to \overline{\mathbf{R}} = (-\infty, +\infty]$ is closed if and only if the set $\operatorname{epi}(f)$ is a closed set in X (i.e. $\{(x_n, \alpha_{n)}\} \subseteq \operatorname{epi}(f)$ with $(x_n, \alpha_n) \to (x, \alpha)$ (strongly) implies $(x, \alpha) \in \operatorname{epi}(f)$).*

It is clear that a function is bounded above on any ball $B_\delta(z) \subseteq \operatorname{dom} f$ i.e. at each point $z \in \operatorname{int} \operatorname{dom} f$.

We also have that:

Lemma 4.0.1 *Any closed convex function possesses an affine minorant and is bounded below on any nonempty bounded set.*

Proof. The strategy is to take a point $x \in \operatorname{dom} f$ and $\xi < f(x)$ and then consider the vector

$$P_{\operatorname{epi} f}(x, \xi) - (x - \xi).$$

The perp of this vector is a supporting hyperplane for $\operatorname{epi} f$ at the point $P_{\operatorname{epi} f}(x, \xi)$, and that hyperplane is the graph of an affine function that minorizes f.

For a more detailed proof, see [2, Section 9.3]. □

Theorem 4.0.2 *Let $f : \operatorname{dom}(f) \subseteq \mathbf{R}^n \to \overline{\mathbf{R}} = (-\infty, +\infty)$ be a convex function. Then f is actually locally Lipschitz continuous around a point z in its domain if and only if it is bounded above on a neighbourhood of z.*

Proof. Suppose f is locally Lipschitz continuous around z. Then f is continuous at z. So by definition of continuity, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that when $x \in B_\delta(z)$ we have $f(x) \in B_\varepsilon(f(z))$ implying $f(x) \leq f(z) + \varepsilon < +\infty$ for all $x \in B_\delta(z)$.

Now suppose f is bounded on the neighbourhood $B_{2\delta}(z)$ (i.e. $M \geq \sup\{f(v) \mid v \in B_{2\delta}(z)\}$) and $x, y \in B_\delta(z)$ with $x \neq y$. Place $\alpha = \|x - y\|$, let $w = y + \delta/\alpha(y - x)$ and note that

$$\begin{aligned}\|w - z\| &= \|y - z + (\delta/\alpha)(y - x)\| \\ &\leq \|y - z\| + \delta\frac{\|y - x\|}{\|y - x\|} \leq 2\delta \quad \text{so } \ w \in B_{2\delta}(z)\end{aligned}$$

Now solving for y (in the equation where we defined w), we obtain

$$\begin{aligned}y &= \frac{1}{1 + \delta/\alpha}(w + (\delta/\alpha)x) \\ &= \frac{\alpha}{\alpha + \delta}w + \frac{\delta}{\alpha + \delta}x,\end{aligned}$$

which is is a convex combination (inside $B_{2\delta}(z)$) and so

$$f(y) \leq \frac{\alpha}{\alpha + \delta}f(w) + \frac{\delta}{\alpha + \delta}f(x) \quad \text{and so}$$

$$f(y) - f(x) \leq \frac{\alpha}{\alpha + \delta}\left(f(w) - f(x)\right) \leq \overbrace{\frac{\alpha}{\delta}}^{\geq \frac{\alpha}{\alpha+\delta}} \underbrace{(\overbrace{M}^{\geq f(z)} - \overbrace{\inf_{x \in B_\delta(z)} f(x)}^{\leq f(x)})}_{=:M'}$$

$$= \frac{1}{\delta}M' \underbrace{\|x - y\|}_{=\alpha}.$$

Here the infimum exists by Lemma 4.0.1. Swapping the roles of x and y, we can likewise show

$$f(x) - f(y) \leq \left(M'/\delta\right)\|x - y\|.$$

Therefore, $|f(y) - f(x)| \leq L\|x - y\|$ with $L = (M'/\delta)$. □

This will be useful for studying sudifferentials, because this Lipschitz property $f(x + td) - f(x) \leq L\|d\|t$ means that differential quotients will be bounded:

$$\frac{f(x + td) - f(x)}{t} \leq L\|d\|.$$

4.1 The Convex Subdifferential

We follow the development in [4].

Definition 4.1.1 *Let $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ be convex. The vector $\lambda \in \mathbf{R}^n$ is a subgradient of f at $\bar{x}$ if it satisfies the following inequality for all $x \in \mathbf{R}^n$*

$$f(x) - f(\bar{x}) \geq \langle \lambda, (x - \bar{x})\rangle \quad \textit{(the subgradient inequality).}$$

Denote by $\partial f(\bar{x})$ the set of all subgradients of f at $\bar{x}$.

An important fact (see [4] Theorem 3.1.8 (finite dimensions) or [20] page 23 (infinite dimensions)) is that the convex subdifferential is non-empty as long as the domain has some kind of "interior point". It is clear that in $\mathbf{R}^n$ when $\nabla f(\bar{x})$ exists it is a subgradient and in this case $\langle \lambda, (x - \bar{x})\rangle = \langle \nabla f(\bar{x}), (x - \bar{x})\rangle$. Geometrically a subgradient has the property that $(-\lambda, 1)$ is the normal vector to a hyperplane that lies below epi(f) but touches epi(f) at $(\bar{x}, f(\bar{x}))$. You can see this in Figure 4.1, or write it down symbolically:

$$0 \geq \begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} x \\ f(x) \end{bmatrix} - \begin{bmatrix} \bar{x} \\ f(\bar{x}) \end{bmatrix} \right) = \langle \lambda, x - \bar{x}\rangle - (f(x) - f(\bar{x})).$$

Draw a few pictures and convince yourself it's true.

Example 4.1.1 *From (3.3) we know that the support function $w \mapsto S(C, w)$ of a closed convex set C in a Euclidean space satisfies*

$$S(C, v) - S(C, w) \geq \langle y, v - w\rangle$$

when y is the closest point in C to x and $w = x - y$. Consequently

$$y \in \partial S(C, w) := \partial S(C, \cdot)(w).$$

It is evident from previous analysis that all elements of $\partial S(C, w)$ achieve the maximum of the linear function $h \mapsto \langle w, h\rangle$ over $h \in C$. Such elements are said to be supported by the hyperplane $H := \{x \in X \mid \langle x, w\rangle = S(C, w)\}$.

Motivated by the classical observation that a differentiable function of several variables must possess a tangent plane at a given point on the surface corresponding to the function's graph, we take our starting point for nonsmooth analysis to be the study of the normal vector to such "tangent planes".

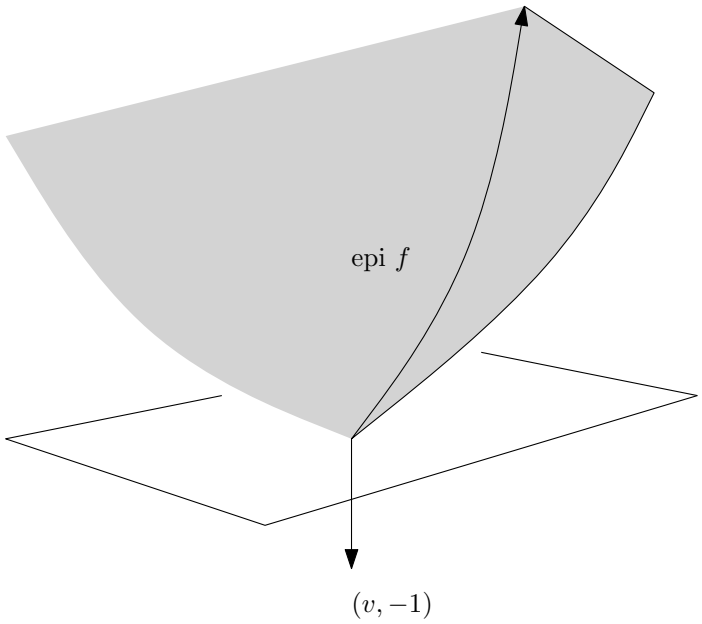


Figure 4.1: A supporting hyperplane gives rise to a subgradient $(v, -1)$

From the figure 4.1 it is clear that f need not be differentiable in order for a subgradient to exist. Note that if $\lambda \in \partial f(\bar{x})$ then we have for all $\delta > 0$ and $d \in \mathbf{R}^n$ that

$$\langle \lambda, (\overbrace{\bar{x} + \delta d}^{y} - \bar{x})\rangle \leq f(\overbrace{\bar{x} + \delta d}^{y}) - f(\bar{x})$$

$$\text{which implies} \quad \langle \lambda, d\rangle \le \frac{1}{\delta}\left(f(\bar{x}+\delta d) - f(\bar{x})\right).$$

Thus when the limit exists we have

$$\langle \lambda, d\rangle \le \lim_{\delta\downarrow 0} \frac{1}{\delta}\left(f(\bar{x}+\delta d) - f(\bar{x})\right) = f'(\bar{x}, d) \quad \text{for all } d \in \mathbf{R}^n, \tag{4.1}$$

where $f'(\bar{x}, d)$ is the (one sided) directional derivative of f at $\bar{x}$ (which exists thanks to Lemma 4.1.1 below). We call it the one sided since we only take the limit $\delta \downarrow 0$ (i.e. $\delta > 0$) not the two side limit. Indeed it is easily shown (let $\delta < 0$ and replace δ by $\bar{\delta} = -\delta$) that

$$\begin{aligned} -f'(\bar{x}, -d) = -\lim_{\bar{\delta}\downarrow 0} \frac{1}{\bar{\delta}}\left(f(\bar{x}+\bar{\delta}(-d)) - f(\bar{x})\right) &= \lim_{\bar{\delta}\downarrow 0}\left(\frac{1}{-\bar{\delta}}\left(f(\bar{x}+\bar{\delta}(-d)) - f(\bar{x})\right)\right) \\ &= \lim_{\delta\uparrow 0} \frac{1}{\delta}\left(f(\bar{x}+\delta d) - f(\bar{x})\right). \end{aligned}$$

In the final term, if the limit as $\delta \uparrow 0$ is equal to the same limit as when $\delta \downarrow 0$, then we have one limit as $\delta \to 0$, in which case the rightmost term is equal to $f'(\bar{x}, d)$. Thus, we have shown the following equivalence:

$$f'(\bar{x}, d) = \lim_{\delta\to 0} \frac{1}{\delta}\left(f(\bar{x}+\delta d) - f(\bar{x})\right) \quad \Longleftrightarrow \quad -f'(\bar{x}, -d) = f'(\bar{x}, d), \quad \text{for all } d \in \mathbf{R}^n. \tag{4.2}$$

Of course, this is certainly the case when f is differentiable. To see why, just notice that when the derivative exists we have the linearity:

$$\begin{aligned} &f'(\bar{x}, d) = \langle \nabla f(\bar{x}), d\rangle = -\langle \nabla f(\bar{x}), -d\rangle = -f'(\bar{x}, -d) \\ &\text{and so} \quad \langle \nabla f(\bar{x}), -d\rangle = f'(\bar{x}, -d) \end{aligned}$$

Since differentiability implies that (4.2) holds for all d, it is clear that $\nabla f(\bar{x})$ satisfies (4.1) with equality throughout, and so $\nabla f(\bar{x}) \in \partial f(\bar{x})$. Note that the condition $\nabla f(\bar{x}) \in \partial f(\bar{x})$ is also clear from the definition of the subdifferential and Theorem 1.1.6(ii).

We say a function $h : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is positively homogeneous when $h(\lambda d) = \lambda h(d)$ for all $\lambda > 0$ and $d \in X$. We say $h : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is subadditive when $h(d_1 + d_2) \le h(d_1) + h(d_2)$ for all d_1, $d_2 \in \mathbf{R}^n$.

Example 4.1.2 *The support function of any closed convex set C is a positively homogeneous function. Note that*

$$\begin{aligned} S(C, d_1 + d_2) &= \sup\{\langle c, d_1 + d_2\rangle \mid c \in C\} \\ &= \sup\{\langle c, d_1\rangle + \langle c, d_2\rangle \mid c \in C\} \\ &\le \sup\{\langle c, d_1\rangle \mid c \in C\} + \sup\{\langle c, d_2\rangle \mid c \in C\} \\ &= S(C, d_1) + S(C, d_2) \end{aligned}$$

and for all $\lambda > 0$ we have

$$S(C, \lambda d) = \sup\{\langle c, \lambda d\rangle \mid c \in C\} = \lambda \sup\{\langle c, d\rangle \mid c \in C\} = \lambda S(C, d).$$

Lemma 4.1.1 *Let $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is convex and $\bar{x} \in \operatorname{dom} f$, For each d, the differential quotient $\delta \mapsto \frac{1}{\delta}\left(f(\bar{x}+\delta d) - f(\bar{x})\right) := \Delta f(\bar{x}, \delta, d)$ is non–decreasing[1] and so $f'(\bar{x}, d)$ exists and we have*

$$f'(\bar{x}, d) = \inf_{\delta > 0} \frac{1}{\delta}\left(f(\bar{x}+\delta d) - f(\bar{x})\right).$$

Moreover, $f'(\bar{x}, \cdot)$ is a positively homogeneous convex function of d, with $f'(\bar{x}, 0) = 0$ and

$$-f'(\bar{x}, -d) \leq f'(\bar{x}, d) \quad \textit{for all } d \in \mathbf{R}^n. \tag{4.3}$$

When

$$-f'(\bar{x}, -d) = f'(\bar{x}, d) \quad \textit{for all } d \in \mathbf{R}^n \tag{4.4}$$

then $d \mapsto f'(\bar{x}, d)$ is linear and so $\nabla f(\bar{x})$ exists.

Proof. Let $h(d) = f(\bar{x}+d) - f(\bar{x})$ and so $\Delta f(\bar{x}, \delta, d) = \delta^{-1} h(\delta d)$ and $h(0) = 0$. In effect this translates $(\bar{x}, f(\bar{x}))$ to the origin $(0, 0)$. Assume now that $(\bar{x}, f(\bar{x})) = (0, 0)$ and $0 < t < s$, then by convexity

$$\begin{aligned} h(tx) = h\left(\frac{t}{s}(sx) + \frac{(s-t)}{s}0\right) &\leq \frac{t}{s}h(sx) + \frac{(s-t)}{s}h(0) = \frac{t}{s}h(sx) \\ \implies \quad t^{-1}h(tx) \leq s^{-1}h(sx) \quad \text{or} \quad \Delta f(\bar{x}, t, d) &\leq \Delta f(\bar{x}, s, d), \text{ for } 0 < t < s. \end{aligned} \tag{4.5}$$

This proves the first assertion (monotonicity). Thus $\lim_{\delta \downarrow 0} \Delta f(\bar{x}, t, d)$ must exist (as it is a monotonic limit, but we will allow the possibility that it is $+\infty$ for now) and $\lim_{\delta \downarrow 0} \Delta f(\bar{x}, t, d) = \inf_{\delta > 0} \Delta f(\bar{x}, t, d)$. Now we'll show positive homogeneity. Let $\lambda > 0$ and

$$\begin{aligned} f'(\bar{x}, \lambda d) &= \inf_{\delta > 0} \Delta f(\bar{x}, t, \lambda d) \\ &= \inf_{\delta > 0} \frac{1}{\delta}\left(f(\bar{x}+\delta\lambda d) - f(\bar{x})\right) = \lambda \inf_{\delta > 0} \frac{1}{\underbrace{\lambda\delta}_{\hat{\delta}}}\left(f(\bar{x} + \underbrace{\delta\lambda}_{\hat{\delta}} d) - f(\bar{x})\right) = \lambda f'(\bar{x}, d). \end{aligned}$$

Now we show convexity. For all $\lambda \in [0, 1]$ and $d_1,\ d_2 \in \mathbf{X}$ we have

$$\begin{aligned} f'(\bar{x}, \lambda d_1 + (1-\lambda)d_2) &= \lim_{\delta > 0} \frac{1}{\delta}\left(f(\bar{x} + \delta\left(\lambda d_1 + (1-\lambda)d_2\right)) - f(\bar{x})\right) \\ &= \lim_{\delta > 0} \frac{1}{\delta}\left(f(\lambda\left(\bar{x} + \delta d_1\right) + (1-\lambda)\left(\bar{x} + \delta d_2\right)) - f(\bar{x})\right) \\ &\leq \lim_{\delta > 0} \frac{1}{\delta}\left(\lambda f(\bar{x} + \delta d_1) + (1-\lambda) f\left(\bar{x} + \delta d_2\right)) - f(\bar{x})\right) \\ \text{(monotonic limit)} \quad &= \lambda \lim_{\delta > 0} \frac{1}{\delta}\left(f(\bar{x} + \delta d_1) - f(\bar{x})\right) + (1-\lambda) \lim_{\delta > 0} \frac{1}{\delta}\left(f\left(\bar{x} + \delta d_2\right) - f(\bar{x})\right) \\ &= \lambda f'(\bar{x}, d_1) + (1-\lambda) f'(\bar{x}, d_2). \end{aligned}$$

That shows the convexity. Now we show sub-additivity. For all $d_1,\ d_2 \in \mathbf{X}$ we have for $\lambda \in (0, 1)$:

$$f'(\bar{x}, d_1 + d_2) = f'\left(\bar{x}, \lambda\left(\frac{1}{\lambda}d_1\right) + (1-\lambda)\left(\frac{1}{(1-\lambda)}d_2\right)\right)$$

[1] i.e. take the slice function $g_{\bar{x},(\bar{x}+d)}$, and fit secant lines to its graph by sampling at pairs $(0, \delta_1)$ and $(0, \delta_2)$ with $\delta_2 > \delta_1$. The secant line fitted for δ_2 will have a greater slope.

$$\begin{aligned}
&\leq \lambda f'\left(\bar{x}, \left(\frac{1}{\lambda}d_1\right)\right) + (1-\lambda) f'\left(\bar{x}, \left(\frac{1}{(1-\lambda)}d_2\right)\right) \quad \text{(convexity)} \\
&= \lambda \frac{1}{\lambda} f'(\bar{x}, d_1) + (1-\lambda)\frac{1}{(1-\lambda)} f'(\bar{x}, d_2) \quad \text{(positive homogeneity)} \\
&= f'(\bar{x}, d_1) + f'(\bar{x}, d_2).
\end{aligned}$$

This shows the sub-additivity.

Now obviously:

$$f'(\bar{x}, 0) = \lim_{\delta \downarrow 0} \frac{1}{\delta}(f(\bar{x} - \delta 0) - f(\bar{x})) = 0.$$

Now we show (4.3). Notice that

$$\frac{1}{2}f'(\bar{x}, -d) + \frac{1}{2}f'(\bar{x}, d) \overset{\text{(convexity)}}{\geq} f'(\bar{x}, \frac{1}{2}(-d) + \frac{1}{2}d) = f'(\bar{x}, 0) = 0.$$

Subtracting $f'(\bar{x}, d)$ from both sides, we obtain $f'(\bar{x}, -d) \geq -f'(\bar{x}, d)$.

Finally, we show the linearity under the assumption (4.4): $-f'(\bar{x}, -d) = f'(\bar{x}, d)$. When this holds, the two sided directional derivative exists and hence

$$\begin{aligned}
f'(\bar{x}, -d_1 - d_2) &\leq f'(\bar{x}, -d_1) + f'(\bar{x}, -d_2) \quad \text{(sub-additivity)} \\
&= -f'(\bar{x}, d_1) - f'(\bar{x}, d_2) \quad \text{(assumption (4.4))}
\end{aligned} \tag{4.6}$$

and so

$$f'(\bar{x}, d_1 + d_2) \overset{\text{assumption (4.4)}}{=} -f'(\bar{x}, -(d_1 + d_2)) \overset{-(4.6)}{\geq} f'(\bar{x}, d_1) + f'(\bar{x}, d_2).$$

To show the reverse inequality, simply notice that

$$\begin{aligned}
f'(\bar{x}, d_1 + d_2) = f'(\bar{x}, \frac{1}{2}2d_1 + \frac{1}{2}2d_2) &\overset{\text{(convexity)}}{\leq} \frac{1}{2}f'(\bar{x}, 2d_1) + \frac{1}{2}f'(\bar{x}, 2d_2) \\
&\overset{\text{(pos-homogeneity)}}{=} f'(\bar{x}, d_1) + f'(\bar{x}, d_2).
\end{aligned}$$

Having shown both directions, we have that

$$f'(\bar{x}, d_1 + d_2) = f'(\bar{x}, d_1) + f'(\bar{x}, d_2).$$

If this holds for all d we have $d \mapsto f'(\bar{x}, d)$ linear and so $f'(\bar{x}, d) = \langle z, d\rangle$ for some z. This occurs only when $z = \nabla f(x)$ exists. □

Lemma 4.1.2 *If $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is convex and $\bar{x} \in \operatorname{dom} f$, then $\lambda \in \partial f(\bar{x})$ iff $\langle \lambda, d\rangle \leq f'(\bar{x}, d)$ for all $d \in \mathbf{R}^n$.*

Proof. We have already shown in (4.1) that $\lambda \in \partial f(\bar{x})$ implies $\langle \lambda, d\rangle \leq f'(\bar{x}, d)$ for all $d \in \mathbf{R}^n$. Now assume $\langle \lambda, d\rangle \leq f'(\bar{x}, d)$ for all $d \in \mathbf{R}^n$. Then as

$$\langle \lambda, d\rangle \leq \inf_{\delta > 0} \frac{1}{\delta}(f(\bar{x} + \delta d) - f(\bar{x})) \leq \frac{1}{\delta}(f(\bar{x} + \delta d) - f(\bar{x})) \quad \text{for all } \delta > 0 \text{ and } d \in X.$$

Multiplying both sides by δ yields

$$\langle \lambda, \delta d\rangle \leq f(\bar{x} + \delta d) - f(\bar{x}).$$

Consequently on placing $y = \bar{x} + \delta d \in \mathbf{R}^n$ we obtain

$$\begin{aligned} &\langle \lambda, (\bar{x} + \delta d) - \bar{x} \rangle \leq f(\bar{x} + \delta d) - f(\bar{x}) \\ &\qquad \Longrightarrow \quad \langle \lambda, y - \bar{x} \rangle \leq f(y) - f(\bar{x}) \quad \text{for all } y \in X. \end{aligned}$$

Thus $\lambda \in \partial f(\bar{x})$. □

A common function in convex analysis is the indicator function of a convex set.

Lemma 4.1.3 *Suppose C is a closed convex set in $\mathbf{R}^n$ and $\bar{x} \in C$ then*

$$\begin{aligned} \partial \delta_C(\bar{x}) &= [\operatorname{cone}(C - \bar{x})]^\circ \\ \textit{where} \quad [\operatorname{cone}(C - \bar{x})]^\circ &:= \{x^* \mid \langle x^*, d \rangle \leq 0, \textit{ for all } d \in \operatorname{cone}(C - \bar{x})\} \\ \textit{and} \quad \operatorname{cone}(C - \bar{x}) &= \cup_{\lambda \geq 0} \lambda (C - \bar{x}). \end{aligned}$$

Proof. First note that if $\bar{x} \in C$ then for $d \neq 0$

$$\begin{aligned} \delta'_C(\bar{x}, d) &= \inf_{t>0} \frac{1}{t} \left(\delta_C(\bar{x} + td) - \delta_C(\bar{x}) \right) \\ &= \begin{cases} 0 & \text{if} \quad \bar{x} + td \in C \quad \text{for some } t > 0 \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

The condition $\delta'_C(\bar{x}, d) = 0$ with $d \neq 0$ is equivalent to (just solve for d above):

$$\begin{aligned} d &\in \lambda (C - \bar{x}) \quad \text{for some } \lambda = \frac{1}{t} > 0 \\ \text{so} \quad d &\in \cup_{\lambda \geq 0} \lambda (C - \bar{x}) := \operatorname{cone}(C - \bar{x}). \end{aligned}$$

When $d = 0$ then clearly $\delta'_C(\bar{x}, 0) = 0$. Thus we have

$$\begin{aligned} \partial \delta_C(\bar{x}) &= \left\{ x^* \mid \langle x^*, d \rangle \leq \delta'_C(\bar{x}, d), \forall d \in \mathbf{X} \right\} \quad \text{(Lemma 4.1.2)} \\ &= \{x^* \mid \langle x^*, d \rangle \leq 0, \text{ for all } d \in \operatorname{cone}(C - \bar{x})\} \\ &= [\operatorname{cone}(C - \bar{x})]^\circ. \end{aligned}$$

□

Note: We have only shown the identity $\partial \delta_C(\bar{x}) = [\operatorname{cone}(C - \bar{x})]^\circ$ when $\bar{x} \in C$. When $\bar{x} \notin C$, then we may not even use Lemma 4.1.2. In fact, it turns out that:

$$\partial \delta_C(\bar{x}) = N_C(\bar{x}) := \begin{cases} [\operatorname{cone}(C - \bar{x})]^\circ & \text{if } \bar{x} \in C \\ \emptyset & \text{otherwise} \end{cases}.$$

The operator N_C is called the *normal cone operator*.

The following result will be useful later on.

Corollary 4.1.4 *Suppose C is a closed convex set with $\operatorname{int} C \neq \emptyset$ and $\bar{x} \in C$. Then*

$$N_C(\bar{x}) = [\operatorname{cone}(C - \bar{x})]^\circ$$

is a pointed cone in the sense that there does not exist nonzero $x^* \in N_C(\bar{x})$ such that $-x^* \in N_C(\bar{x})$.

Proof. Since $\operatorname{int} C \neq \emptyset$ by assumption, we may, without loss of generality, let it contain 0. If there existed nonzero $-x^*, x^* \in N_C(\bar{x})$ then we would have

$$\pm\langle c - \bar{x}, x^* \rangle = \langle c - \bar{x}, \pm x^* \rangle \leq 0 \quad \text{for all } c \in C.$$

It follows that $\langle c - \bar{x}, x^* \rangle = 0$ for all $c \in C$. In particular, since $\bar{x} \in C$, it holds for $c = \bar{x}$. Thus $\langle c - \bar{x}, x^* \rangle = 0$ for all $c \in C$.

But since 0 is an interior point of C, there exists a ball of radius $\delta > 0$ about 0 such that $B_\delta(0) \subset C$. Thus we have $\langle c - \bar{x}, x^* \rangle = 0$ for all $c \in B_\delta(0)$. Thus $x^*(\bar{x} + B_\delta(0)) = 0$, where $\bar{x} + B_\delta(0)$ is an open ball, and this forces x^* to be zero, a contradiction. □

Corollary 4.1.5 *Suppose C is a closed convex set in $\mathbf{R}^n$ and $\bar{x} \in C$. Then*

$$\partial \delta_C(\bar{x}) = [(C - \bar{x})]^\circ . \tag{4.7}$$

Proof. We need only to show that $[C - \bar{x}]^\circ = [\operatorname{cone}(C - \bar{x})]^\circ$. Take any $x^* \in [\operatorname{cone}(C - \bar{x})]^\circ$, and we have that $\langle x^*, y \rangle \leq 0$ for all $y \in \operatorname{cone}(C - \bar{x}) \supset (C - \bar{x})$, and so clearly $x^* \in [C - \bar{x}]^\circ$.

Take any $x^* \in [C - \bar{x}]^\circ$ and take any $y \in \operatorname{cone}[C - \bar{x}]$. Then $y = \lambda(c - \bar{x})$ for some $c \in C, \lambda > 0$. Then $\langle x^*, y \rangle = \lambda \langle x^*, (c - \bar{x}) \rangle \leq 0$. Thus $x^* \in [\operatorname{cone}(C - \bar{x})]^\circ$. □

4.1.1 Notions of interiority

A function $h : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is sublinear if

$$h(\lambda x + \mu y) \leq \lambda h(x) + \mu h(y) \quad \text{for all } \lambda, \mu \geq 0$$

Note we get subadditivity when $\lambda = \mu = 1$.

Definition 4.1.2 *A point $\bar{x} \in \operatorname{core} D$ if for all $y \in X$ there exists $t > 0$ such that $\bar{x} + t(y - \bar{x}) \in D$. Note that t may depend on the choice of y.*

If $\bar{x} \in \operatorname{int} D$ then $B_\delta(\bar{x}) \subseteq D$ for some $\delta > 0$. Clearly $\bar{x} \in \operatorname{int} D$ implies $\bar{x} \in \operatorname{core} D$ (why?). The core may be a large set than the interior. Also note that any point in $\operatorname{core} D$ is also in D (to see why, just choose $y = 0$ in the definition).

Exercise 4.1.1 *Consider the set in $\mathbf{R}^2$*

$$D = \left\{ (x, y) \mid y = 0 \text{ or } |y| \geq x^2 \right\}.$$

Prove that $0 \in \operatorname{core}(D) \cap (\operatorname{int}(D))^c$.

Soln: *Clearly $0 \notin \operatorname{int}(D)$ since $\emptyset \neq B_\delta(0) \cap \left\{ (x, y) \in \mathbf{R}^2 \mid |y| < x^2 \right\} \subseteq D^c$. Now consider $t > 0$ and $(x, y) \in \mathbf{R}^2$ then $t(x, y) \in D$ is equivalent to*

$$t\,|y| \geq t^2 x^2 \quad \Rightarrow \quad |y| \geq t x^2.$$

If $(x, y) \neq 0$ then this occurs for $\frac{|y|}{x^2} \geq t > 0$. If $x = 0$ but $y \neq 0$ then $t(x, y) \in D$ is equivalent to $t\,|y| \geq 0$ which occurs for any $t \geq 0$. If $y = 0$ then $(x, y) \in D$ for all x.

Proposition 4.1.6 *If $f : \mathrm{dom}(f) \subseteq \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is convex, then, for any $\bar{x} \in \mathrm{core}\,(\mathrm{dom}\, f)$, the directional derivative $f'(\bar{x}, d)$ is everywhere finite and sublinear.*

Proof. From (4.5) we have the monotonic property:

$$\Delta f(\bar{x}, t, d) \leq \Delta f(\bar{x}, s, d), \text{ for } 0 < t < s.$$

If $\bar{x} \in \mathrm{core}\,(\mathrm{dom}\, f)$ then for s sufficiently small we have $\bar{x} + sd \in \mathrm{dom}\, f$ implying that $f(\bar{x} + sd) < \infty$, and so we have $\Delta f(\bar{x}, s, d) < +\infty$. Thus

$$(\forall d \in \mathbf{X}) \quad f'(\bar{x}, d) \overset{\text{(Lemma 4.1.1)}}{=} \inf_{t>0} \Delta f(\bar{x}, t, d) \leq \Delta f(\bar{x}, s, d) < +\infty. \tag{4.8}$$

Now use (4.3) with $d = -y$ to deduce that:

$$\begin{gathered} -\infty \overset{(4.8)}{<} -f'(\bar{x}, y) \overset{(4.3)}{\leq} f'(\bar{x}, -y) \overset{(4.8)}{<} +\infty \\ \text{so} \quad f'(\bar{x}, y) > -\infty \quad \text{for } y \in X. \end{gathered}$$

Hence $|f'(\bar{x}, y)| < +\infty$ and by Lemma 4.1.1 $d \mapsto f'(\bar{x}, d)$ is convex and positively homogeneous and sub-additive.
To obtain the sub-linearity of $f'(\bar{x}, \cdot)$, simply notice that if you combine the sub-additivity of a function h with positive homogeneity, you obtain the sub-linearity straight away as follows:

$$h(\lambda x + \mu y) \overset{\text{sub-additivity}}{\leq} h(\lambda x) + h(\mu y) \overset{\text{pos-homogeneity}}{=} \lambda h(x) + \mu h(y),$$

which is the definition of sub-linearity.

□

Definition 4.1.3 *The intrinsic core of A consists of all $\bar{x} \in A$ such that for all*

$$y \in \mathrm{affine}\, A := x + \mathrm{span}\,(A - x)$$

there exists a $t > 0$ such that $\bar{x} + t(y - \bar{x}) \in A$. Denote the set of all such points as $\mathrm{icr}\, A$.

The intrinsic core is just the core applied in the space affine A. Remark 4.1.1 is an important fact (see [4] Theorem 3.1.8 (finite dimensions) or [20] page 23 (infinite dimensions)). The proof from [4] is reproduced in Appendix B. The power of the idea of a subgradient comes from the fact that the subdifferential is often nonempty.

Remark 4.1.1 *Note that when $f(x) = -\infty$ for any $x \in \mathrm{dom}\, f$ then $f \equiv -\infty$. Indeed for any $y \in \mathrm{dom}\, f$ and $\lambda \in (0, 1)$ we have $-\infty = \lambda f(y) + (1-\lambda) f(x) \geq f(\lambda y + (1-\lambda)x)$. Thus we see that it is important to remove the trivial function $f \equiv -\infty$ from consideration. This is why we only consider proper functions, which are defined to satisfy $f(x) > -\infty$ for all $x \in \mathrm{dom}\, f$ and f is not identically $+\infty$.*

Theorem 4.1.7 (nonempty subgradient Theorem) *If $f : \mathrm{dom}(f) \subseteq \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is a proper, convex function, then at any point $\bar{x} \in \mathrm{icr}\, \mathrm{dom}(f)$ we have $\partial f(\bar{x}) \neq \emptyset$ and*

$$f'(\bar{x}, d) = \max \{\langle \lambda, d \rangle \mid \lambda \in \partial f(\bar{x})\}.$$

Proof. See Appendix A. □

Remark 4.1.2 *It can also be shown that when $\nabla f(\bar{x})$ exists then $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$ is unique and so in finite dimensions $f'(\bar{x}, d) = \nabla f(\bar{x})^T d$ which is the classical characterization of the directional derivative. Indeed when $\nabla f(\bar{x})$ does not exist, then $\partial f(\bar{x})$ may contain many values, prompting it to be referred to as a multi-function (i.e. one to many mapping).*

Proof. See Appendix A. □

Potentially in many applications we know $f(x) < +\infty$ but $f(x) = -\infty$ must be ruled out in order to use Theorem 4.1.7.

Lemma 4.1.8 *If the function $f : \mathbf{R}^m \to \mathbf{R}_{+\infty}$ is convex and at some point $\hat{y} \in \operatorname{core}\operatorname{dom}(f)$ we have $f(\hat{y}) > -\infty$, then h never takes the value $-\infty$. Similarly if at some point $\hat{y} \in \operatorname{icr}\operatorname{dom}(f)$ we have $f(\hat{y}) > -\infty$, then f never takes the value $-\infty$ on $\operatorname{affine}\operatorname{dom}(f)$.*

Proof. Suppose that $f(y) = -\infty$ and we will show a contradiction arises. Since $\hat{y} \in \operatorname{core}\operatorname{dom}(f)$ there exists a $t > 0$ such that $\hat{y} + t(\hat{y} - y) \in \operatorname{dom}(f)$. Hence there is a real number r such that $(\hat{y} + t(\hat{y} - y), r) \in \operatorname{epi}(f)$. Now for any s we have $(y, s) \in \operatorname{epi}(f)$ and so

$$\left(\hat{y}, \frac{r + ts}{1 + t}\right) = \frac{1}{1 + t}(\hat{y} + t(\hat{y} - y), r) + \frac{t}{1 + t}(y, s) \in \operatorname{epi}(f).$$

Letting $s \to -\infty$ we arrive at the contradiction $f(\hat{y}) = -\infty$. The extension to $\hat{y} \in \operatorname{icr}\operatorname{dom}(f)$ follows by a similar argument. □

4.1.2 Problem set 6a: Computing subdifferentials

Problem 4.1.9 *Calculate ∂f for the functions $f : \mathbf{R} \to \mathbf{R}$*

1. $f(x) = |x|$
2. $f(x) = \delta_{[0,+\infty)} = \begin{cases} 0 & \text{if } x \in [0, +\infty) \\ +\infty & \text{if } x \notin [0, +\infty) \end{cases}$
3. $f(x) = \begin{cases} -\sqrt{x} & \text{if } x \in [0, +\infty) \\ +\infty & \text{if } x \notin [0, +\infty) \end{cases} = -\sqrt{x} + \delta_{[0,+\infty)}.$

4.1.3 Conjugation and the Subdifferential

We introduce the conjugate and calculate the conjugate of some frequently encountered functions.

Definition 4.1.4 *Let $h : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ then the function $h^* : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is defined by*

$$h^*(x^*) = \sup_x \{\langle x^*, x\rangle - h(x)\}.$$

The biconjugate is given by

$$h^{**}(x) := \sup_{x^*} \{\langle x^*, x\rangle - h(x^*)\}$$

Lemma 4.1.10 *Let f be closed, have a nonempty domain, and possess an affine minorant. Then f^* is convex, proper, and closed.*

Proof. Let's show convex and closed. For any x, the function $x^* \mapsto \langle x^*, x\rangle - f(x)$ is affine and continuous, and so its epigraph is closed and convex. The epigraph of f^* is the intersection of such epigraphs, and so it is also closed and convex.

Now let's show proper. As f is assumed to be minorized by some affine function $x \mapsto \langle z^*, x\rangle + a$, we have

$$f^*(z^*) = \sup_x(\langle z^*, x\rangle - f(x)) \leq \sup_x(\langle z^*, x\rangle - (\langle z^*, x\rangle + a)) = -a < \infty,$$

and so $z^* \in \operatorname{dom} f^*$. Also, as $\operatorname{dom} f \neq \emptyset$, there exists some $\bar{x} \in \operatorname{dom} f$, and so

$$f^*(y^*) = \sup_x \langle y^*, x\rangle - f(x) \geq \langle y^*, \bar{x}\rangle - f(\bar{x}) > -\infty.$$

This shows $f^*(y^*) > -\infty$ for any y^*, and so f^* is proper. □

Note that any proper convex f will have an affine minorant by Lemma 4.0.1.

Theorem 4.1.11 (Biconjugate Theorem [2, Theorem 13.37]) *When $h : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is a convex and proper function then we have $h^{**}(x) = h(x)$ if and only if h is lower semi–continuous at x i.e.*

$$h(x) = \liminf_{x' \to x} h(x') := \lim_{\delta \downarrow 0} \inf_{x' \in B_\delta(x)} h(x'). \tag{4.9}$$

Proof. We won't prove the whole thing, but part of it is easy. Notice that we always have

$$\begin{aligned} f^{**}(x) &= \sup_{x^*} \langle x, x^*\rangle - f^*(x^*) \\ &= \sup_{x^*} \left(\langle x, x^*\rangle - \sup_y (\langle x^*, y\rangle - f(y)) \right) \\ &= \sup_{x^*} \left(\langle x, x^*\rangle + \inf_y (f(y) - \langle x^*, y\rangle) \right) \\ &\leq \sup_{x^*} (\langle x, x^*\rangle - \langle x^*, x\rangle + f(x)) = f(x). \end{aligned}$$

Thus $f^{**} \leq f$. □

Remark 4.1.3 *The condition (4.9) corresponds to the proposition that $(x, h(x)) \in \overline{\operatorname{epi} h}$. When $\overline{\operatorname{epi} h} = \operatorname{epi} h$ we say h is a closed and convex. It can be shown that the only convex function that attains the value $-\infty$ must be identically $-\infty$ (i.e. equals $-\infty$ everywhere).*

In fact, it turns out that the conjugate f^* is always convex, and $\operatorname{epi} f^{**} = \overline{co} \operatorname{epi} f$. The next lemma allows us to prove this rather cheaply.

Lemma 4.1.12 *If $f \geq g$ then $f^* \leq g^*$. If f, g are convex and proper and closed (lsc), then the reverse implication $(f^* \leq g^*) \implies (f \geq g)$ also holds.*

Proof. Note that if $f \geq g$ then

$$f^*(x^*) = \sup_x \{\langle x^*, x\rangle - f(x)\} \leq \sup_x \{\langle x^*, x\rangle - g(x)\} = g^*(x^*). \tag{4.10}$$

Now suppose $f^* \leq g^*$. Then, by the same argument we just used, $f^{**} \geq g^{**}$. If f, g are convex, proper, and closed, then the biconjugate Theorem 4.1.11 tells us that $f = f^{**}$ and $g = g^{**}$, whereupon we are finished. □

When the conditions of Lemma 4.1.12 are satisfied, it is easy to show that $\text{epi}\, f^{**} = \overline{co}\, \text{epi}\, f$. Simply let g be any closed convex function that satisfies $g \leq f$. Then Lemma 4.1.12 guarantees both $f^* \geq g^*$ and $g^{**} \leq f^{**}$, and so

$$g \overset{\text{Theorem 4.1.11}}{=} g^{**} \leq f^{**} \overset{\text{Theorem 4.1.11}}{\leq} f.$$

Thus f^{**} is a convex function and an upper bound on all convex functions that lower bound f.

Exercise 4.1.2 *Show that the conjugate of the exponential function is given by*

$$\exp^*(t) = \begin{cases} t\log t - t & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ +\infty & \text{if } t < 0 \end{cases}.$$

Note that although the exponential function is finite valued its conjugate is not.

Solution 4.1.13 *From definitions*

$$\exp^*(t) = \sup_x \{xt - \exp(x)\}$$

and using calculus we may find the stationary points

$$\frac{d}{dx}\{xt - \exp(x)\} = t - \exp(x) = 0 \quad \Longrightarrow \quad x = \ln(t) \ \text{ if } t > 0.$$

Thus for $t > 0$ we have

$$\exp^*(t) = t\ln(t) - \exp(\ln(t)) = t\log t - t.$$

If $t = 0$ then

$$\exp^*(0) = \sup_x \{0 - \exp(x)\} = -\inf_x \exp(x) = 0.$$

When $t < 0$ we have

$$\exp^*(t) = \sup_x \{xt - \exp(x)\} \geq \lim_{x \downarrow -\infty} (xt - \exp(x)) = \infty.$$

Example 4.1.3 *Support functions occur frequently. Let $C \subseteq X^*$ be closed and convex. The conjugate of the support function of C is the indicator of C. First notice:*

$$S^*(C, \cdot)(x^*) = \sup_x \{\langle x, x^* \rangle - S(C, x)\}.$$

If $x^* \in C$ then

$$S(C, x) = \sup_{c \in C} \langle c, x \rangle \geq \langle x, x^* \rangle, \ \text{ and so } \ S^*(C, \cdot)(x^*) = \sup_x \{\langle x, x^* \rangle - S(C, x)\} \leq 0.$$

We can achieve the supremum of 0 by choosing $x = 0$, and so we have

$$(x^* \in C) \implies S^*(C, \cdot)(x^*) = 0.$$

Now consider the case when $x^* \notin C$. Then we may properly separate x^* from C by a linear functional $\bar{x}$. In other words, $\bar{x}$ satisfies $\langle \bar{x}, x^* \rangle > S(C, \bar{x})$ (or $\langle \bar{x}, x^* \rangle - S(C, \bar{x}) = \delta > 0$) and so for all $\alpha > 0$

$$\begin{aligned} S^*(C, \cdot)(x^*) &= \sup_x \{\langle x, x^* \rangle - S(C, x)\} \\ &\geq \{\langle \alpha \bar{x}, x^* \rangle - S(C, \alpha \bar{x})\} = \alpha \delta \to +\infty \quad \text{as } \alpha \to +\infty. \end{aligned}$$

Hence

$$S^*(C, \cdot)(x^*) = \begin{cases} 0 & \text{if} \quad x^* \in C \\ +\infty & \text{if} \quad x^* \notin C \end{cases} = \delta_C(x^*).$$

Conjugates and subgradients

Let $x \in X$ and $x^* \in X^*$ then we always have the so called Fenchel–Young inequality

$$h(x) + h^*(x^*) = h(x) + \sup_y \{\langle x^*, y \rangle - h(y)\} \geq h(x) + \langle x^*, x \rangle - h(x) = \langle x^*, x \rangle. \tag{FYI}$$

Note that if $x^* \in \partial h(x)$ then by the subgradient inequality

$$\begin{aligned} & h(y) - h(x) \geq \langle x^*, y - x \rangle \quad \text{for all } y \in X \\ \text{i.e.} \quad & \langle x^*, x \rangle - h(x) \geq \langle x^*, y \rangle - h(y) \quad \text{for all } y \in X \\ \text{equivalently} \quad & \langle x^*, x \rangle - h(x) \geq \sup_y \{\langle x^*, y \rangle - h(y)\} = h^*(x^*) \\ & \text{or} \quad \langle x^*, x \rangle \geq h(x) + h^*(x^*). \end{aligned}$$

Thus

$$x^* \in \partial h(x) \iff \langle x^*, x \rangle = h(x) + h^*(x^*).$$

Likewise, if the requirements of the biconjugate theorem 4.1.11 are satisfied (h closed, convex, proper) then if $x \in \partial h^*(x^*)$ holds, the subgradient inequality yields

$$\begin{aligned} & h^*(y^*) - h^*(x^*) \geq \langle x, y^* - x^* \rangle \quad \text{for all } y^* \in X \\ \text{i.e.} \quad & \langle x, x^* \rangle - h^*(x^*) \geq \langle x, y^* \rangle - h^*(y^*) \quad \text{for all } y^* \in X \\ \text{equivalently} \quad & \langle x, x^* \rangle - h^*(x^*) \geq \sup_{y^*} \{\langle x, y^* \rangle - h^*(y^*)\} = h^{**}(x) \overset{\text{Biconjugate theorem 4.1.11}}{=} h(x) \\ & \text{or} \quad \langle x, x^* \rangle \geq h^*(x^*) + h(x). \end{aligned}$$

Thus if h is closed, convex, and proper, we have the equivalences:

$$x^* \in \partial h(x) \iff \langle x^*, x \rangle = h(x) + h^*(x^*) \iff x \in \partial h^*(x^*). \tag{FY}$$

When any of the equivalent conditions in (FY) holds, then we have equality throughout (FYI), and so

$$\begin{aligned} & h(x) + \overbrace{\sup_y \{\langle x^*, y \rangle - h(y)\}}^{h^*(x^*)} \overset{\text{(FY)}}{=} \langle x^*, x \rangle \\ \text{and so} \quad & \sup_y \{\langle x^*, y \rangle - h(y)\} = \langle x^*, x \rangle - h(x). \end{aligned}$$

In other words, x attains the supremum defining $h^*(x^*)$. When, in addition, h is closed, convex, and proper,

$$\overbrace{\sup_{y^*}\{\langle x, y^*\rangle - h^*(y^*)\}}^{h^{**}(x)} + h^*(x^*) \overset{\text{Theorem 4.1.11}}{=} h(x) + h^*(x^*)$$

$$\overset{\text{(FY)}}{=} \langle x^*, x\rangle$$

$$\text{and so} \quad \sup_{y^*}\{\langle x, y^*\rangle - h^*(y^*)\} = \langle x^*, x\rangle - h^*(x^*).$$

And so it is clear that x^* attains the supremum defining $f^{**}(x)$. We say that (FY), the "Fenchel Young equality," characterizes the elements in the subdifferential.

Problem set 6b

Problem 4.1.14 *Compute the conjugates of the functions in Problem 4.1.9.*

Proposition 4.1.15 *Suppose $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is a closed, proper convex function then ∂h has a closed graph i.e. for any $(x_n, x_n^*) \in \text{Graph } \partial f := \{(x, x^*) \mid x^* \in \partial f(x)\}$ with $(x_n, x_n^*) \to (x, x^*)$ we have $(x, x^*) \in \text{Graph } \partial f$ or $x^* \in \partial f(x)$.*

Proof. As f is lower semi-continuous and proper, so also is f^* (Lemma 4.1.10). And so

$$\liminf_n f(x_n) \geq f(x) \quad \text{and} \quad \liminf_n f^*(x_n^*) \geq f^*(x^*)$$

$$\text{implying} \quad \langle x^*, x\rangle = \liminf_n \langle x_n^*, x_n\rangle \overset{\text{(FY)}}{=} \liminf_n [f(x_n) + f^*(x_n^*)] \geq f(x) + f^*(x^*).$$

Here (FY) the Fenchel–Young equality $\langle x_n^*, x_n\rangle = f(x_n) + f^*(x_n^*)$, which holds for all n. Now we always have the Fenchel–Young inequality $f(x) + f(x^*) \leq \langle x^*, x\rangle$. Altogether we have equality $f(x) + f(x^*) = \langle x^*, x\rangle$ and so $x^* \in \partial f(x)$. □

Let us formalize the subgradient lemma, which we proved last time.

Lemma 4.1.16 (Subgradient Lemma) *Suppose $h : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is a closed, proper convex function then*

$$x^* \in \partial h(x) \quad \Leftrightarrow \quad x \in \partial h^*(x^*).$$

Proof. We have that

$$x^* \in \partial h(x) \overset{\text{(FY)}}{\Longleftrightarrow} \langle x^*, x\rangle = h(x) + h^*(x^*)$$

$$\overset{\text{biconjugate Theorem 4.1.11}}{=} h^{**}(x) + h^*(x^*) \overset{\text{(FY)}}{\Longleftrightarrow} x \in \partial h^*(x^*).$$

Here (FY) is the Fenchel–Young characterization of the subdifferential, in the first case for $h(x)$, and in the second case for $h^*(x)$. The equality $f(x) = f^{**}(x)$ (FM) is the Fenchel–Moreau theorem. □

There exists an extensive "calculus" for the conjugate operation.

Proposition 4.1.17 *Suppose h is a closed then*

$f = g^*$	$g = f^*$
$f(x)$	$g(x^*)$
$h(ax) \quad (a \neq 0)$	$h^*(\frac{x^*}{a})$
$h(x + b)$	$h^*(x^*) - \langle b, x^*\rangle$
$ah(x) \quad (a > 0)$	$ah^*(\frac{x^*}{a})$

Proof. Let us show that $f(x) = h(x+b)$ implies $f^*(x^*) = h^*(x^*) - \langle b, x^*\rangle$. This is via direct calculation

$$\begin{aligned} f^*(x^*) &= \sup_x \{\langle x^*, x\rangle - f(x)\} = \sup\{\langle x^*, x\rangle - h(x+b)\} \\ &= \sup_x \{\langle x^*, x+b\rangle - h(x+b)\} - \langle b, x^*\rangle \\ &= \sup_x \{\langle x^*, x\rangle - h(x)\} - \langle b, x^*\rangle = h^*(x^*) - \langle b, x^*\rangle. \end{aligned}$$

Now consider $f(x) := h(ax)$ with $a \neq 0$ then

$$\begin{aligned} f^*(x^*) &= \sup_x \{\langle x^*, x\rangle - f(x)\} \\ &= \sup_x \{\langle x^*, x\rangle - h(ax)\} = \sup_y \left\{\langle x^*, \frac{y}{a}\rangle - h(y)\right\} \quad (\text{with } y = ax) \\ &= \sup_y \left\{\langle \frac{x^*}{a}, y\rangle - h(y)\right\} = h^*\left(\frac{x^*}{a}\right). \end{aligned}$$

Finally for $f(x) = ah(x)$ we have

$$\begin{aligned} f^*(x^*) &= \sup_x \{\langle x^*, x\rangle - f(x)\} \\ &= \sup_x \{\langle x^*, x\rangle - ah(x)\} \\ &= a \sup_y \left\{\langle \frac{x^*}{a}, x\rangle - h(x)\right\} = ah^*\left(\frac{x^*}{a}\right). \end{aligned}$$

□

Lemma 4.1.18 *Let $q_A(v) = \frac{1}{2}v^T Av$ for $A \in \mathcal{P}(n)$. Then*

$$\nabla q_A(v) = Av \quad \textit{and} \quad \nabla^2 q_A(v) = A.$$

Note that this $n \times n$ result is consistent with the case where A is a 1×1 matrix (a real number), as we would expect. To see why it holds, notice that

$$\begin{aligned} \frac{1}{2}v^T Av &= \frac{1}{2}v^T \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n-1} & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n-1} & a_{2,n} \\ \vdots & \dots & \ddots & \vdots & \\ a_{n-1,1} & \vdots & \dots & a_{n-1,n-1} & a_{n-1,n} \\ a_{n,1} & a_{n,2} & \dots & a_{n,n-1} & a_{n,n} \end{pmatrix} \cdot v \\ &= \frac{1}{2}\begin{pmatrix} v_1a_{1,1} & +v_2a_{1,2} & +\dots & +v_{n-1}a_{1,n-1} & +v_na_{1,n} \\ v_1a_{2,1} & +v_2a_{2,2} & +\dots & +v_{n-1}a_{2,n-1} & +v_na_{2,n} \\ \vdots & \dots & \ddots & \vdots & \\ v_1a_{n-1,1} & +\vdots & +\dots & +v_{n-1}a_{n-1,n-1} & +v_na_{n-1,n} \\ v_1a_{n,1} & +v_2a_{n,2} & +\dots & +v_{n-1}a_{n,n-1} & +v_na_{n,n} \end{pmatrix} \cdot v \\ &= \frac{1}{2}\left(v_1\left(v_1a_{1,1} + v_2a_{1,2} + \dots + v_na_{1,n}\right) + \dots + v_n\left(v_1a_{n,1} + v_2a_{n,2} + \dots + v_na_{n,n}\right)\right). \end{aligned}$$

Therefore,

$$\frac{\partial}{\partial v_i}q_A(v) = \frac{1}{2}\left(2v_ia_{i,i} + 2\sum_{j\neq i} v_ja_{j,i}\right) = \sum_j v_ja_{j,i} = \sum_j v_j \overbrace{a_{i,j}}^{symmetric} = (Av)_i$$

$$\textit{and so} \quad (\forall i,j) \quad \frac{\partial}{\partial v_i}\frac{\partial}{\partial v_j} q_A(v) = a_{i,j}.$$

The first derivative $\nabla q_A(v)$ is the gradient: $((Av)_1, \ldots, (Av)_n)$. The second derivative is the hessian built out of the first and second partial derivatives; its (i,j)th component we can see is $a_{i,j}$.

Problem set 7: Convex Conjugates

Problem 4.1.19 *Let $q_A(x) = \frac{1}{2}x^T Ax$ where A is a real, positive definite square matrix.*

1. *Show that q_A is strictly convex.*
2. *Show that $q_A^*(y) = q_{A^{-1}}(y) := y^T A^{-1} y$.*
3. *Finally, show that when $A, B \in \operatorname{int} \mathcal{P}(n)$*

$$A \succeq B \quad \iff \quad B^{-1} \succeq A^{-1}.$$

Hints: These identities should be useful to you.

(a) *For showing strict convexity, Remark 4.1.18 and Lemma 1.1.6 may help.*

(b) *For computing the conjugate, we can find the point that maximizes the function $f : x \mapsto x \cdot y - \frac{1}{2}x^T Ax$ by taking the gradient with respect to x, setting it equal to zero, and solving for x. By plugging the maximizing vector back into f in place of x, we obtain $q_A^*(y)$. The gradient of $x \cdot y$ is y, while the gradient of q_A is in Lemma 4.1.18.*

(c) *Remember that for any convex functions f and g we have $f(x) \geq g(x)$ for all x implies $f^*(y) \leq g^*(y)$ for all y. You may also use the fact that $A \succeq B$ if and only if $x^T Ax \geq x^T Bx$ for all x.*

Problem 4.1.20 *Verify, with a derivation (i.e. don't just cite Remark 4.2.3), that the conjugate of the convex function f is the stated function f^*:*

$$f(x) = \|x\|_2 := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \quad \textit{with}$$
$$f^*(y) = \begin{cases} 0 & \textit{if} \quad \|y\|_2 \leq 1 \\ +\infty & \textit{otherwise} \end{cases} = \delta_{B_1}(y), \quad \textit{where } B_1 := \{y \mid \|y\|_2 \leq 1\}.$$

Hint: You can show f^* is the conjugate of f directly from the definition if you prefer, but there is a shorter way. Because f^* is closed and convex, it actually suffices by the biconjugate Theorem 4.1.11 to show that f is the conjugate of f^*. Remember that by Cauchy–Schwarz, $\langle x, y\rangle \leq \|x\|_2 \|y\|_2$.

Problem 4.1.21 *The infinimal convolution $f \square h$ of two convex functions f and h is defined to be*

$$(f \square h)(x) := \inf_z \left(f(z) + h(x - z)\right). \tag{4.11}$$

1. *Show that*

$$(f \square h)^*(x^*) = f^*(x^*) + h^*(x^*). \tag{4.12}$$

[Hint: You may want to use the fact that $\inf_z f(z) + h(x-z) = -\sup_z -f(z) - h(x-z)$.]

2. *Show that $g(x,z) = f(z) + h(x-z)$ is convex. Then, setting*

$$\phi(x) := \inf_z g(x,z) = (f \square h)(x),$$

use the results of example 1.2.3 to argue that $f \square h$ is convex but possibly not closed. Argue that when $f \square h$ is closed we must have

$$(f^* + h^*)^* (x) = (f \square h)(x).$$

3. *Use the biconjugate identity $f^{**} = f$ and $g^{**} = g$ for closed proper convex functions to establish*

$$(f+h)^* (x^*) = (f^* \square h^*)(x^*) \tag{4.13}$$

under the assumption that $f^* \square h^*$ is closed. [Interesting fact: A condition that ensures $f^* \square h^*$ is closed is the following $0 \in \operatorname{core}(\operatorname{dom} f - \operatorname{dom} g)$. This is proved in appendix E]

4. *Consider the special case of $f(x) = \delta_C(x)$ and $h(x) = \|x\|$ for some closed convex set C. Evaluate the right hand sides of (4.11) and (4.12).*

[Hint: remember from Example 4.1.3 that $\delta_C^* - S(C,\cdot)$ and remember from Problem 4.1.20 that $\|\cdot\|^* = \delta_{B_1}$ where B_1 denotes the unit ball.]

4.2 Fenchel Duality and Minimum Norm Problem

Recall that for arbitrary $x \in X$ and $x^* \in X^*$ we have the so called Fenchel inequality

$$h(x) + h^*(x^*) = h(x) + \sup_y \{\langle x^*, y\rangle - h(y)\} \geq h(x) + \langle x^*, x\rangle - h(x) = \langle x^*, x\rangle.$$

Also recall from (FY) that $x^* \in \partial f(x)$ if and only if $x \in \partial f^*(x^*)$ (i.e. both are equivalent to $\langle x^*, x\rangle = h(x) + h^*(x^*)$). This implies that $x \in \partial f^*(x^*)$ is the x which achieves the maximum in the definition of $f^*(x^*)$. Define the adjoint mapping of a linear function $A : X \to Y$ to be the linear mapping $A^* : Y^* \to X^*$ by

$$\langle Ax, y^*\rangle = \langle x, A^* y^*\rangle$$

for all $x \in X$ and $y^* \in Y^*$.

Theorem 4.2.1 ***(Fenchel duality and convex calculus)*** *Let X and Y be Banach spaces, $f : X \to \mathbf{R}_{+\infty}$ and $g : Y \to \mathbf{R}_{+\infty}$ be functions and $A : X \to Y$ a linear mapping. Let p and d be the primal and dual values defined by:*

$$p = \inf_{x \in X} \{f(x) + g(Ax)\} \quad \text{and}$$
$$d = \sup_{y^* \in Y^*} \{-f^*(A^* y^*) - g^*(-y^*)\}.$$

These values satisfy the weak duality inequality $p \geq d$. If, furthermore f and g are convex and satisfy the condition

$$0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f), \tag{4.14}$$

then the values are equal ($p = d$), and the supremum in the dual is attained if it is finite. If $p = d$ and x achieves this infimum in p and y^* achieves the supremum in d then

$$A^* y^* \in \partial f(x) \quad \text{and} \quad Ax \in \partial g^*(-y^*). \tag{∂FD}$$

Proof. The "weak duality" follows immediately from the Fenchel inequality i.e. for all $x \in X$ and $y^* \in Y^*$

$$\begin{aligned}&(f(x) + g(Ax)) - (-f^*(A^*y^*) - g^*(-y^*))\\ &= (f(x) + f^*(A^*y^*)) + (g(Ax) + g^*(-y^*)) \overset{\text{(FYI) twice}}{\geq} \langle x, A^*y^*\rangle - \langle Ax, y^*\rangle = 0.\end{aligned}$$

This shows

$$(f(x) + g(Ax)) \geq (-f^*(A^*y^*) - g^*(-y^*)).$$

Taking the infimum over all x on the left and supremum over all y^* on the right gives

$$p \geq d \tag{WD}$$

Now let's prove what we claimed in (∂FD) about what happens **if** we have strong duality. Well, if $p = d$ and x achieves this infimum in p and y^* achieves the supremum in d then:

$$f(x) + f^*(A^*y^*) = -(g(Ax) + g^*(-y^*)) \overset{-(\text{FYI})}{\leq} -\langle Ax, -y^*\rangle = \langle x, A^*y^*\rangle \tag{4.15}$$

Thus, by (FY) (subgradient lemma), we know $A^*y^* \in \partial f(x)$. Similarly we may show the second inclusion in (∂FD) by applying the Fenchel inequality to the left hand side of (4.15) like so:

$$g(Ax) + g^*(-y^*) = -(f(x) + f^*(A^*y^*)) \overset{-(\text{FYI})}{\leq} -\langle x, A^*y^*\rangle = \langle Ax, -y^*\rangle.$$

Thus, by (FY), we know $Ax \in \partial g^*(-y^*)$.

To prove the equality $p = d$ we define an optimal value function $h : Y \to [-\infty, +\infty]$ by

$$h(u) = \inf_{x \in E} \{f(x) + g(Ax + u)\}.$$

We first show h is convex. We can do this by showing that the strict epigraph $\operatorname{epi}_s h := \{(u, \alpha) \mid h(u) < \alpha\}$ is convex. Now $(u, \alpha) \in \operatorname{epi}_s h$ implies the existence of $x \in X$ such that $f(x) + g(Ax + u) < \alpha$. Therefore, when we take $(u_i, \alpha_i) \in \operatorname{epi}_s h$ for $i = 1, 2$ and $\lambda \in [0, 1]$, there exists x_i such that $f(x_i) + g(Ax_i + u_i) < \alpha_i$ and so

$$\lambda(f(x_1) + g(Ax_1 + u_1)) + (1 - \lambda)(f(x_2) + g(Ax_2 + u_2)) < \lambda\alpha_1 + (1 - \lambda)\alpha_2$$

and so

$$\begin{aligned}&f(\lambda x_1 + (1 - \lambda)x_2) + g(A(\lambda x_2 + (1 - \lambda)x_2) + (\lambda u_1 + (1 - \lambda)u_2))\\ &(\text{convexity of } f, g) \quad \leq \lambda(f(x_1) + g(Ax_1 + u_1)) + (1 - \lambda)(f(x_2) + g(Ax_2 + u_2))\\ &\qquad\qquad < \lambda\alpha_1 + (1 - \lambda)\alpha_2.\end{aligned}$$

Hence $h(\lambda u_1 + (1 - \lambda)u_2) < \lambda\alpha_1 + (1 - \lambda)\alpha_2$ and $(\lambda u_1 + (1 - \lambda)u_2, \lambda\alpha_1 + (1 - \lambda)\alpha_2) \in \operatorname{epi}_s h$, which is convex. Thus h is convex.

Now we show

$$\operatorname{dom} h = \operatorname{dom} g - A \operatorname{dom} f. \tag{4.16}$$

Let $u \in \operatorname{dom} h$. Then, from the definition of h, we know that for any $\epsilon > 0$, there exists $x \in \operatorname{dom} f$ such that

$$h(u) + \varepsilon \geq f(x) + g(Ax + u) \geq h(u).$$

Hence $Ax + u \in \operatorname{dom} g$, and so $u \in \operatorname{dom} g - Ax$. Combining with the fact that $x \in \operatorname{dom} f$, we obtain $u \in \operatorname{dom} g - A \operatorname{dom} f$. This shows $\operatorname{dom} h \subset \operatorname{dom} g - A \operatorname{dom} f$.

Now take $u \in \operatorname{dom} g - A \operatorname{dom} f$. Then there exists $x \in \operatorname{dom} f$ such that $u + Ax \in \operatorname{dom} g$. Consequently $f(x) + g(Ax + u) < \infty$, and so $u \in \operatorname{dom} h$. This shows $\operatorname{dom} h \supset \operatorname{dom} g - A \operatorname{dom} f$, so we have (4.16).

If $p = -\infty$ then weak duality forces $d = -\infty$, and so $p = d$. Now suppose p is finite valued. Having shown (4.16) is true, and having assumed that (4.14) is true, we have that $0 \in \operatorname{core} \operatorname{dom} h$. Thus $h(0) = p > -\infty$, and so h is proper (Lemma 4.1.8), and Theorem 4.1.7 tells us that there exists $y^* \in Y^*$ such that $-y^* \in \partial h(0)$. Thus for all $u \in Y$ and $x \in X$ we have

$$\begin{aligned} p = h(0) &\leq h(u) + \langle y^*, u - 0 \rangle \quad \text{(subgradient inequality)} \\ &\leq f(x) + g(Ax + u) + \langle y^*, u \rangle \quad \text{(definition of } h\text{)} \\ &= f(x) + g(Ax + u) + \langle y^*, u \rangle + \overbrace{(\langle y^*, Ax \rangle - \langle A^* y^*, x \rangle)}^{=0} \\ &= (f(x) - \langle A^* y^*, x \rangle) + (g(Ax + u) - \langle -y^*, Ax + u \rangle) \,. \end{aligned} \tag{4.17}$$

Notice that

$$\sup_u \{\langle -y^*, Ax + u \rangle - g(Ax + u)\} = \sup_x \{\langle -y^*, x \rangle - g(x)\} = g^*(-y^*). \tag{4.18}$$

From (4.17), it is clear that we may take the infimum over u to obtain:

$$\begin{aligned} p = h(0) &\leq (f(x) - \langle A^* y^*, x \rangle) + \inf_u (g(Ax + u) - \langle -y^*, Ax + u \rangle) \\ &= (f(x) - \langle A^* y^*, x \rangle) - \sup_u (\langle -y^*, Ax + u \rangle - g(Ax + u)) \\ &\overset{(4.18)}{=} (f(x) - \langle A^* y^*, x \rangle) - g^*(-y^*). \end{aligned}$$

Next, we may take the infimum over x to obtain

$$\begin{aligned} p = h(0) &\leq \inf_x (f(x) - \langle A^* y^*, x \rangle) - g^*(-y^*) \\ &= -\sup_x (\langle A^* y^*, x \rangle - f(x)) - g^*(-y^*) \\ &= -f^*(A^* y^*) - g^*(-y^*) \\ &\leq \sup_{x^*} (-f^*(A^* x^*) - g^*(-x^*)) = d \overset{(\mathrm{WD})}{\leq} p. \end{aligned}$$

We therefore have equality throughout, and so the "multiplier" y^* attains the supremum in problem d and $p = d$. □

Remark 4.2.1 *We could have weakened the assumption of this theorem by using the concept of intrinsic core (see Theorem 4.1.7). We then have the qualification assumption*

$$0 \in \operatorname{icr} (\operatorname{dom} g - A \operatorname{dom} f) = \operatorname{relint} (\operatorname{dom} g - A \operatorname{dom} f) \,. \tag{4.19}$$

The second equality holding when $0 \in \operatorname{core} (\operatorname{dom} g - A \operatorname{dom} f)$.

Remark 4.2.2 *A complete space with the inner product induced norm is called a Hilbert space. Letting $x^* \in X^*$ be a continuous linear functional, there exists (Reisz representation theorem) $z^* \in X$ so that*

$$x^*(x) = \langle z^*, x \rangle.$$

We can abuse notation and simply write $z^* = x^*$. Recall by Cauchy–Schwarz that $\langle x, x^* \rangle \leq \|x\| \|x^*\|$. We say that x, x^* are aligned if $\langle x, x^* \rangle = \|x\| \|x^*\|$. In $\mathbb{R}^n$ with the Euclidean norm, the intuitive meaning behind this terminology is clear, because alignment is equivalent to $x = \frac{\|x\|}{\|x^*\|} x^*$.

Example 4.2.1 *One important problem is the minimum norm problem:*

$$\inf\{\|x-r\| \mid x \in M\}$$

where M is a subspace of X, $x \neq r$, and the norm is induced by the inner product. This can be reformulated using Fenchel duality by using the indicator function

$$\delta_M(x) = \begin{cases} 0 & \text{if } x \in M \\ +\infty & \text{if } x \notin M \end{cases}.$$

Then we have $f(x) = \|x-r\|$ and $g(x) = \delta_M(x)$, and want to minimize $f+g$. As far as the interiority conditions go, we always have

$$0 \in \operatorname{core}(\operatorname{dom} g - \operatorname{dom} f) = \operatorname{core} X = X$$

since $\operatorname{dom} f = X$*, and so we have strong duality.*

Recall that $\langle x, x^*\rangle \leq \|x\|\,\|x^*\|$ with equality when x and x^* are aligned (i.e. $\langle x, x^*\rangle = \|x\|\,\|x^*\|$ if and only if x and x^* are aligned). Let $h(x) := \|x\|$ then

$$\begin{aligned} h^*(x^*) &= \sup_x \{\langle x, x^*\rangle - \|x\|\} \\ &= \sup_{\|x\|=K,\ K\in\mathbf{R}} \{\langle x, x^*\rangle - \|x\|\} \\ &= \sup_{K\in\mathbf{R}} \sup_{\|x\|=K} \{\langle x, x^*\rangle - \|x\|\} \\ &= \sup_{K\in\mathbf{R}} \{K\|x^*\| - K\} \\ &\quad \text{since the sup is attained when } x, x^* \text{ are aligned so } \langle x, x^*\rangle = \|x^*\|\|x\| \\ &= \sup_{K\in\mathbf{R}} K(\|x^*\|-1) = \begin{cases} +\infty & \text{if } \|x^*\| > 1 \\ 0 & \text{if } \|x^*\| \leq 1 \end{cases} = \delta_{B_1^*}(x^*) \end{aligned}$$

where $B_1^* = \{x^* \mid \|x^*\| \leq 1\}$. Note that for a given x^* we require x to be aligned with x^* to obtain the supremum defining h^*. Apply the calculus rule

f	f^*
$h(x+b)$	$h^*(x^*) - \langle b, x^*\rangle$

to $f(x) = h(x-r)$ with $b = -r$ to obtain

$$f^*(x^*) = h^*(x^*) - \langle b, x^*\rangle = \delta_{B_1}(x^*) + \langle r, x^*\rangle.$$

Notice that that for a given x^* we need $x-r$ to be aligned with x^* in order for it to obtain the supremum defining f^*. Denote $M^\perp := \{x^* \mid \langle x, x^*\rangle = 0 \text{ for all } x \in M\}$. Then

$$\begin{aligned} g^*(x^*) &= \sup_x \{\langle x, x^*\rangle - \delta_M(x)\} \\ &= \sup_{x\in M} \langle x, x^*\rangle \qquad (4.20) \\ &= S(M, x^*) \overset{(a)}{=} \begin{cases} 0 & \text{if } x^* \in M^\perp \\ +\infty & \text{if } x^* \notin M^\perp \end{cases} = \delta_{M^\perp}(x^*). \qquad (4.21) \end{aligned}$$

The equality (a) is just Problem 3.1.7 applied to the case where $M = \ker A$. If you need your memory jogged, just notice that the $+\infty$ comes about due to the fact that $x \in M$ implies $-x \in M$ so when $x^* \notin M^\perp$ there exists $x \in M$ with $\langle x, x^*\rangle = \delta > 0$. As $\gamma x \in M$

for all $\gamma > 0$ we have $\sup_{x\in M}\langle x, x^*\rangle \geq \sup_{\gamma>0}\gamma\delta = +\infty$. Putting this together (with $A = I$, $A^* = I$ and $Y = X$) we get

$$\begin{aligned}
p :&= \inf\{\|x - r\| \mid x \in M\}\\
&= \inf\{f(x) + g(x)\}\\
&= \sup_{x^*\in X^*}\{-f^*(x^*) - g^*(-x^*)\} \quad (= d)\,\textit{(Strong duality: Theorem 4.2.1)}\\
&= -\min_{x^*\in X^*}\left(\overbrace{\delta_{B_1^*}(x^*) + \langle r, x^*\rangle}^{f^*(x^*)} + \overbrace{\delta_{M^\perp}(-x^*)}^{g^*(-x^*)}\right)\\
&= -\min_{x^*\in B_1\cap M^\perp}\langle r, x^*\rangle \quad (\textit{obviously } \delta_{M^\perp}(x^*) = \delta_{M^\perp}(-x^*))\\
&= \max_{x^*\in B_1\cap M^\perp}\langle r, -x^*\rangle = \max_{x^*\in B_1\cap M^\perp}\langle r, x^*\rangle = d
\end{aligned}$$

When $p = \inf\{\|x - r\| \mid x \in M\}$ is finite then the maximum defining d is achieved at some $x^* \in B_1^* \cap M^\perp$, wherefore this x^* should be of unit length. We have already noted that if x achieves the infimum defining p then $\langle x - r, x^*\rangle = \|x^*\|\,\|x - r\|$, wherefore x^* is aligned with $x - r$.

Let us do a sanity check. We already know, from (∂FD), that when x^* attains the supremum in d and x attains the infimum in p, we must have

$$x^* \in \partial f(x).$$

However, since $x \neq r$, we can differentiate f, and obtain $\nabla f(x) = (x - r)/\|x - r\|$, and so any point in $\nabla f(x)$ will be of unit length and aligned with $x - r$. This is consistent.

Remark 4.2.3 *In the last example, we implicitly proved that the Euclidean norm is dual to δ_B, where B, where B is the unit ball. Because of the biconjugate theorem 4.1.11 and example 4.1.3, it must follow that $\|\cdot\| = S(B, \cdot)$.*

We briefly consider some calculus rules and begin by noting that $y^* \in \partial\left(f(\cdot) + \langle x^*, \cdot\rangle\right)(x)$ is equivalent to

$$\left(f(\cdot) + \langle x^*, \cdot\rangle\right)^*(y^*) + f(x) + \langle x^*, x\rangle \overset{(\mathrm{FY})}{=} \langle y^*, x\rangle. \tag{4.22}$$

Now use the calculus in Proposition 4.1.17 to obtain

$$\left(f + \langle x^*, \rangle\right)^*(y^*) = f^*(y^* - x^*).$$

Thus (4.22) is equivalent to

$$f^*(y^* - x^*) + f(x) = \langle y^* - x^*, x\rangle,$$

which, by (FY), is equivalent to

$$y^* - x^* \in \partial f(x) \quad \text{or} \quad y^* \in \partial f(x) + x^*$$

$$\text{and so we have shown} \quad \partial\left(f(\cdot) + \langle x^*, \cdot\rangle\right)(x) = \partial f(x) + x^*.$$

The Fenchel duality theorem implies the following important calculus rule for the subdifferential. A proof may be found in Appendix B.

Theorem 4.2.2 *Suppose $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ and $g : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ and $A : X \to Y$ be a linear mapping. At any point $x \in X$ the calculus rule*

$$\partial\left(f + g \circ A\right)(x) \supseteq \partial f(x) + A^*\partial g\left(Ax\right)$$

with equality holding when $0 \in \operatorname{core}(\operatorname{dom} g - A \operatorname{dom} f)$.

Proof. Appendix B. □

4.3 Examples of Fenchel Duality

Exercise 4.3.1 *Use the Fenchel duality theorem*

$$\begin{aligned} &\inf_x f(x) \\ &\textit{Subject to } Ax \leq b \quad (A \textit{ is } m \times n) \end{aligned} \tag{Prim}$$

to show that the dual to the Primal problem is:

$$\begin{aligned} &\max_{\eta \in \mathbf{R}^m} \quad \left\{-f^*(A^*(-\eta)) - b^T\eta\right\} \\ &\textit{Subject to} \quad A^*(-\eta) \in \operatorname{dom} f^*, \quad \eta \geq 0. \end{aligned} \tag{Dual}$$

where A^* is the adjoint of the linear mapping A defined by $x \mapsto (\langle a_1, x\rangle, \ldots, \langle a_1, x\rangle)$ from $\mathbf{X}$ into $\mathbf{R}^m$.

Soln: We formulate this as

$$\inf\left\{f(x) + \delta_{b-\mathbf{R}^m_+}(Ax)\right\}$$

and so $g(y) := \delta_{b-\mathbf{R}^m_+}(y)$. Then

$$\begin{aligned} g^*(y^*) &= \sup_y \left\{\langle y, y^*\rangle - \delta_{b-\mathbf{R}^m_+}(y)\right\} \\ &= \sup_{y \in b-\mathbf{R}^m_+} \{\langle y, y^*\rangle\} \\ &= b^T y^* + \sup_{y \in -\mathbf{R}^m_+} \langle y, y^*\rangle \\ &= b^T y^* + S\left(-\mathbf{R}^m_+, y^*\right) \overset{(b)}{=} b^T y^* + \delta_{\mathbf{R}^m_+}(y^*). \end{aligned}$$

Here (b) holds because $\mathbf{R}^m_+ = (-\mathbf{R}^m_+)^\circ$ and when C is a closed convex cone,

$$S(C, y^*) = \sup_{c \in C}\langle c, y^*\rangle = \begin{cases} 0 & \text{if } y^* \in C^\circ \\ \infty & \text{otherwise.} \end{cases} = \delta_{C^\circ}(y^*). \tag{4.23}$$

The Fenchel dual is

$$\sup_{y^*}\{-f^*(A^*(y^*)) - g^*(-y^*)\}$$

Letting $\eta = -y^*$ we can rewrite the dual problem as

$$\sup_\eta \left\{-f^*(A^*(-\eta)) - b^T\eta - \delta_{\mathbf{R}^m_+}(\eta)\right\} = \sup_{\eta \in \mathbf{R}^m_+}\left\{-f^*(A^*(-\eta)) - b^T\eta\right\}.$$

Here, we can see that we must have the implicit constraint that $A^*(-\eta) \in \operatorname{dom} f^*$, because otherwise $-f^*(A^*(y^*)) = -\infty$.

Exercise 4.3.2 *Use Fenchel duality to show that the dual of*

$$\inf_x f(x) \quad \textit{subject to} \quad Ax \leq b \quad (A \textit{ is } m \times n) \quad \textit{and} \quad Bx = d \quad (B \textit{ is } k \times n)$$

is given by

$$\begin{aligned} &\max_{\lambda \in \mathbf{R}^{m+k}} \quad \left\{-f^*(A^*(-\lambda_1) + B^*(-\lambda_2)) - b^T\lambda_1 - d^T\lambda_2\right\} \\ &\textit{Subject to} \quad (A^*(-\lambda_1) + B^*(-\lambda_2)) \in \operatorname{dom} f^*, \quad \lambda_1 \geq 0. \end{aligned}$$

where A^* is the adjoint of the linear mapping of A defined by $x \mapsto (\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle)$ from $\mathbf{X}$ into $\mathbf{R}^m$ and B^* is the adjoint linear mapping of B defined by $x \mapsto (\langle b_1, x\rangle, \ldots, \langle b_k, x\rangle)$ from $\mathbf{X}$ into $\mathbf{R}^k$.

Soln: As before we formulate this as

$$\inf\left\{f(x)+\delta_{\left\{b-\mathbf{R}^m_+\right\}\times\{d\}}(Ax,Bx)\right\} \tag{4.24}$$

and we have $g(y_1,y_2):=\delta_{\left\{b-\mathbf{R}^m_+\right\}\times\{d\}}(y_1,y_2)$. Then

$$\begin{aligned}
g^*(\lambda_1,\lambda_2) &= \sup_{(y_1,y_2)}\left\{\langle(y_1,y_2),(\lambda_1,\lambda_2)\rangle-\delta_{\left\{b-\mathbf{R}^m_+\right\}\times\{d\}}(y_1,y_2)\right\}\\
&= \sup_{(y_1,y_2)\in\left\{b-\mathbf{R}^m_+\right\}\times\{d\}}\{\langle(y_1,y_2),(\lambda_1,\lambda_2)\rangle\}\\
&= \sup_{y_0\in-\mathbf{R}^m_+}\left\{\langle y_0,\lambda_1\rangle+b^T\lambda_1+d^T\lambda_2\right\}\\
&\quad \text{since the sup must have } y_2=d \text{ and } y_1=b-y_0 \text{ for some } y_0\in\mathbf{R}^m_+\\
&= S\left(-\mathbf{R}^m_+,\lambda_1\right)+b^T\lambda_1+d^T\lambda_2 = b^T\lambda_1+d^T\lambda_2+\delta_{\mathbf{R}^m_+}(\lambda_1).
\end{aligned}$$

The final equality uses the identity (4.23) described in the previous example.
We also need the adjoint of the linear mapping $(A,B)\,x: X\to R^{m+k}$ which we find by considering

$$\begin{aligned}
\langle(A,B)\,x,(\lambda_1,\lambda_2)\rangle &= \langle(Ax,Bx),(\lambda_1,\lambda_2)\rangle\\
&= \langle Ax,\lambda_1\rangle+\langle Bx,\lambda_2\rangle\\
&= \langle x,A^*\lambda_1\rangle+\langle x,B^*\lambda_2\rangle\\
&= \langle x,A^*\lambda_1+B^*\lambda_2\rangle
\end{aligned}$$

and so

$$(A,B)^*(\lambda_1,\lambda_2)=A^*\lambda_1+B^*\lambda_2.$$

The Fenchel dual of (4.24) is given by

$$\sup_{(\lambda_1,\lambda_2)}\left\{-f^*\left((A,B)^*(-\lambda_1,-\lambda_2)\right)-g^*(\lambda_1,\lambda_2)\right\}$$

where the adjoint of $x\mapsto(Ax,Bx)\in\mathbf{R}^{m+k}$ is given by $(A,B)^*:\mathbf{R}^{m+k}\to X^*$. Thus our dual problem becomes

$$\sup_{(\lambda_1,\lambda_2)}\left\{-f^*\left(-A^*\lambda_1-B^*\lambda_2\right)-b^T\lambda_1-d^T\lambda_2-\delta_{\mathbf{R}^m_+}(\lambda_1)\right\}$$

This is just

$$\begin{aligned}
&\sup_{(\lambda_1,\lambda_2)}\left\{-f^*\left(-A^*\lambda_1-B^*\lambda_2\right)-b^T\lambda_1-d^T\lambda_2\right\}\\
&\text{Subject to}\quad \left(A^*(-\lambda_1)+B^*(-\lambda_2)\right)\in\operatorname{dom}f^*,\quad \lambda_1\geq 0.
\end{aligned}$$

Problem Set 8

Problem 4.3.1 *Suppose $A:\mathbf{R}^m\to\mathbf{R}^n$. Show that the dual of the following problem:*

$$J:=\inf_{(u,\gamma)}\sum_{k=0}^{m}\gamma_k$$

$$\begin{aligned} \textit{subject to } & u+\gamma \geq 0 \\ & -u+\gamma \geq 0 \\ & Au = c,\ \gamma \geq 0 \end{aligned}$$

is given by (with $e = (1, 1, 1, \ldots, 1)$*)*

$$J^* := \max_{(x,y,z)} \left\{ \langle 0, x\rangle + \langle 0, y\rangle + \langle c, z\rangle \mid \begin{array}{l} x - y - A^* z = 0 \\ x + y \leq e \text{ and } x, y \geq 0 \end{array} \right\}$$

[Hint: Show that $\operatorname{dom} f^* = \{(0, e)\}$ *and use* $\lambda_1 = (x, y, w) \geq 0$ *and* $\lambda_2 = -z$ *only to eliminate* w *in the end.]*

4.4 Lagrangian Duality

We consider the following optimisation problem

$$\inf \{ f(x) \mid g(x) \leq 0,\ x \in E \} \tag{4.25}$$

where f, and the component functions $g_1, g_2, \ldots, g_m : E \subseteq \mathbf{R}^n \to \overline{\mathbf{R}}$ are all convex and satisfy $\emptyset \neq \operatorname{dom} f \subseteq \cap_i \operatorname{dom} g_i$ and $g : x \mapsto (g_1(x), \ldots, g_m(x))$. Define the Lagrangian as

$$L(x, \lambda) = f(x) + \lambda^T g(x) : E \times \mathbf{R}^m_+ \to \overline{\mathbf{R}}$$

and note that if $C = \cap_x \{x \mid g(x) \leq 0\}$ then

$$\begin{aligned} \sup_{\lambda \geq 0} L(x, \lambda) &= \left\{ \begin{array}{cc} f(x) + 0^T g(x) & \text{if } g(x) \leq 0 \\ \lim_{\lambda \to \infty} \left[f(x) + \lambda^T g(x) \right] = \infty & \text{otherwise} \end{array} \right. \\ &= \left\{ \begin{array}{cc} f(x) & \text{if } x \text{ is feasible} \\ +\infty & \text{otherwise} \end{array} \right. = f(x) + \delta_C(x). \end{aligned}$$

Thus the solution to the primal problem is

$$\inf_{x \in E} \sup_{\lambda \geq 0} L(x, \lambda) = \inf_{x \in E} \left[f(x) + \delta_C(x) \right] = p.$$

4.4.1 Intuition behind Lagrange Multipliers

$$\text{minimize}_x\ x^2 \ \text{ subject to } \ x \geq 3$$

Then $\min_x L(x, \lambda) = x^2 + \lambda(3 - x)$ is attained when (differentiating w.r.t. x and solving) $x^* = \lambda/2$. Then $\max_\lambda L(x^* = \lambda/2, \lambda) = (\lambda/2)^2 + \lambda(3 - (\lambda/2)) = -(1/4)\lambda^2 + 3\lambda$ which is maximized at (differentiating and solving) $\lambda^* = 6 \in \partial f(3)$. The function $x^2 + 6(3 - x)$ is minimized (differentiate and solve) $x = 3$, which is $x = \lambda/2$. The idea we are after is essentially that (under certain assumptions):

1. $L(x, \lambda^*)$ (where λ^* is the dual solution) gives us a function whose unconstrained minimum is equal to the minimum of the original constrained problem.

2. λ is called a Lagrange multiplier. In vector calculus, you may have seen the projection x^* of a point w onto an ellipse $C = \{x \mid g(x) = (x_1/a)^2 + (x_2/b)^2 - 1 \leq 0\}$ satisfies $\nabla f(x^*) + \lambda^* \nabla g(x^*) = 0$ where $f = \| \cdot - w\|^2$ and $\lambda^* \geq 0$. Of course, if $w \in C$, then $\lambda^* = 0$. In essence, this new idea of Lagrange multiplier is just an extension of an old idea that you have seen before: there is a function g_i whose gradient along points on the boundary of $\text{lev}_{\leq 0} g$ points directly away from our feasible set.

   - If w is not in C, then λ^* will have (some) nonzero values, and $\lambda_i > 0$ implies that g_i is an "active" constraint (meaning the solution lives on the boundary of that constraint set rather than properly inside it).
   - Often, $g(x) = Ax - b$. For example, if $C = \mathbb{R}^2_+$ is the non-negative orthant, then $x \in C \iff Ax \leq 0$ with $A = -I, b = 0$. Then $Ax = [-x_1, -x_2] = [g_1(x), g_2(x)]$. The projection of $(1, -1)$ onto C minimizes $f(x) = \|x - (1, -1)\|^2$ subject to $g(x) \leq 0$. We have $x^* = (1, 0)$ and $\lambda^* = (0, 2)$ which satisfies $\underbrace{\nabla_x f(x^*)}_{=(0,2)} + \underbrace{\nabla_x(\lambda_1^* g_1 + \lambda_2^* g_2)(x^*)}_{=\nabla_x[2(-x_2)](x^*)=(0,-2)} = 0$.

4.4.2 The Lagrangian dual problem

This make us consider the dual problem

$$d = \sup_{\lambda \geq 0} \inf_{x \in E} L(x, \lambda),$$

where $d \in [-\infty, +\infty]$ is called the value of the dual problem. Let us show that $d \leq p$. First set

$$\Phi(\lambda) := \inf_{x \in E} L(x, \lambda).$$

Clearly, for any $\lambda \geq 0$, we have

$$\begin{aligned} & & L(x, \lambda) &\leq \sup_{\lambda \geq 0} L(x, \lambda) \\ \text{and so} & & \Phi(\lambda) = \inf_{x \in E} L(x, \lambda) &\leq \inf_{x \in E} \sup_{\lambda \geq 0} L(x, \lambda) = p \\ \text{and so} & \quad d = \sup_{\lambda \geq 0} \Phi(\lambda) = & \sup_{\lambda \geq 0} \inf_{x \in E} L(x, \lambda) &\leq \inf_{x \in X} \sup_{\lambda \geq 0} L(x, \lambda) \leq p, \end{aligned}$$

the so called weak duality. To study the duality gap $p - d$, we need to study the optimal value function

$$v(b) := \inf \{f(x) \mid x \in X \text{ such that } g(x) \leq b\} : \mathbf{R}^m \to \overline{\mathbf{R}}. \tag{4.26}$$

We will be interested in the λ^* such that $-\lambda^* \in \partial v(0)$. In essence, this $-\lambda^*$ tells us how our optimal value changes as our resource constraints change.

Value functions (intuition). At the optimal primal solution, my (instantaneous) change in my optimal value with a unit change in resource x_i is λ_i^*. For example, if minimizing x^2 subject to $-x + 3 \leq 0 = b$ (i.e. $x \geq 3$) then $v(1)$ gives us the optimal solution with the constraint $x \geq 2$, and so $v(0) = 9, v(1) = 4, v(2) = 1, v(3) = 0, v(4) = 0$, and $v(b) = \text{dist}^2_{\{u \mid u \geq 3\}}(b)$. And $\partial v(0) = -6 = -\lambda^*$, which tells me that at the optimal primal solution 3, my (instantaneous) change in my optimal value with a unit change in resource x_1 is $\lambda_1^* = 6$. Thus, the dual problem is, in a sense, to value the resources $x_i, \ldots, x_n$ according to their ability to contribute (in an instantaneous sense) to the optimal profit.

Lemma 4.4.1 *The function $v : \mathbf{R}^m \to \overline{\mathbf{R}}$ as defined in (4.26) is convex.*

Proof. We show that $\operatorname{epi}(v)$ is a convex subset of $\mathbf{R}^{m+1}$. Let $\varepsilon > 0$ be arbitrary and $(b_1, \alpha_1), (b_2, \alpha_2) \in \operatorname{epi} v$ then we have the existence of $x_1, x_2 \in X$ such that for $j = 1, 2$ we have: $g_i(x_j) \leq (b_j)_i$ for each i, and $f(x_j) \leq v(b_j) + \varepsilon \leq \alpha_j + \varepsilon$. That is $(x_j, \alpha_j + \varepsilon) \in \operatorname{epi}(f)$ and $(x_j, (b_j)_i) \in \operatorname{epi}(g_i)$ for $j = 1, 2$ and $i = 1, \ldots, m$. As $f, g_1, \ldots, g_m$ are convex we have for any $\lambda \in [0, 1]$ that

$$\begin{aligned} &\lambda(x_1, \alpha_1 + \varepsilon) + (1 - \lambda)(x_2, \alpha_2 + \varepsilon) \\ &= (\lambda x_1 + (1 - \lambda)x_2, \lambda\alpha_1 + (1 - \lambda)\alpha_2 + \varepsilon) \in \operatorname{epi}(f) \end{aligned}$$

and

$$\begin{aligned} &\lambda(x_1, (b_1)_i) + (1 - \lambda)(x_2, (b_2)_i) \\ &= (\lambda x_1 + (1 - \lambda)x_2, \lambda(b_1)_i + (1 - \lambda)(b_2)_i) \in \operatorname{epi}(g_i) \end{aligned}$$

for $i = 1, \ldots, m$. Hence

$$\begin{aligned} f(\lambda x_1 + (1 - \lambda)x_2) &\leq \lambda\alpha_1 + (1 - \lambda)\alpha_2 + \varepsilon \\ \text{and } g_i(\lambda x_1 + (1 - \lambda)x_2) &\leq \lambda(b_1)_i + (1 - \lambda)(b_2)_i \quad \text{for } i = 1, \ldots, m. \end{aligned}$$

Thus since $\lambda x_1 + (1 - \lambda)x_2$ is feasible (satisfies $g(\ldots) \leq \lambda b_1 + (1 - \lambda)b_2$):

$$\begin{aligned} v(\lambda b_1 + (1 - \lambda)b_2) &\leq f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda\alpha_1 + (1 - \lambda)\alpha_2 + \varepsilon \\ \text{and so } v(\lambda b_1 + (1 - \lambda)b_2) &\leq \lambda\alpha_1 + (1 - \lambda)\alpha_2 + \varepsilon \quad \text{for all } \varepsilon > 0. \end{aligned}$$

Taking ε to be arbitrarily small yields

$$v(\lambda b_1 + (1 - \lambda)b_2) \leq \lambda\alpha_1 + (1 - \lambda)\alpha_2.$$

Thus $(\lambda b_1 + (1 - \lambda)b_2, \lambda\alpha_1 + (1 - \lambda)\alpha_2) \in \operatorname{epi}(v)$. This shows the convexity. □

Note that

$$v(0) = \inf\{f(x) \mid g(x) \leq 0\} = p.$$

Theorem 4.4.2 (Dual optimal value) *For the function*

$$v(b) = \inf_x \{f(x) \mid g(x) \leq b\},$$

the following hold.

(i) The primal optimal value p is $v(0)$.

(ii) The conjugate of the value function satisfies

$$v^*(-\lambda) = -\Phi(\lambda) + \delta_{\mathbf{R}^m_+}(\lambda) = \begin{cases} -\Phi(\lambda) & \text{if} \quad \lambda \geq 0 \\ +\infty & \text{otherwise} \end{cases} \tag{4.27}$$

(iii) The dual optimal value d is $v^{**}(0)$.

(iv) Suppose $v(0) = p < \infty$. Then $p = d$ iff $b \mapsto v(b)$ is lower semi–continuous at $b = 0$. In this case the set of optimal dual solutions λ is given by $-\partial v(0)$.

Proof.

(i): This is clear.

(ii): Notice

$$\begin{aligned}
v^*(-\lambda) &= \sup_{b\in\mathbf{R}^m}\{-\lambda^T b - v(b)\} \\
&= \sup_{b\in\mathbf{R}^m}\{-\lambda^T b + \sup_{g(x)\le b}(-f(x))\} \\
&= \sup\left\{-\lambda^T b - f(x) \mid g(x)+z=b,\, x\in\operatorname{dom} f,\, b\in\mathbf{R}^m,\, z\in\mathbf{R}^m_+\right\} \\
&= \sup\left\{-\lambda^T(g(x)+z) - f(x) \mid x\in\operatorname{dom} f,\, z\in\mathbf{R}^m_+\right\} \\
&= -\underbrace{\inf\left\{f(x)+\lambda^T g(x) \mid x\in\operatorname{dom} f\right\}}_{\Phi(\lambda)} + \underbrace{\sup\left\{-\lambda^T z \mid z\in\mathbf{R}^m_+\right\}}_{\delta_{\mathbf{R}^m_+}(\lambda)} \\
&= \begin{cases} -\Phi(\lambda) & \text{if} \quad \lambda\ge 0 \\ +\infty & \text{otherwise} \end{cases}.
\end{aligned}$$

(iii): Observe that

$$\begin{aligned}
d \overset{\text{def}}{=} \sup_{\lambda\in\mathbf{R}^m_+}\Phi(\lambda) &= -\inf_{\lambda\in\mathbf{R}^m_+}\{-\Phi(\lambda)\} \\
&= -\inf_{\lambda}(-\Phi(\lambda)+\delta_{\mathbf{R}^m_+}(\lambda)) \quad =: \text{“dual problem”} \\
&= -\inf_{\lambda} v^*(-\lambda) \quad \text{(using (ii))} \\
&= \sup_{\lambda}\{\underbrace{-\lambda^T 0}_{\text{adding zero}} - v^*(-\lambda)\} = v^{**}(0).
\end{aligned}$$

Thus $p=d$ exactly when $v(0)=v^{**}(0)$.

(iv): By the biconjugate (Fenchel–Moreau) Theorem 4.1.11 this condition occurs when $\lambda\mapsto v(\lambda)$ is lower semi–continuous at $\lambda=0$. The dual problem is to minimize $-\Phi(\lambda)+\delta_{\mathbf{R}^m_+}(\lambda)$. A characterization of the minimizer of a convex function is that the subgradient of the function at the minimizer contains zero, because the subgradient inequality becomes

$$(\forall x)\quad h(x)-h(0)\ge\langle 0, x-0\rangle = 0.$$

Draw a picture and make sure you understand. With this in mind, the minimizing λ for our dual problem should satisfy this condition:

$$0\in\partial\left(-\Phi+\delta_{\mathbf{R}^m_+}\right)(\lambda) = \partial v^*(-\lambda).$$

We call this a *stationarity condition*. By the subgradient Lemma 4.1.16, we can swap this condition for the equivalent one:

$$-\lambda\in\partial v(0). \tag{4.28}$$

□

When we have no duality gap (i.e. $p=d$) how does this help us solve the original problem (4.9)?

Theorem 4.4.3 *Suppose $\bar\lambda\ge 0$ solves the dual problem then if $\bar x\in E$ solves the original problem (4.9) then $\bar x$ also solves $\min_x L(x,\bar\lambda)$. In particular when $0\in\operatorname{core}(\operatorname{dom} g-\operatorname{dom} f)$ then $\bar x$ satisfies*

$$0\in\partial f(\bar x)+\bar\lambda^T\partial g(\bar x). \tag{4.29}$$

Proof. First note that

$$f(x) \geq \inf\{f(z) \mid g(z) \leq g(x)\} = v(g(x)). \tag{4.30}$$

We have assumed that $\bar{\lambda}$ solves the dual problem, and so $-\bar{\lambda} \in \partial v(0)$ holds (as described above (4.28)). Thus we have the subgradient inequality:

$$v(g(x)) - v(0) \geq \langle -\bar{\lambda}, g(x) - 0 \rangle = -\bar{\lambda}^T g(x) \tag{4.31}$$

For all x it follows that:

$$f(x) - p \geq \overbrace{v(g(x))}^{f(x) \geq \text{ by (4.30)}} - \overbrace{v(0)}^{=p} \overset{(4.31)}{\geq} -\bar{\lambda}^T g(x)$$

$$\text{or} \qquad f(x) + \bar{\lambda}^T g(x) \geq v(0) = p \overset{(a)}{=} f(\bar{x}). \tag{4.32}$$

The final equality (a) is because $\bar{x}$ is assumed to solve the primal problem.

If we take the above inequality and choose $x = \bar{x}$, then the f terms cancel each other and we have $\bar{\lambda}^T g(\bar{x}) \geq 0$. However, as $\bar{\lambda} \geq 0$ and $g(\bar{x}) \leq 0$ we must also have $\bar{\lambda}^T g(\bar{x}) \leq 0$. Therefore, identically:

$$\bar{\lambda}^T g(\bar{x}) = 0.$$

Consequently, we may add it to the right hand side in (4.32) and obtain

$$f(x) + \bar{\lambda}^T g(x) \geq f(\bar{x}) + \underbrace{\bar{\lambda}^T g(\bar{x})}_{=0}. \tag{4.33}$$

This is just

$$L(x, \lambda) \overset{\text{def}}{=} f(x) + \bar{\lambda}^T g(x) \geq f(\bar{x}) + \bar{\lambda}^T g(\bar{x}) = L\left(\bar{x}, \bar{\lambda}\right)$$

and so $\bar{x}$ solves $\min_x L\left(x, \bar{\lambda}\right)$. Thus $0 \in \partial L\left(\cdot, \bar{\lambda}\right)(\bar{x})$ and when $\bar{x} \in \operatorname{core}(\operatorname{dom} g - \operatorname{dom} f)$ we can use the calculus rule (Theorem 4.2.2) to obtain

$$0 \in \partial L\left(\cdot, \bar{\lambda}\right)(\bar{x}) = \partial f(\bar{x}) + \bar{\lambda}^T \partial g(\bar{x}).$$

□

Remark 4.4.1 *of course when all functions f, g_i $i = 1, \ldots, m$ are differentiable then (4.29) is equivalent to*

$$0 = \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x}).$$

Also $0 \in \operatorname{core}(\operatorname{dom} g - \operatorname{dom} f)$ is implied by

$$\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g \neq \emptyset. \tag{4.34}$$

Indeed if $x \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} g$ then

$$0 \in \operatorname{int}(\operatorname{dom} g - x) \subseteq \operatorname{int}(\operatorname{dom} g - \operatorname{dom} f) \subseteq \operatorname{core}(\operatorname{dom} g - \operatorname{dom} f).$$

In order to show $p = d$ we need assume a kind of interiority condition for the constraint set.

Definition 4.4.1 *The Slater constraint qualification holds for (P) if there exists $\hat{x} \in \operatorname{dom}(f)$ with $g_i(\hat{x}) < 0$ for all $i = 1, \ldots, m$.*

The Slater CQ implies (4.34) when all g_i are continuous at $\hat{x}$. To see why, notice that the continuity of each g_i at $\hat{x}$ implies by definition that each of the g_i are bounded above on a neighborhood of $\hat{x}$, whereupon $\hat{x} \in \operatorname{int} \operatorname{dom} g$.

Theorem 4.4.4 (Dual Attainment) *Assume that the Slater constraint qualification holds. Then $p = d$ and the value is finite.*

Proof. If $p = -\infty \geq d$ then $p = d$. Thus we may assume now that p is finite. Next note that as there exists $g_i(\hat{x}) < 0$ with $f(\hat{x}) < +\infty$ we deduce that for any $b \in \mathbf{R}^m$ and t sufficiently small we have $tb \geq g_i(\hat{x})$ and so

$$v(tb) = \inf_{g(x) \leq tb} f(x) \leq f(\hat{x}) < +\infty.$$

Thus $t0 + t(b - 0) = tb \in \operatorname{dom}(v)$. That is $0 \in \operatorname{core} \operatorname{dom}(v)$ so Lemma 4.1.8 applies, so we know that v never takes the value $-\infty$. We may now apply the *nonempty subgradient* Theorem 4.1.7 to deduce that $\partial v(0) \neq \emptyset$. Consequently every $-\lambda \in \partial v\,(0)$ is an optimal solution of the dual problem and hence v is finite and (in view of Theorems 4.4.2 and 4.4.3) lower semi–continuous at 0. Thus $p = d$. □

The last result assures us that the dual problem has an optimal solution, but we do not yet know if the primal solution is attained.

Definition 4.4.2 (Level set) *We denote the r-level set:*

$$\operatorname{lev}_r h := \{x \mid h(x) \leq r\}.$$

Theorem 4.4.5 *Suppose that all functions f, $g_1, \ldots, g_m$ are all closed convex functions of $\mathbf{R}^m$ and that there exists a vector of non-negative multipliers $\left(\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_m\right)$ such that for all r we have*

$$\left\{x \mid \hat{\lambda}_0 f\,(x) + \lambda^T g\,(x) \leq r\right\} = \operatorname{lev}_r \left(\hat{\lambda}_0 f + \lambda^T g\right)$$

is bounded. Then v is closed and the infimum

$$v(b) = \inf\{f(x) \mid x \in X,\ g(x) \leq b\}$$

is attained when $v\,(b)$ is finite. Consequently when all the functions are convex and $d > -\infty$ then $p = d$ and the primal solution is attained.

Proof. To show v is closed take $(b_k, s_k) \in \operatorname{epi} v$ with $(b_k, s_k) \to (b, s)$. Then for every k there is an $x_k \in \operatorname{dom} f \cap \{\bigcap_{i=1}^m \operatorname{dom} g_i\}$ and $\delta_k \to 0$ with

$$f\,(x_k) \leq s_k + \frac{1}{k} \quad \text{and} \quad g\,(x_k) \leq b_k \leq b + \delta_k(1, \ldots, 1).$$

Since $s_k \to s$, it is a bounded sequence, and so $\max_k(s_k)$ is finite. Consequently, for all k,

$$\hat{\lambda}_0 f\,(x_k) + \lambda^T g\,(x_k) \leq \hat{\lambda}_0 \left(s_k + \frac{1}{k}\right) + \lambda^T b_k \leq \hat{\lambda}_0 \left(\max_k(s_k) + 1\right) + \hat{\lambda}^T b + \hat{\lambda}^T(1, \ldots, 1) =: r.$$

Thus $x_k \in \operatorname{lev}_r \left(\hat{\lambda}_0 f + \lambda^T g\right)$. By our assumptions, this subset of $\mathbf{R}^m$ is bounded. Thus we may extract a convergent subsequence. As f and each of the g_i are closed and $\left(x_k, s_k + \frac{1}{k}\right) \in \operatorname{epi} f$ and $(x_k, b_k) \in \operatorname{epi} g$ we have $(x, s) \in \operatorname{epi} f$ and $(x, b) \in \operatorname{epi} g$ implying

$$v\,(b) = \inf\{f\,(u) \mid g\,(u) \leq b\} \leq f\,(x) \leq s$$

or $(b,s) \in \operatorname{epi} v$ (i.e. $v(b) \leq s$). Thus v is closed. We already know from Lemma 4.4.1 that it is convex.

Having shown that $v(b) \leq s$, let's consider what happens when they are equal. If we take $s = v(b)$ we obtain from the previous argument that there exists x such that $v(b) \leq f(x) \leq v(b)$ and so the infimum is attained. Finally we note that when $d = +\infty$ we have from $p \geq d$ that $p = d = +\infty$. When d is finite $d = v^{**}(0)$ is finite and as v is closed $v = v^{**}$ (as it is lower semi–continuous everywhere). Thus $p = v(0) = v^{**}(0) = d$. □

Example 4.4.1 *Calculate the Lagrangian dual of the following problems (given $a^1, a^2, \ldots, a^m$ and c in $\mathbf{R}^n$)*

1. *The linear program*

$$\inf_{x \in \mathbf{R}^n} \left\{ \langle c, x \rangle \mid \langle a^i, x \rangle \leq b_i, \text{ for } i = 1, \ldots, m \right\}.$$

Soln: *The Lagrangian is*

$$L(x, \lambda) = \langle c, x \rangle + \sum_{i=1}^{m} \lambda_i \left(\langle a^i, x \rangle - b_i \right) = \langle c + \sum_{i=1}^{m} \lambda_i a^i, x \rangle - \sum_{i=1}^{m} \lambda_i b_i.$$

We need to find first $\inf_x L(x, \lambda)$. Note that for the infimum to be finite, we will need to have

$$c + \sum_{i=1}^{m} \lambda_i a^i = 0 \quad \text{which implies} \quad \sum_{i=1}^{m} \lambda_i a^i = -c. \tag{4.35}$$

Therefore, when we calculate the maximum over all λ in the dual problem $\sup_{\lambda \geq 0} \inf_x L(x, \lambda)$, the maximizing λ must satisfy (4.35). Thus we have a Lagrangian dual of the form

$$\begin{aligned} \sup_{\lambda \geq 0} \inf_x L(x, \lambda) &= \sup_{\lambda \geq 0} \inf_x \langle c + \sum_{i=1}^{m} \lambda_i a^i, x \rangle - \sum_{i=1}^{m} \lambda_i b_i \\ &\sup_{\lambda \geq 0} \left\{ -\sum_{i=1}^{m} \lambda_i b_i \mid \sum_{i=1}^{m} \lambda_i a^i = -c \right\} \\ &= -\inf_{\lambda \geq 0} \left\{ \sum_{i=1}^{m} \lambda_i b_i \mid \sum_{i=1}^{m} \lambda_i a^i = -c \right\}. \end{aligned}$$

2. *Another linear program*

$$\inf_{x \in \mathbf{R}^n} \left\{ \langle c, x \rangle + \delta_{\mathbf{R}^n_+}(x) \mid \langle a^i, x \rangle \leq b_i, \text{ for } i = 1, \ldots, m \right\}.$$

Soln: *The Lagrangian is*

$$L(x, \lambda) = \langle c, x \rangle + \delta_{\mathbf{R}^n_+}(x) + \sum_{i=1}^{m} \lambda_i \left(\langle a^i, x \rangle - b_i \right) = \langle c + \sum_{i=1}^{m} \lambda_i a^i, x \rangle + \delta_{\mathbf{R}^n_+}(x) - \sum_{i=1}^{m} \lambda_i b_i$$

and for $\inf_x L(x, \lambda)$ to be finite (a condition implicit in the maximization over λ in the dual) we require $c + \sum_{i=1}^{m} \lambda_i a^i \geq 0$ so that $\langle c + \sum_{i=1}^{m} \lambda_i a^i, x \rangle \geq 0$ for all $x \in \mathbf{R}^n_+$. With this condition imposed, we have

$$\inf_x \langle c + \sum_{i=1}^{m} \lambda_i a^i, x \rangle + \delta_{\mathbf{R}^m_+}(x) = \langle c + \sum_{i=1}^{m} \lambda_i a^i, 0 \rangle + 0 = 0$$

$$\textit{and so} \qquad \inf_x L(x,\lambda) = -\sum_{i=1}^m \lambda_i b_i.$$

Thus we have a Lagrangian dual of the form (taking out the minus sign again)

$$\inf_{\lambda \geq 0} \left\{ \sum_{i=1}^m \lambda_i b_i \mid \sum_{i=1}^m \lambda_i a^i \geq -c \right\}.$$

3. *The quadratic program (for C positive definite and hence invertible)*

$$\inf_{x \in \mathbf{R}^n} \left\{ \frac{1}{2} x^T C x \mid \langle a^i, x \rangle \leq b_i, \textit{ for } i = 1, \ldots, m \right\}.$$

Soln: *The Lagrangian is*

$$L(x,\lambda) = \frac{1}{2} x^T C x \ + \sum_{i=1}^m \lambda_i \left(\langle a^i, x \rangle - b_i \right)$$

and for $\inf_x L(x,\lambda)$ to be finite we require

$$0 = \nabla L(x,\lambda) = \overbrace{Cx}^{\textit{Lemma 4.1.18}} + \sum_{i=1}^m \lambda_i a^i$$

$$\textit{or} \quad x = -C^{-1} \left(\sum_{i=1}^m \lambda_i a^i \right). \tag{4.36}$$

Thus

$$\begin{aligned}
\inf_x L(x,\lambda) &= \inf_x (L(x,\lambda) - \overbrace{x^T C x + x^T C x}^{0}) \\
&= \left(\frac{1}{2} x^T C x - x^T C x \right) + \left(\sum_{i=1}^m \lambda_i (\langle a^i, x \rangle - b_i) + x^T C x \right) \\
&= -\frac{1}{2} x^T C x + \left(\left\langle \sum_{i=1}^m \lambda_i a^i x \right\rangle - \sum_{i=1}^m b_i + \langle Cx, x \rangle \right) \\
&= -\frac{1}{2} x^T C x + \langle \underbrace{Cx + \sum_{i=1}^m \lambda_i a^i}_{=0 \textit{ by } (4.36)}, x \rangle - \sum_{i=1}^m \lambda_i b_i \\
&= -\sum_{i=1}^m \lambda_i b_i - \frac{1}{2} x^T C x \\
&= -\sum_{i=1}^m \lambda_i b_i - \frac{1}{2} \left(\sum_{j=1}^m \lambda_j a^j \right)^T C^{-1} \left(\sum_{i=1}^m \lambda_i a^i \right) \qquad \begin{array}{c} \textit{(substituting (4.36)} \\ \textit{for } x \textit{ again)} \end{array}
\end{aligned}$$

(Using that $C^{-1/2} = (C^{-1/2})^T$ since C is positive definite and so also is its inverse). Define $A^T = [a^1, \ldots, a^m]$ (the a_i are the column vectors). Then we can rewrite this as

$$L(x,\lambda) = -b^T \lambda - \frac{1}{2} \left(A^T \lambda \right)^T C^{-1} \left(A^T \lambda \right)$$

$$= -\left(b^T\lambda + \frac{1}{2}\lambda^T AC^{-1}A^T\lambda\right).$$

then the Lagrangian dual is

$$\min_{\lambda\geq 0}\left[b^T\lambda + \frac{1}{2}\lambda^T AC^{-1}A^T\lambda\right]. \tag{4.37}$$

4. *For given matrices $A_1, A_2, \ldots, A_m$ and C in $\mathcal{S}(n)$ calculate the dual of the semi-definite program*

$$\inf_{X\in\mathcal{S}(n)}\left\{\langle C, X\rangle + \delta_{\mathcal{P}(n)}(X) \mid \langle A_i, X\rangle \leq b_i, \text{ for } i = 1, \ldots, m\right\}.$$

Soln: *The Lagrangian is*

$$\begin{aligned} L(X,\lambda) &= \langle C, X\rangle + \delta_{\mathcal{P}(n)}(X) + \sum_{i=1}^m \lambda_i\left(\langle A_i, X\rangle - b_i\right) \\ &= \langle C + \sum_{i=1}^m \lambda_i A_i, X\rangle + \delta_{\mathcal{P}(n)}(X) - \sum_{i=1}^m \lambda_i b_i. \end{aligned}$$

Now unless we have the condition $0 \preceq C + \sum_{i=1}^m \lambda_i A_i$ (or $C + \sum_{i=1}^m \lambda_i A_i \in \mathcal{P}(n)$), we will end up with $\inf_{\lambda\geq 0} L(X,\lambda) = -\infty$. Therefore, when we compute the supremum over λ, we must have the condition. The infimum is then attained by $X = 0$, so

$$\inf_{X\in\mathcal{P}(n)} L(X,\lambda) = -\sum_{i=1}^m \lambda_i b_i.$$

Thus the Lagrangian dual becomes

$$\min_{\lambda\geq 0}\left\{\sum_{i=1}^m \lambda_i b_i \mid 0 \preceq C + \sum_{i=1}^m \lambda_i A_i\right\}.$$

We finish this section by observing the connection that Lagrangian duality has to saddle points:

Theorem 4.4.6 (Karush–Kuhn–Tucker) *The pair (x^*, λ^*) are the optimal primal dual pair for the convex optimisation problem*

$$\inf\{f(x) \mid g(x) \leq 0,\ x \in E\}$$

if and only if the pair (x^*, λ^*) with $x^* \in E$, $\lambda^* \geq 0$ is a saddle point of the Lagrangian i.e.

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*) \tag{4.38}$$

for all $x \in E$, $\lambda \geq 0$, or equivalently: $\inf_{x\in E}\sup_{\lambda\geq 0} L(x,\lambda) = \sup_{\lambda\geq 0}\inf_{x\in E} L(x,\lambda) = L(x^*,\lambda^*)$. This gives rise in the differentiable case to the so-named Karush–Kuhn–Tucker (KKT) conditions, which are necessary and sufficient for (x^*,λ^*) to be a saddle point:

$$\begin{aligned} &\nabla_x L(x^*,\lambda^*) = 0 \\ &\nabla_\lambda L(x^*,\lambda^*) \in \mathbb{R}^n_- \end{aligned} \tag{KKT}$$

Note*: In the case when inequality constraints are all replaced with equality constraints (for example, the constraints $g_i(x) \leq 0$ AND $-g_i(x) \leq 0$ together force $g_i(x) = 0$), then the latter condition becomes $\nabla_\lambda L(x^*,\lambda^*) = 0$ (this will be the case when we work with ADMM later). Since $\nabla_\lambda L(x^*,\lambda^*) = g(x)$, we would not expect the gradient to be exactly zero in an inequality constrained case.*

Proof. Let $p = d$ where (x^*, λ^*) is the optimal primal dual pair. Then

$$d = \sup_{\lambda \geq 0} \inf_{x \in E} L(x, \lambda) = \inf_{x \in E} L(x, \lambda^*) \leq L(x, \lambda^*) \quad \text{for any } x$$

and we also have

$$p = \inf_{x \in E} \sup_{\lambda \geq 0} L(x, \lambda) = \sup_{\lambda \geq 0} L(x^*, \lambda) \geq L(x^*, \lambda) \quad \text{for any } \lambda.$$

Thus (x^*, λ^*) are the optimal primal primal dual pair implies

$$L(x^*, \lambda) \leq p = L(x^*, \lambda^*) = d \leq L(x, \lambda^*).$$

Now suppose (x^*, λ^*) is a saddle point of the Lagrangian in the sense of (4.38). Then it is immediate that $g(x^*) \leq 0$, because otherwise we would have $\sup_{\lambda \geq 0} L(x^*, \lambda) = +\infty$. Then it is also clear that $\sup_{\lambda \geq 0} \lambda^T g(x^*) = 0$, because in taking the supremum we will have $\lambda_i = 0$ whenever $g_i(x^*) < 0$. Thus:

$$\begin{aligned} f(x^*) &= \sup_{\lambda \geq 0} f^*(x^*) + \overbrace{\lambda^T g(x^*)}^{=0} \\ &= \sup_{\lambda \geq 0} L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq \inf_{x \in E} L(x, \lambda^*) = \Phi(\lambda^*) \leq p \leq f(x^*) \end{aligned}$$

giving equality throughout. Thus $f(x^*) = p$, and so we know that x^* is a primal optimal solution. As

$$\Phi(\lambda^*) \leq \sup_{\lambda \geq 0} \Phi(\lambda) \overset{\text{(def)}}{=} d \overset{\text{(always)}}{\leq} p \overset{\text{(equality throughout above)}}{=} \Phi(\lambda^*),$$

we obtain $d = p$, and so $\lambda^* \geq 0$ is a dual optimal solution. □

The following Lemma might help you with Problem 4.4.9.

Lemma 4.4.7 *Let $A \in \mathcal{S}(n)$. Then $\nabla_X \langle A, X \rangle = A$.*

Problem Set 9

Use the Lagrangian optimality condition to solve the following problems.

Problem 4.4.8 *Given strictly positive reals $a_1, a_2, \ldots, a_n, c_1, c_2, \ldots, c_n$ and b*

$$\inf \left\{ \sum_{i=1}^n \frac{c_i}{x_i} \mid \sum_{i=1}^n a_i x_i \leq b,\ x > 0 \right\}.$$

Problem 4.4.9 *For a positive definite matrix A and a real $b > 0$*

$$\inf \{ -\log \det X \mid \langle A, X \rangle \leq b,\ X > 0 \}.$$

Hint: Example 3.2.1 and Lemma 4.4.7 may simplify your life.

4.5 Penalty Methods and Primal Functions

One practical way of handling constraints computationally is to penalize infeasible points. Consider the optimisation problem

$$\begin{aligned} &\min \quad f(x) \\ &\text{Subject to} \\ &\qquad x \in X \subseteq \mathbf{R}^n, \quad g_j(x) \leq 0,\ j = 1, \ldots, r \end{aligned} \tag{Prim}$$

which has a feasible set

$$\mathcal{F} := \{x \in X \mid g_j(x) \leq 0,\ j = 1, \ldots, r\}.$$

Now consider a function $P : \mathbf{R}^r \to \mathbf{R}$ satisfying

$$\begin{aligned} &P(u) = 0, \quad u \leq 0, \\ &P(u) > 0, \quad \text{if } u_j > 0 \text{ for some } j = 1, \ldots, r. \end{aligned}$$

Then

$$P(g_1(x), \ldots, g_r(x)) = \begin{cases} \text{something} > 0 & \text{if} \quad g_j(x) > 0 \text{ for some } j \text{ or } x \notin \mathcal{F} \\ 0 & x \in \mathcal{F} \end{cases}$$

then we can try performing an unconstrained minimization

$$\min_{x \in X} \quad f(x) + P(g_1(x), \ldots, g_r(x))$$

expecting the minimization to occur at $x \in \mathcal{F}$ if $P(g_1(x), \ldots, g_r(x))$ is sufficiently large for $x \notin \mathcal{F}$. Of course, if P returns values that are not on the same order as f, then we could have errors in floating point arithmetic, harming the accuracy of our solutions. Some examples of penalty functions are:

$$\begin{aligned} P(u) &:= \frac{c}{2} \sum_{j=1}^{r} (\max\{0, u_j\})^2 \quad \text{or} \\ P(u) &:= \frac{c}{2} \max\{0, u_1, \ldots, u_r\}. \end{aligned}$$

We will study this in some detail using the tools we have developed.

One object that is of use in this the primal value function introduced in section 4.4:

$$v(u) = \inf_{\substack{x \in X,\ g_j(x) \leq u_j \\ j = 1, \ldots, r}} f(x)$$

Note that $v(0) = \inf\{f(x) \mid x \in X,\ g_j(x) \leq 0, j = 1, \ldots, r\}$ the primal optimal value. Let $\operatorname{dom} v$ denote the domain of v i.e. the set of u for which the constraint set

$$\{x \in X \mid g_j(x) \leq u_j,\ j = 1, \ldots, r\}$$

is nonempty, and consequently $v(u) < +\infty$. There are some interesting connections to dual functions: when $\mu \geq 0$ we have

$$\Phi(\mu) = \inf_{x \in X} \left\{ f(x) + \sum_{j=1}^{r} \mu_j g_j(x) \right\}$$

$$= \inf_{\{(u,x)|x\in X,\ g_j(x)\le u_j,\ j=1,\ldots,r\}} \left\{ f(x) + \sum_{j=1}^{r} \mu_j g_j(x) \right\}$$

$$= \inf_{\{(u,x)|x\in X,\ g_j(x)\le u_j,\ j=1,\ldots,r\}} \left\{ f(x) + \sum_{j=1}^{r} \mu_j u_j \right\}$$

(as $\mu_j \ge 0$ and $g(x)$ is one of the possible u and $g(x)$ lower bounds u)

$$= \inf_{u\in \operatorname{dom} v} \inf_{\substack{x\in X,\ g_j(x)\le u_j \\ j=1,\ldots,r}} \left\{ f(x) + \sum_{j=1}^{r} \mu_j u_j \right\}$$

$$= \inf_{u\in \operatorname{dom} v} \left\{ \inf_{\substack{x\in X,\ g_j(x)\le u_j \\ j=1,\ldots,r}} f(x) + \mu^T u \right\}$$

$$= \inf_{u\in \operatorname{dom} v} \left\{ v(u) + \mu^T u \right\} = -\sup_u \left\{ (-\mu)^T u - v(u) \right\} = -v^*(-\mu).$$

Notice that we have seen this before, in Theorem 4.4.2(ii).

We suppose v is proper (i.e. $v(u) > -\infty$ for all $u \in \operatorname{dom} v$) then v is convex when each of f, $g_j(x) \le 0$, $j = 1, \ldots, r$ are convex (Lemma 4.4.1). Now it is the structure of v around $u = 0$ that determines the existence of multipliers. We proved in Section 4.4 that:

$$-\mu^* \in \partial v(0) \quad \Longleftrightarrow \quad \mu^* \text{ is a Lagrange multiplier for the problem (Prim)}$$

See (4.28) in the proof of Theorem 4.4.2(iv).

We will need the dual of P and so note that if $\mu_j < 0$ for some j, then for $u_j < 0$:

$$P^*(\mu) = \sup_u \left\{ \mu^T u - P(u) \right\} \ge u_j \mu_j - P(0, \ldots 0, u_j, 0 \ldots, 0)$$
$$= u_j \mu_j \to +\infty \quad \text{as } u_j \to -\infty.$$

On the other hand, if $\mu \ge 0$, then:

$$P^*(\mu) = \sup_u \left\{ \mu^T u - P(u) \right\} \ge -P(0) = 0$$

Thus

$$P^*(\mu) = \sup_u \left\{ \mu^T u - P(u) \right\} = \begin{cases} +\infty & \text{if } \mu_j < 0 \\ \text{something} \ge 0 & \text{otherwise} \end{cases}$$

Now consider the penalised problem

$$\min_{x\in X} \quad f(x) + P(g(x)). \tag{P-penalty}$$

Let us now relate it to the primal value functions

$$\min_{x\in X} \{ f(x) + P(g(x)) \} = \min_u \min_{\substack{x\in X,\ g_j(x)\le u_j \\ j=1,\ldots,r}} \{ f(x) + P(g(x)) \} \tag{4.39}$$

(Note that $P(g(x)) \le P(u)$ and $P(u) = P(g(x))$ if $u = g(x)$)

$$= \min_u \left\{ \left[\min_{\substack{x\in X,\ g_j(x)\le u_j \\ j=1,\ldots,r}} f(x) \right] + P(u) \right\} = \min_u \{ v(u) + P(u) \}.$$

We now use Fenchel duality

$$\min_u \{v(u) + P(u)\} \overset{\text{Theorem 4.2.1}}{=} \max_{-\mu \geq 0} \{-v^*(\mu) - P^*(-\mu)\}$$
$$= \max_{\mu \geq 0} \{-v^*(-\mu) - P^*(\mu)\}. \tag{4.40}$$

The Fenchel duality holds due to the fact that

$$\operatorname{dom} P = \mathbf{R}^r \quad \text{and so}$$
$$0 \in \operatorname{int}\{\operatorname{dom} v - \operatorname{dom} P\} = \mathbf{R}^r.$$

Consequently combining (4.39) and (4.40) we have

$$\min_{x \in X} \{f(x) + P(g(x))\} = \min_u \{v(u) + P(u)\} = \max_{\mu \geq 0} \{-v^*(-\mu) - P^*(\mu)\}. \tag{4.41}$$

Theorem 4.5.1 *Consider the primal problem (Prim) and its penalised version (P-Penalty).*

(i) The problems (Prim) and (P-Penalty) have equal optimal values if and only if there exists a Lagrange multiplier μ^* such that

$$(\mu^*)^T u \leq P(u), \quad \textit{for all } u. \tag{4.42}$$

(In practice, this means the penalty parameter c needs to be big enough)

(ii) For the problems (Prim) and (P-Penalty) to have the same set of solutions it is sufficient that there exists a Lagrange multiplier μ^* such that

$$(\mu^*)^T u < P(u), \quad \textit{for all } u \textit{ with } u_j > 0 \textit{ for some } j. \tag{4.43}$$

Proof. (i): First notice that (4.42) is equivalent to

$$P^*(\mu^*) = \sup_u \langle \mu^*, u \rangle - P(u) \leq 0.$$

Combining with the fact that $P^* \geq 0$ (clear from the definition of P), we actually have that (4.42) is equivalent to

$$P^*(\mu^*) = 0.$$

Therefore, it suffices to show that this is equivalent to (Prim) and (P-Penalty) having equal optimal values. This is what we will do.

Notice that as $P(g(x)) = 0$ for all feasible solutions, we have

$$f^{opt} = v(0) = \min_{x \in \mathcal{F}} f(x) = \min_{x \in \mathcal{F}, g(x) \leq g(x)} f(x) + P(g(x)) = \min_{x \in \mathcal{F}} v(g(x)) + P(g(x))$$

so

$$f^{opt} \geq \min_u \{v(u) + P(u)\} \overset{(4.41)}{=} \max_{\mu \geq 0} \{-v^*(-\mu) - P^*(\mu)\}. \tag{4.44}$$

Now suppose (Prim) and (P-Penalty) have equal optimal values. In other words:

$$f^{opt} = \min_u \{v(u) + P(u)\}. \tag{4.45}$$

Using (4.44), we have that (4.45) is equivalent to:

$$f^{opt} = \max_{\mu \geq 0} \{-v^*(-\mu) - P^*(\mu)\}$$

$$= \{-v^*(-\mu^*) - P^*(\mu^*)\} \quad (\forall\ \mu^* \in \arg\max\{-v^*(-\mu) - P^*(\mu)\}) \tag{4.46}$$

Recall that by Lagrangian duality (see (4.27))

$$-v^*(-\mu) = \Phi(\mu) = \inf_{x\in X} L(x,\mu) \leq f^{opt} \tag{4.47}$$

with equality when $\mu = \mu^*$ is a Lagrange multiplier. Thus we have

$$f^{opt} \overset{(4.46)}{=} -v^*(-\mu^*)\overbrace{-P^*(\mu^*)}^{\leq 0} \leq -v^*(-\mu^*) \overset{(4.47)}{=} f^{opt},$$

The equality throughout forces $P^*(\mu^*) = 0$, and so we have shown that (4.46) is equivalent to $P^*(\mu^*) = 0$. This is what we needed to show.

(ii): Suppose (4.43) holds. Let x^* be a optimal solution to (Prim). Then x^* is feasible, in which case

$$f^{opt} = f(x^*) = f(x^*) + \overbrace{P(g(x^*))}^{=0} \geq \min_{x\in X}\{f(x) + P(g(x))\} \geq \min_u \{v(u) + P(u)\} \overset{(a)}{=} f^{opt}$$

To see why (a) holds, notice that the condition (4.43) implies the condition (4.42), and so the equality (a) holds by (i). Notice that we have equality throughout the above equation. Thus x^* is also an optimal solution to (P-Penalty).

Now suppose x^* is an optimal solution of the penalised problem (P-Penalty).

Case: Suppose x^* is feasible. Then it must also be a solution of (Prim), because

$$\begin{aligned} f^{opt} &:= \inf_{\substack{x\in X,\ g_j(x)\leq 0 \\ j=1,\ldots,r}} f(x) = \inf_{\substack{x\in X,\ g_j(x)\leq 0 \\ j=1,\ldots,r}} \{f(x) + P(g(x))\} \\ &\geq \min_{x\in X}\{f(x) + P(g(x))\} = f(x^*) + \underbrace{P(g(x^*))}_{=0} = f(x^*)\ . \end{aligned}$$

Here the equality $P(g(x^*))$ uses our assumption that x^* is feasible.

Case: Suppose x^* is not feasible and so $g_j(x^*) > 0$ for some j. Since (4.43) holds for all u, it certainly holds for $g(x^*)$, and so there exists $\varepsilon > 0$ such that

$$g(x^*)^T \mu^* + \varepsilon < P(g(x^*))\,. \tag{4.48}$$

Let $\tilde{x}$ be a feasible vector such that $f(\tilde{x}) \leq f^{opt} + \varepsilon$. Since $P(g(\tilde{x})) = 0$ and $f^{opt} = \min_{x\in X}\left\{f(x) + g(x)^T\mu^*\right\}$ (where μ^* is the optimal Lagrange multiplier) we have

$$\begin{aligned} f(\tilde{x}) + \overbrace{P(g(\tilde{x}))}^{=0} = f(\tilde{x}) \leq f^{opt} + \varepsilon &= \min_{x\in X}\left\{f(x) + g(x)^T\mu^*\right\} + \varepsilon \\ &\leq f(x^*) + g(x^*)^T\mu^* + \varepsilon \\ &\overset{(4.48)}{<} f(x^*) + P(g(x^*))\,. \end{aligned}$$

Thus $\tilde{x}$ is a strictly better solution to (P-Penalty) than x^*. This is a contradiction, and so any optimal x^* must be feasible. □

Primal and penalised optimality for specific penalty functions

We finish this chapter by considering what the conditions (4.42) and (4.43) look like for two commonly used penalty terms. Consider the exact penalty

$$P(u) := c\sum_{j=1}^{r} \max\{0, u_j\}$$

where $c > 0$ is to chosen. Now (4.42) corresponds to

$$\sum_{j=1}^{r} \mu_j^* u_j \leq c\sum_{j=1}^{r} \max\{0, u_j\} \quad \text{for all } u$$

which is equivalent to $\mu_j^* \leq c$, for all $j = 1, \ldots, r$ (by choosing $u_j = 1$ in turn).

Likewise, the condition (4.43) corresponds to

$$\mu_j^* < c, \text{ for all } j = 1, \ldots, r.$$

Now consider a different penalty:

$$P(u) := c \max\{0, u_1, \ldots, u_r\}$$

the condition (4.42) corresponds to

$$\sum_{j=1}^{r} \mu_j^* u_j \leq c \max\{0, u_1, \ldots, u_r\} \quad \text{for all } u \tag{4.49}$$

$$\text{which clearly implies} \quad \sum_{j=1}^{r} \mu_j^* \leq c \quad \text{(by choosing all } u_j = 1\text{)}. \tag{4.50}$$

Indeed when this last inequality holds we have

$$\sum_{j=1}^{r} \mu_j^* u_j \leq \max\{0, u_1, \ldots, u_r\} \left(\sum_{j=1}^{r} \mu_j^*\right) \leq c \max\{0, u_1, \ldots, u_r\} \quad \text{for all } u$$

and so (4.49) and (4.50) are equivalent. Similarly condition (4.43) in this case corresponds to

$$\sum_{j=1}^{r} \mu_j^* < c.$$

Problem Set 10

Problem 4.5.2 *Consider the optimisation problem*

$$\begin{aligned} \min_x \quad & f(x) \\ & \textit{Subject to} \\ & g_j(x) \leq 0, \quad i = 1, \ldots, r. \end{aligned}$$

Introduce slack variable $z_j, \geq 0$ $j = 1, \ldots, r$ *and a penalty so that we have now have an equivalent equality constrained problem*

$$\begin{aligned} \min_{x,z} \quad & f(x) + \frac{c}{2} \sum_{j=1}^{r} |g_j(x) + z_j|^2 \\ & \textit{Subject to} \\ & g_j(x) + z_j = 0, \quad i = 1, \ldots, r. \end{aligned}$$

Form the augmented Lagrangian (a Lagrangian with a penalty on the infeasibility of the constraint)

$$L_c(x, z, \mu) := f(x) + \sum_{j=1}^{r} \mu_j (g_j(x) + z_j) + \frac{c}{2} \sum_{j=1}^{r} |g_j(x) + z_j|^2 .$$

The Lagrangian dual problem is of the form

$$\sup_{\mu \geq 0} \min_{x, z \geq 0} L_c(x, z, \mu) .$$

1. *Perform the minimization over z first*

$$L_c(x, \mu) := \min_{z \geq 0} \left\{ f(x) + \sum_{j=1}^{r} \mu_j (g_j(x) + z_j) + \frac{c}{2} \sum_{j=1}^{r} |g_j(x) + z_j|^2 \right\} \tag{4.51}$$

   to show that $z_j = \max(0, \hat{z}_j)$ *is the constrained* ($z \geq 0$) *minimum of (4.51) where* $\hat{z}_j$ *is the unconstrained minimum i.e. solves*

$$\min_{z} \left\{ \sum_{j=1}^{r} \mu_j (g_j(x) + z_j) + \frac{c}{2} \sum_{j=1}^{r} |g_j(x) + z_j|^2 \right\} .$$

2. *Give an argument that shows that*

$$L_c(x, \mu) = f(x) + \frac{1}{2c} \sum_{j=1}^{r} \left\{ (\max\{0, \mu_j + c g_j(x)\})^2 - \mu_j^2 \right\} .$$

3. *Now consider the penalty function*

$$P_{c^k}\left(u, \mu^k\right) := \frac{1}{2c^k} \sum_{j=1}^{r} \left\{ \left(\max\left\{0, \mu_j^k + c u_j\right\} \right)^2 - \left(\mu_j^k\right)^2 \right\}$$

   Show that the conjugate with respect to u *is given (as a function of a dual variable* μ*) by*

$$P^*_{c^k}\left(\mu, \mu^k\right) = \begin{cases} \frac{1}{2c^k} \left\| \mu - \mu^k \right\|^2 & \textit{if } \mu \geq 0 \\ +\infty & \textit{otherwise} \end{cases}$$

   [Hint: Try and calculate

$$P^*_{c^k}\left(\mu, \mu^k\right) = \sup_{u} \left\{ \mu^T x - \inf_{z \geq 0} \left\{ \sum_{j=1}^{r} \mu_j^k (u_j + z_j) + \frac{c^k}{2} \sum_{j=1}^{r} |u_j + z_j|^2 \right\} \right\}].$$

4. *Deduce that the dual formulation given in (4.41) is given by*

$$\min_x L_{c^k}\left(x,\mu^k\right)=\max_{\mu\geq 0}\left\{-v^*(-\mu)-\frac{1}{2c^k}\left\|\mu-\mu^k\right\|^2\right\}$$

where

$$v\left(u\right)=\min_{\substack{x:\ g_j(x)\leq u_j\\ j=1,\ldots,r}} f\left(x\right)$$

5. *Now we observe that if x^k attains the minimum $\min_x L_{c^k}\left(x,\mu^k\right)$ then $u^k:=g\left(x^k\right)$ attains the minimum in $\min_u\left\{p\left(u\right)+P_{c^k}\left(u,\mu^k\right)\right\}$. Then the generalised Lagrangian optimality conditions (∂FD) corresponds to (taking into account the change of sign)*

$$u^k\in\partial P^*_{c^k}\left(\cdot,\mu^k\right)\left(\mu^{k+1}\right)\quad\Longleftrightarrow\quad\mu^{k+1}\in\arg\max_{\mu\geq 0}\left\{\mu^T u^k-P^*_{c^k}\left(\mu,\mu^k\right)\right\}. \tag{4.52}$$

Show that (4.52) corresponds to the multiplier iteration

$$\mu_j^{k+1}=\max\left\{0,\mu_j^k+c^k u_j^k\right\}=\max\left\{0,\mu_j^k+c^k g_j\left(x^k\right)\right\}.$$

4.5.1 More properties of subdifferentials

Proposition 4.5.3 *For any convex function f and point x, the set $\partial f(x)$ is closed and convex.*

Proof.
Closed: For any x, the set

$$\Gamma_y:=\{\lambda\mid f(y)-f(x)\geq\langle\lambda,y-x\rangle\}$$

is closed. To see why, take any convergent sequence $\lambda_n\in\Gamma_y$, and the continuity of the inner product will guarantee that its limit will still clearly belong to Γ_y. Then

$$\partial f(x)=\{\lambda\mid f(y)-f(x)\geq\langle\lambda,y-x\rangle\quad\forall y\}=\bigcap_y\partial\Gamma_y$$

is an intersection of closed sets, and so it is closed.

Convex: If $\partial f(x)=\emptyset$, then it vacuously contains all of the convex combinations of its elements. If it is nonempty, then any $x_1^*,x_2^*\in\partial f(x)$ will satisfy the subgradient inequality, whereupon it is straightforward to see that $\lambda x_1^*+(1-\lambda)x_2^*$ will satisfy the subgradient inequality for any $\lambda\in[0,1]$. □

Proposition 4.5.4 *If g_i is differentiable for $i=1,\ldots,k$, then*

$$\begin{aligned}f\left(x\right)&=\max\left\{g_1,\ldots,g_k\right\}\quad\textit{satisfies}\\ \partial f\left(x\right)&=\operatorname{co}\left\{\nabla g_i\left(x\right)\mid i\in I\left(x\right)\right\}\ \textit{ where } I\left(x\right):=\left\{i\mid g_i\left(x\right)=f\left(x\right)\right\}.\end{aligned}$$

Proof. We know from Theorem 4.1.7 that that the directional derivative in a direction d of f has to be the support function of the subdifferential.

$$f'(x,d)=S(\partial f(x),d)=\sup_{z\in\partial f(x)}\langle z,d\rangle. \tag{4.53}$$

And also have for any $i \in I(x)$ that

$$f'(x,d) = \lim_{t\downarrow 0} \frac{1}{t}\left(f(x+td) - f(x)\right) \geq \lim_{t\downarrow 0} \frac{1}{t}(\overbrace{g_i(x+td)}^{f(x+td)\geq} - \overbrace{g_i(x)}^{f(x)=}) = g_i'(x,d) = \langle \nabla g_i(x), d\rangle. \tag{4.54}$$

Now (4.53) and (4.54) together force

$$\begin{aligned} & S(\partial f(x), d) \geq \langle \nabla g_i(x), d\rangle \quad (\forall i \in I(x)) \\ \text{or equivalently} \quad & S(\partial f(x), d) \geq S(\operatorname{co}\left\{\nabla g_i\left(x\right) \mid i \in I\left(x\right)\right\}, d). \end{aligned} \tag{4.55}$$

Let $t_n \downarrow 0$ and consider the sequence of index sets: $I(x + t_n d)$. We have a finite set of indices and an infinite sequence. By the pigeonhole principle, at least one index j must occur infinitely often. We can restrict to a subsequence such that $j \in I(x + t_n d)$ for all n. Then

$$f'(x,d) = \lim_{t\downarrow 0} \frac{1}{t}\left(f(x+td) - f(x)\right) = \lim_{t\downarrow 0} \frac{1}{t}(\overbrace{g_j(x+t_nd)}^{f(x+t_nd)=} - \overbrace{g_j(x)}^{f(x)=}) = g_j'(x,d) = \langle \nabla g_j(x), d\rangle. \tag{4.56}$$

Combining (4.55) and (4.56), clearly

$$S(\partial f(x), d) = S(\operatorname{co}\left\{\nabla g_i\left(x\right) \mid i \in I\left(x\right)\right\}, d).$$

Since it is true for any d, we must have that

$$S(\partial f(x), \cdot) = S(\operatorname{co}\left\{\nabla g_i\left(x\right) \mid i \in I\left(x\right)\right\}, \cdot).$$

As the two support functions for the two closed and convex sets are equal ($\partial f(x)$ is closed and convex by Proposition 4.5.3), their conjugates must be equal, but the conjugates are the indicator functions for the two sets (Example 4.1.3). Thus the two sets are equal. $\square$

Part II

Numerical Methods in Nonsmooth Optimisation

Chapter 5

Simple Descent Methods

Here we will begin to study the problem of providing effective algorithms for minimizing a nonsmooth function. As is usually the case we begin by considering the unconstrained case and then extend to include constraints. This is less problematical than in smooth optimisation as there is a real sense in which the minimization of an extended real valued function also encapsulates all constrained optimisation problems in that

$$x \in \arg\min \left\{ f(x) \quad \text{subject to } x \in F^{-1}(\Omega) \right\} \quad \Longleftrightarrow \quad x \in \arg\min \left\{ f(x) + \delta_{F^{-1}(\Omega)}(x) \right\},$$

where Ω can be a cone like $\mathbb{R}^n_+$ or some other closed convex set like $\mathbb{R}^n_+ \times \{0\}$. We begin by considering the difficulties arising when f is not differentiable and the failure of some simple strategies. In this first part we will consider only the minimization of convex functions.

5.1 Why Special Methods

Most descent algorithmic optimisation methods for smooth functions are based on the following basic approach:

General Descent Method: Initially we start with $x^0 \in \mathbf{R}^n$ and $k = 0$.
While x^k is not 'almost stationary' **do** iteration k: We terminate on an 'almost stationary' point, which frequently means $\|\nabla f(x^k)\| < \varepsilon$.

1. Given x^k choose d^k as a descent direction for f at x^k (i.e. $\nabla f(x^k)^T d^k < 0$).
2. Choose a step length $t^k > 0$ (via some line search procedure) such that

$$f(x^k + t^k d^k) < f(x^k).$$

Let $x^{k+1} = x^k + t^k d^k$ and $k \leftarrow k + 1$.
End While:

Problems for nonsmooth functions with simple descent:

1. The stopping condition $\|g\| \leq \varepsilon$ for $g \in \partial f(x)$ may never occur. For example, consider $f(x) = |x|$. The necessary condition $0 \in \partial f(x)$ needs the construction or at least an approximation of the set $\partial f(x)$!

2. One cannot use forward, backward or central differences to approximate subgradients. This is only valid if $d \mapsto f'(x,d)$ is linear in d (i.e. f is differentiable). For example, recall (Proposition 4.5.4) that for a max function

$$\begin{aligned} f(x) &= \max\{g_1,\ldots,g_k\} \quad \text{then} \\ \partial f(x) &= \operatorname{co}\{\nabla g_i(x) \mid i \in I(x)\} \text{ where } I(x) := \{i \mid g_i(x) = f(x)\}. \end{aligned}$$

Consider $f(x) = \max\{x_1, x_2, x_3\}$. For any point on the diagonal $0 < x_1 = x_2 = x_3$, I can use the above formula to calculate $\partial f(x) = \left\{x \in \mathbf{R}^3 \mid x_i \geq 0 \text{ and } \sum_{i=1}^3 x_i = 1\right\}$. The forward, backwards and central differences used by numerical techniques at the point $p = (5,5,5)$ are (respectively)

$$\begin{aligned} \text{(forward)} \qquad & (1,1,1) = (d(p,e_1), d(p,e_2), d(p,e_3)) \\ \text{(backward)} \qquad & (0,0,0) = (d(p,-e_1), d(p,-e_2), d(p,-e_3)) \\ \text{(central)} \quad & \left(\frac{1}{2},\frac{1}{2},\frac{1}{2}\right) = \text{average(forward,backward)}. \end{aligned}$$

None of these numerical approximations for the subgradient are actually in $\partial f(p)$.

3. Since $x \mapsto \partial f(x)$ is only upper semi-continuous, small changes in x_k leads to large changes in $g_k \in \partial f(x_k)$ and so produce very different subsequent iterates x_{k+1}. Thus running the same problem and code on a different computer can lead to different round off error and uncomparable performance.

One way that to overcome some of these problems is to choose a special subgradient, namely that of smallest norm from $\partial f(x)$. At least in this case we must find $0 \in \partial f(x)$ if one exists. This is conceptually valid, but we will find runs into problems when dealing with numerical issues again. In a perfect world, a descent direction d^k at x^k should ideally guarantee a decrease in that $f\left(x^k + t^k d^k\right) < f\left(x^k\right)$. This implies the condition

$$f'\left(x^k, d^k\right) = \inf_{t>0} \frac{1}{t}(f(x^k + td^k) - f(x^k)) < 0 \quad \text{or} \quad \sup\{\langle z.d^k\rangle \mid z \in \partial f\left(x^k\right)\} < 0.$$

For all non-optimal x^k we have $0 \notin \partial f\left(x^k\right)$, where the latter is closed, and so there exists a strict separation of the sets $\{0\}$ and $\partial f\left(x^k\right)$. Note that the descent directions in Figure 5.1 all make an obtuse angle to all elements of $\partial f\left(x^k\right)$.

This leads us to a modification of the smooth descent method.

General Nonsmooth Descent Method: Initially we start with $\mathbf{x}^0 \in \mathbf{R}^n$ and $k = 0$.
While $0 \notin \partial f\left(x^k\right)$ **do** iteration k:

1. Given x^k choose d^k as a descent direction for f at $\mathbf{x}^k$ (i.e. $g^T d^k < 0 \quad$ for all $g \in \partial f\left(x^k\right)$).

2. Choose a step length $t^k > 0$ (via some line search procedure) such that

$$f(x^k + t^k d^k) < f(x^k).$$

Let $x^{k+1} = x^k + t^k d^k$ and $k \leftarrow k+1$.
End While:

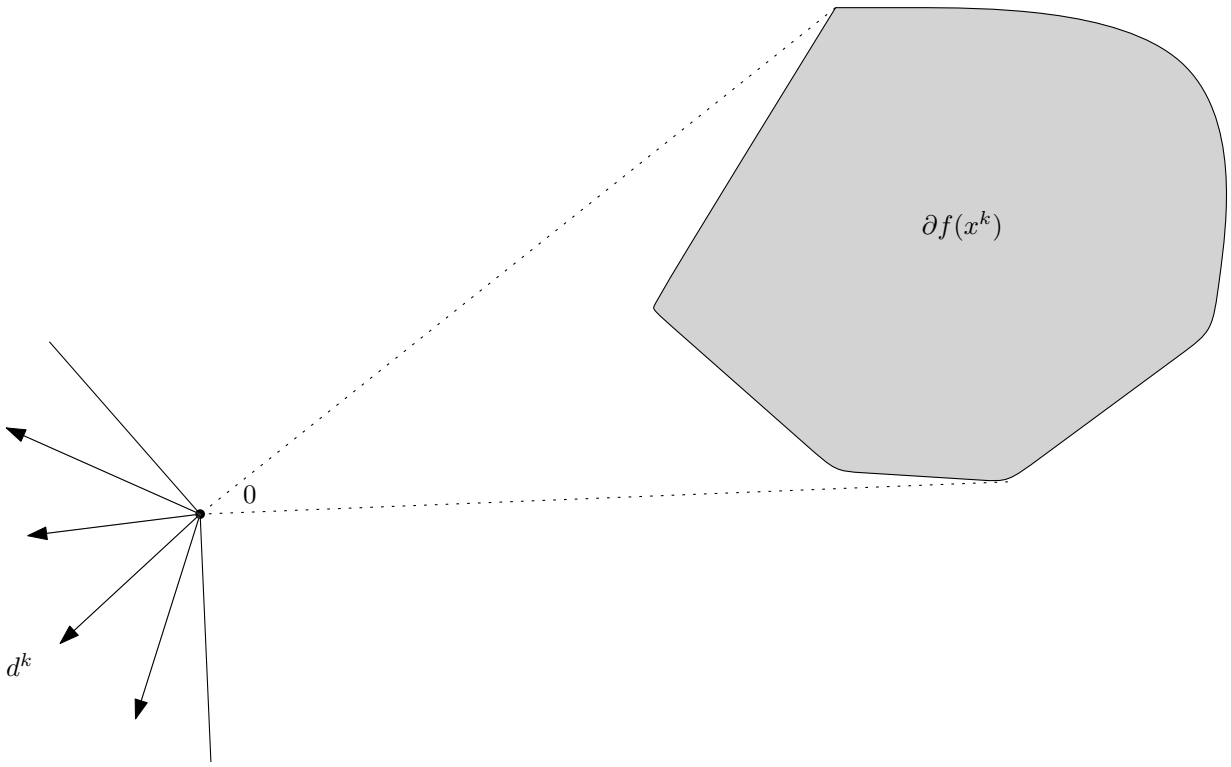


Figure 5.1: Descent directions

5.2 Steepest Descent

For the time being we will consider the simplest case of a convex function. We will progress to the nonconvex case later. On simplistic way to try and extended the steepest descent method is to calculate the following direction

$$d^k \in \arg\min_{\|d\|=1} f'\left(x^k, d\right) = \arg\min_{\|d\|=1} \max_{g \in \partial f\left(x^k\right)} \langle g, d \rangle.$$

Now via a theorem (the minmax theorem) we can reverse the order of the min max to a max min to get

$$\begin{aligned}
\min_{\|d\|=1} f'\left(x^k, d\right) &= \min_{\|d\|=1} \max_{g \in \partial f\left(x^k\right)} \langle g, d \rangle \\
&= \max_{g \in \partial f\left(x^k\right)} \min_{\|d\|=1} \langle g, d \rangle \\
&= \max_{g \in \partial f\left(x^k\right)} \left[- \|g\| \right] \quad \text{(Cauchy-Schwarz and alignment)} \\
&= - \min_{g \in \partial f\left(x^k\right)} \|g\| := - \| \Pr_{\partial f\left(x^k\right)}(0) \| \qquad (5.1)
\end{aligned}$$

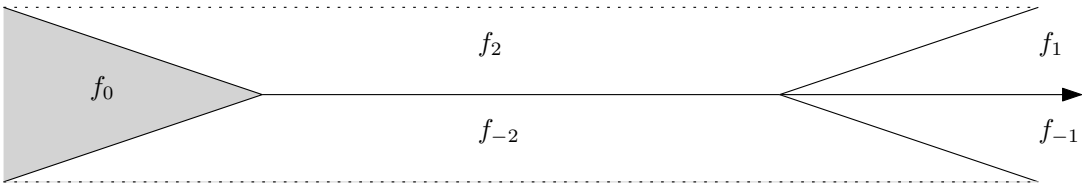
where the last operator denotes the projection of the origin onto the subdifferential. Sometimes we know how to compute this, but often we do not. If we do know it, then we can choose $d^k = -\frac{P_{\partial f(x^k)}(0)}{\left\| P_{\partial f(x^k)}(0) \right\|}$ and use some kind of line search procedure to find the step length t^k in $x^{k+1} = x^k + t^k d^k$. This seems appealing until one notes that we still need to know the whole of $\partial f\left(x^k\right)$ in order to calculate it. Just like steepest descent in the

smooth case the iterates oscillate a lot and even worse they can prematurely converge to a nonstationary point.

Example 5.2.1 *Consider an exact line search and steepest descent applied to*

$$f(x) = \max\{f_0(x), f_{-1}(x), f_{-2}(x), f_1(x), f_2(x)\},$$

where $f_0(x) = -100$, $f_{\pm 1}(x) = 3x_1 \pm 2x_2$, $f_{\pm 2}(x) = 2x_1 \pm 5x_2$. *The optimal value is* $f^* = -100$ *and* $\arg\min f = \{(x_1, x_2) \mid x_1 \leq -50 \text{ and } |x_2| \geq 0.4x_1 + 20\}$ *where* f_0 *is active.*



Starting the steepest descent at $x^1 = (9, -3)$ *we have* f_{-1} *and* f_{-2} *active so*

$$\partial f(9,-3) = \text{co}\{\overbrace{(3,-2)}^{\partial f_{-1}(9,-3)}, \overbrace{(2,-5)}^{\partial f_{-2}(9,-3)}\} \quad \text{so } \Pr_{\partial f(9,-3)}(0) = (3,-2).$$

Omitting the normalization the steepest descent direction is $d^1 = (-3, 2)$. *We then have to search along* $x^1 + td^1 = (9,-3) + t(-3,2) = (9-3t, -3+2t)$ *and find that*

$$t^1 \in \arg\min_{t \geq 0} \{f(9-3t, -3+2t)\}.$$

There are two possible solution.
The line search function has two kinks $x^1 + \frac{3}{2}d^1$ *and* $x^1 + 2d^1$*:*

$$f\left(x^1 + td^1\right) = f(9-3t, -3+2t) = \begin{cases} 33 - 13t & \text{if } 0 \leq t \leq \frac{3}{2} \\ 21 - 5t & \text{if } \frac{3}{2} \leq t \leq 2 \\ 3 + 4t & \text{if } 2 \leq t \leq 3 \end{cases}$$

The exact line search chooses the smallest value which is $t^1 = 2$ *and so* $x^2 = x^1 + 2d^1 = (9-6, -3+4) = (3,1)$. *A similar analysis gives the iterate sequence to be* $x^k = \left(3^{3-k}, (-1)^k 3^{2-k}\right)_{k \geq 1}$ *which converges very slowly to the non-optimal point* $x^\infty = (0,0)$. *From Figure 5.2 one can see the way these iterates bounce off the barrier provided by the discontinuity in the derivative. It is worth a small pause to contemplate what is so different here from smooth optimisation. It is true that we may arbitrarily closely approximate this convex optimisation problem with a smooth one (the Moreau envelope produces a* $C^{1,1}$ *approximation is an example). These arbitrarily closely approximating smooth functions would admit convergent iterates when steepest descent was applied! We would get the kind of cumulative deviation we flagged before that was associated with numerical round off. A very different iterate would occur on each of these approximations. In actual fact we only need to cross the boundaries of these "kinks" in order to over come this problem and get a method that converges to an optimal solution. In fact, this is what occurs with these smooth approximations.*

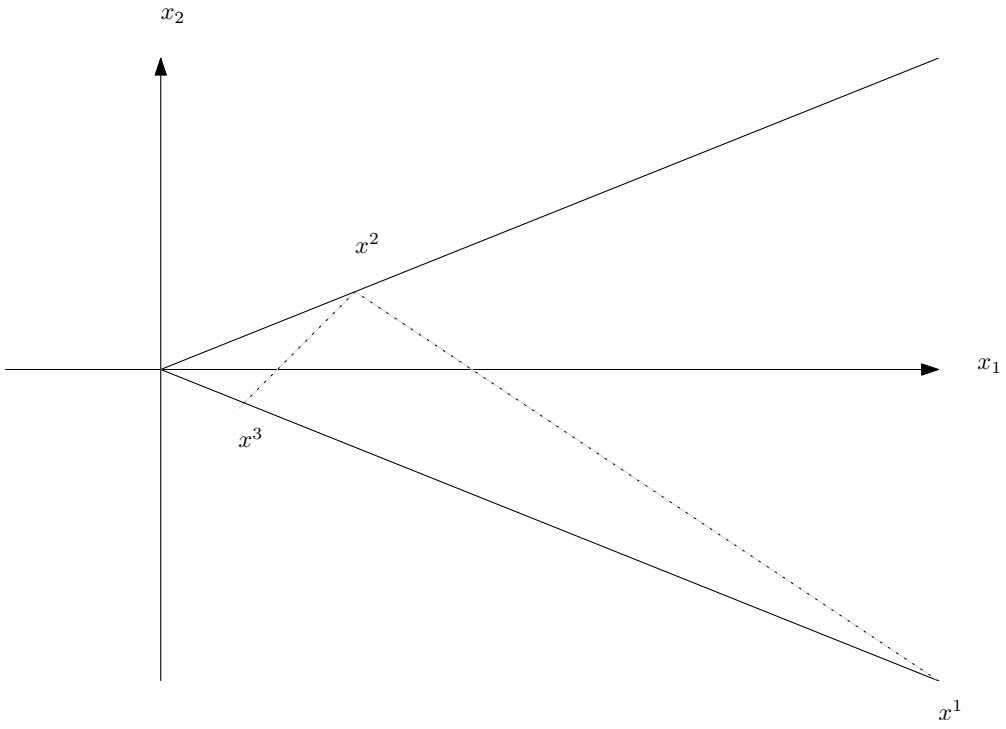


Figure 5.2: Zigzagging of trajectory

5.2.1 Stablisation via the ε-subdifferential

Descent methods try to achieve the following list of requirements:

1. The sequence $\left\{x^k\right\}$ produces $\left\{f\left(x^k\right)\right\}_{k\geq 1}$ that is strictly decreasing
2. the sequence $\left\{x^k\right\}_{k\geq 1}$ is a minimizing sequence in that

$$\liminf f\left(x^k\right) = f(\bar{x}) \leq f(x) \quad \text{for all } x$$

   where $\bar{x}$ is a cluster point of $\left\{x^k\right\}_{k\geq 1}$.
3. We require $\bar{x} \in \arg\min f$.

It is debatable whether it is desirable to seek all this list of stringent requirements. In particular monotonicity $\left\{f\left(x^k\right)\right\}_{k\geq 1}$ could be a a very detrimental requirement. The rejection of this requirements leads to non-monotone methods. To understand the failure of convergence in Example 5.2.1 we need to consider the sequence of distance functions

$$\left\{D^k := d\left(0, \partial f\left(x^k\right)\right) = \left\|g^k\right\|\right\}$$

where $g^k = P_{\partial f(x^k)}(0)$. Now the multi-function $x \mapsto \partial f(x)$ has a closed graph i.e. $g^k \in \partial f\left(x^k\right)$ then

$$\text{if } \left(x^k, g^k\right) \to (\bar{x}, \bar{g}) \quad \text{we have} \quad \bar{g} \in f(\bar{x}). \quad \text{(upper semicontinuity of multifunction)}$$

Thus *if* $D^k \to 0$ then we have $0 \in \partial f(\bar{x})$ but our example generates a sequence where $g^k = \left(3, (-1)^k 2\right)$ so $D^k \geq \sqrt{13} > 0$.

This is a challenging aspect of steepest descent in nonsmooth optimisation as we actually require some kind of continuity of $x \mapsto \partial f(x)$ in order to ensure $D^k \to 0$. This would require the additional assumption that the $x \mapsto \partial f(x)$ satisfies the following lower semi-continuity assumption (which does not hold in general):

$$\text{if } x^k \to \bar{x} \quad \text{and } \bar{g} \in \partial f(\bar{x}) \quad \text{then there exists } g^k \in \partial f\left(x^k\right) \quad \text{with } g^k \to \bar{g}.$$
$$\text{(lower semicontinuity of multifunction)}$$

(Consider $f(x) = |x|$ and $\bar{x} = 0$ then $\left|g^k\right| = 1$ for all k and cannot converge to $0 \in \partial f(0)$).

If one is to pursue this approach then a viscosity parameter that smears the subdifferential needs to be introduced. In the convex case we use the following "enlargement" called the ε-subdifferential

$$\partial_\varepsilon f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x\rangle - \varepsilon \text{ for all } y \in \mathbf{R}^n\}.$$

Now $g \in \partial_\varepsilon f(x)$ iff

$$\begin{aligned} f(y) &\geq f(x) + \langle g, y - x\rangle - \varepsilon \text{ for all } y \in \mathbf{R}^n \\ \langle g,\ x\rangle - f(x) &\geq \langle g, y\rangle - f(y) - \varepsilon \text{ for all } y \in \mathbf{R}^n \\ \text{or} \quad \langle g,\ x\rangle - f(x) &\geq f^*(g) - \varepsilon \\ \text{or} \qquad \langle g, x\rangle &\geq f^*(g) + f(x) - \varepsilon. \end{aligned} \tag{5.2}$$

Thus the ε-subdifferential loosens the Fenchel–Young inequality (FYI).

Note that $0 \in \partial_\varepsilon f(x)$ implies $f(y) \geq f(x) - \varepsilon$ for all $y \in \mathbf{R}^n$ and so this is a kind of approximate stationarity, called ε-optimal.

Proposition 5.2.1 *Suppose $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is a closed, proper convex function then $(x, \varepsilon) \mapsto \partial_\varepsilon f(x)$ has a closed graph i.e. for any $(x_n, g_n) \in \operatorname{Graph} \partial_{\varepsilon_n} f := \{(x, g) \mid g \in \partial_{\varepsilon_n} f(x)\}$ and $\varepsilon_n \to \varepsilon$ with $(x_n, g_n) \to (x, g)$ we have $(x, g) \in \operatorname{Graph} \partial_\varepsilon f$ or $g \in \partial_\varepsilon f(x)$.*

Proof. Note that both f and f^* are lower semi-continuous, proper functions and so

$$\liminf_n f(x_n) \geq f(x) \quad \text{and} \quad \liminf_n f^*(g_n) \geq f^*(g)$$

$$\text{implying} \quad \langle g, x\rangle = \liminf_n \langle g_n, x_n\rangle \overset{(5.2)}{\geq} \liminf_n \left[f(x_n) + f^*(g_n) - \varepsilon_n\right] \geq f(x) + f^*(g) - \varepsilon.$$

and so $g \in \partial_\varepsilon f(x)$. □

In variational analysis we have terminology for this kind of limit:

$$\partial_\varepsilon f(x) \subseteq \limsup_{\substack{x' \to x \\ \varepsilon' \to \varepsilon}} \partial f_{\varepsilon'}(x'),$$

where the lim sup is formalized in the following definition.

Definition 5.2.1 *Given a set valued mapping $\Gamma : \mathbb{R}^n \to \mathbb{R}^n$ we have*

$$\limsup_{x' \to x} \Gamma\left(x'\right) := \{y \in \mathbb{R}^n \mid \exists x_n \to x \text{ and } \exists y_n \in \Gamma(x_n) \text{ with } y_n \to y\}.$$

That is the set of all accumulation points of images locally around x.

Since f is convex we can assume it is Lipschitz continuous (if finite) then one can show that we have for any $\varepsilon > 0$ some $\delta > 0$ such that for $\left\|x^k - \bar{x}\right\| \leq \delta$ implies $\partial_\varepsilon f\left(x^k\right) \subseteq \partial_\varepsilon f\left(\bar{x}\right) + B_\delta\left(0\right)$. This is an outer approximation of $\partial f\left(x\right)$ in that $\partial f\left(x\right) \subseteq \partial_\varepsilon f\left(x\right)$ for all $\varepsilon \geq 0$. Now suppose x^k is not ε-optimal. Notice that the condition $g \in \partial_\varepsilon f\left(x^k\right)$ is equivalent to

$$f(x^k + td) - f(x^k) + \varepsilon \geq \langle g, td\rangle \quad (\forall (t,d) \in (0,\infty) \times X)$$

$$\text{or} \quad f'_\varepsilon\left(x^k, d\right) := \inf_{t>0} \frac{1}{t}\left(f\left(x^k + td\right) - f\left(x^k\right) + \varepsilon\right) \geq \langle g, d\rangle.$$

Suppose we can find $g^k \in \partial_\varepsilon f\left(x^k\right)$ and a direction d^k that attain the following minimum:

$$f'_\varepsilon\left(x^k, d^k\right) = \min_{\|d\|=1} f'_\varepsilon\left(x^k, d^k\right) = \langle g^k, d^k\rangle < 0.$$

This is just

$$\inf_{t>0} \frac{1}{t}\left(f(x^k + td^k) - f(x^k) + \varepsilon\right) = \langle g^k, d^k\rangle < 0.$$

Thus, for some $t^k > 0$, we have the strict descent:

$$f\left(x^k + t^k d^k\right) - f\left(x^k\right) + \varepsilon \leq 0 \quad \text{or} \quad f\left(x^k + t^k d^k\right) \leq f\left(x^k\right) - \varepsilon < f\left(x^k\right).$$

Moreover, we will show in Proposition 5.2.2 that if we have $0 \in \partial_\varepsilon f\left(x\right)$ then there exists $g^k \in \partial_\varepsilon f\left(x^k\right)$ with $g^k \to 0$ so

$$f'_\varepsilon\left(x^k, d^k\right) = \min_{\|d\|=1} f'_\varepsilon\left(x^k, d^k\right) \overset{(5.1)\ \varepsilon\text{-version}}{=} -\min_{g \in \partial_\varepsilon f\left(x^k\right)} \|g\| \to 0.$$

Proposition 5.2.2 *Suppose $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is a closed, proper convex function then $\left(x, f\left(x\right), \varepsilon\right) \mapsto \partial_\varepsilon f\left(x\right)$ is lower semi-continuous on $\operatorname{dom} f \times \mathbf{R}_+$ i.e. if $\left(x^k, f\left(x^k\right)\right) \to \left(x, f\left(x\right)\right)$ and $\varepsilon^k \to \varepsilon > 0$ with $g \in \partial_\varepsilon f\left(x\right)$ then there exists $g^k \in \partial_{\varepsilon^k} f\left(x^k\right)$ with $g^k \to g$.*

Proof. **Case 1**: We will show that if

$$f\left(y\right) > f\left(x\right) + \langle g, y - x\rangle - \varepsilon \text{ for all } y \in \mathbf{R}^n \tag{5.3}$$

then we have $g \in \partial_{\varepsilon^k} f\left(x^k\right)$ for k sufficiently large. By (5.3) we have a $\delta > 0$ such that

$$f\left(y\right) \geq f\left(x\right) + \langle g, y - x\rangle - \varepsilon + \delta \text{ for all } y \in \mathbf{R}^n \tag{5.4}$$

and as $f\left(x^k\right) \to f\left(x\right)$, $\varepsilon^k \to \varepsilon$ and $x^k \to x$ we have that for k sufficiently large:

$$\begin{aligned}
&|\overbrace{f\left(x^k\right) + \langle g, y - x^k\rangle - \varepsilon^k}^{=:F^k} - \overbrace{\left(f\left(x\right) + \langle g, y - x\rangle - \varepsilon\right)}^{-\delta + RHS((5.4))}| \\
&= \left|f\left(x^k\right) - f\left(x\right) + \langle g, x - x^k\rangle + \left(\varepsilon - \varepsilon^k\right)\right| \\
&\leq \left|f\left(x^k\right) - f\left(x\right)\right| + \|g\|\left\|x - x^k\right\| + \left|\varepsilon - \varepsilon^k\right| \leq \delta \quad \text{(triangle and C-S)}
\end{aligned}$$

Consequently,

$$0 \geq f(x^k) + \langle g, y - x^k\rangle - \varepsilon^k - (f(x) + \langle g, y - x\rangle - \varepsilon) - \delta \tag{5.5}$$

Adding (5.4) to (5.5) yields

$$f(y) \geq f\left(x^k\right) + \langle g, y - x^k\rangle - \varepsilon^k \text{ for all } y \in \mathbf{R}^n.$$

This shows that $g \in \partial_{\varepsilon^k} f\left(x^k\right)$ for k sufficiently large.

Case 2: Now consider $g \in \partial_\varepsilon f(x)$ on the boundary of $\partial_\varepsilon f(x)$ so that $f'_\varepsilon(x,d) = \langle g, d\rangle$ for some d. There exists $g_n \to g$ with $g_n \in \partial_\varepsilon f(x)$ so that
$f(y) > f(x) + \langle g_n, y - x\rangle - \varepsilon - \delta_n$ (for any $\delta_n > 0$ with $\delta_n \downarrow 0$) for all $y \in \mathbf{R}^n$.
As $x^k \to x$ for each n there exists k_n such that $\left\|x - x^{k_n}\right\| \leq \frac{\delta_n}{3\|g_n\|}$, $\left|f(x) - f\left(x^{k_n}\right)\right| \leq \frac{\delta_n}{3}$, $\left|\varepsilon + \delta_n - \left(\varepsilon^{k_n} + \delta_n\right)\right| \leq \frac{\delta_n}{3}$ and hence—by our argument for case 1—we have
$g_n \in \partial_{\varepsilon^k + \delta_n} f\left(x^k\right)$ for $k \geq k_n$. Thus

$$d\left(g, \partial_{\varepsilon^k+\delta_n} f\left(x^k\right)\right) \leq \|g - g_n\| \quad \text{for } k \geq k_n$$

$$\text{implying} \qquad \lim_{k,n\to\infty} d\left(g, \partial_{\varepsilon^k+\delta_n} f\left(x^k\right)\right) = 0,$$

and so we have the existence of $g^{k,n} \in \partial_{\varepsilon^k+\delta_n} f\left(x^k\right)$ with $g^{k,n} \to g$ for any $\delta_n \downarrow 0$. For each k we may find $g^{k,n} \to_{n\to\infty} g^k \in \partial_{\varepsilon^k} f\left(x^k\right) = \bigcap_{\delta>0} \partial_{\varepsilon^k+\delta} f\left(x^k\right)$ and so we have $\|g^k - g\| \leq \|g^k - g^{k,n}\| + \|g^{k,n} - g\| \to 0$. □

Problem Set 11:

Problem 5.2.3 *Consider $f(x) = |x|$ at $x = 0$ and $0 \in \partial f(0)$. Show that there exists $g_n \in \partial_{1/2} f(x_n)$ for any $x_n \to 0$ with $g_n \to 0$.*

Problem 5.2.4 ***(Von Neumann's Minmax theorem)***
Suppose Y is a Euclidean space. Suppose that the sets $C \subseteq X$ and $D \subseteq Y$ are nonempty and convex with D closed and that the map $A : X \to Y$ is linear.

1. *By considering the Fenchel problem*

$$\inf_{x\in X} \{\delta_C(x) + \delta_D^*(Ax)\}$$

prove that under the assumption

$$0 \in \operatorname{core}\left(\operatorname{dom}\delta_D^* - AC\right) \tag{5.6}$$

it holds that

$$\inf_{x\in C} \sup_{y\in D} \langle y, Ax\rangle = \max_{y\in D} \inf_{x\in C} \langle y, Ax\rangle$$

where the max *is attained if finite.*

2. *Prove that (5.6) holds in either of two cases.*
   (a) *D is bounded, or*
   (b) *A is onto (surjective) and $0 \in \operatorname{int} C$.*
   [Hint: Use the fact that the image of an open set under a continuous linear mapping is also an open set.]

3. *Suppose both C and D are closed and bounded, and that X is also a Euclidean space. Prove that in the finite case ($p = d < \infty$):*

$$\min_{x\in C} \max_{y\in D} \langle y, Ax\rangle = \max_{y\in D} \min_{x\in C} \langle y, Ax\rangle.$$

More about ϵ-subdifferentials

When

$$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

is the maximum of a finite number of convex functions then one can construct a $g \in \partial_\varepsilon f(x)$ from any selection $s_i \in \partial f_i(x)$. Indeed we claim

$$\partial_\varepsilon f(x) = \left\{ \sum_{i=1}^m \lambda_i s_i \mid \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i = 1 \text{ and } f(x) \leq \sum_{i=1}^m \lambda_i f_i(x) + \varepsilon \right\}$$

Note that all functions and subgradients are involved in this calculation (unlike the case of a maximum of a finite number of differential functions and the ordinary subdifferential). We will show only the direction $\supseteq$. Take any $\lambda \geq 0$ such that

$$\sum_{i=1}^m \lambda_i = 1 \quad \text{and} \quad f(x) \overset{(a)}{\leq} \sum_{i=1}^m \lambda_i f_i(x) + \varepsilon.$$

Now notice that

$$\begin{aligned}
& & f_i(y) &\geq f_i(x) + \langle s_i, y - x\rangle \quad \text{for all } i \text{ and } y \\
& \text{then} & \sum_{i=1}^m \lambda_i f_i(y) &\geq \sum_{i=1}^m \lambda_i f_i(x) + \langle \sum_{i=1}^m \lambda_i s_i, y - x\rangle \quad \text{for all } y \\
\text{so} \quad f(y) = \sum_{i=1}^m \lambda_i \max_{j=1,\ldots,m} f_j(y) &\geq \sum_{i=1}^m \lambda_i f_i(y) + \langle \sum_{i=1}^m \lambda_i s_i, y - x\rangle & & \\
& \overset{\text{using (a)}}{\geq} f(x) + \langle \sum_{i=1}^m \lambda_i s_i, y - x\rangle - \varepsilon \quad \text{for all } y & & \\
& \text{hence} \quad \sum_{i=1}^m \lambda_i s_i \in \partial_\varepsilon f(x). & &
\end{aligned}$$

In fact one can show that the set of all such ε- subdifferentials constitute the whole of $\partial_\varepsilon f(x)$. In our Example 5.2.1 we have

$$\partial_\varepsilon f(9,-3) = \operatorname{co}\{(3,-2), (2,-5), (3,2)\}$$

and so

$$d^1 = \Pr_{\partial_\varepsilon f(9,-3)}(0) = \left(\frac{140}{47}, -\frac{46}{47}\right).$$

In this case, our first update would be as in Figure 5.3

The main issue with steepest descent even when we refine it using the ε-subdifferential is that we require information about the whole of $\partial_{\varepsilon^k} f\left(x^k\right)$ for each k in order to calculate a descent direction. Moreover we then have to solve an optimisation problem to find this descent direction. Thus we do not pursue this method any further here.

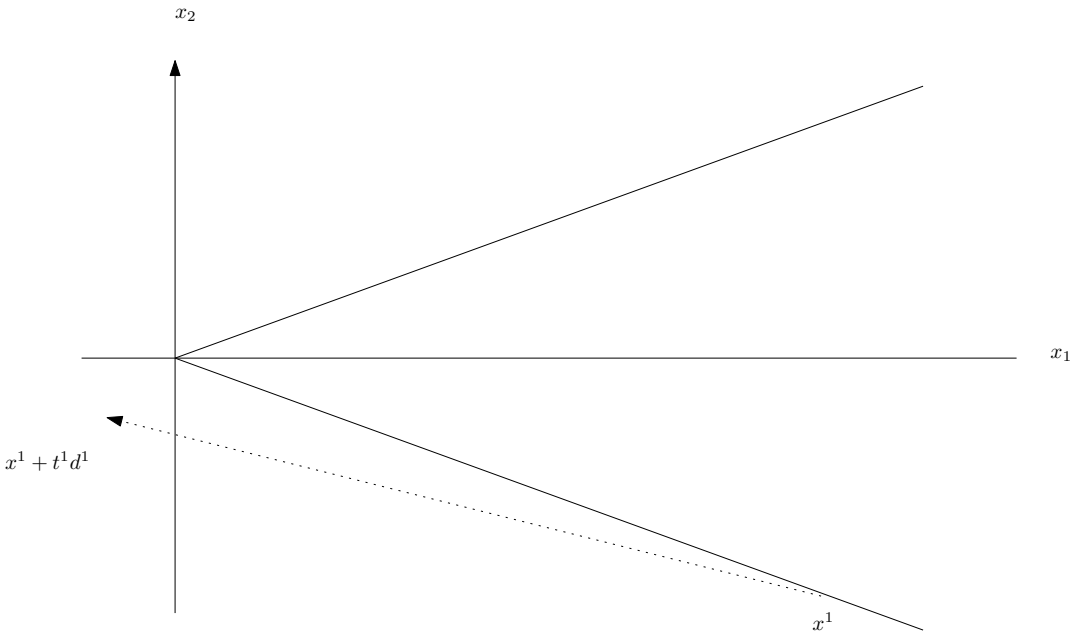


Figure 5.3: The ε-steepest descent

Chapter 6

Differentiability, Convexity and Approximation

In this section we will use the tools we have developed in previous sections to study a bridge from classical differentiability to the subdifferential of nonsmooth analysis.
Black-Box Evaluation:
From now on we will only consider methods that assume we can sample one subgradient $g^k \in \partial f\left(x^k\right)$ at each point x^k. Thus we assume the user supplies a black box subgradient evaluator: The black box is a routine that could be a simulation, a numerical estimator

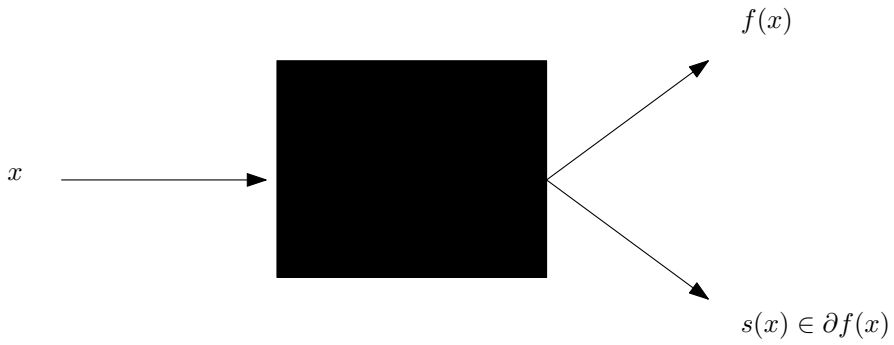


Figure 6.1: The internal mechanics of evaluating a subgradient are ignored.

of a analytic functional evaluation. We give some techniques to implement this section.

6.1 Connections to Differentiability

We will from time to time use the following very important results. We state this without proof.

Theorem 6.1.1 (Rademacher) *A locally Lipschitz function f between two Euclidean spaces is differentiable almost everywhere (in particular this implies dense differentiability).*

We know that convex functions are locally Lipschitz in any ball contained in the interior of $\operatorname{dom} f$. That is, in a ball in which f is bounded above. Can we use these point of differentiability to get the subdifferential for convex functions? Denote
$S_1(f) := \{x' \mid \nabla f(x') \text{ exists}\}$.

Remember the definition (Definition 5.2.1) of the limsup of a set-valued mapping is:

$$\limsup_{x' \to x} \Gamma\left(x'\right) := \left\{y \in \mathbb{R}^n \mid \exists x_n \to x \text{ and } \exists y_n \in \Gamma\left(x_n\right) \text{ with } y_n \to y\right\}.$$

That is the set of all accumulation points of images locally around x. We have the following theorem.

Theorem 6.1.2 *Suppose $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is convex and $x \in \operatorname{int} \operatorname{dom} f$ then*

$$\partial f\left(x\right) = \overline{\operatorname{co}}\left[\limsup\left\{\nabla f\left(x'\right) \mid x' \in S_1\left(f\right) \text{ and } x' \to x\right\}\right] = \limsup_{\substack{x' \to x \\ f(x') \to f(x)}} \partial f\left(x'\right).$$

Moreover the limit is insensitive to the removal of null sets from $S_1\left(f\right)$.

Proof. To prove the result, we show the following inclusions:

$$\begin{aligned}
\partial f(x) &\supseteq \overline{\operatorname{co}}\left[\limsup\left\{\nabla f\left(x'\right) \mid x' \in S_1\left(f\right) \text{ and } x' \to x\right\}\right] && \text{(6.1a)}\\
&\supseteq \limsup_{\substack{x' \to x \\ f(x') \to f(x)}} \overline{\operatorname{co}}\left[\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1\left(f\right) \text{ and } x'' \to x'\right\}\right] && \text{(6.1b)}\\
&\supseteq \limsup_{\substack{x' \to x \\ f(x') \to f(x)}} \partial f\left(x'\right) && \text{(6.1c)}\\
&\supseteq \partial f(x). && \text{(6.1d)}
\end{aligned}$$

(6.1d): To see why this holds, just choose $x'_n = x$ for all n.
(6.1a): By Remark 4.1.2, when $x' \in S_1(f)$, we know $\left\{\nabla f\left(x'\right)\right\} = \partial f\left(x'\right)$. Moreover by Proposition 4.1.15 we have Graph ∂f closed. From these two facts, (6.1a) clearly follows.
(6.1b and 6.1c): It suffices to show that

$$\overline{\operatorname{co}}\left[\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1\left(f\right) \text{ and } x'' \to x'\right\}\right] \supseteq \partial f\left(x'\right),$$

when $x' \in \operatorname{int} \operatorname{dom} f$. If we can show this, then a composite upper limits lead to

$$\begin{aligned}
&\overline{\operatorname{co}}\left[\limsup\left\{\nabla f\left(x'\right) \mid x' \in S_1\left(f\right) \text{ and } x' \to x\right\}\right]\\
&\supseteq \limsup_{\substack{x' \to x \\ f(x') \to f(x)}} \overline{\operatorname{co}}\left[\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1\left(f\right) \text{ and } x'' \to x'\right\}\right] \supseteq \limsup_{\substack{x' \to x \\ f(x') \to f(x)}} \partial f\left(x'\right),
\end{aligned}$$

which shows (6.1b) and (6.1c).
Suppose $x' \in \operatorname{int} \operatorname{dom} f$ For any h such that $\left\{t \mid x' + th \in S_1\left(f\right)\right\}$ has full measure (and there is a set of full measure of such hs), Fubini's theorem yields

$$f\left(x' + th\right) - f\left(x'\right) = \int_0^t \left\langle \nabla f\left(x' + \tau h\right), h\right\rangle d\tau.$$

Then as $f'\left(x', h\right) = \liminf_{t \downarrow 0} \frac{1}{t}\left(f\left(x' + th\right) - f\left(x'\right)\right)$ we have

$$\begin{aligned}
f'\left(x', h\right) &= \liminf_{t \downarrow 0} \frac{1}{t}\left(f\left(x' + th\right) - f\left(x'\right)\right) && \text{(6.2)}\\
&= \liminf_{t \downarrow 0} \frac{1}{t} \int_0^t \left\langle \nabla f\left(x' + \tau h\right), h\right\rangle d\tau\\
&\leq \frac{1}{t} \int_0^t S\left(\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1\left(f\right) \text{ and } x'' \to x'\right\}, h\right) d\tau && \text{(6.3)}
\end{aligned}$$

$$\leq \left[\frac{1}{t}\int_0^t d\tau\right] S\left(\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1(f) \text{ and } x'' \to x'\right\}, h\right)$$
$$= S\left(\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1(f) \text{ and } x'' \to x'\right\}, h\right) \quad (6.4)$$
$$= S\left(\overline{\text{co}}\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1(f) \text{ and } x'' \to x'\right\}, h\right). \quad (6.5)$$

Using the fact that $S\left(\partial f\left(x'\right), h\right) = f'\left(x', h\right)$ (Theorem 4.1.7) together with (6.5), we have

$$S\left(\partial f\left(x'\right), h\right) \leq S\left(\overline{\text{co}}\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1(f) \text{ and } x'' \to x'\right\}, h\right) \quad (6.6)$$

for a dense set of h. Now we need only use the density of that set of hs and the fact that closed convex sets C_1 and C_2 satisfy

$$S(C_1, \cdot) \leq S(C_2, \cdot) \quad \Longleftrightarrow \quad C_1 \subseteq C_2.$$

to conclude. When $h \mapsto S(C, h)$ is a densely defined proper convex function its domain must then be the whole space (being convex) and hence $h \mapsto S(C, h)$ is finite and Lipschitz continuous. Hence we may claim (6.6) holds for all h (extended value limits obtained using continuity). Thus

$$\partial f\left(x'\right) \subseteq \overline{\text{co}}\limsup\left\{\nabla f\left(x''\right) \mid x'' \in S_1(f) \text{ and } x'' \to x'\right\}.$$

Note that this argument only requires a set of full measure S on which ∇f exists and hence we can exclude a null set from $S_1(f)$ and still obtain equality. □

It is clear that for any locally Lipschitz functions we can form a convex hull of the limit of gradients that exists locally (by Rademacher's theorem).

The main idea behind some methods is to build an approximation of the subdifferential on the run. Given a sequence $g^k \in \partial f\left(x^k\right)$, at step n we can take $\text{co}\left\{g^k \mid k \in \{n, n-1, \ldots, n-l\}\right\}$ (for some l) as an estimate for $\partial f\left(x^n\right)$. Moreover, when seek an element $g^k \in \partial f\left(x^k\right)$ clearly we can take $g^k = \nabla f\left(x^k\right)$ if $x^k \in S_1(f)$, an option that exists with high probablility.

Problem Set 12:

Problem 6.1.3 *Use Theorem 6.1.2 to calculate $\partial f(0)$ where f is the function:*

1. $f(x) = \|x\|_2$
2. $f(x) : \mathbb{R}^2 \to \mathbb{R} : x \mapsto \|x\|_1 = |x_1| + |x_2|$.
3. $f : \mathbb{R}^2 \to \mathbb{R} : x \mapsto \|x\|_\infty = \max\{|x_1|, |x_2|\}$.

6.2 Subgradient Methods

Given our discussion of the pitfalls of steepest descent, an even more simplistic approach would hardly seem worthy of investigation. However, subgradient methods have proven to be very useful in areas such the solution of Lagrangian relaxation problems in integer programming. These apply only to the case of convex functions f. Thus they deserve discussion. Analogously to smooth steepest descent, these methods seek descent in the direction

$$d^k := -\frac{s\left(x^k\right)}{\|s\left(x^k\right)\|}.$$

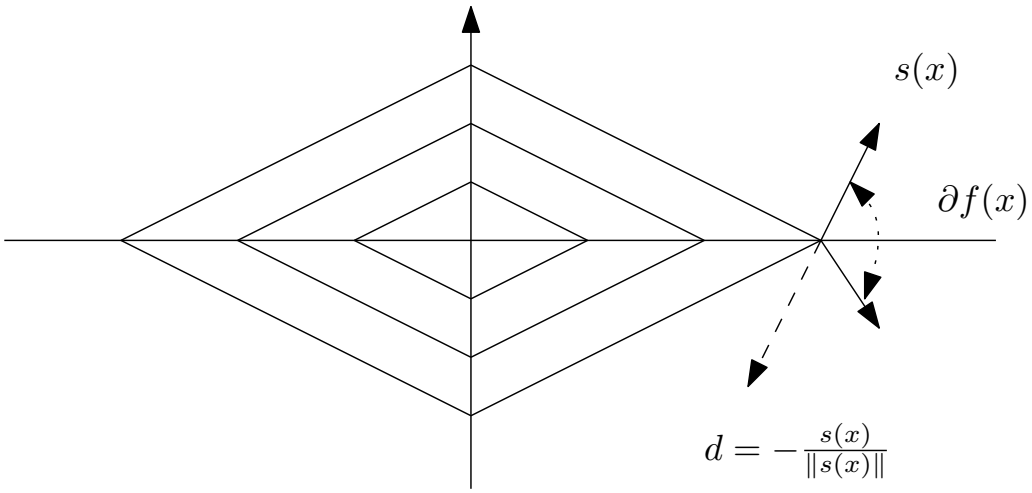


Figure 6.2: The negative subgradient is not necessarily a descent direction.

Clearly we are not assured descent (e.g. Figure 6.2).

Their characteristics are:

1. They are applied directly to nondifferentiable problems
2. They are simple to code and require low levels of memory to implement (an advantage for large scale problems)
3. They do not assure descent at every step (and often the function values do increase during some iterates).
4. They use a predetermined step length (no line search is used).
5. When combined with primal-dual decomposition methods, it is sometime possible simple to use distributed algorithms for a problem (and thus are then quite useful).

We will try and solve the following problem:

$$\min f(x) \quad \text{subject to } x \in C,$$

where f is a finite, proper, convex function and C is closed and convex (and so admits a tractable projection P_C onto C).
We follow the following algorithmic pattern:

General Subgradient Method: Initially we start with $\mathbf{x}^0 \in \mathbf{R}^n$ and $k = 0$.
While $0 \notin \partial f\left(x^k\right)$ **do** iteration k:

1. Given x^k call the Backbox to obtain $s\left(x^k\right) \in \partial f\left(x^k\right)$ and choose $d^k = -s\left(x^k\right)$ as a search direction for f at x^k.
2. Choose a step length $t^k > 0$ (via a pre-specified rule), and *if possible* (not enforced here) satisfying the condition

$$f(x^k + t^k d^k) < f(x^k).$$

Let $x^{k+1} = P_C\left(x^k + t^k d^k\right)$ and $k \leftarrow k+1$.
End While:

As this is not a descent method (i.e. descent is not guaranteed) the best function value and point found so far are usually kept in the ledger:

$$f_{best}^k := \min\left\{f\left(x^1\right), \ldots, f\left(x^k\right)\right\}$$

and x_{best}^k is the value that achieves this best function value.

6.2.1 Step Size Rules

At step k let $g^k := s\left(x^k\right)$ and place $x^{k+1} = P_C\left(x^k - t^k g^k\right)$.
Note that this g^k has not been normalized. There are many different step size rules used but the main "golden rule" is to not prematurely decrease the step length (and thus stall convergence).

1. Constant step size: $t^k = \alpha > 0$ for all k.
2. Constant step length (scaled by subgradient): $t^k = \frac{\gamma}{\left\|g^k\right\|_2}$ where $\gamma > 0$. So $x^{k+1} = P_C\left(x^k - \gamma \frac{g^k}{\left\|g^k\right\|_2}\right) = P_C(x^k + \gamma d^k)$ where d^k is a normalized descent direction).
3. Square summable but not summable: We use
$$t^k > 0, \quad \sum_{k=1}^{\infty}\left(t^k\right)^2 < +\infty \quad \text{and} \sum_{k=1}^{\infty} t^k = \infty.$$
[Typically one can use $t^k = \frac{a}{b+k}$ where a, $b > 0$.]
4. Nonsummable rule: Use
$$t^k > 0, \quad t^k \to 0 \quad \text{and} \sum_{k=1}^{\infty} t^k = \infty.$$
[One can use $t^k = \frac{a}{\sqrt{k}}$ where $a > 0$.]
5. Nonsummable diminishing step lengths: Use the size $t^k = \frac{\gamma^k}{\left\|g^k\right\|}$ where (- effectively $x^{k+1} = P_C\left(x^k - \gamma^k \frac{g^k}{\left\|g^k\right\|_2}\right) = P_C\left(x^k + \gamma^k d^k\right)$, essentially the non-summable rule with the added normalization of the g_k).
$$\gamma^k > 0, \quad \gamma^k \to 0 \quad \text{and} \sum_{k=1}^{\infty} \gamma^k = \infty.$$
6. Polyak's step length: Assume the optimal objective value is known f^* then use:
$$t^k = \frac{\alpha\left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|^2}.$$

(Effectively $x^{k+1} = P_C\left(x^k - \left[\frac{\alpha(f(x^k)-f^*)}{\|g^k\|_2}\right]\frac{g^k}{\|g^k\|_2}\right)$, and so we have $x^{k+1} = P_C(x^k + \gamma^k d^k)$, $(d^k := -\frac{g^k}{\|g^k\|_2}$) with $\gamma^k = \frac{\alpha(f(x^k)-f^*)}{\|g^k\|^2}$).

In practice, we don't always know f^*, but people instead substitute a known lower bound on f^*. For example, if f is a nonnegative function, 0 is a known lower bound. If you substitute $f^* = 0$, $C = X$, and $\alpha = 1$, this step size rule may remind you of Newton–Raphson method, which is $x^{k+1} = x^k - f(x^k)f'(x^k)/f'(x^k)^2$.

6.3 Convergence of Subgradient Methods.

We will assume f is finite valued and hence locally Lipschitz. Thus the subgradients are bounded by the Lipschitz constant L. We will also assume there is a number R such that $R \geq d\left(x^1, X^*\right)$ where

$$X^* := \arg\min\left\{f\left(x\right) \mid x \in C\right\}.$$

We seek results that either establish the existence of an accumulation point of $\left\{x^k\right\}$ that lies in the optimal set X^* (i.e. for some subsequence $\left\{x^{k_m}\right\}$ we have $x^{k_m} \to x^* \in X^*$) or that the whole sequence converges i.e. $x^k \to x^* \in X^*$).

Proposition 6.3.1 *The projection operator (onto a closed convex set) is a nonexpansive mapping i.e.* $\left\|P_C\left(x\right) - P_C y\right\| \leq \left\|x - y\right\|$.

Proof. Let $z \in C$ and $0 < \lambda < 1$ so that $z_\lambda := \lambda z + (1-\lambda) P_C\left(x\right) \in C$ (by convexity). Thus

$$\begin{aligned}\left\|x - P_C\left(x\right)\right\|^2 &\leq \left\|x - z_\lambda\right\|^2 \quad \text{(since } P_C(x) \text{ is the closest point in } C \text{ to } x\text{)}\\ &= \left\|\left(x - P_C\left(x\right)\right) - \lambda\left(z - P_C\left(x\right)\right)\right\|^2\\ &= \left\|x - P_C\left(x\right)\right\|^2 - \lambda\langle x - P_C\left(x\right), z - P_C\left(x\right)\rangle + \lambda^2\left\|z - P_C\left(x\right)\right\|^2\end{aligned}$$

Subtracting $\|x - P_C(x)\|^2$ from both sides and dividing by λ yields

$$0 \leq -\left\langle x - P_C\left(x\right), z - P_C\left(x\right)\right\rangle + \lambda\left\|z - P_C\left(x\right)\right\|^2 \to_{\lambda\to 0} -\left\langle x - P_C\left(x\right), z - P_C\left(x\right)\right\rangle.$$

Thus

$$\left\langle x - P_C\left(x\right), z - P_C\left(x\right)\right\rangle \leq 0. \tag{6.7}$$

In other words, the angle between these two difference vectors is obtuse (we have seen a version of this inequality before when studying support functions). When $z = P_C\left(y\right)$ we have

$$\left\langle x - P_C\left(x\right), P_C\left(y\right) - P_C\left(x\right)\right\rangle \leq 0. \tag{6.8}$$

As x is arbitrary in (6.7), we know that a similar inequality will hold if we replace the point x by the point y and choose $z = P_C\left(x\right)$:

$$\left\langle y - P_C\left(y\right), P_C\left(x\right) - P_C\left(y\right)\right\rangle \leq 0. \tag{6.9}$$

Adding (6.7) and (6.9) we have

$$\begin{aligned}0 &\geq \left\langle x - P_C\left(x\right) - \left(y - P_C\left(y\right)\right), P_C\left(y\right) - P_C\left(x\right)\right\rangle\\ &= \left\langle x - y - \left(P_C\left(x\right) - P_C\left(y\right)\right), P_C\left(y\right) - P_C\left(x\right)\right\rangle\end{aligned}$$

$$= -\langle x-y, P_C(x) - P_C(y)\rangle + \|P_C(y) - P_C(x)\|^2$$

Thus

$$\|P_C(y) - P_C(x)\|^2 \leq \langle x-y, P_C(x) - P_C(y)\rangle \overset{\text{C-S}}{\leq} \|x-y\| \|P_C(x) - P_C(y)\|$$
$$\text{or} \quad \|P_C(y) - P_C(x)\| \leq \|x-y\|.$$

□

For simple constraint sets like a sphere, a simplex, or a box, one can write down an analytic expression for projections. For a set defined by linear constraints $Ax \leq d$ then we can formulate the projection as a quadratic programming problem:

$$\min_y \frac{1}{2}\|y-x\|_2^2$$
$$\text{s.t. } Ay \leq d.$$

When $Ax \leq d$ then the projection of x onto $C := \{y \mid Ay \leq d\}$ is just x. Otherwise we can place $z = y - x$ and rewrite our problem as

$$\min_y \frac{1}{2}\|z\|_2^2$$
$$\text{Subj to } Az \leq d - Ax := b.$$

where $P_C(x) = y = z + x$. This can be reformulated via Lagrangian duality (see (4.37) as:

$$\min_{\lambda \geq 0} \left[\lambda^T b + \frac{1}{2}\lambda^T \left[AA^T\right]\lambda\right]$$

which is a differentiable optimisation problem. We will discuss methods to solve this later in the course. Once an optimal Lagrange multiplier $\lambda^* \geq 0$ is found, the primal solution is the solution to the equations given by $\nabla L(x, \lambda^*) = 0$ i.e.

$$\nabla_z \left(\frac{1}{2}\|z\|_2^2 + \lambda^{*T}(Az - (d - Ax))\right)$$
$$= z + A^T\lambda^* = 0 \quad \text{or} \quad P_C(x) = z + x = x + A^T\lambda^*.$$

The optimal Lagrange multiplier determines the active constraints. In other words, $\lambda_i^* > 0$ implies that the projection resides in the face $\{x \mid a_i x = d_i\}$ of the polyhedral set C. When C is a linear subspace, then we have the exact solution $b + \left[AA^T\right]\lambda^* = 0$ or $\lambda^* = -\left[AA^T\right]^{-1} b$ with $P_C(x) = z + x = x + A^T\left[AA^T\right]^{-1}(d - Ax)$. Then $AP_C(x) = Ax + AA^T\left[AA^T\right]^{-1}(d - Ax) = d$.

Theorem 6.3.2 *Suppose the subgradient method uses any one of the following rules:*

1) Square summable but not summable

2) Nonsummable

3) Nonsummable with diminishing step lengths

Then we have

$$\liminf_k f\left(x^k\right) = f^*, \tag{6.10}$$

when $\left\{x^k\right\}$ *is bounded.*

Proof. First consider the following sequence of inequalities: Let $\bar{x} \in C$ then

$$\begin{aligned}
\left\|x^{k+1} - \bar{x}\right\|^2 &= \left\|P_C\left(x^k - t^k g^k\right) - P_C \bar{x}\right\|_2^2 \\
\text{(nonexpansivity)} \quad &\leq \left\|x^k - t^k g^k - \bar{x}\right\|^2 \\
&= \left\|x^k - \bar{x}\right\|^2 - 2t^k \langle g^k, x^k - \bar{x}\rangle + \left(t^k\right)^2 \left\|g^k\right\|^2 \\
\text{(subgradient inequality)} \quad &\leq \left\|x^k - \bar{x}\right\|^2 - 2t^k \left(f\left(x^k\right) - f\left(\bar{x}\right)\right) + \left(t^k\right)^2 \left\|g^k\right\|^2 \qquad (6.11)
\end{aligned}$$

as the subgradient inequality gives

$$f\left(\bar{x}\right) - f\left(x^k\right) \geq \langle g^k, \bar{x} - x^k \rangle.$$

Applying this inequality recursively we generate a series

$$0 \leq \left\|x^{k+1} - \bar{x}\right\|^2 \leq \left\|x^1 - \bar{x}\right\|^2 \underbrace{-2\sum_{i=1}^{k} t^i \left(f\left(x^i\right) - f\left(\bar{x}\right)\right)}_{=(a)} + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2. \qquad (6.12)$$

Using $\left\|x^1 - \bar{x}\right\|^2 \leq R^2$ and adding term (a) to both sides yields

$$2\sum_{i=1}^{k} t^i \left(f\left(x^i\right) - f\left(\bar{x}\right)\right) \leq R^2 + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2. \qquad (6.13)$$

Combining this with

$$\sum_{i=1}^{k} t_i \left(f\left(x^i\right) - f\left(\bar{x}\right)\right) \geq \left(\sum_{i=1}^{k} t_i\right) \min_{i=1,\ldots,k} \left(f\left(x^i\right) - f\left(\bar{x}\right)\right) = \left(\sum_{i=1}^{k} t_i\right) \left(f_{best} - f\left(\bar{x}\right)\right),$$

we have the inequality

$$f_{best} - f\left(\bar{x}\right) = \min_{i=1,,k} \left(f\left(x^i\right) - f\left(\bar{x}\right)\right) \leq \frac{R^2 + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2}{2\left(\sum_{i=1}^{k} t_i\right)}. \qquad (6.14)$$

As $\left\{x^k\right\}$ is bounded we have $\left\|g^i\right\| \leq L$ (as f is finite valued and so Lipschitz continuous on bounded sets) we have

$$f_{best} - f\left(\bar{x}\right) \leq \frac{R^2 + L^2 \sum_{i=1}^{k} \left(t^i\right)^2}{2\left(\sum_{i=1}^{k} t_i\right)}. \qquad (6.15)$$

Note that since this holds for any feasible $\bar{x}$, it certainly holds for any optimal $\bar{x} \in X^*$. From this inequality we may verify the convergence results, starting with:

1) Square summable but not summable

We have $\|t\|_2^2 = \sum_{i=1}^{\infty} \left(t^i\right)^2$ so for $\bar{x} \in X^*$ we have $f\left(\bar{x}\right) = f^*$

$$f_{best} - f^* \overset{(6.15)}{\leq} \frac{R^2 + L^2 \|t\|_2^2}{2\left(\sum_{i=1}^{k} t_i\right)} \to_{k\to\infty} 0$$

as $\sum_{i=1}^{k} t_i \to \infty$.

2) Nonsummable

This is harder and corresponds to the "classical result": Choose N_1 such that $(i \geq N_1) \implies t^i \leq \varepsilon/L^2$ and choose N_2 such that

$$(i \geq N_2) \quad \implies \quad \sum_{i=1}^{N_2} t^i \geq \frac{1}{\varepsilon}\left(R^2 + L^2 \sum_{i=1}^{N_1} \left(t^i\right)^2\right),$$

which we can do because $\sum_{i=1}^{\infty} t^i = +\infty$. Let $N = \max\{N_1, N_2\}$ then for $k \geq N$ we have

$$(\text{since } N \geq N_2) \quad \sum_{i=1}^{k} t^i \geq \frac{1}{\varepsilon}\left(R^2 + L^2 \sum_{i=1}^{N_1} \left(t^i\right)^2\right) \tag{6.16}$$

and

$$(\text{since } N \geq N_1) \quad \sum_{i=N_1+1}^{k} (t^i)^2 \leq \left(\frac{\varepsilon}{L^2}\right) \sum_{i=N_1+1}^{k} t^i \tag{6.17}$$

Combining these facts, we have

$$\begin{aligned}
f_{best} - f^* &\overset{(6.15)}{\leq} \frac{R^2 + L^2 \sum_{i=1}^{k} \left(t^i\right)^2}{2\left(\sum_{i=1}^{k} t_i\right)} = \frac{R^2 + L^2 \sum_{i=1}^{N_1} (t^i)^2 + L^2 \sum_{i=N_1+1}^{k} \left(t^i\right)^2}{2\sum_{i=1}^{N_1} t_i + 2\sum_{i=N_1+1}^{k} t_i} \\
&\leq \frac{R^2 + L^2 \sum_{i=1}^{N_1} \left(t^i\right)^2}{2\sum_{i=1}^{k} t_i} + \frac{L^2 \sum_{i=N_1+1}^{k} \left(t^i\right)^2}{2\sum_{i=N_1+1}^{k} t_i} \\
&\leq \frac{R^2 + L^2 \sum_{i=1}^{N_1} \left(t^i\right)^2}{(2/\varepsilon)\left(R^2 + L^2 \sum_{i=1}^{N_1} \left(t^i\right)^2\right)} + \frac{L^2 \left(\frac{\varepsilon}{L^2}\right) \sum_{i=N_1+1}^{k} t^k}{2\sum_{i=N_1+1}^{k} t_i} \\
&= \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

use (6.16) for the left term and (6.17) for the right term

Thus, we again have that $f_{best} - f^*$ goes to zero.

3) Nonsummable diminishing step lengths

When $t^k = \gamma^k / \left\|g^k\right\|$ we have

$$\begin{aligned}
f_{best} - f^* &\overset{(6.14)}{\leq} \frac{R^2 + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2}{2\left(\sum_{i=1}^{k} t_i\right)} \\
&= \frac{R^2 + \sum_{i=1}^{k} \left(\gamma^i\right)^2}{2\left(\sum_{i=1}^{k} \gamma_i / \left\|g^k\right\|\right)} \\
&\leq \frac{R^2 + \sum_{i=1}^{k} \left(\gamma^i\right)^2}{(2/L)\left(\sum_{i=1}^{k} \gamma_i\right)} \to_{k\to\infty} 0
\end{aligned}$$

via a similar argument to (2) "nonsummable case" applied to γ instead of t.

□

The convergence result (6.10) is rather weak. So we explore further to determine more about rates of convergence and how $\left\{x^k\right\}$ behaves. Recall that a sequence $\left\{x^k\right\}$ converges linearly when

$$\frac{\left\|x^{k+1} - x^*\right\|}{\left\|x^k - x^*\right\|} \leq c < 1.$$

Theorem 6.3.3 *Suppose we use the "nonsummable diminishing step lengths rule" (i.e.* $x^{k+1} = P_C(x^k - \gamma^k \frac{g^k}{\|g^k\|_2})$ *with* $\gamma^k > 0, \quad \gamma^k \to 0 \quad$ *and* $\sum_{k=1}^{\infty} \gamma^k = \infty$*), with the addition that* $\sum_{k=1}^{\infty} \left(\gamma^k\right)^2 < +\infty$ *then* $x^k \to x^* \in X^*$*. In such a case, the convergence rate may only be sublinear.*

Proof. Using the same basic arithmetic argument we used in (6.11) we have that for $\bar{x} \in X^* \subseteq C$ and $t^k := \frac{\gamma^k}{\|g^k\|}$ we have

$$\begin{aligned}
\left\|x^{k+1} - \bar{x}\right\|^2 &= \|P_C(x^k - \gamma^k \frac{g^k}{\|g^k\|}) - P_C(\bar{x})\|^2 \\
\text{(nonexpansivity)} \quad &\leq \|x^k - \gamma^k \frac{g^k}{\|g^k\|} - \bar{x}\|^2 \\
&= \left\|x^k - \bar{x}\right\|^2 - 2\gamma^k \langle \frac{g^k}{\|g^k\|}, x^k - \bar{x}\rangle + \left(\gamma^k\right)^2 \\
\text{(subgradient inequality)} \quad &\leq \left\|x^k - \bar{x}\right\|^2 + 2\frac{\gamma^k}{\|g^k\|}\left(f\left(\bar{x}\right) - f\left(x^k\right)\right) + \left(\gamma^k\right)^2 \\
&\leq \left\|x^k - \bar{x}\right\|^2 + \left(\gamma^k\right)^2 \qquad (6.18)
\end{aligned}$$

because

$$0 \geq f\left(\bar{x}\right) - f\left(x^k\right) \geq \langle g^k, \bar{x} - x^k\rangle.$$

Subtracting $\|x^k - \bar{x}\|$ from both sides, we realize that we can have

$$\|x^{k+1} - \bar{x}\|^2 - \|x^k - \bar{x}\|^2 \leq (\gamma^k)^2,$$

and so we can telescope and use the fact that $\sum_{k=1}^{\infty} \left(\gamma^k\right)^2 < +\infty$. We obtain that for all m:

$$\begin{aligned}
\|x^m - \bar{x}\|^2 - \left\|x^1 - \bar{x}\right\|^2 &= \sum_{k=1}^{m} \left[\left\|x^{k+1} - \bar{x}\right\|^2 - \left\|x^k - \bar{x}\right\|^2\right] \\
&= \sum_{k=1}^{m} (\gamma^k)^2 \leq \sum_{k=1}^{\infty} \left(\gamma^k\right)^2 < +\infty.
\end{aligned}$$

Thus $\{x^m\}$ remains bounded. By the local boundedness of $\partial f\left(x^k\right)$ we must also have $\left\{g^k\right\}$ bounded. But by the basic convergences results we have $\liminf_k f\left(x^k\right) = f^*$ together with the boundedness of $\left\{x^k\right\}$ implies there exists an accumulation point $\bar{x}$ in X^*. Thus we have that one of the subsequences converges to the accumulation point $\bar{x} \in X^*$. Now to show the convergence of the main sequence $x^k \to \bar{x} \in X^*$, note that the tail $\sum_{i=k}^{\infty} \left(\gamma^i\right)^2 \to 0$ as $k \to \infty$. Applying (6.18) iteratively we have for all $m < k$ that

$$\left\|x^{k+1} - \bar{x}\right\|^2 \leq \|x^m - \bar{x}\|^2 + \sum_{i=m+1}^{k} \left(\gamma^i\right)^2 \leq \|x^m - \bar{x}\|^2 + \sum_{i=m+1}^{\infty} \left(\gamma^i\right)^2 .$$

Take a subsequence $x^{m_l} \to \bar{x}$ so that $\lim_l \|x^{m_l} - \bar{x}\|^2 = 0$, we have

$$\left\|x^{k+1} - \bar{x}\right\|^2 \leq \|x^{m_l} - \bar{x}\|^2 + \sum_{i=m_l+1}^{\infty} \left(\gamma^i\right)^2$$

$$\text{and so} \qquad \lim_{k\to\infty} \left\|x^{k+1} - \bar{x}\right\|^2 \leq \lim_{l} \left(\|x^{m_l} - \bar{x}\|^2 + \sum_{i=m_l+1}^{\infty} \left(\gamma^i\right)^2 \right) = 0$$

(using the fact that the tail of a convergent series goes to zero). Hence $x^k \to \bar{x} \in X^*$. Now we will show that the convergence rate may only be sublinear. Let $C = X$ and suppose for a contradiction that the convergence has linear rate, so $\left\|x^k - \bar{x}\right\| \leq c\theta^k$ for some $0 < \theta < 1$. Since $x^{k+1} = P_C(x^k - \gamma^k g^k/\|g^k\|)$, we have that

$$\begin{aligned} \|x^{k+1} - x^k\|^2 &= \|P_C(x^k - \gamma^k g^k/\|g^k\|) - x^k\|^2 \quad \text{(by definition)} \\ &= \|x^k - \gamma^k g^k/\|g^k\|^2 - x^k\|^2 \quad \text{(equality since } C = X \text{ so } P_C = I) \\ &= (\gamma^k)^2\|g^k\|^2/\|g^k\|^2 = (\gamma^k)^2. \end{aligned} \tag{6.19}$$

Thus we have that

$$\begin{aligned} \gamma^k &\overset{(6.19)}{=} \left\|x^{k+1} - x^k\right\| \leq \left\|x^{k+1} - \bar{x}\right\| + \left\|x^k - \bar{x}\right\| \leq c\theta^{k+1} + c\theta^k \leq 2c\theta^k \\ &\implies \qquad \sum_k \gamma^k \leq \sum_k 2c\theta^k < +\infty \quad \text{(geometric series } \textstyle\sum \theta^k \text{ converges).} \end{aligned}$$

which cannot happen, because we assumed the series γ^k does not converge. □

[Note that better step lengths for faster convergence is an area of ongoing research. Some methods for computing step lengths are given in these notes. In fact, Yura Malitsky gave an interesting talk on such research just last week. Methods based on gradient descent (specifically stochastic gradient descent) motivate the way that neural networks are trained. And so the questions are far from resolved.]

The above theorem uses the diminishing step lengths. Other rules (like the nonsummable rule and the square summable but nonsummable rule) do not normalise the search direction $-g^k$. In such cases, the iteration $x^{k+1} = P_C(x^k - t^k g^k)$ can fail to generate a convergent sequence due to the unboundedness of $\left\{g^k\right\}$. If one assumes boundedness of $\left\{g^k\right\}$, then the following theorem gives us similar convergence results to those in Theorem 6.3.3.

Theorem 6.3.4 *Suppose that the subgradient optimisation is applied, using the square summable but nonsummable rule: $x^{k+1} = P_C(x^k - t^k g^k)$ with $t^k > 0, \quad t^k \to 0$ and $\sum_{k=1}^{\infty} t^k = \infty$, where in addition $\sum_{k=1}^{\infty} \left(t^k\right)^2 < +\infty$. If the sequence $\left\{g^k\right\}$ is bounded we have $x^k \to x^* \in X^*$.*

Proof. By assumption we have $\left\|g^k\right\| \leq \Delta$ for all k.

Case 1: Suppose that $\left\|g^k\right\| \geq \delta > 0$ for all k. We can then define $\bar{t}^k := t^k \left\|g^k\right\|$ and we'll have that

$$x^{k+1} = P_C\left(x^k - t^k g^k\right) = P_C\left(x^k - \bar{t}^k \frac{g^k}{\|g^k\|}\right).$$

Clearly $\bar{t}^k \in \left[\delta t^k, \Delta t^k\right]$ for all k, and so for all N,

$$\sum_{k=1}^{N} \bar{t}^k \in \left[\delta \sum_{k=1}^{N} t^k, \Delta \sum_{k=1}^{N} t^k\right] \quad \text{and} \quad \sum_{k=1}^{N} (\bar{t}^k)^2 \in \left[\delta^2 \sum_{k=1}^{N} (t^k)^2, \Delta^2 \sum_{k=1}^{N} (t^k)^2\right],$$

whereupon $\bar{t}^k$ is both square summable and not summable. Then we may apply Theorem 6.3.3 for $\gamma^k = \bar{t}^k$ to get $x^k \to x^* \in X^*$.

Case 2: Suppose there is a subsequence of g^k that converges to zero. Let $x^* \in X^*$. Just as we argued in (6.18) in Theorem 6.3.3, we still have

$$\left\|x^{k+1} - x^*\right\| \overset{(6.18)}{\leq} \left\|x^k - x^*\right\| + \left(t^k\right)^2 \left\|g^k\right\|^2 \overset{\text{(telescoping)}}{\leq} \sum_{i=1}^{k} (t^i)^2 \|g^i\|^2 \overset{\text{(our assumptions)}}{<} \infty.$$

This shows that $\left\{x^k\right\}$ is bounded. Thus there is a *further* subsequence—of the subsequence of x^k such that $\left\{g^k\right\}$ converges to zero—that converges to some x^*. We pass to that subsequence and notice that as ∂f has closed graph and $g^k \to 0$, we must have $0 \in \partial f\left(x^*\right)$, which guarantees the optimality: $x^* \in X^*$. Convergence of the entire sequence follows in a similar way to Theorem 6.3.3. □

Remark 6.3.1 *A stopping criterion is available for the methods in Theorem 6.3.2. Letting $R \geq \left\|x^1 - x^*\right\|$ we may re-arrange (6.13) from*

$$2\sum_{i=1}^{k} t^i \left(f\left(x^i\right) - f\left(x^*\right)\right) \leq R^2 + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2$$

to the equivalent

$$l^k := \frac{2\sum_{i=1}^{k} t^i f\left(x^i\right) - R^2 - \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2}{2\sum_{i=1}^{k} t^i} \leq f\left(x^*\right) := f^*$$

which can be computed at the kth step. The sequence $\left\{l^k\right\}$ is non-monotonic so we keep track of the best lower bound so far

$$l_{best} := \max\left\{l^1, \ldots, l^k\right\}.$$

We can terminate when we have $f^k_{best} - l^k_{best}$ smaller than a threshold. In practise this goes slowly to zero, so often the subgradient method is used without such a formal stopping criterion. We usually stop when the improvement slows.

6.3.1 The Polyak Step Length

The rate of convergence of the subgradient method is disappointing and certainly motivates further research (even for convex functions). A better rate can be obtained by using Polyak's step length

$$t^k = \frac{\alpha\left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|^2} \quad (\geq 0)$$

and so $x^{k+1} = P_C\left(x^k - \left[\frac{\alpha\left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|_2}\right] \frac{g^k}{\left\|g^k\right\|_2}\right)$. We still have (6.13) as before. Applying it with $f\left(\bar{x}\right) = f^*$:

$$\begin{aligned}\left\|x^{k+1} - \bar{x}\right\|^2 &\overset{(6.11)}{\leq} \left\|x^k - \bar{x}\right\|^2 + \left(t^k\right)^2 \left\|g^k\right\|^2 - 2t^k\left(f\left(x^k\right) - f\left(\bar{x}\right)\right) \\ &\leq \left\|x^k - \bar{x}\right\|^2 + \alpha^2\left(\frac{\left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|^2}\right)^2 \left\|g^k\right\|^2 - 2\alpha\left[\frac{\left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|}\right]^2\end{aligned}$$

$$= \left\|x^k - \bar{x}\right\|^2 + \alpha\,(1-2\alpha)\left[\frac{\left(f\left(x^k\right) - f^*\right)}{\|g^k\|}\right]^2.$$

which gives a better reduction of error at each iteration when $\alpha > \frac{1}{2}$. Most implementation of this replaces f^* by a lower bound estimate. In fact it can be shown that the t^k chosen here minimizes the right hand side of (6.12) as a function of t when $\alpha = 1$. Remember that we have (6.13):

$$2\sum_{i=1}^{k} t^i\left(f\left(x^i\right) - f\left(\bar{x}\right)\right) \le R^2 + \sum_{i=1}^{k}\left(t^i\right)^2\left\|g^i\right\|^2.$$

Placing our Polyak's step length (with $\alpha = 1$) into this inequality yields

$$2\sum_{i=1}^{k}\frac{\left(f\left(x^i\right) - f^*\right)^2}{\left\|g^i\right\|^2} = 2\sum_{i=1}^{k} t^i\left(f\left(x^i\right) - f\left(x^*\right)\right) \le R^2 + \sum_{i=1}^{k}\left(\frac{\left(f^* - f\left(x^i\right)\right)}{\left\|g^i\right\|^2}\right)^2\left\|g^i\right\|^2$$

$$= R^2 + \sum_{i=1}^{k}\frac{\left(f^* - f\left(x^i\right)\right)^2}{\left\|g^i\right\|^2}$$

$$\text{and so} \quad \sum_{i=1}^{k}\frac{\left(f\left(x^i\right) - f^*\right)^2}{\left\|g^i\right\|^2} \le R^2.$$

If we assume the boundedness $\left\|g^i\right\| \le L$, then we get

$$k(f_{best} - f^*)^2 \overset{\text{(obvious)}}{\le} \sum_{i=1}^{k}\left(f\left(x^i\right) - f^*\right)^2 \overset{(\|g^i\| \le L)}{\le} R^2L^2.$$

We conclude that $f\left(x^i\right) - f^* \to 0$ and also observe that

$$\frac{(f_{best} - f^*)^2}{R^2L^2} \le \frac{1}{k},$$

which means an accuracy of ε is assured after $k = \left(\frac{LR}{\varepsilon}\right)^2$ steps. In other words, $f_{best} - f^* > \varepsilon$ implies $k < \left(\frac{LR}{\varepsilon}\right)^2$.

Problem Set 13

Problem 6.3.5 (Projections onto linear subspaces) *Let X be an inner product space and M a closed linear subspace of X. Define $M^\perp = \{x \in X \mid \langle x, e\rangle = 0$ for all $e \in M\}$. Assume that there is a unique solution to the minimum norm problem of finding $y \in M$ such that*

$$\|y - x\| = \min_{\hat{y} \in M}\|\hat{y} - x\|.$$

Let $e \in M$ and $\lambda > 0$. Place $z = x - y$ where y is the unique solution mentioned above.

1. *Explain why*
$$\|x - (y + \lambda e)\|^2 \ge \|z\|^2$$
and use this to deduce that $\lambda[\lambda\|e\|^2 - 2\langle z, e\rangle] \ge 0$.

2. *Deduce that $\langle z, e\rangle = 0$ for all $e \in M$.*

3. *Deduce that for all $x \in X$ there exists a* unique *pair $y \in M$ and*

$$z \in M^{\perp} := \{x \mid \langle x, m\rangle = 0 \text{ for all } m \in M\}$$

such that $x = y + z$.

4. *Show that $\|x\|^2 = \|y\|^2 + \|z\|^2$ and verify that $\langle Px, x\rangle = \|Px\|^2$.*

Problem 6.3.6 *A monotone operator T is a (potentially set valued) mapping satisfying the property that when $y \in T(x)$ and $u \in T(v)$ we have*

$$\langle x - v, y - u\rangle \geq 0.$$

[Note: We may assume via (6.7) we have for all $z \in C$ and any $x \notin C$ that for a closed convex set that $y = P_C(x)$ statisfies $\langle z - y, x - y\rangle \leq 0$.]

1. *Show that P_C satisfies the property that*

$$\langle P_C x - P_C y, x - y\rangle \geq \|P_C x - P_c y\|^2 \geq 0.$$

[Hint: We showed earlier in the course that any $e \in C$ satisfies $\langle e - P_C u, u - P_C u\rangle \leq 0$ for any u. Plug in $x = u$ and $e = P_C y$. Then plug in $y = u$ and $e = P_C x$. Add the two resulting equations together.]

2. *Prove the following identity: for any $\alpha \in \mathbb{R}$ we have*

$$\|\alpha x + (1-\alpha) y\|^2 + \alpha(1-\alpha)\|x - y\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2.$$

3. *Letting $\alpha := 2$ and $\mathbf{x} := Tx - Ty$ and $\mathbf{y} := x - y$, use the identity you just proved to verify that*

$$\|Rx - Ry\|^2 = 2\|Tx - Ty\|^2 - \|x - y\|^2 + 2\|(I - T)x - (I - T)y\|^2.$$

where $R := 2T - I$.

4. *A mapping is called nonexpansive when $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in X$. A mapping T is also called firmly non-expansive when*

$$\|Tx - Ty\|^2 + \|(I - T)x - (I - T)y\|^2 \leq \|x - y\|^2 \quad \text{for all} \quad x, y \in X.$$

Show the following are equivalent: for all $x, y \in X$ we have:

   (a) *T is firmly non-expansive:*

   (b) *The "reflection operator" $2T - I$ is nonexpansive*

   (c) *We have*

$$\langle Tx - Ty, x - y\rangle \geq \|Tx - Ty\|^2 \geq 0$$

   and hence T is monotone.

   [Hint: first show that 4a $\Longleftrightarrow$ 4b. To show that 4b $\Longleftrightarrow$ 4c, expand the term $\|(I - T)x - (I - T)y\|^2$ in the definition of firmly quasinonexpansive.]

6.4 Optimizing the Lagrangian Dual in Linear Programming

Consider the linear program (LP)

$$\begin{aligned} \min \quad & c^T x \\ \text{Subject to} \quad & Ax \geq b, \text{ and } Bx \geq d, \end{aligned}$$

where $X := \{x \mid Bx \geq d\}$ is compact. Assume the LP optimal solution set is non-empty, and for any $x^* \in X^*$ (where X^* still denotes the set of minimizers) we have $c^T x^* = v^*$. When an LP is too large to solve directly (these still exist) then we can obtain an lower bound on the optimal value of the LP by absorbing some of the hard constraints (say $Ax \geq b$) into the objective. Let $y \geq 0$ denote the multiplier for the constraint $Ax \geq b$, and consider solving the "relaxation" given by

$$\begin{aligned} f : \mathbb{R}^m_+ : y \mapsto \min_{x \in X} \quad & c^T x + (b - Ax)^T y \\ = \min_{x \in X} \quad & c^T x - x^T A^T y + b^T y \\ = \min_{x \in X} \quad & c^T x - \left[A^T y\right]^T x + b^T y \\ = \min_{x \in X} \quad & b^T y + \left(c - A^T y\right)^T x \end{aligned}$$

Let $F(x, y) := \left(c - A^T y\right)^T x$. Define the Lagrangian dual function as

$$f(y) = b^T y + \min_{x \in X} F(x, y).$$

In applications of interest the structure of X is such that the computation of $f(y)$ for a given y is much easier than solving the LP (in some cases trivial solutions exist). The Lagrangian dual is then the best lower bound

$$\max_{y \geq 0} f(y). \tag{LD}$$

It is clear that $f(y)$ is a concave function, because it is the infimum of a family of linear functions.

Proposition 6.4.1 *Let $X^*(y) = \left\{x \in X \mid f(y) = b^T y + F(x, y)\right\} \neq \emptyset$ (i.e. $X^*(y)$ is the set of x's such that their affine functions $u \mapsto b^T u + \left(c - A^T u\right)^T x$ are equal to f at y; it is the set of x's that minimize $F(x, y)$). Let X be compact (e.g. X is a closed, bounded subset of $\mathbb{R}^n$, a.k.a. the polytope generated by B). Then*

(i) $y \mapsto X^*(y)$ has a closed graph.

(ii) the super-differential of f at y is then $\partial^+ f(y) = \overline{\text{co}}\{b - Ax \mid x \in X^*(y)\}$.

Proof. (i): First we show that $y \mapsto X^*(y)$ has a closed graph. Take $x^k \in X^*\left(y^k\right)$ with $\left(x^k, y^k\right) \to (x', y^*)$. Then for each k we have $x^k \in X^*\left(y^k\right)$ and so

$$f\left(y^k\right) = b^T y^k + F\left(x^k, y^k\right).$$

As X is compact we have $y \mapsto f(y)$ finite. To see why, pick any $(u_j)_j \subset X$ such that

$$b^T y + (c - A^T y)^T u_j \to_{j\to\infty} \inf_{x\in X} b^T y + (c - A^T y)^T x = f(y).$$

Compactness guarantees that a subsequence of the u_j converges to some $u \in X$, whereupon continuity of the linear operator guarantees $b^T y + (c - A^T y)^T u = f(y)$.

Thus the function f is finite and concave, and so it is (Lipschitz) continuous over bounded sets in $\mathbb{R}^n$, while $\{y^k\}_k$ is bounded since $y^k \to y^*$. As $x^k \in X := \{x \mid Bx \geq d\}$ for all k we have $\lim_k Bx^k = Bx' \geq d$. As $y^k \to y^*$ (by assumption) and we have $\lim_k f\left(y^k\right) = f(y^*)$ it follows that

$$\begin{aligned} f\left(y^k\right) &= b^T y^k + F\left(x^k, y^k\right) \\ &= b^T y^k + \left(c - A^T y^k\right)^T x^k \\ &\to b^T y^* + \left(c - A^T y^*\right)^T x' \overset{\text{(continuity)}}{=} f(y^*) = b^T y^* + \min_{x\in X} F(x, y^*). \end{aligned}$$

Subtracting $b^T y$ from both sides of the final equality yields

$$\left(c - A^T y\right)^T x' = \min_{x\in X} F(x, y), \quad \text{which is just} \quad x' \in X^*(y^*).$$

This shows that $X^*(\cdot)$ has closed graph, as claimed.

(ii): Now we prove what we claimed about the super-differential.

($\supseteq$): We will prove $\supseteq$. Take $z \in \{b - Ax \mid x \in X^*(y)\}$ then there is a $x \in X^*(y)$ that satisfies $z = b - Ax$, and we have for any other y' that

$$\begin{aligned} f\left(y'\right) - f(y) &\leq \overbrace{b^T y' + \left(c - A^T y'\right)^T x}^{f(y')\leq} - \overbrace{\left[b^T y + \left(c - A^T y\right)^T x\right]}^{f(y)} \\ &= (b - Ax)^T \left(y' - y\right) = z^T \left(y' - y\right). \end{aligned}$$

That is $z \in \partial^+ f(y)$. This shows the first inclusion $\supseteq$, and therefore

$$S(\partial^+ f(y), \cdot) \geq S(\overline{\text{co}}\{b - Ax \mid x \in X^*(y)\}, \cdot).$$

Now take $x_t \in X^*(y + td)$. By definition, this x_t satisfies
$f(y + td) = b^T(y + td) + \left(c - A^T(y + td)\right) x_t$. Then

$$\begin{aligned} \frac{1}{t}\left(f(y + td) - f(y)\right) &\geq \frac{1}{t}\left[\overbrace{b^T(y + td) + \left(c - A^T(y + td)\right)^T x_t}^{=f(y+td)} - \overbrace{\left[b^T y + \left(c - A^T y\right)^T x_t\right]}^{\geq f(y)}\right] \\ &= \frac{1}{t}(b^T(td) - (A^T(td))^T x_t) \\ &= (b - Ax_t)^T d. \end{aligned}$$

Letting $t \downarrow 0$, we obtain

$$f'(x, d) \geq (b - Ax)^T d \qquad \text{where } x_t \to x \in X^*(y),$$

where we are passing to a convergent x_t because we always can do so, because of what we showed in (i). We also know from (ii)($\supseteq$) that since $x \in X^*(y)$, it satisfies $(b - Ax)^T d \in \partial^+ f(y)$ and so $f'(x,d) \leq (b - Ax)$. Altogether, have shown the equality

$$f'(x,d) = (b - Ax)^T d$$

for this specific $x \in X^*(y)$. Hence

$$S\left(\overline{\text{co}}\left\{b - Ax \mid x \in X^*(y)\right\}, d\right) = f'(x,d) = S(\partial^+ f(y), d).$$

Thus we have shown that for an arbitrary d, the two support functions are equal. Since they are equal everywhere, we have equality of the two closed convex sets: $\partial^+ f(y) = \text{co}\left\{b - Ax \mid x \in X^*(y)\right\}$. □

The above result "almost" follows immediately Proposition 4.5.4 in the sense that $-f(y) = \max_{x \in X}\{g_x(y)\}$ with $g_x(y) = -b^T y - (c - A^T y)^T x$ and $\nabla_y g_x(y) = -b + Ax$. Here $X^*(y)$ is simply $I(y)$: the index set that tells us which affine functions satisfy $g_x(y) = -f(y)$. The only reason that Proposition 6.4.1 does not follow as an immediate consequence of Proposition 4.5.4 is because we no longer assume a finite number of g_i, and so rely on compactness instead of the pigeonhole principle.

Building a projected supergradient method for the Lagrangian relaxation problem

Note that $\partial f^+(y)$ is bounded (and f is Lipschitz) when X is compact. If the subgradient ("supergradient") method is applied to maximize the Lagrangian dual problem f in (LD) with an appropriate nonsummable step lenths rule, our results (such as Theorems 6.3.3 and 6.3.4) guarantee that $\left\{y^k\right\}$ will converge to some $y^* \in Y^*$, but the corresponding primal sequence $\left\{x^k\right\}$ is not assured to converge, or even accumulate, in general! Our goal is to make $\left\{x^k\right\}$ accumulate to $x^* \in X^*$ (the primal solutions). The first approach uses weighted averages of the form

$$\bar{x}^k = \sum_{i=1}^{k} \lambda_i^k x^i \quad \text{where} \quad \lambda_i^k = \frac{t^i}{\sum_{j=1}^k t^j}, \quad \text{for } 1 \leq i \leq k. \tag{6.20}$$

Note that for any k and i we have $\sum_{i=1}^k \lambda_i^k = 1$ with $\lambda_i^k \geq 0$. Thus $\bar{x}^k \in X$ and when X is compact there exists a convergent subsequence. In fact it converges to the optimal solution.

In what follows, we ensure non-negativity of the multiplier sequence y^k by letting our set C be the nonnegative orthant. The projection onto $C = \mathbb{R}^m_+$ is then

$$[y]^+ := P_{\mathbb{R}^m_+}(y) = (\max\{0, y_1\}, \ldots, \max\{0, y_m\}).$$

Theorem 6.4.2 *Suppose the supergradient method* $y^{k+1} = P_{\mathbf{R}^n_+}\left(y^k + t^k d^k\right) = \left[y^k + t^k\left(b - Ax^k\right)\right]^+$ *is applied to (LD) using* $t^k > 0$, $t^k \to 0$ *and* $\sum_{k=1}^{\infty} t^k = \infty$ *and* $\bar{x}^k$ *is as defined in (6.20). Assume that* $y^k \to y^* \in Y^*$. *Then any accumulation point of* $\{\bar{x}_k\}$ *is in* X^* *(at least one exists since* X *is compact).*

Proof. Let x^* be an accumulation point of $\left\{\bar{x}^k\right\}$. To prove $x^* \in X^*$ we need to show that $Ax^* \geq b$ and $c^T x^* = v^*$. By construction

$$y^{k+1} = \left[y^k + t^k\left(b - Ax^k\right)\right]^+ \geq y^k + t^k\left(b - Ax^k\right) \tag{6.21}$$

where $[y]^+$ is the projection onto $\mathbf{R}^n_+$ (i.e. $[y]^+_i = \max\{0, x_i\}$). Notice that

$$\begin{aligned}\sum_{i=1}^{k} t^i A x^i &= A\left(\sum_{i=1}^{k} t^i x^i\right)\\ &= \left(\sum_{l=1}^{k} t^l\right) A\left(\frac{\sum_{i=1}^{k} t^i x^i}{\sum_{l=1}^{k} t^l}\right)\\ &= \left(\sum_{l=1}^{k} t^l\right) A\left(\sum_{i=1}^{k} \lambda_i^k x^i\right)\\ &= \left(\sum_{l=1}^{k} t^l\right)\left(A\bar{x}^k\right). \end{aligned} \tag{6.22}$$

Using this identity, we have that

$$\begin{aligned} y^{k+1} - y^1 &\geq \sum_{i=1}^{k} t^i\left(b - Ax^i\right) \quad \text{(telescoping (6.21))} \qquad (6.23)\\ &= \left(\sum_{l=1}^{k} t^l\right)\left(b - A\bar{x}^k\right)\\ \text{so} \quad \frac{y^{k+1} - y^1}{\sum_{l=1}^{k} t^k} &\geq b - A\bar{x}^k. \qquad (6.24)\end{aligned}$$

In the left hand side of this inequality the numerator converges to $y^* - y^1$ but $\sum_{i=1}^{k} t^k$ goes to infinity. So, letting $k \to \infty$ we get $0 \geq b - Ax^*$. Since $\mathbb{R}^m_+$ is closed and $y^k \in \mathbb{R}^m_+$ for all k, we obviously must have we have $y^* \geq 0$. Now consider the case when $y^*_j > 0$. Then for all sufficiently large k (say for all $k > K$), we have that $y^k_j > 0$, which forces $y^k_j = [y^k_j]^+$ and therefore $y^{k+1}_j - y^k_j = t^k\left(b - Ax^k\right)_j$. Then we argue that $\left(b - Ax^*\right)_j = 0$ as follows:

$$\begin{aligned} y^k_j - y^K_j &= \sum_{l=K}^{k} t^l\left(b - Ax^l\right)_j\\ &= \left(\sum_{l=1}^{k} t^l\left(b - Ax^l\right) - \sum_{l=1}^{K-1} t^l\left(b - Ax^l\right)\right)_j\\ &= \left(\sum_{l=1}^{k} t^l\left(b - A\bar{x}^k\right) - \sum_{l=1}^{K-1} t^l\left(b - A\bar{x}^{K-1}\right)\right)_j \quad \text{using (6.22)}\\ \text{and so} \quad y^k_j - y^K_j + \left(\sum_{l=1}^{K-1} t^l\left(b - A\bar{x}^{K-1}\right)\right)_j &= \left(\sum_{l=1}^{k} t^l\left(b - A\bar{x}^k\right)\right)_j\\ \text{and so} \quad \frac{1}{\left(\sum_{l=1}^{k} t^l\right)}\left(y^k_j - y^K_j + \left(\sum_{l=1}^{K-1} t^l\left(b - A\bar{x}^{K-1}\right)\right)_j\right) &= \left(b - A\bar{x}^k\right)_j. \end{aligned}$$

Taking the limit as $k \to \infty$, $y^k_j \to y^*_j$ and $\bar{x}^k \to x^*$ while the denominator of the left side goes to infinity. Thus we have that $(b - Ax^*)_j = 0$. Having shown that this is the case for any j such that $y^*_j > 0$, we have that $(b - Ax^*)^T y^* = 0$ (called *complementarity*).

Finally $x^k \in X^*\left(y^k\right)$ and $y^k \to y^*$ together imply that $x^* \in X^*\left(y^*\right)$ (because $y \mapsto X^*(y)$ has closed graph by Proposition 6.4.1). Thus

$$v^* := f\left(y^*\right) = b^T y^* + \left(c - A^T y^*\right)^T x^* = c^T x^* + \overbrace{\left(b - Ax^*\right)^T y^*}^{=0} = c^T x^*. \tag{6.25}$$

That is: the maximum of the Lagrangian function (achieved by y^*) equals the primal objective value ($c^T x^*$) of a feasible solution (x^*). Since any feasible solution must return an objective value that upper bounds the maximum of the Lagrangian function, and x^* achieves that lowest bound, it is optimal for the original problem. □

Note that we only need $y^k \to y^*$ in order to prove optimality. Feasibility only requires $\left\{y^k\right\}$ bounded. Let's put all the pieces together for one example in the following corollary.

Corollary 6.4.3 *Suppose that the supergradient method is applied to LD using the square summable but not summable rule i.e. for $t^k \geq 0$*

$$y^{k+1} = \left[y^k + t^k\left(b - Ax^k\right)\right]^+ \quad \textit{where} \quad \sum_{k=1}^{\infty} t^k = +\infty \textit{ and } \sum_{k=1}^{\infty}\left(t^k\right)^2 < +\infty. \tag{6.26}$$

Let $\left\{\bar{x}^k\right\}$ be constructed by (6.20). Then $y^k \to y^* \in Y^*$ and any accumulation point of $\left\{\bar{x}^k\right\}$ is in X^*.

Proof. We proved in Theorem 6.3.4 that when t is square summable but not summable, the sequence $\left\{y^k\right\}$ generated by the projected subgradient (in this case supergradient) method converges. Since it converges, we need only apply Theorem 6.4.2 to obtain our results about the sequence $\bar{x}^n$. □

Now that we have considered the case of weighted averages (6.20), Let's consider the sequence of simple averages:

$$\bar{x}^k := \frac{1}{k}\sum_{i=1}^{k} x^i. \tag{6.27}$$

Theorem 6.4.4 *Suppose the supergradient optimisation (6.26) is applied with $t^k = \frac{1}{k}$ with $\left\{\bar{x}^k\right\}$ defined as in (6.27). Then $y^k \to y^* \in Y^*$ and any accumulation point of $\left\{\bar{x}^k\right\}$ is in X^*.*

Proof. From Theorem 6.3.4 we have $y^k \to y^* \in Y^*$ and if $\bar{x}^{k_m} \to \bar{x}$ then clearly $B\bar{x} = \lim_m B\bar{x}^{k_m} \geq d$. To see why, just notice that since X is convex and $\bar{x}^k$ is a convex combination of points in X, we must also have $\bar{x}^k \in X$. The limiting characterization follows because X is closed. Thus $\bar{x} \in X$. What remains is to show that $\bar{x}$ is feasible with respect to the matrix A. For the particular choice of step length $t^k = \frac{1}{k}$ we use the following identity:

$$\begin{aligned}\sum_{i=1}^{k} i\left(y^{i+1} - y^i\right) &= y^2 - y^1 + 2\left(y^3 - y^2\right) + 3\left(y^4 - y^3\right) + 4\left(y^5 - y^4\right) + \cdots + k\left(y^{k+1} - y^k\right) \\ &= -y^1 + y^2 - 2y^2 + 2y^3 - 3y^3 + 3y^4 - 4y^4 + 4y^5 + \cdots - ky^k + ky^{k+1} \\ &= -\left(\sum_{i=1}^{k} y^i\right) + ky^{k+1} = ky^{k+1} - \sum_{i=1}^{k} y^i,\end{aligned}$$

and so we have

$$\begin{aligned}
y^{k+1} - \frac{1}{k}\sum_{i=1}^{k} y^i &= \frac{1}{k}\left(ky^{k+1} - \sum_{i=1}^{k} y^i\right) \\
&= \frac{1}{k}\left(\sum_{i=1}^{k} i\left(y^{i+1} - y^i\right)\right) \\
&= \frac{1}{k}\left(\sum_{i=1}^{k} i\left(\left[y^i + t^i\left(b - Ax^i\right)\right]^+ - y^i\right)\right) \quad \text{with } t^i = \frac{1}{i} \\
&\geq \frac{1}{k}\left(\sum_{i=1}^{k} i\left(y^i + t^i\left(b - Ax^i\right) - y^i\right)\right) \\
&= \frac{1}{k}\left(\sum_{i=1}^{k} i\frac{1}{i}\left(b - Ax^i\right)\right) = b - A\left(\frac{1}{k}\sum_{i=1}^{k} x^i\right) = b - A\bar{x}^k. \qquad (6.28)
\end{aligned}$$

When $y^k \to y^*$ then there exists a N such that for $i \geq N$ we have $\left\|y^i - y^*\right\| \leq \varepsilon/2$ and as $\left\{y^k\right\}$ is bounded there is a constant B such that $\left\|y^i\right\| \leq B/2$ for all i so

$$\begin{aligned}
\left\|y^* - \frac{1}{k}\sum_{i=1}^{k} y^i\right\| &= \frac{1}{k}\left\|\sum_{i=1}^{k}\left(y^* - y^i\right)\right\| \\
&\leq \frac{1}{k}\sum_{i=1}^{N}\underbrace{\left\|y^* - y^i\right\|}_{\leq B} + \frac{1}{k}\sum_{i=N}^{k}\underbrace{\left\|y^* - y^i\right\|}_{\leq \varepsilon/2} \\
&\leq \frac{N}{k}B + \frac{1}{k}\sum_{i=N}^{k}\varepsilon/2 = \frac{N}{k}B + \frac{\varepsilon}{2}\left(\frac{k-N}{k}\right) \leq \varepsilon
\end{aligned}$$

for k sufficiently large so that $\frac{N}{k}B \leq \frac{\varepsilon}{2}$. Thus $\frac{1}{k}\sum_{i=1}^{k} y^i \to y$ and (6.28) yields $0 \geq \lim_m \left(b - A\bar{x}^{k_m}\right) = b - A\bar{x}$ for any accumulation point $\bar{x}^{k_m} \to \bar{x}$. Thus $\bar{x}$ is feasible. We can conclude this proof in a similar way to how we concluded Theorem 6.4.2. Namely, we show that $(b - A\bar{x})^T y = 0$, and then we have that $\bar{x}$ is optimal by establishing that (6.25) again holds. □

The update (6.27) is better than (6.20) as it puts less weight on earlier iterates. One can periodically restart the averaging process to further diminish the effect of earlier iterates. The so called "volume method" is essentially the subgradient method with the Polyak step length and the primal updates defined by

$$\bar{x}^1 = x^1, \quad \bar{x}^k = \rho x^k + (1-\rho)\,\bar{x}^{k-1}, \text{ for } k > 1.$$

This makes an weighted average with later iterates getting more weight. We will look at this method again later when discussing signal processing and sparse optimisation.

6.4.1 Feasibility Problems

We consider the problem of finding a common point in a sequence of convex sets i.e.

$$\text{Find } x \in C_1 \cap C_2 \cap \cdots \cap C_m$$

One approach to this problem is to minimize the function

$$f(x) = \max\left\{d_{C_1}(x), d_{C_2}(x), \ldots, d_{C_m}(x)\right\}$$

which is convex when each C_i is convex. If $f(x) = 0$ we are finished otherwise we need a subgradient of a max function. Let $I(x) := \{i \in \{1, \ldots, m\} \mid d(x, C_i) = f(x)\}$ then

$$\partial f(x) = \operatorname{co}\{\partial d_{C_i}(x) \mid i \in I(x)\}.$$

Now since $c_t = \operatorname{proj}_{C_i}(x + td) \to c = \operatorname{proj}_{C_i} x \in C_i$ as $t \downarrow 0$ have by the mean value theorem (of the smooth function $\|\cdot - c_t\|$)

$$\begin{aligned}
d'_{C_i}(x, d) &= \lim_{t\downarrow 0} \frac{1}{t}\left(d_{C_i}(x + td) - d_{C_i}(x)\right) \\
&\geq \lim_{t\downarrow 0} \frac{1}{t}(\overbrace{\|x + td - c_t\|}^{=d_{C_i}(x+td)} - \overbrace{\|c_t - x\|}^{\geq d_{C_i}(x)}) \\
\text{(using mean value theorem)} \quad &= \lim_{t\downarrow 0} \frac{1}{t}\langle \nabla \|\cdot - c_t\|(x_{t'}), td\rangle \qquad \text{for } x_{t'} \in [x, x + td] \\
&\lim_{t\downarrow 0} \frac{1}{t}\langle \nabla \|\cdot\|(x_{t'} - c_t), td\rangle x, x + td] \\
&= \langle \nabla \|\cdot\|(x - c), d\rangle.
\end{aligned}$$

Moreover $\nabla \|\cdot\|(x - c) = \frac{x-c}{\|x-c\|}$, for the Euclidean norm. Thus the fact that

$$d'_{C_i}(x, d) \geq \langle \nabla \|\cdot\|(x - c), d\rangle \quad \text{for all } d \quad \text{and } \nabla \|\cdot\|(x - c) = \frac{x - c}{\|x - c\|}$$

we deduce that $\frac{x - \operatorname{proj}_{C_i} x}{\left\|x - \operatorname{proj}_{C_i} x\right\|} \in \partial d_{C_i}(x)$ and we can take our candidate subgradient as $g^k = \frac{x^k - \operatorname{proj}_{C_i} x^k}{\left\|x^k - \operatorname{proj}_{C_i} x^k\right\|}$ if $i \in I\left(x^k\right)$. Now $\left\|g^k\right\| = 1$ so we will take $\alpha = 1$ and $f^* = 0$ (where f^* again is our lower bound on the objective function, and is tight) so the Polyak step length is

$$t^k = \frac{\alpha\left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|^2} = f\left(x^k\right) = d_{C_i}(x^k) = \|x^k - \operatorname{proj}_{C_j} x^k\|.$$

Let $j \in I\left(x^k\right)$ then we have

$$\begin{aligned}
x^{k+1} &= x^k - t^k g^k \\
&= x^k - \overbrace{\left\|\operatorname{proj}_{C_j} x^k - x^k\right\|}^{t^k} \frac{x^k - \operatorname{proj}_{C_j} x^k}{\left\|x^k - \operatorname{proj}_{C_j} x^k\right\|} \\
&= \operatorname{proj}_{C_j} x^k.
\end{aligned}$$

Thus at each step we take the most violated set and we project onto it. We iterate this way. This is known as the project-project algorithm. It is most interesting when we can decompose a complicated constraint into simple components that can be individually projected onto in a simple way. These include

- Affine sets
- non-negative orthant
- half slabs

- box constraints
- unit simplex
- Euclidean ball
- ellipsoid
- and other structures in the space of symmetric matrices.

We can improve on convergence rates for other forms of projection algorithms for feasibility problems and will discuss the so called Douglas-Rachford algorithm.

Problem Set 14

Problem 6.4.5 *Let's have fun with projection algorithms.*

1. *The reflection operator is defined in terms of the projection P_X as:*

$$\begin{aligned} R_X x &:= (2P_X - I)\, x \\ &\equiv x + 2(P_X - x). \end{aligned}$$

There are numerous algorithms for the feasibility problem $x \in A \cap B$ is based on the Picard iteration

$$x_{n+1} \in T_{A,B} x_n$$

for various choices of the operator $T_{A,B}$. The Douglas-Rachford uses the following operator based in "Reflect-Reflect and average"

$$T_{A,B} x = \frac{R_A R_B x + x}{2}.$$

Alternating projections is based on the same Picard scheme but uses the operator:

$$T_{A,B} x = P_A P_B x.$$

Notice that in the case where there are only two sets, alternating-projections vacuously coincides with the scheme we described above where we project onto the set C such that $d_C(x^k) = \max_{C \in \{A,B\}} \{d_C(x^k)\}$.

Write small computer programs to execute these two algorithms. In particular

   (a) *First state with reasoning what the feasible sets are.*
   (b) *Describe how one executes the projection and reflection steps in each example.*
   (c) *Run the programs and plot the trajectories of the algorithms trying a few starting points (some close to the feasible point and some further away). State whether they converge or not and whether the proximity effects the result.*
      - *Note: If you are not sure what to use, Geogebra and Cinderella have straightforward interfaces and are easy to learn.*
   (d) *Estimate the rate of convergence, categorizing it as R-linear if $x_n \to x^* \in A \cap B$ with*

$$\lim_{k \to \infty} \frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|} = r < 1$$

R-superlinear if

$$\lim_{k \to \infty} \frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|} = 0,$$

and sublinear if neither happens.

2. *Problems:*

   (a) *Two lines in* $\mathbb{R}^2$

   $$A = \{(x_1, x_2) \mid x = 0\},$$
   $$B = \{(x_1, x_2) \mid x_1 = x_2\}.$$

   (b) *A line and a ball in* $\mathbb{R}^2$

   $$A = \{(x_1, x_2) \mid x_2 = 0\},$$
   $$B = \left\{(x_1, x_2) \mid x_1^2 + (x_2 - 1)^2 \leq 1\right\}.$$

   (c) *A cross and a subspace in* $\mathbb{R}^2$

   $$A = \mathbb{R} \times \{0\} \cup \{0\} \times \mathbb{R}$$
   $$B = \{(x_1, x_2) \mid x_1 = x_2\}.$$

   This problem relates to sparse-signal recovery. The set A is not convex.

   (d) *A circle and a line*

   $$A = \left\{(x_1, x_2) \mid x_2 = \sqrt{2}/2\right\}$$
   $$B = \left\{(x_1, x_2) \mid x_1^2 + x_2^2 = 1\right\}.$$

   Again the set B is non-convex and the problem relates to phase retrieval problem.

6.5 Constrained Optimisation with Structure

Here we consider the following optimisation problem

$$\begin{aligned} & \min f_0(x) \\ \text{Subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

where f_i are all finite convex and

$$C = \{x \mid f_i(x) \leq 0, \quad i = 1, \ldots, m\}$$

is compact. The algorithm takes the form

$$x^{k+1} = x^k - t^k g^k$$

where $t^k > 0$ is a step size and g^k is a subgradient taken according to the rule

$$g^k \in \begin{cases} \partial f_0(x) & f_i\left(x^k\right) \leq 0 \ \ \forall i = 1, \ldots, m \\ \partial f_j\left(x^k\right) & f_j\left(x^k\right) > 0. \end{cases}$$

In other words we choose objective descent if we are feasible; otherwise we try and return to feasibility. This is a predictor corrector approach. We improve when feasible and correct when infeasible. Unlike the projected subgradient method, iterates can often be infeasible and often are. We keep track of the best solution found so far via

$$f_{best}^k = \min\left\{f_0\left(x^i\right) \mid x^i \text{ is feasible}, \ i \in \{1, \ldots, k\}\right\}.$$

Let us assume the Slater constraint qualification holds; i.e. there exists $\tilde{x}$ such that $f_i(\tilde{x}) < 0$ for all i. We assume the problem has an optimal point x^*, and $\|x^1 - x^*\| \leq R$. We want to establish convergence of objective values under the usual diminishing non-convergent sum rule. Suppose for a contradiction that $f^k_{best} \geq f^* + \varepsilon$ for some $\varepsilon > 0$ for all k. If we can deduce a contradiction, then we will be able to show that there exists a feasible convergent $x^{k_n} \to x^* \in X^*$ such that $f^{k_n}_{best} \to f^*$. First, suppose for a contradiction that $f\left(x^k\right) \geq f^* + \varepsilon$ for all x^k that are feasible. We show that the latter leads to a contradiction.

We will first find a point $\bar{x}$ and a positive number μ such that

$$f_0(\bar{x}) \leq f^* + \frac{\varepsilon}{2}, \quad f_1(\bar{x}) \leq -\mu, \ \ldots, \ f_m(\bar{x}) \leq -\mu. \tag{6.29}$$

Such a point is an $\frac{\varepsilon}{2}$-optimal solution that satisfies the constraints to a margin of $-\mu$. We can take $\bar{x} = (1-\theta)x^* + \theta\tilde{x}$, where $\theta \in (0,1)$ and $x^* \in X^*$. This gives us

$$f_0(\bar{x}) \leq (1-\theta) f_0(x^*) + \theta f_0(\tilde{x}) = f^* + \theta\left(f_0(\tilde{x}) - f^*\right).$$

So if we choose $\theta = \min\left\{1, (\varepsilon/2)/\left(f_0(\tilde{x}) - f^*\right)\right\}$ we have $f_0(\bar{x}) \leq f^* + \frac{\varepsilon}{2}$ (which is needed in the bound below) and

$$f_i(\bar{x}) \leq (1-\theta)\overbrace{f_i(x^*)}^{\leq 0} + \theta f_i(\tilde{x}) \leq \theta f_i(\tilde{x}) \leq -\mu \quad \text{choosing} \quad \mu := -\theta \min_i f_i(\tilde{x}) > 0.$$

Altogether, with these choices, $\bar{x}$ satisfies (6.29). Consider the index $i \in \{1, \ldots, k\}$ for which x^i is feasible. Then we have $g^i \in \partial f_0\left(x^i\right)$, and $f_0\left(x^i\right) \geq f^* + \varepsilon$. Since $\bar{x}$ is $\frac{\varepsilon}{2}$-optimal and we have assumed (for a contradiction) that any feasible x^i satisfies $f_0(x^i) \geq f^* + \varepsilon$, we have

$$f_0\left(x^i\right) - f_0(\bar{x}) \geq \frac{\varepsilon}{2}. \tag{6.30}$$

Thus

$$\begin{aligned}
\left\|x^{i+1} - \bar{x}\right\|^2 &= \left\|x^i - t^i g^i - \bar{x}\right\|^2 \\
&= \left\|x^i - \bar{x}\right\|^2 - 2t^i\langle g^i, x^i - \bar{x}\rangle + \left(t^i\right)^2 \left\|g^i\right\|^2 \\
\text{(subgradient inequality)} \quad &\leq \left\|x^i - \bar{x}\right\|^2 - 2t^i\left(f_0\left(x^i\right) - f_0(\bar{x})\right) + \left(t^i\right)^2 \left\|g^i\right\|^2 \\
\text{(Using (6.30))} \quad &\leq \left\|x^i - \bar{x}\right\|^2 - 2t^i\frac{\varepsilon}{2} + \left(t^i\right)^2 \left\|g^i\right\|^2 \\
&= \left\|x^i - \bar{x}\right\|^2 - t^i\varepsilon + \left(t^i\right)^2 \left\|g^i\right\|^2
\end{aligned}$$

where the subgradient inequality gave us

$$f_0(\bar{x}) - f_0\left(x^i\right) \geq \langle g^i, \bar{x} - x^i\rangle.$$

Now suppose that i is such that x^i is infeasible, and $g^i \in \partial f_p\left(x^i\right)$, where $f_p\left(x^i\right) > 0$ (is infeasible). Since $f_p(\bar{x}) \leq -\mu$ we have

$$f_p\left(x^i\right) - f_p(\bar{x}) \geq \mu. \tag{6.31}$$

Therefore

$$\begin{aligned}
\left\|x^{i+1} - \bar{x}\right\|^2 &= \left\|x^i - \bar{x}\right\|^2 - 2t^i\langle x^i - \bar{x}, g^i\rangle + \left(t^i\right)^2 \left\|g^i\right\|^2 \\
\text{(subgradient inequality)} \quad &\leq \left\|x^i - \bar{x}\right\|^2 - 2t^i\left(f_p\left(x^i\right) - f_p(\bar{x})\right) + \left(t^i\right)^2 \left\|g^i\right\|^2
\end{aligned}$$

$$\text{Using (6.31)} \quad \leq \left\|x^i - \bar{x}\right\|^2 - 2t^i\mu + \left(t^i\right)^2 \left\|g^i\right\|^2.$$

In both cases we have

$$\left\|x^{i+1} - \bar{x}\right\|^2 \leq \left\|x^i - \bar{x}\right\|^2 - t^i\delta + \left(t^i\right)^2 \left\|g^i\right\|^2$$

where $\delta = \min\{\varepsilon, 2\mu\} > 0$. Applying this recursively we have get

$$\left\|x^{k+1} - \bar{x}\right\|^2 \leq \left\|x^1 - \bar{x}\right\|^2 - \sum_{i=1}^{k} t^i\delta + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2$$

and so it follows that

$$\begin{aligned} 0 \leq \left\|x^{k+1} - x\right\|^2 &\leq \left\|x^1 - x\right\|^2 - \sum_{i=1}^{k} t^i\delta + \sum_{i=1}^{k} \left(t^i\right)^2 \left\|g^i\right\|^2 \\ &\leq R^2 - \delta\sum_{i=1}^{k} t^i + L^2\sum_{i=1}^{k} \left(t^i\right)^2, \end{aligned}$$

$$\text{so we have} \quad \delta\sum_{i=1}^{k} t^i \leq R^2 + L^2\sum_{i=1}^{k} \left(t^i\right)^2 \quad \text{or} \quad \delta \leq \frac{R^2 + L^2\sum_{i=1}^{k} \left(t^i\right)^2}{\sum_{i=1}^{k} t^i}.$$

We cannot have this true for k large as

$$\frac{R^2 + L^2\sum_{i=1}^{k} \left(t^i\right)^2}{\sum_{i=1}^{k} t^i} \to 0 \quad \text{as } k \to \infty.$$

Thus we have our contradiction, and so f_{best}^k decreases monotonically to f^*. Note that if $f_{best}^k = f^*$ for some k, then the corresponding feasible x^k is both feasible and optimal, meaning we have finite convergence and are done. Otherwise, the non-finite convergence of f_{best}^k to f^* guarantees the existence of a non-finite feasible subsequence of x^k with $f(x^k) = f_{best}^k$. Using our assumption that the feasible set is compact, this sequence will possess a convergent subsequence x^{k_n}. Using continuity of f_0 and closedness of the contraint set C, we have that the limit of x^{k_n} is in X^*.

There are many possible variations on this basic rule. For example, there is the "over projection" step (via $\varepsilon > 0$)

$$t^k = \begin{cases} \left(f_0\left(x^k\right) - f^*\right) / \left\|g^k\right\|^2 & \text{if } x^k \text{ is feasible} \\ \left(f_i\left(x^k\right) + \varepsilon\right) / \left\|g^k\right\|^2 & \text{if } x^k \text{ is in feasible} \end{cases}$$

where ε is a small positive margin, and i is the index of the most violated inequality in the case when x^k is infeasible.

6.6 Speeding up the Subgradient Method

One method is to take a search direction that is a convex combination of the current direction and last search direction i.e.

$$x^{k+1} = x^k - t^k g^k + \beta^k\left(x^k - x^{k-1}\right)$$

where β^k is a positive constant ("momentum term"). Polyak refers to this class of algorithm as the "heavy ball method;" others call it the "momentum method." Conjugate gradient methods have a similar form. We will briefly describe two approaches:

1. Take
$$x^{k+1} = P_C\left(x^k - t^k s^k\right), \quad t^k = \frac{f\left(x^k\right) - f^*}{\left\|s^k\right\|^2}$$
where s^k is the filtered subgradient
$$s^k = (1-\beta)\, g^k + \beta s^{k-1}.$$
Here $\beta \in (0,1)$ is a constant.

2. A more sophisticated approach is due to Camerini, Fratta and Maffioli [7]. We take
$$x^{k+1} = P_C\left(x^k - t^k s^k\right), \quad t^k = \frac{f\left(x^k\right) - f^*}{\left\|s^k\right\|^2}$$
plus
$$\begin{aligned} & s^k = g^k + \beta^k s^{k-1} \\ \text{where} \qquad & \beta^k = \max\left\{0, -\gamma^k \left(s^{k-1}\right)^T g^k / \left\|s^{k-1}\right\|^2\right\} \end{aligned}$$
where $\gamma^k \in [0,2]$ (they recommend $\gamma^k = 1.5$). They show that
$$\frac{\left(x^k - x^*\right)^T \left(-s^k\right)}{\left\|s^k\right\|^2} \leq \frac{\left(x^k - x^*\right)^T \left(-g^k\right)}{\left\|s^k\right\|^2}$$
so that this search direction makes a smaller angle towards the optimal set than the negative subgradient.

6.7 Smooth Convex Optimisation and Best Complexity

If we assume that $f : \mathbf{R}^n \to \mathbf{R}$ is convex, differentiable with a Lipschitz gradient then we can design descent methods with much faster convergence estimates. These are based on similar ideas as to those suggested as improvements in the last section. These draw inspiration from momentum methods. Again we consider the problem:
$$\min_x f(x) \quad \text{subj to } x \in C \text{ (a convex closed set).}$$
We assume X^* is the set of minimizers of f over C and $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. The method has the form:
$$\begin{aligned} y_k &= x_k + \beta_k\left(x_k - x_{k-1}\right), \qquad \text{(extrapolation)} \\ x_{k+1} &= P_C\left(y_k - \alpha \nabla f\left(y_k\right)\right) \quad \text{(gradient projection step),} \end{aligned} \tag{6.32}$$
where P_C is the projection onto C, $x_{-1} \equiv x_0$ and $\beta_k \in (0,1)$. To get good convergence rates, we use the following
$$\beta_k = \frac{\theta_k\left(1 - \theta_{k-1}\right)}{\theta_{k-1}}, \quad k = 0, 1, 2, \ldots \tag{6.33}$$

where $\{\theta_k\}$ satisfies $\theta_0 = \theta_{-1} \in (0, 1]$ and

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{6.34}$$

One can use the following specific choice:

$$\beta_k = \begin{cases} 0 & \text{if } k = 0 \\ \frac{k-1}{k+2} & \text{if } k = 1, 2, \ldots \end{cases} \quad \text{with} \quad \theta_k = \begin{cases} 1 & \text{if } k = -1 \\ \frac{2}{k+2} & \text{if } k = 0, 1, 2, \ldots \end{cases}. \tag{6.35}$$

Initially we will assume that $\alpha = \frac{1}{L}$, but, as L can only be estimated in generaly, we will extend an adaptive step size rule. When we do not know L, we start with one choice

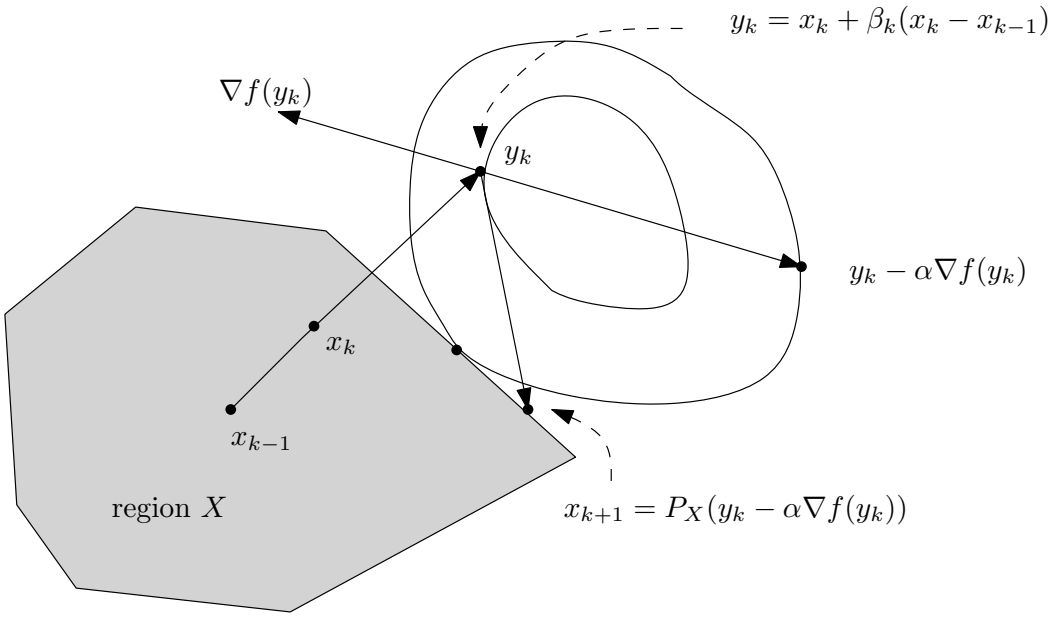


Figure 6.3: Illustration of the two step method with extrapolation

$\alpha > 0$ and continue to use this choice for α as long as

$$f(x_{k+1}) \leq l(x_{k+1}; y_k) + \frac{1}{2\alpha} \|x_{k+1} - y_k\|^2$$
$$\text{where } l(x; y_k) := f(y_k) + \nabla f(y_k)^T (x - y_k).$$

As soon as this condition is violated, we reduce α by some fraction, and we repeat this as many times as is necessary to satisfy the condition. We can always terminate such a process finitely, because this condition will be satisfied as soon as $\alpha \leq \frac{1}{L}$, which occurs after a finite number of iterations. If we did not use the interpolation step, then usually straight gradient descent would use a line search like the Amijo-Wolfe inexact line search (that is detailed in the appendix for comparision). This requires a lot more function evaluations.

Lemma 6.7.1 (Descent Lemma) *Let $f : \mathbf{R}^n \to \mathbf{R}$ is convex, differentiable with a Lipschitz gradient and associated Lipchitz constant $L > 0$ and X a closed convex set. Then for all $x, y \in X$ we have*

$$f(y) \leq l(x; y) + \frac{L}{2} \|y - x\|^2. \tag{6.36}$$

Proof. Let $g(t) := f(x + t(y - x))$ for $t \in [0, 1]$. The fundamental theorem of calculus gives

$$\begin{aligned}
f(y) - f(x) = g(1) - g(0) &= \int_0^1 g'(t)\, dt \\
\text{(chain rule)} \quad &= \int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt \\
&= \int_0^1 \nabla f(x + t(y - x))^T (y - x) + (\overbrace{f(x)^T(y - x) - f(x)^T(y - x)}^{=0})\, dt \\
&\le \int_0^1 \nabla f(x)^T (y - x)\, dt + \left| \int_0^1 [\nabla f(x + t(y - x)) - \nabla f(x)]^T (y - x)\, dt \right| \\
\text{(C-S)} \quad &\le \nabla f(x)^T (y - x) \left[t|_{t=0}^{t=1} \right] + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \, \|y - x\|\, dt \\
\text{(Lipschitz)} \quad &\le \nabla f(x)^T (y - x) + \|y - x\| \int_0^1 Lt \, \|y - x\|\, dt \\
&= \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2 .
\end{aligned}$$

so

$$f(y) \le \left[f(x) + \nabla f(x)^T (y - x) \right] + \frac{L}{2} \|y - x\|^2 .$$

□

Proposition 6.7.2 (Descent Properties) *Let $f : \mathbf{R}^n \to \mathbf{R}$ is convex, differentiable with a Lipschitz gradient and associated Lipschitz constant $L > 0$ and C a closed convex set. Then for all $x_k \in C$ and $\alpha > 0$ we have:*

(i) If $x_k(\alpha) := P_C(x_k - \alpha \nabla f(x_k)) \neq x_k$ then $x_k(\alpha) - x_k$ is a feasible descent direction at x_k in that

$$\nabla f(x_k)^T (x_k(\alpha) - x_k) \le -\frac{1}{\alpha} \|x_k(\alpha) - x_k\|^2, \quad \text{for all } \alpha > 0. \tag{6.37}$$

(ii) If $x_k(\alpha) = x_k$ for some $\alpha > 0$ then x_k satisfies the optimality condition:

$$\nabla f(x_k)^T (x - x_k) \ge 0, \quad \text{for all } x \in C. \tag{6.38}$$

Proof. From (6.7) we have

$$\langle z - P_C(z), x - P_C(z) \rangle \le 0, \quad \text{for all } x \in C$$

so (with $z = x_k - \alpha \nabla f(x_k)$, $x_k(\alpha) = P_C(x_k - \alpha \nabla f(x_k))$),

$$(\overbrace{x_k - \alpha \nabla f(x_k)}^{z} - \overbrace{x_k(\alpha)}^{P_C(z)})^T (x - \overbrace{x_k(\alpha)}^{P_C(z)}) \le 0, \quad \text{for all } x \in C \tag{6.39}$$

(i): By setting $x = x_k$, we obtain

$$\begin{aligned}
(x_k - \alpha \nabla f(x_k) - x_k(\alpha))^T (x_k - x_k(\alpha)) &\le 0 \\
\text{or equivalently} \quad \frac{1}{\alpha} \|x_k - x_k(\alpha)\|^2 + \nabla f(x_k)^T (x_k(\alpha) - x_k) &\le 0,
\end{aligned}$$

which shows (6.37).
(ii): If $x_k(\alpha) = x_k$ then (6.39) becomes (6.38). □
The condition that $\bar{x} \in C$ satisfies

$$\nabla f(\bar{x})^T (x - \bar{x}) \geq 0, \quad \text{for all } x \in C,$$

is actually an optimality condition. We rewrite this in terms of a normal cone to C at $\bar{x}$:

$$-\nabla f(\bar{x}) \in N_C(\bar{x}) := \left\{ y \mid \langle y, x - \bar{x} \rangle = y^T (x - \bar{x}) \leq 0, \text{ for all } x \in C \right\}$$

and so we have (by Theorem 4.2.2) that

$$0 \in \nabla f(\bar{x}) + N_C(\bar{x}) = \partial \left(f(\cdot) + \delta_C(\cdot) \right)(\bar{x}).$$

This is called a variational inequality and it tells us that

$$\bar{x} \in \arg\min (f + \delta_C)$$

and solves $\min \{ f(x) \mid x \in C \}$.

Proposition 6.7.3 *Let $f : \mathbf{R}^n \to \mathbf{R}$ is convex, differentiable with a Lipschitz gradient and associated Lipschitz constant $L > 0$. Consider the gradient projections step*

$$x_{k+1} = P_C(x_k - \alpha \nabla f(x_k))$$

with constant step length $\alpha \in (0, \frac{2}{L})$. Then any limit point $\overline{x}$ of $\{x_k\}$ satisfies the (optimality) condition

$$\nabla f(\bar{x})^T (x - \bar{x}) \geq 0, \quad \textit{for all } x \in C.$$

Proof. By (6.36) we have

$$\begin{aligned} f(x_{k+1}) &\leq l(x_{k+1}; x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \end{aligned}$$

and by equation (6.37) we have

$$\nabla f(x_k)^T (x_{k+1} - x_k) \leq -\frac{1}{\alpha} \|x_{k+1} - x_k\|^2.$$

Combining these two relations we have

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{1}{\alpha} - \frac{L}{2} \right) \|x_{k+1} - x_k\|^2, \tag{6.40}$$

and as $\alpha \in (0, \frac{2}{L})$, we have $1/\alpha \in (L/2, \infty)$ and so $1/\alpha - L/2 > 0$, so the method reduces the cost function at each iteration. Let $\bar{x}$ be an accumulation point of $\{x_k\}_{k \in K}$. Then we must have $f(x_k) \downarrow f(\bar{x})$ (by continuity). By re-arranging equation (6.40),

$$\liminf_{k \in C} \|x_{k+1} - x_k\|^2 \leq \liminf_{k \in K} \frac{f(x_k) - f(x_{k+1})}{\left(\frac{1}{\alpha} - \frac{L}{2}\right)} \to 0.$$

Hence (using continuity, triangle inequality, etc.):

$$\|P_C(\bar{x} - \nabla f(\bar{x})) - \bar{x}\| = \liminf_{k \in K} \|x_{k+1} - x_k\| = 0$$

and so $\bar{x} = P_C(\bar{x} - \nabla f(\bar{x}))$ and Proposition 6.7.2 applies to guarantee $\bar{x}$ satisfies the optimality condition. □

Proposition 6.7.4 *Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable function and C a closed convex set. The for all $x \in C$ and $\alpha > 0$*

$$P_C\left(x - \alpha \nabla f\left(x\right)\right)$$

is the unique vector that attains the minimum in

$$\min_{y \in C} \left\{ l\left(y; x\right) + \frac{1}{2\alpha} \left\| y - x \right\|^2 \right\}.$$

Proof. Using the definition of l, we have for all x, y and $\alpha > 0$ that

$$\begin{aligned}
& l\left(y; x\right) + \frac{1}{2\alpha} \left\| y - x \right\|^2 \\
&= f\left(x\right) + \nabla f\left(x\right)^T \left(y - x\right) + \frac{1}{2\alpha} \left\| y - x \right\|^2 \\
&= f(x) + \frac{1}{2\alpha} \|y - x\|^2 + \frac{1}{\alpha} \langle y - x, \alpha \nabla f(x) \rangle \\
&= f(x) + \frac{1}{2\alpha} \left(\|y - x\|^2 + 2\langle y - x, \nabla f(x) \rangle + \alpha^2 \|\nabla f(x)\|^2 \right) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \\
&= f\left(x\right) + \frac{1}{2\alpha} \left\| y - \left(x - \alpha \nabla f\left(x\right)\right) \right\|^2 - \frac{\alpha}{2} \left\| \nabla f\left(x\right) \right\|^2 .
\end{aligned}$$

This trick of completing the square is used often in optimisation. The gradient projection $P_C\left(x - \alpha \nabla f\left(x\right)\right)$ minimizes the right-hand side over C, and hence also minimises the left-hand side over C. □

The following result replaces the inequality (6.11) for subgradient methods.

Proposition 6.7.5 *Let $f : \mathbf{R}^n \to \mathbf{R}$ is convex, differentiable function and C a closed convex set. Consider*

$$x_k\left(\alpha\right) = P_C\left(x_k - \alpha \nabla f\left(x_k\right)\right), \quad \alpha > 0.$$

Then for all $y \in \mathbb{R}^n$ and $\alpha > 0$, we have

$$\left\| x_k\left(\alpha\right) - y \right\|^2 \leq \left\| x_k - y \right\|^2 - 2\alpha \left(l\left(x_k\left(\alpha\right); x_k\right) - l\left(y; x_k\right) - \delta_C(y) \right) - \left\| x_k - x_k\left(\alpha\right) \right\|^2 . \tag{6.41}$$

Proof. By Proposition 6.7.4 we have

$$\begin{aligned}
x_k\left(\alpha\right) &\in \arg\min_x \left(l\left(x; x_k\right) + \delta_C\left(x\right) + \frac{1}{2\alpha} \left\| x - x_k \right\|^2 \right) \\
&= \arg\min_{x \in C} \left(l\left(x; x_k\right) + \frac{1}{2\alpha} \left\| x - x_k \right\|^2 \right).
\end{aligned} \tag{6.42}$$

We also have

$$\begin{aligned}
\left\| x_k - y \right\|^2 &= \left\| x_k - x_k\left(\alpha\right) + x_k\left(\alpha\right) - y \right\|^2 \\
&= \left\| x_k - x_k\left(\alpha\right) \right\|^2 + 2 \underbrace{\left(x_k - x_k\left(\alpha\right) \right)^T \left(x_k\left(\alpha\right) - y \right)}_{=:(a)} + \left\| x_k\left(\alpha\right) - y \right\|^2 .
\end{aligned} \tag{6.43}$$

Because we know (6.42), we have the optimality condition

$$0 \in \partial_x \left(l\left(x; x_k\right) + \delta_C\left(x\right) + \frac{1}{2\alpha} \left\| x - x_k \right\|^2 \right) |_{x = x_k(\alpha)}$$

$$= \partial \left(l\left(x; x_k\right) + \delta_C\left(x\right)\right)|_{x=x_k(\alpha)} + \frac{1}{\alpha}\left(x_k\left(\alpha\right) - x_k\right)$$

which yields

$$\frac{1}{\alpha}\left(x_k - x_k\left(\alpha\right)\right) \in \partial \left(l\left(x; x_k\right) + \delta_C\left(x\right)\right)|_{x=x_k(\alpha)}.$$

So we have that the subgradient inequality gives for all $y \in \mathbb{R}^n$

$$\frac{1}{\alpha}\overbrace{\left(x_k - x_k\left(\alpha\right)\right)^T\left(x_k\left(\alpha\right) - y\right)}^{=(a)} \geq l\left(x_k\left(\alpha\right); x_k\right) + \overbrace{\delta_C(x_k(\alpha))}^{=0} - \left(l\left(y; x_k\right) + \delta_C(y)\right).$$

Substituting this lower bound for (a) into (6.43) gives

$$\|x_k - y\|^2 \geq \|x_k - x_k\left(\alpha\right)\|^2 + 2\alpha\left(l\left(x_k\left(\alpha\right); x_k\right) - l\left(y; x_k\right) - \delta_C(y)\right) + \|x_k\left(\alpha\right) - y\|^2$$

and provides (6.41) on re-arranging. □

We are now in a position where we can prove fast convergence of this methods.

Theorem 6.7.6 (Nemirovski Method) *Let $f : \mathbb{R}^n \to \mathbb{R}$ is convex, differentiable function and C a closed convex set. Assume ∇f satisfies the Lipschitz condition with Lipschitz constant L and that $X^* \neq \emptyset$ is the set of optimal solutions for minimising f over C. Let $\{x_k\}$ be a sequence generated by the gradient descent method (6.32) with $\alpha = \frac{1}{L}$ and β_k satisfying (6.33)-(6.34) using (6.35). Then for $f^* = f\left(x^*\right)$ (for $x^* \in X^*$) we have $\lim_{k\to\infty} d\left(x_k, X^*\right) = 0$ and*

$$f\left(x_k\right) - f^* \leq \frac{2L}{\left(k+1\right)^2} d^2\left(x_0, X^*\right), \quad \text{for } k = 1, 2, \ldots$$

Proof. Define the auxiliary sequence

$$z_k = x_{k-1} + \theta_{k-1}^{-1}\left(x_k - x_{k-1}\right), \quad k = 0, 1, 2, \ldots \tag{6.44}$$

where $x_{-1} = x_0$, so that $z_0 = x_0$. Using (6.33) we have $\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}$ and also $y_k = x_k + \beta_k\left(x_k - x_{k-1}\right)$ so

$$\begin{aligned} z_k &= x_k + x_{k-1} - x_k + \theta_{k-1}^{-1}\left(x_k - x_{k-1}\right) \\ &= x_k + \left(\theta_{k-1}^{-1} - 1\right)\left(x_k - x_{k-1}\right) \\ &= x_k + \left(\theta_{k-1}^{-1} - 1\right)\frac{1}{\beta_k}\left(y_k - x_k\right) \\ &= x_k + \ \theta_k^{-1}\left(y_k - x_k\right) \end{aligned} \tag{6.45}$$

Fix $k \geq 0$ and $x^* \in X^*$ and let

$$y^* = \left(1 - \theta_k\right)x_k + \theta_k x^*.$$

Since $x_k, x^* \in C$ and C is convex, it is clear that $y^* \in C$, and so

$$\delta_C(y^*) = 0. \tag{6.46}$$

Using (6.36) we have

$$f\left(x_{k+1}\right) \leq l\left(x_{k+1}; y_k\right) + \frac{L}{2}\left\|x_{k+1} - y_k\right\|^2. \tag{6.47}$$

Since x_{k+1} is the projection of $y_k - \alpha \nabla f(y_k) = y_k - \frac{1}{L}\nabla f(y_k)$ onto C, it minimises

$$l(y; y_k) + \frac{L}{2}\|y - y_k\|^2$$

over $y \in C$, so by Proposition 6.7.5:

$$\begin{aligned} l(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 &\leq l(y^*; y_k) + \frac{L}{2}\|y^* - y_k\|^2 - \frac{L}{2}\|y^* - x_{k+1}\|^2 + \overbrace{\delta_C(y^*)}^{=0 \text{ by } (6.46)} \\ &= l(y^*; y_k) + \frac{L}{2}\|y^* - y_k\|^2 - \frac{L}{2}\|y^* - x_{k+1}\|^2. \end{aligned}$$

Combining with (6.47), we have

$$\begin{aligned} f(x_{k+1}) &\leq l(y^*; y_k) + \frac{L}{2}\|y^* - y_k\|^2 - \frac{L}{2}\|y^* - x_{k+1}\|^2 \\ &= l((1-\theta_k)x_k + \theta_k x^*; y_k) + \frac{L}{2}\|(1-\theta_k)x_k + \theta_k x^* - y_k\|^2 \\ &\quad - \frac{L}{2}\|(1-\theta_k)x_k + \theta_k x^* - x_{k+1}\|^2 \\ &= l((1-\theta_k)x_k + \theta_k x^*; y_k) + \frac{\theta_k^2 L}{2}\|x^* + \overbrace{\theta_k^{-1}(x_k - y_k) - x_k}^{=-z_k \text{ by } (6.45)}\|^2 \\ &\quad - \frac{\theta_k^2 L}{2}\|x^* + \overbrace{\theta_k^{-1}(x_k - x_{k+1}) - x_k}^{=z_{k+1} \text{ by } (6.44)}\|^2 \\ &= l((1-\theta_k)x_k + \theta_k x^*; y_k) + \frac{\theta_k^2 L}{2}\|x^* - z_k\|^2 \\ &\quad - \frac{\theta_k^2 L}{2}\|x^* - z_{k+1}\|^2 \\ &\overset{\text{(convexity)}}{\leq} (1-\theta_k)\, l(x_k; y_k) + \theta_k l(x^*; y_k) + \frac{\theta_k^2 L}{2}\|x^* - z_k\|^2 \\ &\quad - \frac{\theta_k^2 L}{2}\|x^* - z_{k+1}\|^2 \end{aligned}$$

Using the inequality

$$l(x_k; y_k) \leq f(x_k) \quad \text{(implicitly the subgradient inequality)}$$

we have

$$f(x_{k+1}) \leq (1-\theta_k) f(x_k) + \theta_k l(x^*; y_k) + \frac{\theta_k^2 L}{2}\|x^* - z_k\|^2 - \frac{\theta_k^2 L}{2}\|x^* - z_{k+1}\|^2.$$

Re-arranging terms and adding $0 = f^* - f^*$ (where f^* denotes the optimal value over C), we have

$$\begin{aligned} &\frac{1}{\theta_k^2}(f(x_{k+1}) - f^*) + \frac{L}{2}\|x^* - z_{k+1}\|^2 \\ &\qquad \leq \frac{1-\theta_k}{\theta_k^2}(f(x_k) - f^*) + \frac{L}{2}\|x^* - z_k\|^2 - \frac{f^* - l(x^*; y_k)}{\theta_k} \end{aligned}$$

and adding these for $k = 0, 1, 2, 3, \ldots$ we can use the inequality

$$\frac{1-\theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2},$$

to telescope so that we obtain

$$\frac{1}{\theta_k^2}\left(f\left(x_{k+1}\right)-f^*\right)+\sum_{k=0}^{k}\frac{f^*-l\left(x^*;y_i\right)}{\theta_i}\leq\frac{L}{2}\left\|x^*-z_0\right\|^2.$$

Using the fact that $x_0=z_0$, and $f^*-l\left(x^*;y_i\right)\geq 0$, and $\theta_k\leq\frac{2}{k+1}$, $\theta_{-1}=1$ and taking the minimum over all $x^*\in X^*$ we obtain

$$f\left(x_{k+1}\right)-f^*\leq\frac{2L}{\left(k+2\right)^2}\left[d\left(x_0,X^*\right)\right]^2$$

giving the desired result. □

6.8 Cutting Plane Methods

These are most effective for convex functions, but potential does exists to extend them to the non-convex case. Here we briefly discuss other methods where a model of the function is built by collecting subgradients and forming their corresponding supporting hyperplanes to the epi f. This process is sometimes referred to as a cutting plane method. Given x^i and the corresponding $f^i:=f\left(x^i\right)$ and $g^i=s\left(x^i\right)\in\partial f\left(x^i\right)$ we can form a supporting plane

$$z=f^i+\langle g^i,x-x^i\rangle.$$

As each of these affine functions minorises the convex function f we may form a lower approximation

$$f\left(x\right)\geq\max_{i=1,\ldots,k}\left[f^i+\langle g^i,x-x^i\rangle\right]:=\widetilde{f}_k\left(x\right).$$

Note that for all k

$$\widetilde{f}_k\leq\widetilde{f}_{k+1}\quad\text{and}\quad\widetilde{f}_k\leq f$$

and $\widetilde{f}_k(x^i)=f(x^i)$ at each $i=1,\ldots,k$.
To construct such an approximation we proceed as follows:

General Cutting Plane Method: Set a tolerance $\varepsilon>0$, and a compact convex subset S that contains our desired minimum. Initially we start with $x^0\in S$ and $i=0$. **While** $\delta^k>\varepsilon$ **do** iteration i:

1. Call the black box to obtain $g^i\in\partial f\left(x^i\right)$ and compute

$$\widetilde{f}_i\left(x\right):=\max\left[f(x^i)+\langle g^i,x-x^i\rangle,\widetilde{f}_{i-1}\left(x\right)\right]$$
$$\text{and}\qquad\delta^i:=f\left(x^i\right)-\widetilde{f}_{i-1}\left(x^i\right).$$

2. Solve the following problem (at least approximately)

$$d^i\in\arg\min_{x^i+d\in S}\widetilde{f}_i\left(x^i+d\right).$$

3. Choose a step length: We take $t^k=1$

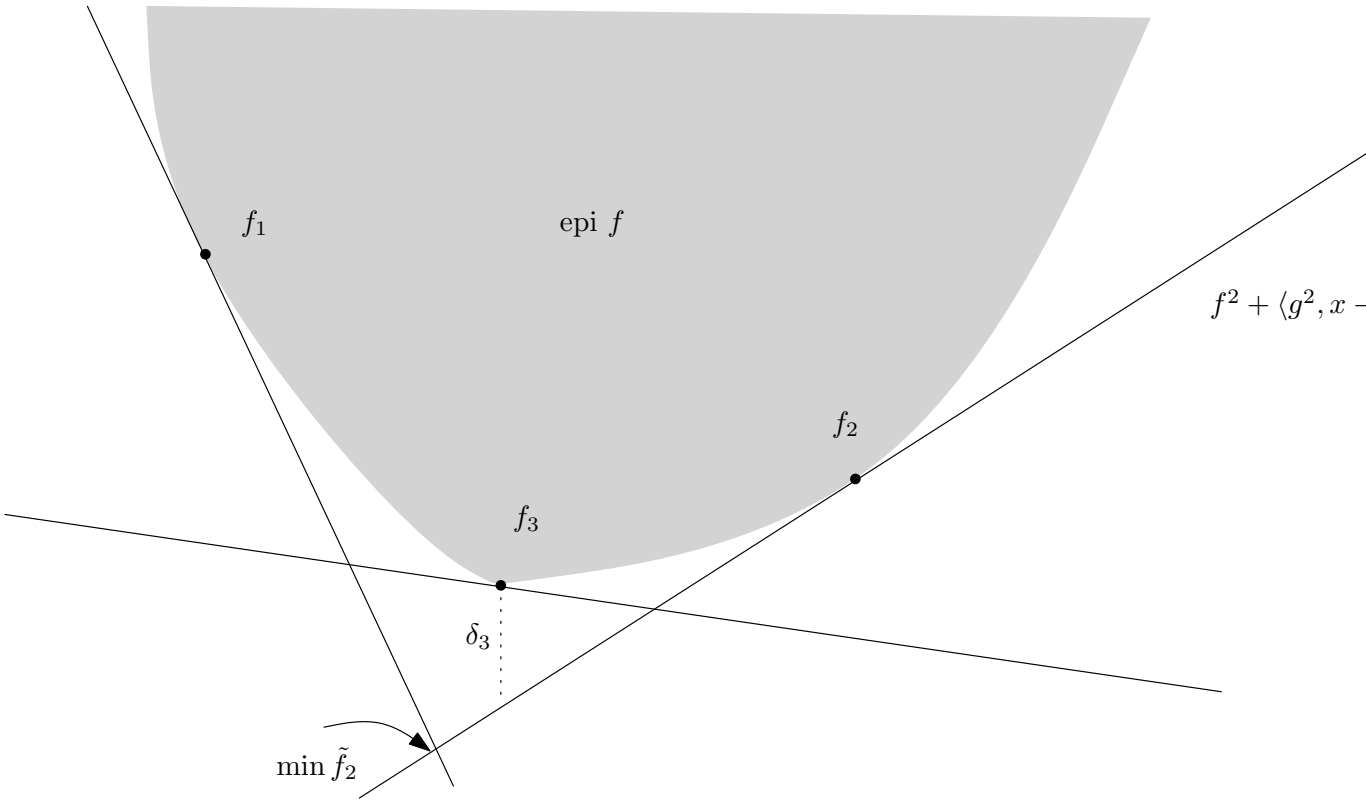


Figure 6.4: Three iterates of a cutting plane proceedure

4. Let $x^{i+1} = x^i + t^i d^i$ and $i \leftarrow i + 1$.

End While:

There is a price to be obtained for having to minimize a "model" function over a bounded region S. Suppose S is a box constraint then we must solve in step 2 an LP (linear program):

$$\begin{aligned} &\min_{(d,z)} \quad z \\ &\text{Subject to} \\ &\qquad z \geq f^j + \langle g^j, x^i - x^j \rangle + \langle g^i, d \rangle, \quad j = 1, \ldots, k_i \\ &\qquad\quad x^i + d \in S. \end{aligned}$$

The iterates will not provide a monotonically decreasing sequence. That is the model can predict a decrease and we find that when we evaluate the real function we get an increase. This is more likely when we get close to the minimizing point as we the are dealing will almost horizontal cutting planes, See figure 6.5. To over come this other methods use strategies to force a decrease in f. We may prove convergence of

$$\bar{f}_k := \min \left\{ f^1, \ldots, f^k \right\}$$

that is the best function value found. As S is compact we have the value $\left\{ \bar{f}_k \right\}$ finite and bounded below and the sequence of iterates $\left\{ x^k \right\}$ are bounded. Let f^* denote the $\min_{x \in S} f(x)$.

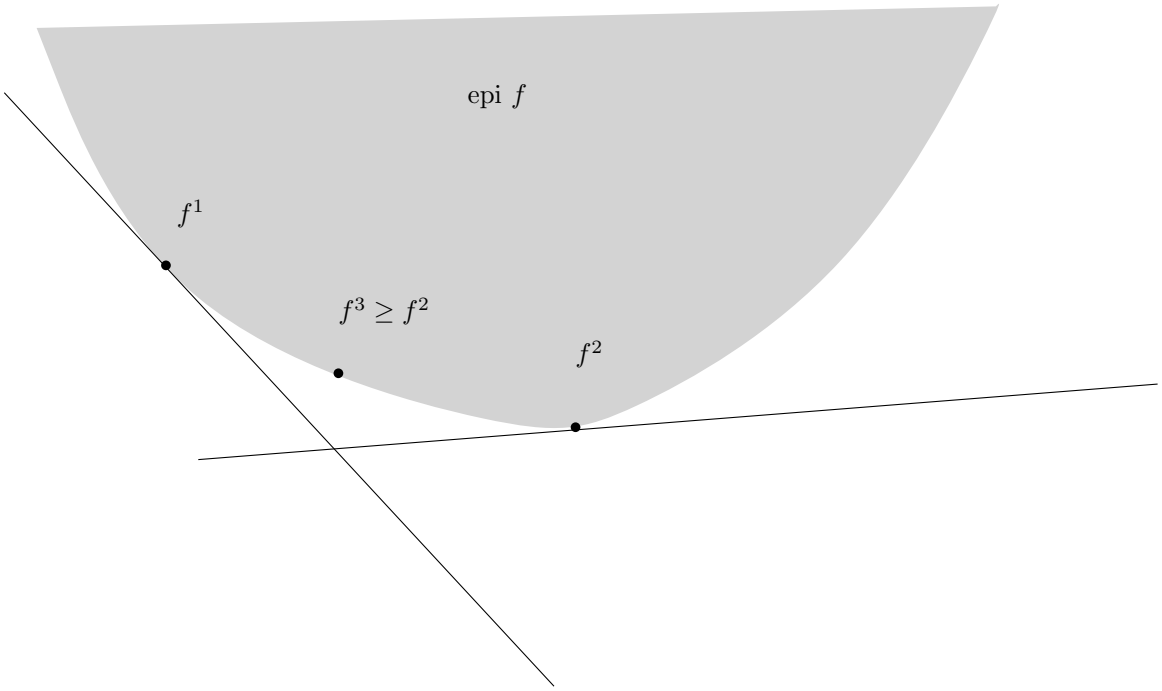


Figure 6.5: Non-monotonic behaviour of cutting plane method

Theorem 6.8.1 *Let f be a finite valued convex function on S a compact set. Consider a sequence $\{x^k\}$ generated by the cutting plane method $x^k := \operatorname{argmin} \widetilde{f}_{k-1}$.*

1. *If the algorithm runs indefinitely we have*

$$\lim_{k\to\infty} \widetilde{f}_{k-1}\left(x^k\right) = f^* = \liminf_{k\to\infty} f\left(x^k\right).$$

   Thus the stopping criteria $\delta^k := f\left(x^k\right) - \widetilde{f}_{k-1}\left(x^k\right) \leq \varepsilon$ will be triggered after a finite number of iterations.

2. *The last iterate $x^{k_{last}}$ is ε-optimal i.e*

$$f\left(x^{k_{last}}\right) \leq f^* + \varepsilon.$$

Proof. For the first part we begin by noting that $\left\{\widetilde{f}_{k-1}\left(x^k\right) = \min_z \widetilde{f}_{k-1}(z)\right\}$ is nondecreasing, because each $\widetilde{f}_{k-1}$ bounds the previous $\widetilde{f}_{k-2}$ from above. Moreover, the $\left\{\widetilde{f}_{k-1}\left(x^k\right)\right\}$ are bounded above by f^*, because $\widetilde{f}_k \leq f$ for all k. Therefore $\lim_k \widetilde{f}_{k-1}\left(x^k\right)$ exists and equals $f^* - C$ for some $C \geq 0$. So it holds that

$$\widetilde{f}_{k-1}\left(x^k\right) \leq f^* - C.$$

The assumption that $C > 0$ will yield a contradiction. As S is compact, take a convergent subsequence $\left\{x^{k_l}\right\}$ of the bounded sequence $\left\{x^k\right\} \subseteq S$. As f is finite valued, it is globally Lipschitz on S, so let M be an upper bound on $\left\{\|g\| \mid g \in \partial f\left(x^{k_l}\right), l = 0, \dots\right\}$. Take k_l sufficiently large so that so that $\left\|x^{k_{l-1}} - x^{k_l}\right\| \leq C/\left(2M\right)$. The desired contradiction follows from the following inequalities

$$f^* - C \geq \widetilde{f}_{k_l-1}\left(x^{k_l}\right)$$

$$
\begin{aligned}
&\geq \widetilde{f}_{k_{(l-1)}}\left(x^{k_l}\right) \quad \text{since } k_l - 1 \geq k_{l-1} \\
&\geq f_{k_{l-1}}\left(x^{k_{l-1}}\right) + \langle g^{k_{l-1}}, x^{k_l} - x^{k_{l-1}}\rangle \quad \text{by the definition of } \widetilde{f}_{k_{l-1}} \\
&= f\left(x^{k_{l-1}}\right) + \langle g^{k_{l-1}}, x^{k_l} - x^{k_{l-1}}\rangle \quad \text{since } f_{k_{l-1}}(x^{k_{l-1}}) = f(x^{k_{l-1}}) \\
&\geq f\left(x^{k_{l-1}}\right) - M\left\|x^{k_{l-1}} - x^{k_l}\right\| \\
&\geq f^* - MC/(2M) = f^* - \frac{C}{2}.
\end{aligned}
$$

This is the contradiction we sought. We now show that

$$\liminf_k f\left(x^k\right) = f^*.$$

Note that this will automatically imply that the decreasing sequence $\bar{f}_k := \min\left\{f(x^1), \ldots, f(x^k)\right\}$ must converge to f^*. Suppose for a contradiction that there is a constant $C > 0$ such that $f\left(x^k\right) \geq f^* + C$ for all k. Once again take a convergent subsequence $\left\{x^{k_l}\right\}$ of the bounded sequence $\left\{x^k\right\} \subseteq S$. Let L be the common Lipschitz constant of f and $\widetilde{f}_k$ (recall that we use the subgradients of f to construct $\widetilde{f}_k$, and so the Lipschitz constant for f is certainly also a Lipschitz constant for $\widetilde{f}_k$). Take k large enough so that $\left\|x^{k_{l+1}} - x^{k_l}\right\| \leq \frac{C}{2L}$. We arrive at a contradiction via the chain of inequalities:

$$
\begin{aligned}
f^* + C &\leq f\left(x^{k_l}\right) \quad \text{by assumption} \\
&= \widetilde{f}_{k_l}\left(x^{k_l}\right) \quad \text{since } \widetilde{f}_{k_l}\left(x^{k_l}\right) = \widetilde{f}\left(x^{k_l}\right) \\
&= \widetilde{f}_{k_l}\left(x^{k_l}\right) - \widetilde{f}_{k_l}\left(x^{k_{l+1}}\right) + \widetilde{f}_{k_l}\left(x^{k_{l+1}}\right) \\
&\leq \widetilde{f}_{k_l}\left(x^{k_l}\right) - \widetilde{f}_{k_l}\left(x^{k_{l+1}}\right) + \widetilde{f}_{(k_{(l+1)}-1)}\left(x^{k_{l+1}}\right) \quad \text{as } k_l \leq (k_{(l+1)} - 1) \\
&= \widetilde{f}_{k_l}\left(x^{k_l}\right) - \widetilde{f}_{k_l}\left(x^{k_{l+1}}\right) + \min_z \widetilde{f}_{(k_{(l+1)}-1)}(z) \\
&\leq \widetilde{f}_{k_l}\left(x^{k_l}\right) - \widetilde{f}_{k_l}\left(x^{k_{l+1}}\right) + f^* \\
&\leq L\left\|x^{k_l} - x^{k_{l+1}}\right\| + f^* \qquad \text{as } \widetilde{f}_{k_l} \text{ is Lipschitz} \\
&\leq L\frac{C}{2L} + f^* = f^* + \frac{C}{2} \qquad \text{a contradiction.}
\end{aligned}
$$

As there exists a subsequence with $f\left(x^{k_l}\right) \to f^*$ we must have $\delta^{k_l} := f\left(x^{k_l}\right) - \widetilde{f}_{k_l - 1}\left(x^{k_l}\right) \to f^* - f^* = 0$. Thus the stopping criteria is activated eventually. Let $\delta^{k_{last}} \leq \varepsilon$ and

$$
\begin{aligned}
f\left(x^{k_{last}}\right) &\leq \widetilde{f}_{k_{last}-1}\left(x^{k_{last}}\right) + \varepsilon \\
&\leq \widetilde{f}_{k_{last}-1}(x) + \varepsilon \quad \text{for all } x \in S \\
&\leq f(x) + \varepsilon \quad \text{for all } x \in S.
\end{aligned}
$$

Thus $f\left(x^{k_{last}}\right) \leq f^* + \varepsilon$. □

For a polyhedral convex functio, cutting plane methods can have good convergence properties but otherwise they exhibit instability and can be thrown far from the optimal solution. To overcome this, we need to control how far we can jump from iterate to iterate. This gives rise to the trust region idea and also the step length controls in bundle

methods. Another very serious drawback of cutting plane methods is the growth of the associated LP (linear program) generated in the method. Not only does it grow in size without bound, but it also becomes extremely difficult to solve due to ill conditioning caused by many very similar constraints or cutting planes. Thus one often deletes the less active constraints as we go. This is done in a rigorous way in bundle methods.

6.8.1 Bundle Methods

This class of algorithm has been until recently the gold plated standard, especially for convex functions where they still are quite competitive. They try and incorporate the best of subgradient and cutting plane methods together while eliminating the worst aspects of both approaches. Subgradient methods under-utilize information provided by the black box. They do not try and estimate the whole subdifferential from which we could calculate and approximation of a steepest descent direction, thus avoiding ascent steps. On the other hand cutting plane methods use all the past information, which is also problematic as this leads to large, ill-possed LPs as well as instability in the iterates (especially close to the optimal solution).

6.8.2 The Bundle as a Stabilizer

We collect information from our black box as we go. At step k we have

$$\left\{f^i := f\left(y^i\right),\, y^i,\, g^i := s\left(y^i\right),\, i = 1, \ldots, k\right\}$$

along with x^k and $f^k := f(x^k)$, the best function value and the point that generated it. When we calculate the $\widetilde{f}_k$ in the cutting plane method we, in effect, construct at each point a polyhedral approximation of the subdifferential given by $\partial \widetilde{f}_k\left(y^k\right)$. In bundle methods one does not necessarily accept an update at each iteration (as in subgradient optimisation). When a candidate point does not yield "sufficient descent" it is rejected and more effort is put into improving the model function we have (or getting a better estimate of the subdifferential). This is usually referred to as a "null step" where we accumulate more $\left\{y^i, g^i\right\}$. When sufficient descent is obtained we then make a "serious step" and update the subsequence $\left\{x^k\right\}$ of primal iterates. Various algorithms can be developed which depend on the following ingredients:

1. The model function φ_k that approximates f (for example $\varphi_k = \widetilde{f}_k$).
2. The choice of stabilization center (the x^k we look for descent from).
3. The method to dampen big oscillation in the iterates (usual depends on a normalization $\|\cdot\|_k$)

Essentially, most methods modify the cutting plane approach to prevent steps from moving "too far" from x^k (stabilization). We only step away from x^k when we obtain sufficient descent. In other words, set $x^{k+1} = y^{k+1}$ when

$$f\left(y^{k+1}\right) \leq f\left(x^k\right) - m\delta_{k+1}$$

where $m \in [0, 1]$ and δ_{k+1} is a nominal decrease as was predicted by the model function i.e $\delta_{k+1} = -\left(\widetilde{f}_k\left(y^{k+1}\right) - \widetilde{f}_k\left(x^k\right)\right) = -\left(\widetilde{f}_k\left(y^{k+1}\right) - f\left(x^k\right)\right) > 0$. In other words, set $x^{k+1} = y^{k+1}$ when

$$f\left(y^{k+1}\right) - f\left(x^k\right) \leq m\left(\widetilde{f}_k\left(y^{k+1}\right) - \widetilde{f}_k\left(x^k\right)\right) < 0.$$

General Bundle Method: Let $tol > 0$ and $m \in [0,1]$ be given parameters. Initially we start with $x^0 \in \mathbf{R}^n$ and $k = 0$ and construct the model function $\widetilde{f}_1$ and $\|\cdot\|$.
While $\delta_k > tol$ **do** iteration k: We terminate on an 'almost stationary' point, which frequently means $d(0, \partial f(x^k)) < \varepsilon$

1. Solve

$$y^{k+1} \in \arg\min \ \widetilde{f}_k(\cdot) + \frac{\rho^k}{2}\|\cdot\|^2 \quad \text{(a stabilized model minimization) and}$$

define $\delta_{k+1} = -\left(\widetilde{f}_k\left(y^{k+1}\right) - \widetilde{f}_k\left(x^k\right)\right) > 0$ (predicted model descent).

2. Call the black box at $x = y^{k+1}$ to obtain $\left(f\left(y^{k+1}\right), g^{k+1} := s\left(y^{k+1}\right)\right)$ and do

$$f(x^k) - f(y^{k+1}) \geq m\delta_{k+1}? \text{ then } \begin{cases} \text{Yes} & x^{k+1} = y^{k+1} \text{ (descent or serious step)} \\ \text{No} & x^{k+1} = x^k \qquad \text{(Null step)} \end{cases}$$

3. (Improving the model) Append y^{k+1} to the model and calculate φ_{k+1} and update $\|\cdot\|_{k+1}$.

End While:
Such methods are called "trust region" methods, in the sense that I "trust" my surrogate model if the descent that is predicted by the model is m-proportional of the descent that I actually get for the function f (case: Yes). When I stop "trusting" my surrogate model $\widetilde{f}_k$ (case: No), I keep updating my model until I "trust" it again (case: Yes), while I only ever trust my model on some region around where I currently am.
The update at y^{k+1} is

$$\widetilde{f}_{k+1}(y) := \max\left\{\widetilde{f}_k(y), f\left(y^{k+1}\right) + \langle g^{k+1}, y - y^{k+1}\rangle\right\}.$$

We reduce the $\left\{f^i := f\left(y^i\right),\ y^i,\ g^i := s\left(y^i\right),\ i = 1,\ldots,k\right\}$ from time to time to clean out redundant "old" information. The set of indices K_s at which serious steps are taken, generates an associated sequence $\{\delta_k\}_{k\in K_s}$.

Lemma 6.8.2 *Consider the General Bundle Method. Suppose it loops forever ($k \to \infty$). Use the notation $f^* := \lim_{k\in K_s} f\left(x^k\right)$ and assume $f^* > -\infty$. Then*

$$0 \leq \sum_{k\in K_s} \delta_k \leq \frac{f^0 - f^*}{m}$$

and so $\delta_k \to 0$.

Proof. Assume we iterate forever with the nominal decrease $\delta_k \geq tol > 0$ for all k. Take a $k \in K_s$. Since the descent test is satisfied, $x^{k+1} = y^{k+1}$ and

$$f\left(x^k\right) - f\left(x^{k+1}\right) = f\left(x^k\right) - f\left(y^{k+1}\right) \geq m\delta_{k+1}.$$

Let k' be the index following k in K_s so between these steps we have only null steps; i.e. $x^{k+j} = x^{k+1}$ for all $j = 2,\ldots,k'-k$. The descent test at k' gives

$$f\left(x^{k+1}\right) - f\left(x^{k'+1}\right) \geq m\delta_{k'+1}.$$

Hence, for any $k'' \in K_s$, (recalling $x^{k+j} = x^{k+1}$ for non-serious steps)

$$m \sum_{k \in K_s}^{k''} \delta_k \leq \sum_{k \in K_s}^{k''} f\left(x^k\right) - f\left(x^{k+1}\right) = \sum_{k=0}^{k''} \left[f\left(x^k\right) - f\left(x^{k+1}\right)\right] = f^0 - f^{k''} \leq f^0 - f^*.$$

Now let $k'' \to \infty$ gives the desired result. □

This result is used to show convergence whenever the algorithm generates an infinite sequence of serious steps. Namely, since the sequence $\left\{f\left(x^k\right)\right\}_{k \in K_s}$ is strictly decreasing, either $f\left(x^k\right) \downarrow -\infty$ or f is bounded below in which case $f\left(x^k\right) \downarrow f^*$. In this case we note that $\sum_{k \in K_s} \delta_k$ is convergent and so $\delta_k \to 0$ (triggering the stopping criteria). To show that we must generate an infinite sequence of serious steps require much more analysis, and we leave this to another course.

Problem Set 15

Problem 6.8.3 *1. Verify the conditions (6.33) and (6.34) for the specific example (6.35).*

2. Consider the fundamental descent finding problem for the bundle method: Denote $\bar{x} := x^i$ (current point or center) then we get d from the problem:

$$\begin{aligned} &\min_{(d,z) \in \mathbb{R}^m \times \mathbb{R}} \quad z + \frac{\rho^k}{2} \|\bar{x} + d\|^2 \\ &\textit{Subject to} \\ &\qquad z \geq f^j + \langle g^j, \bar{x} - x^j \rangle + \langle g^i, d \rangle, \quad j \in \{1, \ldots, k_i\} := J_k. \end{aligned} \tag{6.48}$$

(a) Rewrite the objective cutting plane model $f_k(\bar{x}, d) := \max_{j \in J_k} \left\{f^j + \langle g^j, \bar{x} + d - x^j \rangle\right\}$ as

$$\begin{aligned} f_k(\bar{x}, d) &= \max_{j \in J_k} \left\{\langle g^j, d \rangle - \alpha_{j,k} + f(\bar{x})\right\}, \quad \textit{where} \\ \alpha_{k,j} &:= f(\bar{x}) - \left(f\left(x^j\right) + \langle g^j, \bar{x} - x^j \rangle\right) \ \textit{(the linearisation errors).} \end{aligned}$$

(b) Why is $\alpha_{k,j} \geq 0$?

(c) Show that you can write the problem (6.48) as finding

$$\begin{aligned} d &\in \arg\min_{d'} \left\{f_k(\bar{x}, d' - \bar{x}) + \frac{\rho^k}{2} \left\|d'\right\|^2\right\} \\ &= \arg\min_{d'} \left\{v + \frac{\rho^k}{2} \left\|d'\right\|^2 \mid v \geq \langle g^i, d' \rangle - \alpha_{j,k}\right\} \end{aligned} \tag{6.49}$$

(d) Use Lagrangian duality to show that the solution to the problem (6.49) is given as below, where the dual variable for each constraint are denoted by $\lambda_j \geq 0$ for $i \in J_k$

$$\begin{aligned} 0 &= \lambda_j \left(-v + \langle g^j, d \rangle - \alpha_{j,k}\right) \quad \textit{for all } j \in J_k \\ d &= -(\rho^k)^{-1} \bar{g}^k \quad \textit{where } \bar{g}^k := \sum_{j \in J_k} \lambda_j g^j \quad \textit{and} \end{aligned}$$

$$v = \rho^k \|d\|^2 - \sum_{j \in J_k} \lambda_j \alpha_{k,j}.$$

Hint: remember your primal variable is (d, v) and not just d; use complementarity and the KKT conditions.

(e) Show that when there exists $\lambda_j \geq 0$ for $i \in J_k$ for which

$$\left\| \bar{g}^k \right\| = \left\| \sum_{j \in J_k} \lambda_j g^j \right\| \leq \varepsilon \quad \textit{and} \quad \sum_{j \in J_k} \lambda_j \alpha_{k,j} \leq \varepsilon$$

Then $\bar{x}$ is $\varepsilon > 0$ optimal in that

$$f(\bar{x}) \leq f(x) + \varepsilon \|x - \bar{x}\| + \varepsilon \quad \textit{for all } x \in \mathbb{R}^n.$$

Hint: use the definition of $\alpha_{j,k}$ and the fact that $\bar{x} - x^j = (\bar{x} - x) + (x - x^j)$, and the fact that the linearization minorizes f.

Chapter 7

Introduction to Machine Learning and Stochastic Gradient

We will discuss how descent method apply to problems on Data mining and machine learning where they have had significant impact in recent years. The "learning" part of machine learning revolves around the estimation of parameters in statistical models based on data. The function that is used to set such parameters is an "empirical risk" function that needs to minimised by choosing the "optimal" parameters. There are many questions that need answering to execute such an approach. What is the set of functions we allow the predictor to take? How do we measure how well the predictor performs on the training data? How do we construct predictors from only training data that performs well on unseen test data? We will discuss some of these questions and look at some regression problems and finally discuss support vector machines and classification problems.

7.1 Empirical Risk and Regression

Given a set of data (x_1, y_1), (x_2, y_2) ,..., (x_N, y_N) we would like to estimate a predictor f that depends on a parameter θ i.e.

$$f(x_j, \theta) \approx y_j, \quad \text{for } j = 1, \ldots, N.$$

Let us call $\hat{y}_i = f(x_j, \theta)$ is the predicted value given the parameter θ. Our goal is to use data to get the best estimate for the parameters θ. To do this we usually assume the training data is identically and independently distributed (so that the empirical sample mean unbiasedly estimates the population mean). Depending on the problems nature, the data and underlying distribution we need to introduce a "loss functions" which measures the lack of agreement between the data values y_i and the predicted values $\hat{y}_i$. For example for least-squares regression, we specify that we measure the cost of making an error during training using the squared error:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

we then minimise the empirical risk $R_{\text{emp}}(x, f, y) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, \hat{y}_i)$ for the given data set i.e.

$$\lim_{\theta} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_j, \theta))^2 .$$

When we have a linear model we would put $f(x_j, \theta) = \theta^T x_j$ which gives

$$\lim_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \theta^T x_j\right)^2 = \lim_{\theta} \frac{1}{N} \|y - X\theta\|^2 .$$

To measure how well the choice of parameters (and model) works we usually divide the available data into two sets. We use on set to fit the parameters and the other set to test how well the prediction that are made from out fitted model work. This is referred to as *supervised learning*. This leads us to the idea of *over fitting* of data. This occurs when we have a highly parametrised (complicated model) to explain the available data. The model fits the given data well (or even exactly) but does not "generalise" its predictions to new data. The problem is some what related to the "Occam's Razor". We should use the simplest explanation for the given data. To avoid this one should only use a subset of the data on hand to "train the model" i.e. estimate the parameters and use the other portion to "validate the mode;" i.e. make predictions and see how that pans out. There are more complicated statistical methods of dividing data into multiple subsets and cyclically the training of the model on all but one of these set, finally using the last set to test validation. All of these topic are out of scope of this course, belonging to a course devoted to machine learning. Here we will focus on the parameter fitting aspect only.

Usually in machine learning one assumes the data is generated via "noisy sampling". That is, with x the inputs and y the noise observation we observe data as

$$y_i = f(x_i) + \varepsilon_i$$

where $\varepsilon_i \sim$Probability Distribution. Let us consider a regression problem with a likelihood function

$$p(y \mid x) = N\left(y \mid f(x), \sigma^2\right)$$

where we are assuming $\varepsilon \sim N\left(0, \sigma^2\right)$ are identically independently normally distributed with mean 0 and variance σ^2. Our objective is to find, as close as possible, a representation of the unknown function f. The linear regression problem can be expressed as:

$$\begin{aligned} p(y \mid x, \theta) &= N\left(y \mid \theta^T x, \sigma^2\right) \\ y &= f(x) + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2\right). \end{aligned} \tag{7.1}$$

The likelihood in (7.1) is the probability density function of y evaluated at x. The only presence of uncertainty is due to the noise introduce by $\varepsilon \sim N\left(0, \sigma^2\right)$. Later we will easily generalise this to the situation where the model is linear only in the parameters θ i.e. $y = \phi(x)^T \theta$. Let continue and discuss the specifics of this model when we are given a training set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ with inputs $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ and observation $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$. Then

$$\begin{aligned} p(\mathcal{Y} \mid \mathcal{X}, \theta) &= p(\{y_1, y_2, \ldots, y_N\} \mid \{x_1, x_2, \ldots, x_N\}, \theta) \\ &= \Pi_{i=1}^{N} p(y_i \mid x_i, \theta) = \Pi_{i=1}^{N} N\left(y_i \mid \theta^T x_i, \sigma^2\right). \end{aligned}$$

Our goal is to find the parameter θ that solve

$$\theta^* \in \arg\max_{\theta} p(\mathcal{Y} \mid \mathcal{X}, \theta)$$

(here $\arg\max_\theta$ denotes the argument (the $\theta's$) that maximise the objective.) To turn this into a minimisation problem we take the negative log of this to obtain the "maximum likelihood problem" i.e.

$$\theta^* \in \arg\min_\theta \left[-\log p\left(\mathcal{Y} \mid \mathcal{X}, \theta\right)\right]$$

where $N(y, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$ and

$$\begin{aligned} -\log p\left(\mathcal{Y} \mid \mathcal{X}, \theta\right) &= -\sum_{i=1}^{N} \log\left(N\left(y_i \mid \ \theta^T x_i, \sigma^2\right)\right) \\ &= -\sum_{i=1}^{N} \frac{1}{2\sigma^2}\left(y_i - \theta^T x_i\right)^2 + \text{constants}. \end{aligned}$$

Thus the negative log-likelihood (ignoring constants) is given by

$$\mathcal{L}\left(\theta\right) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \theta^T x_i\right)^2 = \frac{1}{2\sigma^2}\left\|y - X\theta\right\|^2,$$

where the design matrix is $X = [x_1, x_2, \ldots, x_N]$ and training inputs $y = [y_1, y_2, \ldots, y_N]^T$. In this case (and one of the few cases) we can actually write down equations for the solution of this problem. We set the derivative to zero

$$\begin{aligned} \frac{d}{d\theta}\frac{1}{2\sigma^2}\left\|y - X\theta\right\|^2 &= \frac{d}{d\theta}\frac{1}{2\sigma^2}\left(y - X\theta\right)^T\left(y - X\theta\right) \\ &= \frac{d}{d\theta}\frac{1}{\sigma^2}\left[\frac{1}{2}y^T y - \left(X^T y\right)\theta + \frac{1}{2}\theta^T X^T X\theta\right] \\ &= \frac{1}{\sigma^2}\left[-X^T y + X^T X\theta\right] = 0 \end{aligned}$$

giving the set of equations

$$X^T X\theta = X^T y \quad \text{or} \quad \theta = \left(X^T X\right)^{-1} X^T y.$$

Here we assume the inverse $\left(X^T X\right)^{-1}$ exists yet this does not always hold and in those cases we would have to solve the optimisation problem $\max_\theta \left\|y - X\theta\right\|^2$ directly.

7.1.1 Linear on a Nonlinear Basis Regression

We generalise linear regression to the case when there we want to approximate the function using a basis of nonlinear functions $\left\{\phi_1\left(x\right), \phi_2\left(x\right), \ldots, \phi_N\left(x\right)\right\}$. One can view is as being similar to Taylor series approximation of Fourier series. We truncate to a finite length and try to best fit a finite set of functions using a linear sum. We then consider the inclusion of prosteriori distribution for parameters. This changes the optimisation problems solved. We stress again that despite being able to write these problems as equations the resulting systems can be highly singular and so the direct solution of the associate optimisation problem becomes the most viable way to obtain a best solution. Here we assume the parameters are still expressed linearly in the model i.e. $\hat{y}_i = \phi\left(x_i\right)^T\theta$ where ϕ is a vector of nonlinear functions i.e.

$$p\left(y \mid x, \theta\right) = N\left(y \mid \phi(x)^T\theta, \sigma^2\right)$$

$$y = \phi(x)^T\theta + \varepsilon = \sum_{i=0}^{K-1} \theta_i \phi_i(x) + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2\right).$$

For instance we could use $\phi_i(x) = x^i$ for $i = 1, \ldots, N$ and so $f(x) = \sum_{i=0}^{k-1} \theta_i x^i$ is a polynomial. As before we can write

$$\begin{aligned}\Phi(x) &= \begin{bmatrix} \phi_0(x_1) & \phi_2(x_1) & \ldots & \phi_{K-1}(x_1) \\ \phi_0(x_2) & \phi_2(x_2) & \ldots & \phi_{K-1}(x_2) \\ \vdots & \ddots & \vdots & \vdots \\ \phi_0(x_N) & \phi_2(x_N) & \ldots & \phi_{K-1}(x_N) \end{bmatrix} \\ &= \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \ldots & x_1^{K-1} \\ 1 & x_2 & x_2^2 & x_2^3 & \ldots & x_2^{K-1} \\ \vdots & \ddots & x_3^2 & x_4^3 & \ldots & x_3^{K-1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \ldots & \ldots & x_N^{K-1} \end{bmatrix}\end{aligned}$$

and so get $f(x) = \Phi(x)\theta = \sum_{i=0}^{k-1} \theta_i x^i$. Then again we have a log likelihood function

$$\begin{aligned}-\log p(\mathcal{Y} \mid \mathcal{X}, \theta) &= -\sum_{i=1}^{N} \log\left(N\left(y_i \mid \ \phi(x_i)^T\theta, \sigma^2\right)\right) \\ &= \sum_{i=1}^{N} \frac{1}{2\sigma^2}\left(y_i - \ \phi(x_i)^T\theta\right)^2 + \text{constants}.\end{aligned}$$

By a similar calculation as before, placing the derivative of this with respect to θ to zero, this implies $\max_\theta \|y - \Phi(x)\theta\|^2$ is attained for

$$\theta = \left[\Phi(x)\Phi(x)^T\right]^{-1}\Phi(x)\, y$$

assuming we have a nonsingular matrix $\left[\Phi(x)\Phi(x)^T\right]$ (otherwise we must solve this as an optimisation problem). Once we have gotten the model, by finding the optimal parameters θ we can get a measure of fit using the root mean square error

$$\text{RMSE} = \sqrt{\frac{1}{N}\|y - \Phi(x)\theta\|^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \phi(x_i)^T\theta)^2}.$$

Since we find the parameters θ by minimising this we expect the measure to decrease as K (the polynomial size) increases. But when we calculate this for a sample that we did not use in the parameter fitting process the story can be the opposite. Generally we want fewer parameters than data points we use to fit the parameters to avoid over fitting. Over fitting refers to the phenomena where the RMSE decrease to small values for the training data but is very large for data we did not use in the training. We want our model to "generalise" to unseen data. That is have RMSE small for a new set of data generate form the same source as the training data.

Example: Let us fits both of the following functions to:

t (sec)	0	0.3	0.8	1.1	1.6	2.3	3
v (volts)	0	0.6	1.28	1.5	1.7	1.75	1.8

and determine which fits best by plotting the residuals

$$v = a_1 + a_2 e^{-3t/T} \tag{7.2}$$

$$v = b_1 + b_2 e^{-3t/T} + b_3 t e^{-3t/T}. \tag{7.3}$$

where T is the time required for $v(t)$ to reach 95% its final value. You may take $T = 3$.
Soln: We fit the data values by trying to demand that

$$v\left(t_i\right) = v_i \quad \text{for all data values } \left(t_i, v_i\right).$$

This means for (7.2) we try and fit the model by solving the system of equations:

$$\begin{pmatrix} 1 & e^{-3t_1/T} \\ 1 & e^{-3t_2/T} \\ \vdots & \vdots \\ 1 & e^{-3t_7/T} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_7 \end{pmatrix}$$

for (a_1, a_2) and for model (7.3) we solve the system

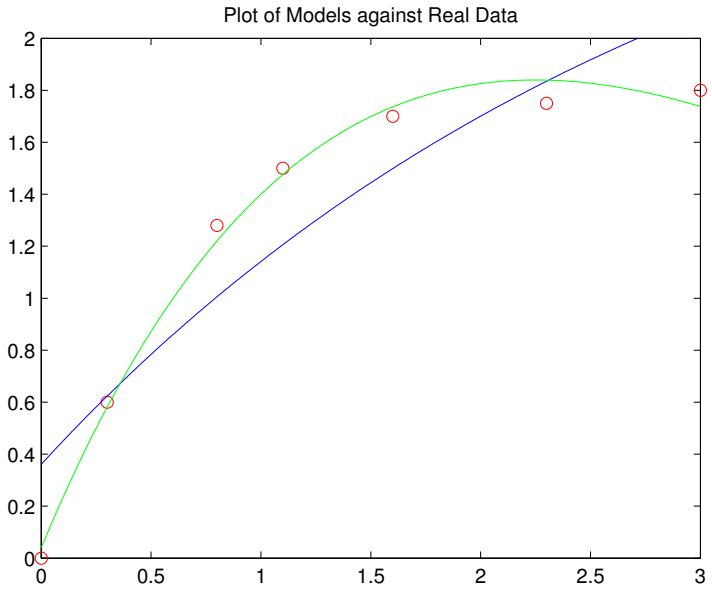
$$\begin{pmatrix} 1 & e^{-3t_1/T} & t_1 e^{-3t_1/T} \\ 1 & e^{-3t_2/T} & t_2 e^{-3t_2/T} \\ \vdots & \ddots & \vdots \\ 1 & e^{-3t_7/T} & t_7 e^{-3t_7/T} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_7 \end{pmatrix}$$

for (b_1, b_2, b_3). We need to solve this over determined system in the least squares sense. We can run a script like the following:

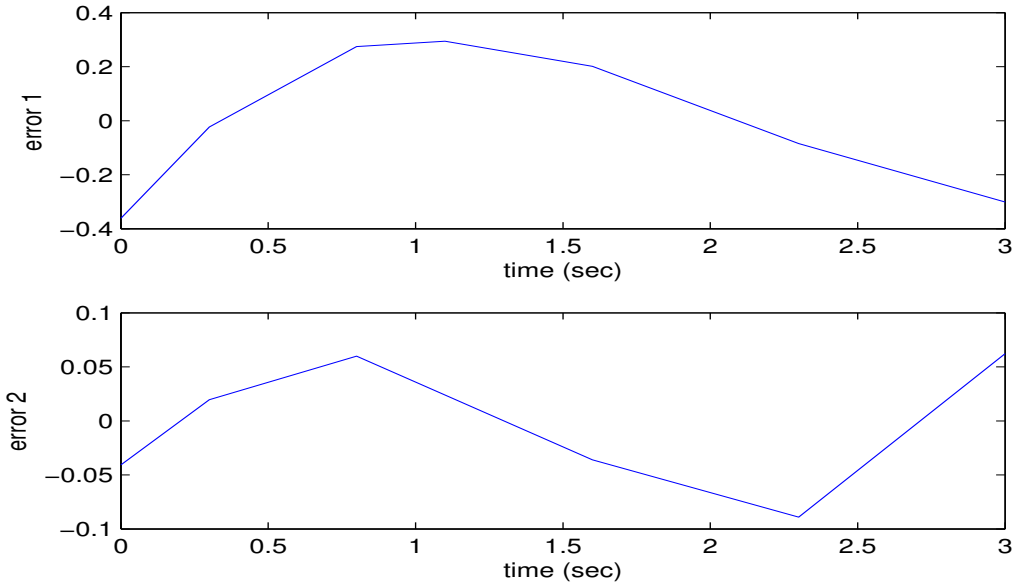
```
clear
x = [0,0.3,0.8,1.1,1.6,2.3,3];
y=[0,0.6,1.28,1.5,1.7,1.75,1.8];
X1 =  [ones(size(x')), exp(-x'/3)];
a = X1\y';

%or could use a = pinv(X1)*y'

X2=[ones(size(x')), exp(-x'/3), x'.*exp(-x'/3)];
b = X2\y';

%or could use b = pinv(X2)*y'

plot(x,y,'ro');
axis([0 3 0 2]);
hold on
u =[0:0.01:3]';
v1=[ones(size(u)) , exp(-u/3)]*a;
v2=[ones(size(u)),exp(-u/3), u.*exp(-u/3)]*b;
plot(u,v1,'b-',u,v2,'g-');
title('Plot of Models against Real Data')
hold off
figure
res1 = y' - [ones(size(x')) , exp(-x'/3)]*a;
```

```
res2 = y'-[ones(size(x')),exp(-x'/3),x'.*exp(-x'/3)] *b;
subplot(2,1,1)
plot(x,res1),xlabel('time (sec)'),ylabel('error 1')
subplot(2,1,2)
plot(x,res2),xlabel('time (sec)'),ylabel('error 2')
```

The `figure` command generates a new figure window so that the first plot is not over written. We get two plots:



and



Clearly the second set of residual display a more random behaviour.

In general we try and fit a n basis functions to the data $\{(y_i, x_i)\}_{i=1}^m$

$$y \approx \beta_1 \phi_1(x) + \cdots + \beta_n \phi_n(x)$$

by demanding the (sometimes impossible) conditions that

$$y_i = \beta_1 \phi_1(x_i) + \cdots + \beta_n \phi_n(x_i) := \Phi(x_i) \quad \text{for } i = 1, \ldots, m.$$

This may be written in matrix form as

$$\begin{pmatrix} \phi_1(x_1) & \ldots & \phi_n(x_1) \\ \phi_1(x_2) & \ldots & \phi_n(x_2) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \ldots & \phi_n(x_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

The solution does not exist so Matlab solves the problem of finding the which minimizes the sum of squares of the residuals i.e. for $\phi(x_i)^T = \begin{bmatrix} \phi_1(x_i) & \ldots & \phi_n(x_i) \end{bmatrix}$

$$J = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \phi(x_i)^T \beta \right)^2.$$

The fit can be measured using

$$r^2 = 1 - \frac{J}{S} \quad \text{where} \quad S = \frac{1}{N} \sum_{i=1}^{m} (y_i - \bar{y})^2 \text{ and } \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

As value of $r = 1$ or $J = 0$ is a "perfect fit". We always have $0 < r < 1$. We may investigate the fit found in our example using the commands:

```
>> format long; mu=mean(y);
>> r1=1-sum(res1.^2)/sum((y-mu).^2)
r1 =

0.84612864014987

>> r2=1-sum(res2.^2)/sum((y-mu).^2)
r2 =

0.99310817827039
```

If $r^2 > 0.99$ the model is considered that the fit is good.

7.1.2 Maximum A Posteriori Estimation

Here we wish to see how information about what parameter are "likely" for the model effects the optimisation problem for regression. We assume we know that $\theta \sim N\left(0, b^2\right)$. Then by Bayesian probability we have

$$p(\theta \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\theta, \mathcal{X}, \mathcal{Y})}{p(\mathcal{X}, \mathcal{Y})} = \frac{p(\theta, \mathcal{X}, \mathcal{Y}) / p(\mathcal{X})}{p(\mathcal{X}, \mathcal{Y}) / p(\mathcal{X})} = \frac{p(\theta, \mathcal{X}, \mathcal{Y}) / p(\mathcal{X})}{p(\mathcal{Y} \mid \mathcal{X})}$$

$$= \left[\frac{p(\theta, \mathcal{X}, \mathcal{Y})}{p(\mathcal{X})\, p(\theta)}\right] \frac{p(\theta)}{p(\mathcal{Y} \mid \mathcal{X})} = \frac{p(\mathcal{Y} \mid \mathcal{X}, \theta)\, p(\theta)}{p(\mathcal{Y} \mid \mathcal{X})}$$

from which we find that

$$\begin{aligned} -\log p(\theta \mid \mathcal{X}, \mathcal{Y}) &= -\log p(\mathcal{Y} \mid \mathcal{X}, \theta) - \log p(\theta) + \text{constants} \\ &\propto \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \phi(x_i)^T \theta\right)^2 + \frac{1}{2b^2} \|\theta\|^2 \\ &= \frac{1}{2\sigma^2} \left[\sum_{i=1}^{N} \left(y_i - \phi(x_i)^T \theta\right)^2 + \frac{\sigma^2}{b^2} \|\theta\|^2\right]. \end{aligned}$$

The nature of the optimisation problem depends on the factor $\frac{\sigma^2}{b^2}$. We see that as the variance of the θ distribution increases the more the problem look like the original least square problem. As $b \to \infty$ the less information we have about $\theta \sim N\left(0, b^2\right)$. We can estimate σ^2 from the sample data using

$$J_{LMS} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \phi(x_i)^T \theta\right)^2$$

but the value of b^2 comes from other sources. One way to think of these problems in optimisation (and also in engineering in signal processing) is as a "regularised" least squares problem, where the second term involves a regularising parameter (say) $\lambda > 0$ that can be varied in order to obtain a better fit. We then solve the optimisation problem $\min_\theta L(\theta)$ where

$$\begin{aligned} L(\theta) &:= \frac{1}{N} \|y - \Phi(x)\, \theta\|^2 + \lambda \|\theta\|^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \phi(x_i)^T \theta\right)^2 + \lambda \sum_{i=1}^{N} \theta_i^2. \end{aligned}$$

The nature of the distribution $\theta \sim N\left(0, b^2\right)$ is given in the $\|\theta\|^2$ term and if we assume a different distribution this term could change. It turns out that for terms of the form $\|\theta\|^p$ the smaller the value of p is the more we consider the distribution to be "sparse". For $p = 1$ we get the LASSO (the Least Absolute Shrinkage and Selector Operator) which modeling a very sparse distribution of parameters (which is common in signal processing). These problems are common in any areas off science and engineering and are best solved using methods we will discuss in coming chapter (such as ADMM - Augmented direction method of multipliers).

7.1.3 Stochastic Gradient Descent

Here we consider approaches to solving the optimisation problem

$$L(\theta) = \sum_{i=1}^{N} L_i(\theta) \equiv -\sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)$$

when N is *very large.* If carry out a descent method using the whole data set then the calculation of the gradient

$$\nabla L\left(\theta^k\right) = \sum_{i=1}^{N} \nabla L_i\left(\theta^k\right)$$

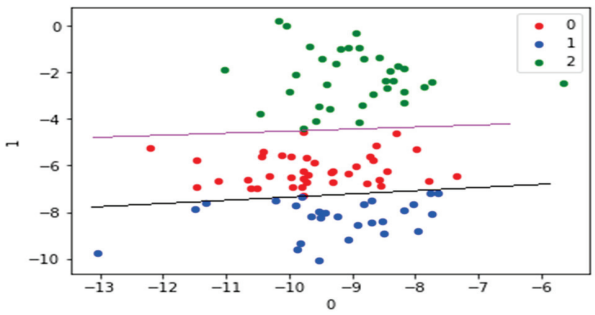
will requires the use of the full sample of size N and when N is large this will be prohibitive, slowing down computations. As we take

$$\theta^{k+1} = \theta^k - \gamma^k \sum_{i=1}^{N} \nabla L_i \left(\theta^k\right)$$

each iterate involves expensive evaluation of gradients for each loose function ∇L_i. It turns out that in order of the descent method to converge (in probability or almost surely to the minimum) we only need to sample subsets of the gradients $\left\{\nabla L_i \left(\theta^k\right)\right\}_{i=1}^N$ as long as the we use an unbiased estimator of the true full gradient. In fact the full estimate $\sum_{i=1}^N \nabla L_i \left(\theta^k\right)$ is itself an estimator of the "true gradient" of the expected value of the loss function. So one strategy is to randomly sample from the set $\left\{\nabla L_i \left(\theta^k\right)\right\}_{i=1}^N$ a smaller sample of moderate size to estimate the gradient at each iteration. The larger the sample to smaller the variance of the estimator will be and this results in a more stable descent. On the other hand small samples lead to the introduction of some random behaviour that can have the beneficial side effect of allowing the descent method to escape local minima and descend to a global minimiser. These methods are widely used in machine learning.

7.2 Support Vector Machines and Classification Problems

We begin by discussing some basic ideas behind support vector machines and classifications problems. Given a set of raw data we want to separate the data in separate groups based on a set of characteristics. The data is represented a vector in which of values x_i for $i = 1, \ldots, N$. We suppose that we want to classify the data set into two separate groups i.e. important emails and junk email. For argument sake, suppose we have a perfect classification based in a binary variable $y_i \in \{1, -1\}$ so we can argument the variable as (x_i, y_i) for $i = 1, \ldots, N$. In general we might have more than two classes and so this may be a ternary variable (for there classes). We want to draw a hyperplanes $z = \langle w, x \rangle + b$ so that all the data of each group falls on separate side of the plane i.e. for three classes



The mathematical way to write down this perfect classification for two classes is to say that:

$$\text{for all } i = 1, \ldots, N \quad y_i \left(\langle w, x_i \rangle + b\right) > 0.$$

How do we classify which value of (w, b) work better i.e. separate the data more convincingly.

Proposition 7.2.1 *The distance x is from the hyperplane $z = \langle w, x\rangle + b$ where $\|w\| = 1$ is given by $d = |\langle w, x\rangle + b|$.*

Proof. Clearly the minimum distance occur perpendicular to the plane which is given by its (unit) normal w. When $\langle w, x_0\rangle + b = 0$ (is on the plane) then $v = x - x_0$ projected onto w will be perpendicular and have the require distance as its magnitude. That is

$$\begin{aligned} d &= \|proj_w(x - x_0)\| \\ &= \|[(x - x_0)\cdot w]\, w\| = |\langle w, x\rangle - \langle w, x_0\rangle| = |\langle w, x\rangle + b\rangle.| \end{aligned}$$

□

The hard SVP training rule would be maximise the separation formed via the choice of w and b. That is, we choose

$$(w, b) \in \arg\max_{\|w\|=1, b\in\mathbb{R}} \left\{ \min_i |\langle w, x_i\rangle + b| : y_i(\langle w, x_i\rangle + b) > 0, \forall i = 1, \ldots, N \right\} \tag{7.4}$$

where the "argmax" denotes the arguments of the functions that maximise it. We claim that this can be solved using the quadratic program (QP):

$$\begin{array}{ll} \text{Input} & (x_i, y_i) \text{ for } i = 1, \ldots, N. \\ \text{Solve} & (w_0, b_0) \in \arg\min \left\{ \|w\|^2 : y_i(\langle w, x_i\rangle + b) > 1, \forall i = 1, \ldots, N \right\} \\ \text{Output} & \hat{w} = \frac{w_0}{\|w_0\|} \text{ and } \hat{b} = \frac{b_0}{\|w_0\|}. \end{array} \tag{7.5}$$

We will explain why via the following argument. Let (w^*, b^*) solve (7.4) and place

$$\gamma^* := \max_{\|w\|=1, b\in\mathbb{R}} \left\{ \min_i |\langle w, x_i\rangle + b| : y_i(\langle w, x_i\rangle + b) > 0, \forall i = 1, \ldots, N \right\}.$$

Note that since (w^*, b^*) solve (7.4) we have $\min_i |\langle w^*, x_i\rangle + b^*| = \gamma^*$ so

$$\begin{aligned} & y_i[\langle w^*, x_i\rangle + b^*] \geq \gamma^*, \quad \forall i = 1, \ldots, N \\ \text{implying} \quad & y_i\left[\langle \frac{w^*}{\gamma^*}, x_i\rangle + \frac{b^*}{\gamma^*}\right] \geq 1, \quad \forall i = 1, \ldots, N. \end{aligned}$$

Hence $\left(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*}\right)$ is feasible for the QP given in (7.5) and consequently, via the optimality of (w_0, b_0), we have

$$\|w_0\| \leq \left\|\frac{w^*}{\gamma^*}\right\| = \frac{1}{\gamma^*}.$$

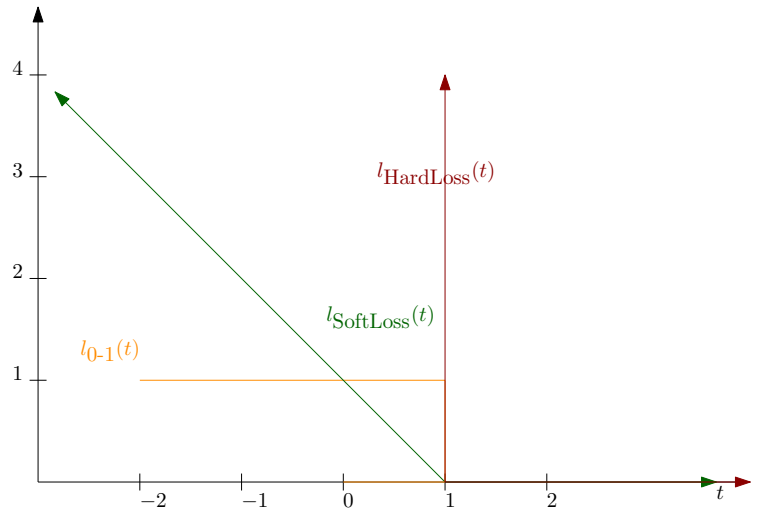
It then follows that for all $i = 1, \ldots, N$ (using the constraints of (7.5) that $y_i[\langle w_0, x_i\rangle + b_0] \geq 1$)

$$\left|\langle \hat{w}, x_i\rangle + \hat{b}\right| = y_i\left[\langle \hat{w}, x_i\rangle + \hat{b}\right] = \frac{1}{\|w_0\|} y_i[\langle w_0, x_i\rangle + b_0] \geq \frac{1}{\|w_0\|} \geq \gamma^* > 0.$$

Since $\|\hat{w}\| = 1$ we have $\left(\hat{w}, \hat{b}\right)$ not only feasible for (7.4) but optimal for (7.4) because

$$\min_{i=1,\ldots,N} \left|\langle \hat{w}, x_i\rangle + \hat{b}\right| \geq \gamma^* = \max_{\|w\|=1, b\in\mathbb{R}} \left\{ \min_i |\langle w, x_i\rangle + b| : y_i(\langle w, x_i\rangle + b) > 0, \forall i = 1, \ldots, N \right\}.$$

We will see later that (7.5) is tractably solved as a quadratic programming problem. But we will also later how we might solve this as an unconstrained problem.



7.2.1 Soft Margin SVM and Loss Function Formulation

We now modify this QP to allow for errors in the classification fit. As the hard constraint $y_i(\langle w, x_i\rangle + b) > 1$ may not be able to be force to hold true for all i we introduce "slack variables" $\xi_i \geq 0$ to write

$$y_i(\langle w, x_i\rangle + b) > 1 - \xi_i, \quad \forall i = 1, \ldots, N.$$

We then try and minimise the average value of the slack variables:

$$\begin{array}{lll} \text{Input} & & (x_i, y_i) \text{ for } i = 1, \ldots, m \text{ and parameter } \lambda > 0 \\ \text{Solve} & (w_0, b_0) \in & \arg\min \left\{ \lambda \|w\|^2 + \frac{1}{N}\sum_{i=1}^{N} \xi_i \right. \\ & & \left. : y_i(\langle w, x_i\rangle + b) > 1 - \xi_i, \quad \xi_i \geq 0, \forall i = 1, \ldots, N \right\} \\ \text{Output} & & w_0 \text{ and } b_0 \, . \end{array} \tag{7.6}$$

We wish to reinterpret this problem as one that arises from the minimisation of a loss function (like the regression problems we discussed). We need to penalise the distance we are from satisfaction of the constraint $y_i(\langle w, x_i\rangle + b) > 1$. To this end let $z_i = (\langle w, x_i\rangle + b)$ and consider the inequality $yt \geq 1$. We consider the following associated *soft margin* loss function

$$l_{\text{SoftLoss}}(t) := \max\{0, 1 - t\}$$

which has the following graph where we compare with the hard loss function:

$$l_{\text{HardLoss}}(t) = \begin{cases} 0 & \text{if } t \geq 1 \\ +\infty & \text{if } t < 0 \end{cases} .$$

We now have a regularised loss function for the training set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ given by

$$\begin{aligned} L(w, b) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} l_{\text{SoftLoss}}(y_i(\langle w, x_i\rangle + b)) \\ &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max\{0, 1 - (y_i(\langle w, x_i\rangle + b))\} . \end{aligned} \tag{7.7}$$

Usually the regularisation (that arises from an assumption of normality of the a prioristori distribution on the parameter w) would be weighted but instead the weights have been moved onto the loss function. Once can reformulate the loss problem $\min_w L(w)$ in the following way. Introduce new "slack" variables to be defined as:

$$\begin{aligned} \xi_i &\geq 1 - (y_i (\langle w, x_i \rangle + b)) \\ \text{or} \quad y_i (\langle w, x_i \rangle + b) &\geq 1 - \xi_i \end{aligned}$$

and then minimise

$$L(w, b) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

Subject to

$$\begin{aligned} y_i (\langle w, x_i \rangle + b) &\geq 1 - \xi_i, \quad i = 1, \ldots, N \\ \xi_i &\geq 0, \quad i = 1, \ldots, N \end{aligned}$$

The minmisation in $\xi_i \geq 0$ will result in $\xi_i = 1 - y_i (\langle w, x_i \rangle + b)$ when $\xi_i = 1 - y_i (\langle w, x_i \rangle + b) \geq 0$ and $\xi_i = 0$ if $1 - y_i (\langle w, x_i \rangle + b) < 0$. That is

$$\xi_i = \max \{0, 1 - (y_i (\langle w, x_i \rangle + b))\}$$

as required. This is the same as (7.6) when $C = \frac{1}{\lambda N}$. In this way we see that the situation here is no different to regression. That we have an optimisation problem with the structure

$$\begin{aligned} &\min_{w,b} R(w, b) + C \; L(w, b \mid \mathcal{X}, \mathcal{Y}) \\ &= \min_{w,b} R(w, b) + C \sum_{i=1}^{N} L(w, b \mid (x_i, y_i)) \end{aligned}$$

Some have observed that the parameter b play a crucial role in this model (especial in unbalanced data sets). One way of handling this is to add the regularisation term $R(w, b) = \frac{1}{2} \left(\|w\|^2 + b^2 \right)$ which introduces mote convexity in the parameter b adding convergence of optimisation algorithms.

In that case we see that we may introduce nonlinearity via a mapping $\phi(x) : \mathbb{R}^n \to \mathcal{H}$ a "feature space". For instance we may set $w = \phi(x)^T \beta = \sum_{i=1}^{N} \phi(x_i) \beta_i$, then the optimisation problem corresponds to an L_2-SVM with regularisation term

$$\begin{aligned} &\frac{1}{2} \beta^T \phi(x) \phi(x)^T \beta + C \sum_{i=1}^{N} \max \left\{ 0, 1 - \left(y_i \left(\langle \phi(x)^T \beta, \phi(x_i) \rangle + b \right) \right) \right\} \\ &= \frac{1}{2} \beta^T \mathbf{K} \beta + C \sum_{i=1}^{N} \max \left\{ 0, 1 - \left(y_i \left(\beta^T \phi(x) \phi(x_i) + b \right) \right) \right\} \\ &= \frac{1}{2} \beta^T \mathbf{K} \beta + C \sum_{i=1}^{N} \max \left\{ 0, 1 - \left(y_i \left(\beta^T \mathbf{K}_i + b \right) \right) \right\} \end{aligned}$$

where

$$\mathbf{K} = \phi(x) \phi(x)^T = \begin{bmatrix} \phi(x_1)^2 & \phi(x_1) \phi(x_2) & \ldots & \phi(x_1) \phi(x_N) \\ \phi(x_2) \phi(x_1) & \phi(x_2)^2 & \vdots & \phi(x_2) \phi(x_N) \\ \phi(x_3) \phi(x_1) & \phi(x_3) \phi(x_2) & & \phi(x_3) \phi(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(x_N) \phi(x_1) & \ldots & \ddots & \phi(x_N)^2 \end{bmatrix}$$

is the "kernel" associated with the mapping ϕ and $\phi(x)\,\phi(x_i)$ is the i th column of $\mathbf{K}$. We would then minimise over the variables (β, b). These optimisation problems are not differentiable and so one may also consider a version that is using the loss function (say)

$$L(w,b) = \frac{1}{2}\|w\|^2 + C\ \sum_{i=1}^{N} [\max\{0, 1-(y_i(\langle w, x_i\rangle + b))\}]^2.$$

To solve the actual SVM we need descent methods that can handle nonsmooth convex functions (we discuss convergence of these in the final section coming next). An simple and effective stochastic subgradient descent method has been proposed to solve the minimisation of the loss function in (7.7) i.e.

$$(w,b) \in \arg\min_{(w,b)} \frac{\lambda}{2}\left(\|w\|^2 + b^2\right) + \frac{1}{N}\ \sum_{i=1}^{N} [\max\{0, 1-(y_i(\langle w, x_i\rangle + b))\}].$$

The vector w is initially set to zero i.e. $w^0 = 0$. At the k th iteration we randomly sample a pair $(x_{t_k}, y_{t_k}) \in \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ in the training set and the objective $L(w,b)$ is approximated by

$$l((w,b), i_k) = \frac{\lambda}{2}\left(\|w\|^2 + b^2\right) + \max(0, 1-(y_{i_k}(\langle w, x_{i_k}\rangle + b))).$$

We require a "gradient" at all points and the function $l_{\text{SoftLoss}}(t) = \max\{0, 1-t\}$ is not differentiable at $t = 1$? We already know from Theorem 1.1.4 that when a convex function g is differentiable at $\bar{t}$ we have

$$g(t) - g(\bar{t}) \geq g'(\bar{t})\,(t - \bar{t}).$$

We keep this inequality true for $l_{\text{SoftLoss}}(t)$ at $t = 0$ i.e.

$$\begin{aligned} &l_{\text{SoftLoss}}(t) - 0 \geq g(t-0) \quad \text{or} \\ &\max\{0, 1-t\} \geq gt. \end{aligned}$$

One can verify that this requires $g \leq 0$ (consider $t > 1$) and $g \geq -1$ (consider $t < 0$ so that $\frac{1-t}{t} \leq g$ with $t \to -\infty$). So any $g \in [-1, 0]$ would do. If we (arbitrarily) take $g = 0$ then the "chain rule" would imply we could consistently take

$$\nabla_{(w,b)} l\left((w^k, b^k), i_k\right)^T = [\lambda w^k - \mathbf{1}\left(1 - y_{i_k}[\langle w^k, x_{i_k}\rangle + b]\right) y_{i_k} x_{i_k},\ \lambda b^k - \mathbf{1}\left(1 - [y_{i_k}\langle w^k, x_{i_k}\rangle + b]\right)].$$

where $\mathbf{1}(u) = 0$ if $u \leq 0$ and 1 if $u > 0$. When the perform an update using

$$\begin{aligned} w^{k+1} &= w^k - \eta^k \nabla_w l\left((w^k, b^k), i_k\right) \\ &= w^k - \eta^k [\lambda w^k - \mathbf{1}\left(1 - y_{i_k}[\langle w^k, x_{i_k}\rangle + b]\right) y_{i_k} x_{i_k}] \\ \text{and} \quad b^{k+1} &= b^k - \eta^k \nabla_b l\left((w^k, b^k), i_k\right) \\ &= b^k - \eta^k \left[\lambda b^k - \mathbf{1}\left(1 - [y_{i_k}\langle w^k, x_{i_k}\rangle + b]\right)\right]. \end{aligned}$$

The step length is the key to the success of the method. We can could choose it to be $\eta^k = \frac{1}{\lambda k}$ say (see next section).

Part III

Alternating Direction Method of Multipliers

Chapter 8

Basic Problem Formulation for Application of ADMM

8.0.1 Method of Multipliers

Consider the problem

$$\begin{aligned} &\min \ f(y) \\ &\text{Subject to } \ Ay = b \end{aligned}$$

Since the condition $Ay = b$ is equivalent to the pair of conditions $Ay - b \leq 0$ and $b - Ay \leq 0$, the Lagrangian reformulation is

$$\max_{(\lambda_1,\lambda_2)\in\mathbb{R}^m_+} \min_y f(y) + \lambda_1^T(Ay - b) + \lambda_2(b - Ay),$$

which, upon setting $\lambda = \lambda_1 - \lambda_2$ is

$$\max_{\lambda\in\mathbb{R}^m} \min_y f(y) + \lambda^T(Ay - b).$$

Since the optimal solution will satisfy complementarity, which translates into $Ay - b = 0$, this is equivalent to

$$\max_{\lambda\in\mathbb{R}^m} \min_y f(y) + \lambda^T(Ay - b) + \frac{\rho}{2}\|Ay - b\|^2.$$

The Lagrangian with this added penalty term $\|Ay - b\|$ is said to be *augmented*, and so this is said to be an *augmented Lagrangian*.

Solving the optimisation problem is akin to finding a saddle point of the augmented Lagrangian. For this purpose, we would like to iteratively update y and λ. We can choose an updated y^{k+1} to be a solution to the system

$$0 \in \partial_y \left(f(y) + (\lambda^k)^T(Ay - b) + \frac{\rho}{2}\|Ay - b\|^2 \right)(y^{k+1}).$$

Updating λ^{k+1} is more complicated, because whenever $Ay^{k+1} - b$ is not equal to zero, the maximization over λ fails to be finite. So, instead, we maximize a quadratic approximation:

$$0 \in \partial_\lambda \left((\lambda - \lambda^k)^T(Ay - b) - \frac{1}{2\rho}\|\lambda - \lambda^k\|^2 \right).$$

The quadratic term ensures that the maximizer is obtained finitely. It also give an explicit update formula:

$$\lambda^{k+1} = \lambda^k + \rho(Ay^{k+1} - b).$$

8.0.2 ADMM

The Alternating Direction Method of Multipliers ADMM is an algorithm that has had renewed interest in recent years. The reason for this is that it enables one to introduce a degree of parallel computing into problems that initially are not separable. It does so by introducing separability in an artificial way.

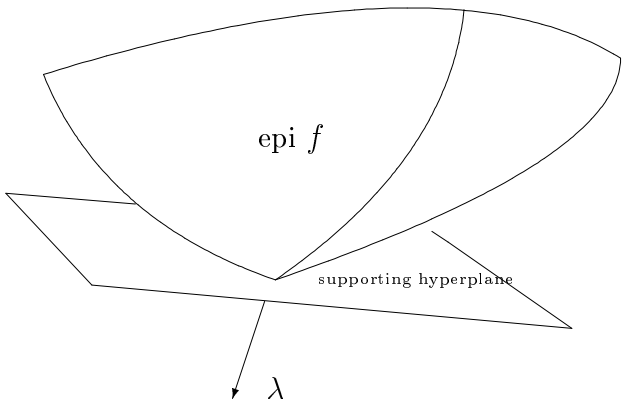
ADMM is an algorithm that is intended to blend the decomposability of dual ascent with the superior convergence properties of the method of multipliers. The algorithm solves problems in the form:

$$\begin{aligned} &\min{}_{y=(x,z)}\, f(x)+g(z) \\ &\text{Subject to } \; Ax+Bz=c \end{aligned} \tag{8.1}$$

where A and B are matrices of compatible dimensions to accommodate $x \in \mathbf{R}^n$ and $z \in \mathbf{R}^m$ i.e. A is $p \times n$ and B is $p \times m$. We assume f and g are convex functions but not

necessarily differentiable. When f lacks differentiability we can replace the derivative with the subderivative

$$\partial f(z) := \left\{ \lambda \in \mathbb{R}^m \mid f\left(z'\right) - f(z) \geq \lambda^T \left(z'-z\right), \text{ for all } z' \in \mathbf{R}^m \right\}.$$



A little review of subdifferentials

When f is differentiable we know that $\lambda := \nabla g(z)$ satisfies this subgradient inequality. Indeed in this case, placing $z' = z + \delta d$ in this inequality, we get

$$\begin{aligned} \lim_{\delta \to 0} \frac{f(z+\delta d)-f(z)}{\delta} = \nabla f(z)^T d \geq \lambda^T d \quad &\text{for all } d \text{ and so} \\ \left(\nabla f(z)-\lambda\right)^T d \geq 0 \quad &\text{for all } d. \end{aligned}$$

Choosing $d = -\left(\nabla f(z)-\lambda\right)$ we conclude that $\|\nabla f(z)-\lambda\|^2 \leq 0$ or $\lambda = \nabla f(z)$ is the only element of $\partial f(z)$. That is $\partial f(z) = \{\nabla f(z)\}$. This can be a larger convex set. Take the example of the indicator function of a convex set $C \subseteq \mathbb{R}^m$ given by:

$$\delta_C(z) := \begin{cases} 0 & \text{if } z \in \mathbb{C} \\ +\infty & \text{otherwise} \end{cases}$$

Then we have $w \in \partial\delta_C(y)$ for $y \in C$ if and only if

$$\delta_C(y') - \delta_C(y) \geq w^T(y' - y), \text{ for all } y' \in \mathbb{R}^m$$

and this inequality only makes a meaningful restriction when $y' := y + \alpha z \in C$, in which case:

$$0 \geq w^T z, \text{ for all } y + \alpha z \in C.$$

That is

$$\begin{aligned}\partial\delta_C(y) &= \left\{w \in \mathbb{R}^m \mid 0 \geq w^T z, \text{ for all } y + \alpha z \in C\right\} \\ &:= N_C(y)\end{aligned}$$

which we call the normal cone to C at $z \in C$. See the figure

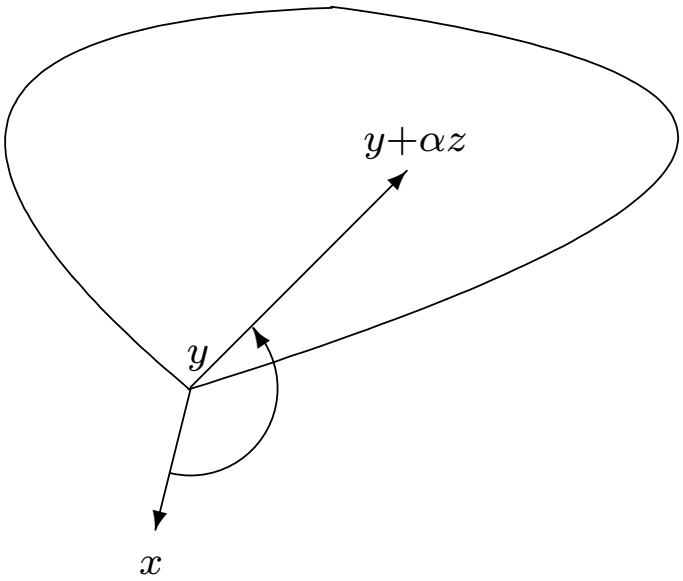


Figure 8.1: When $\mathbf{y}$ is closest point to $\mathbf{x}$ in C then $w := \mathbf{x} - \mathbf{y} \in N_C(\mathbf{y})$.

Remark 8.0.1 *Recall that we have*

$$\partial f(x) = \overline{\text{co}}\left\{z = \lim_{k\to\infty} \nabla f\left(x^k\right) \mid \forall x^k \to x \text{ with } \nabla f\left(x^k\right) \text{ existing and } \left\{\nabla f\left(x^k\right)\right\}_{k=1}^{\infty} \text{ convergent}\right\} \tag{8.2}$$

where $\text{co}\, C$ *is the smallest convex set containing* C *or equivalently the set of all finite convex combinations i.e.* $z \in \text{co}\, C$ *if and only if* $z = \sum_{i=1}^{n+1} \lambda_i z_i$ *for* $\mathbf{z}_i \in C$, $\sum_{i=1}^{n+1} \lambda_i = 1$ *and* $\lambda_i \geq 0$. *Clearly* $\partial f(x)$ *contains the right hand side of (8.2) as for each* k *we have*

$$f\left(x'\right) - f\left(x^k\right) \geq \nabla f\left(x^k\right)^T \left(x' - x^k\right) \quad \text{for all } x'$$

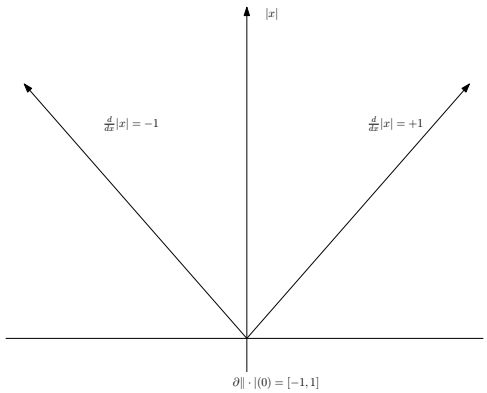
and as $x^k \to x$ *we have* $f\left(x^k\right) \to f(x)$ *and* $z_i = \lim_{k\to\infty} \nabla f\left(x^k\right)$ *so*

$$f\left(x'\right) - f(x) \geq z_i^T\left(x' - x\right) \quad \text{for all } x'$$

and so for any $\sum_{i=1}^{n+1} \lambda_i = 1$ *and* $\lambda_i \geq 0$ *we have*

$$f\left(x'\right) - f\left(x\right) \geq \left(\sum_{i=1}^{n+1} \lambda z_i\right)^T \left(x' - x\right) \quad \textit{for all } x'.$$

For example we have $\partial\left|\cdot\right|\left(0\right) = [-1, 1]$. *On the left of* 0 *we have* $|x|$ *derivative equal to* -1 *while to the right we have its derivative equaling* -1. *Thus at zero we must have* $\partial|\cdot|(0) = \text{co}\{-1, 1\} = [-1, 1]$.



The ADMM

We denote the optimal value of our problem by

$$p^* := \min\left\{f(x) + g(z) \mid Ax + Bz = c\right\}.$$

The associated augmented Lagrangian is given by the usual Lagrangian plus the additional penalty term for the constraints:

$$L_\rho\left(x, z, \lambda\right) := f(x) + g(z) + \overbrace{\lambda^T\left(Ax + Bz - c\right)}^{\text{usual Lagrangian}} + \overbrace{\frac{\rho}{2}\left\|Ax + Bz - c\right\|^2}^{\text{penalty term for constraints}}.$$

The basic iteration for the method of multipliers in the case would be:

$$y^{k+1} = \left(x^{k+1}, z^{k+1}\right) \in \underset{(x,z)}{\text{argmin}} L_\rho\left(x, z, \lambda^k\right)$$
$$\text{and} \qquad \lambda^{k+1} = \lambda^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right).$$

On the other hand, in ADMM, x and z are updated in an alternating or sequential fashion, which accounts for the name "alternating direction." ADMM can be viewed as a version of the method of multipliers where a single Gauss-Seidel pass over x and z is used. Separating the minimization over x and z into two steps is precisely what allows for decomposition when f or g are separable (and the constraints allow the augmented term to be separable):

$$x^{k+1} \in \arg\min_x L_\rho\left(x, z^k, \lambda^k\right) \tag{8.3a}$$
$$z^{k+1} \in \arg\min_z L_\rho\left(x^{k+1}, z, \lambda^k\right) \tag{8.3b}$$

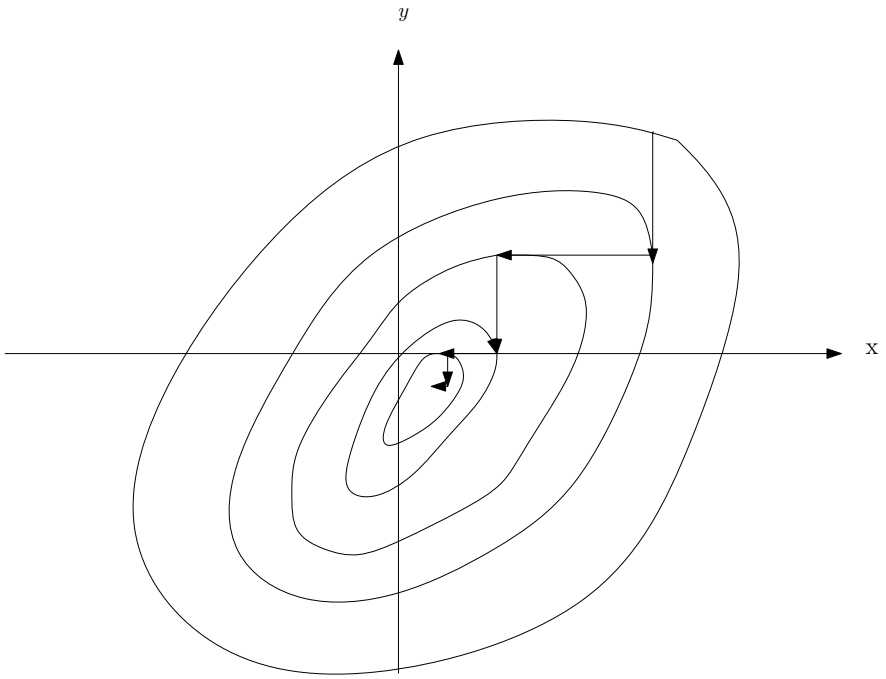


Figure 8.2: Gauss-Seidel alterates between variables performing a sequence of alternate minimizations.

$$\lambda^{k+1} = \lambda^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right). \tag{8.3c}$$

This method can be shown to converge in the following sense:

- Primal residuals converge:

$$r^{k+1} := Ax^{k+1} + Bz^{k+1} - c \to 0.$$

- Objective convergence:

$$f\left(x^k\right) + g\left(z^k\right) \to p^*.$$

- Dual convergence:

$$\lambda^k \to \lambda^* \quad \text{(optimal dual variable)}.$$

- Convergence of the primal variables: x^k and z^k. While we will not show this in class, it can be shown via a complicated Gauss-Seidel argument or through a connection through Fenchel duality with Douglas–Rachford method [10, 12].

Note that we already have the canonical sum-calculus rule in Theorem 4.2.2. What follows is a lighter-weight version.

Lemma 8.0.1 *Suppose f and h are convex and finite valued at $x \in \mathbf{R}^n$ with f not necessarily differentiable and $h \in C^1(\mathbf{R}^n)$. Then we have*

$$\partial(f+h)(x) = \partial f(x) + \{\nabla h(x)\} := \{w + \nabla h(x) \mid w \in \partial f(x)\}.$$

Consider the general quadratic function of several variables

$$q(x) = \frac{1}{2}x^T Qx + b^T x + c = \frac{1}{2}x^T Qx + x^T b + c$$

where b is an n column vector, c a real constant and Q an $n \times n$ real (usually symmetric) matrix. If we differentiate with respect to $\mathbf{x}$ what do we obtain? We get $Qx + b$, as we saw before in Lemma 4.1.18.

Example 8.0.1 *Consider the quadratic function in two variables*

$$\begin{aligned} q(x_1, x_2) &= \frac{1}{2}(x_1, x_2)\begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (3,1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 6 \\ &= \frac{1}{2}(x_1, x_2)\begin{pmatrix} 3x_1 + 4x_2 \\ 2x_1 + x_2 \end{pmatrix} + 3x_1 + x_2 + 6 \\ &= \frac{1}{2}\left(x_1(3x_1 + 4x_2) + x_2(2x_1 + x_2)\right) + 3x_1 + x_2 + 6 \\ &= \frac{3}{2}x_1^2 + 3x_1x_2 + \frac{1}{2}x_2^2 + 3x_1 + x_2 + 6 \\ &= \frac{1}{2}(x_1, x_2)\begin{pmatrix} 3 & 3 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1, x_2)\begin{pmatrix} 3 \\ 1 \end{pmatrix} + 6. \end{aligned}$$

Example 8.0.2 *One may show the last equality by expanding once again. Thus even when Q is not symmetric it may be replaced by a matrix which is symmetric. Now differentiate with respect to x_1 and x_2 to obtain*

$$\begin{aligned} \nabla q(x_1, x_2) &= \begin{pmatrix} 3x_1 + 3x_2 + 3 \\ 3x_1 + x_2 + 1 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} \\ &= Qx + b. \end{aligned}$$

Differentiating again to obtain the Hessian

$$\nabla^2 q(x_1, x_2) = \begin{pmatrix} 3 & 3 \\ 3 & 1 \end{pmatrix} = Q.$$

From the previous example we see that we may formally differentiate as follows

$$\begin{aligned} \nabla_x q(x) &= \nabla_x \left(\frac{1}{2}x^T Qx + x^T b + c\right) = Qx + b \\ \text{and} \quad \nabla_x^2 q(x) &= \nabla_x \left(Qx + b\right) = Q. \end{aligned}$$

These formula are used often. We may now prove convergence after deriving the following inequalities. Before doing this, we note that the (KKT) condition for our problem states that

$$(x, z) \mapsto L_0\left(x, z, \lambda^*\right) = f\left(x\right) + g\left(z\right) + \lambda^{*T}\left(Ax + Bz - c\right)$$

is minimized at $\left(x^*, z^*\right)$, and that $Ax^* + Bz^* = c$, so the (primal) KKT condition is

$$0 \in \partial_{(x,z)} L_0(\cdot, \cdot, \lambda^*)(x^*, z^*) = \partial_{(x,z)}\left[f(x) + g(z) + (\lambda^*)^T(Ax + Bz - c)\right](x^*, z^*).$$

This splits cleanly into the two inclusions:

$$0 \in \partial_x L_0\left(\cdot, z^*, \lambda^*\right)\left(x^*\right) = \partial f\left(x^*\right) + A^T\lambda^* \tag{8.4}$$

$$\text{and} \quad 0 \in \partial_z L_0\left(x^*, \cdot, \lambda^*\right)\left(z^*\right) = \partial g\left(z^*\right) + B^T\lambda^*. \tag{8.5}$$

In terms of the iterates (8.3a) to (8.3c) we may also state that

$$0 \in \partial L_\rho\left(\cdot, z^k, \lambda^k\right)\left(x^{k+1}\right) = \partial f\left(x^{k+1}\right) + A^T\lambda^k + \rho\overbrace{A^T}^{(*)}\left(Ax^{k+1} + Bz^k - c\right) \tag{8.6}$$

$$\text{and} \quad 0 \in \partial L_\rho\left(x^{k+1}, \cdot, \lambda^k\right)\left(z^{k+1}\right) = \partial g\left(z^{k+1}\right) + B^T\lambda^k + \rho B^T\left(Ax^{k+1} + Bz^{k+1} - c\right). \tag{8.7}$$

Here the term $(*)$ comes from the chain rule (likewise for the B^T in the same place on the next line). We may write (8.7) as

$$\begin{aligned} &0 \in \partial g\left(z^{k+1}\right) + B^T\lambda^k + \rho B^T r^{k+1} \\ &\text{where } r^{k+1} := Ax^{k+1} + Bz^{k+1} - c. \end{aligned} \tag{8.8}$$

We may write (8.6) (adding and subtracting the $\dagger = A^TB\rho z^{k+1}$ term) as

$$\begin{aligned} 0 &\in \partial f\left(x^{k+1}\right) + A^T\left(\lambda^k + \rho\big(Ax^{k+1} + B\overbrace{z^{k+1}}^{(\dagger)} - c\big)\right) + \rho A^T B\big(z^k - \overbrace{z^{k+1}}^{-(\dagger)}\big) \\ &= \partial f\left(x^{k+1}\right) + A^T\big(\lambda^{k+1}\big) + \rho A^T B\left(z^k - z^{k+1}\right) \\ \text{or} \quad s^{k+1} &\in \partial f\left(x^{k+1}\right) + A^T\lambda^{k+1} \quad \text{where } s^{k+1} := \rho A^T B\left(z^{k+1} - z^k\right). \end{aligned} \tag{8.9}$$

We may view s^{k+1} as the dual residual as it measures how far we are from satisfying (8.4) while $r^{k+1} := Ax^{k+1} + Bz^{k+1} - c$ is the primal residual as it measures the violation of (8.5). Convergence requires us to ensure $\left(r^k, s^k\right) \to (0, 0)$.

Proposition 8.0.2 *Suppose f and g are convex functions and we iterate (8.3a) to (8.3c). Then, setting p^* to be the optimal objective value, and defining $p^{k+1} := f(x^{k+1}) + g(z^{k+1})$, we have*

$$p^* - p^{k+1} \le \lambda^{*T} r^{k+1} \quad \textit{and} \tag{8.10a}$$

$$p^{k+1} - p^* \le -\left(\lambda^{k+1}\right)^T r^{k+1} - \rho\left(B\left(z^{k+1} - z^k\right)^T\left(-r^{k+1} + B\left(z^{k+1} - z^*\right)\right)\right). \tag{8.10b}$$

Proof. To show (8.10a) we use the fact that $(x, z) \mapsto L_0\left(x, z, \lambda^*\right)$ is minimized at (x^*, z^*), where complementarity holds: $\lambda^{*T}\left(Ax^* + Bz^* - c\right) = 0$. Here, remember that L_0 is the ordinary Lagrangian, since $\rho = 0$ means the quadratic penalty term disappears. Thus we always have

$$L_0\left(x^*, z^*, \lambda^*\right) \le L_0\left(x^{k+1}, z^{k+1}, \lambda^*\right) \quad \text{which is just}$$

$$\underbrace{f(x^*) + g(z^*)}_{p^*} + \underbrace{\lambda^{*T}\left(Ax^* + Bz^* - c\right)}_{=0 \text{ (complementarity)}} \le \underbrace{f(x^{k+1}) + g(z^{k+1})}_{p^{k+1}} + \lambda^{*T}\underbrace{\left(Ax^{k+1} + Bz^{k+1} - c\right)}_{r^{k+1}}$$

This verifies the first assertion (8.10a).

Now we show (8.10b). First notice the equalities:

$$\begin{aligned} A^T\lambda^{k+1} = \nabla_x\left((A^T\lambda^{k+1})^T x\right) &= \nabla_x((\lambda^{k+1})^T Ax) \\ \text{and} \qquad \rho A^T B(z^{k+1}-z^k) &= \nabla_x\left((\rho A^T B(z^{k+1}-z^k))^T x\right) \\ &= \nabla_x\left(\left(\rho B(z^{k+1}-z^k)\right)^T Ax\right), \\ \text{So altogether} \quad A^T\lambda^{k+1} - \rho A^T B(z^{k+1}-z^k) &= \nabla_x\left(\left[\lambda^{k+1} - \rho B\left(z^{k+1}-z^k\right)\right]^T Ax\right). \end{aligned}$$

Using this last identity, we have that

$$\begin{aligned} 0 &\overset{(8.9)}{\in} \partial f\left(x^{k+1}\right) + A^T\lambda^{k+1} - \rho A^T\ B\left(z^{k+1}-z^k\right) \\ \text{(Lemma 8.0.1)} \quad &= \partial_x\left[f(x) + \left[\lambda^{k+1} - \rho B\left(z^{k+1}-z^k\right)\right]^T Ax\right]\left(x^{k+1}\right) \end{aligned}$$

This means that x^{k+1} minimizes $x \mapsto f(x) + \left[\lambda^{k+1} - \rho B\left(z^{k+1}-z^k\right)\right]^T Ax$. Following a similar argument, we obtain from (8.8) that:

$$0 \in \partial_z\left[g(z) + \left[\lambda^k + \rho r^{k+1}\right]^T Bz\right]\left(z^{k+1}\right).$$

This means z^{k+1} minimizes $z \mapsto g(z) + \left[\lambda^k + \rho r^{k+1}\right]^T Bz = g(z) + \left[\lambda^{k+1}\right]^T Bz$. Since x^{k+1} and z^{k+1} are minimizers of these respective functions, they admit smaller function values for them than x^* and z^* respectively. This gives me the following two inequalities:

$$\begin{aligned} f\left(x^{k+1}\right) + \left[\lambda^{k+1} - \rho B\left(z^{k+1}-z^k\right)\right]^T Ax^{k+1} &\le f(x^*) + \left[\lambda^{k+1} - \rho B\left(z^{k+1}-z^k\right)\right]^T Ax^* \\ \text{and} \quad g\left(z^{k+1}\right) + \left[\lambda^{k+1}\right]^T Bz^{k+1} &\le g(z^*) + \left[\lambda^{k+1}\right]^T Bz^*. \end{aligned}$$

Adding these two inequalities, we get

$$\begin{aligned} p^{k+1} + \left[\lambda^{k+1}\right]^T\left[Ax^{k+1} + Bz^{k+1}\right] - \rho\left[B\left(z^{k+1}-z^k\right)\right]^T Ax^{k+1} \\ \le p^* + \left[\lambda^{k+1}\right]^T[Ax^* + Bz^*] - \rho\left[B\left(z^{k+1}-z^k\right)\right]^T Ax^*. \end{aligned}$$

Substituting the identities $Ax^* + Bz^* = c$ and $Ax^{k+1} + Bz^{k+1} - c = r^{k+1}$ we have

$$\begin{aligned} p^{k+1} + \left[\lambda^{k+1}\right]^T\left[c + r^{k+1}\right] - \rho B\left(z^{k+1}-z^k\right)^T Ax^{k+1} \\ \le p^* + \left[\lambda^{k+1}\right]^T c - \rho\left[B\left(z^{k+1}-z^k\right)\right]^T Ax^*. \end{aligned}$$

Collecting terms and again using the identities $Ax^* + Bz^* = c$ and $Ax^{k+1} + Bz^{k+1} - c = r^{k+1}$, we have

$$\begin{aligned} p^{k+1} - p^* &\le -\left[\lambda^{k+1}\right]^T r^{k+1} - \rho\left[B\left(z^{k+1}-z^k\right)\right]^T\left[-Ax^{k+1} + Ax^*\right] \\ &= -\left[\lambda^{k+1}\right]^T r^{k+1} - \rho\left[B\left(z^{k+1}-z^k\right)\right]^T\left[-r^{k+1} + Ax^* + Bz^{k+1} - c\right] \\ &= -\left[\lambda^{k+1}\right]^T r^{k+1} - \rho\left[B\left(z^{k+1}-z^k\right)\right]^T\left[-r^{k+1} + B\left(z^{k+1}-z^*\right)\right]. \end{aligned}$$

This shows (8.10b).

□

Corollary 8.0.3 *When we have* $s^{k+1} = B\left(z^{k+1} - z^k\right) \to 0$ *and* $r^{k+1} \to 0$ *then*

$$\lim_{k\to\infty} p^k = p^*.$$

Proof. We have (8.10a) implying

$$p^* \leq \liminf_{k\to\infty} p^{k+1}$$

and (8.10b) implying

$$\limsup_{k\to\infty} p^{k+1} \leq p^*.$$

As $\liminf_{k\to\infty} p^{k+1} \leq \limsup_{k\to\infty} p^{k+1}$ we have $\liminf_{k\to\infty} p^{k+1} = \limsup_{k\to\infty} p^{k+1} = p^*$ and so the limit exists. □

The main result that requires proving is the following that shows the "Potential Function" V^k is reduced by the scheme.

Theorem 8.0.4 *Suppose* f *and* g *are convex functions and we iterate (8.3a) to (8.3c). Define*

$$V^{k+1} = \left(\frac{1}{\rho}\right)\left\|\lambda^k - \lambda^*\right\|^2 + \rho\left\|B\left(z^k - z^*\right)\right\|^2.$$

Then

$$V^{k+1} \leq V^k - \rho\left\|r^{k+1}\right\|^2 - \rho\left\|B\left(z^k - z^*\right)\right\|^2. \tag{8.11}$$

Proof. See Appendix D. □

Corollary 8.0.5 *Suppose* f *and* g *are convex functions and we iterate (8.3a) to (8.3c). Then* $\left\{\lambda^k\right\}$ *and* $\left\{Bz^k\right\}$ *are bounded and* $B\left(z^{k+1} - z^k\right) \to 0$ *and* $r^{k+1} \to 0$ *implying* $\lim_{k\to\infty} p^{k+1} = p^*$.

Proof. The inequality (8.11) implies $V^{k+1} < V^k$ and in particular $V^k \leq V^0$ which implies both $\left\{\left\|\lambda^k - \lambda^*\right\|\right\}_{k=1}^{\infty}$ and $\left\{s^k = \left\|B\left(z^k - z^*\right)\right\|\right\}_{k=1}^{\infty}$ are bounded. Summing (8.11) and using a telescoping sequence we have

$$\rho\sum_{k=1}^{\infty}\left[\left\|r^{k+1}\right\|^2 + \left\|B\left(z^k - z^*\right)\right\|^2\right] \leq \sum_{k=1}^{\infty}\left[V^k - V^{k+1}\right] = V^0 < +\infty$$

and by the nth term test we have $\left\|r^{k+1}\right\| \to 0$ and $\left\|B\left(z^k - z^*\right)\right\| \to 0$. Now apply Corollary 8.0.3. □

Typically we stop when

$$\begin{aligned} &\left\|r^k\right\| \leq \sqrt{\dim c}\,\epsilon^{\text{abs}} + \epsilon^{\text{rel}}\max\left\{\left\|Ax^k\right\|, \left\|Bz^k\right\|, \|c\|\right\} \\ \text{and} \quad &\left\|s^k\right\| \leq \sqrt{\dim x}\,\epsilon^{\text{abs}} + \epsilon^{\text{rel}}\left\|A^T\lambda^k\right\|, \end{aligned}$$

where $\epsilon^{\text{rel}} = 10^{-3}$ or 10^{-4} and ϵ^{abs} is user chosen.

The Alternating Direction Multiplier Method (ADMM): We assume that $f : \mathbf{R}^n \to \mathbf{R}$ and $g : \mathbf{R}^m \to \mathbf{R}$ are convex.

Initialization:

1. The function $f : \mathbf{R}^n \to \mathbf{R}$ needs to be supplied.
2. Initial starting point x^0, initial $\rho^0 > 0$, values of τ^{inc}, $\tau^{\text{dec}} > 1$ (usually 2) and a value of $\mu (= 10$, say).
3. The tolerances $\epsilon^{\text{abs}}, \epsilon^{\text{rel}} > 0$.
4. Let $i = 0$.

While: x^i is unsatisfactory i.e. either $\left\|r^k\right\| > \sqrt{\dim c}\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\left\{\left\|Ax^k\right\|, \left\|Bz^k\right\|, \|c\|\right\}$ or $\left\|s^k\right\| > \sqrt{n}\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \left\|A^T\lambda^k\right\|$ **do**

Step 1: Let

$$\begin{aligned} x^{k+1} &\in \arg\min_x L_{\rho^k}\left(x, z^k, \lambda^k\right) \\ z^{k+1} &\in \arg\min_z L_{\rho^k}\left(x^{k+1}, z, \lambda^k\right) \\ \lambda^{k+1} &= \lambda^k + \rho^k\left(Ax^{k+1} + Bz^{k+1} - c\right). \end{aligned}$$

Step 2: Set

$$\rho^{k+1} = \begin{cases} \tau^{\text{inc}}\rho^k & \text{if } \left\|r^k\right\| > \mu\left\|s^k\right\| \\ \rho^k/\tau^{\text{dec}} & \text{if } \left\|s^k\right\| > \mu\left\|r^k\right\| \\ \rho^k & \text{otherwise} \end{cases}$$

End While

The changing ρ value is used to encourage $\left\|r^{k+1}\right\|$ and $\left\|s^k\right\|$ to converge to zero at the same rate.

8.1 ADMM and Constrained Convex Programming

One of the main ways that ADMM can be used is to solve the following general convex optimisation problem:

$$\begin{aligned} \min \quad & f(x) \\ \text{Subject to } & x \in C \end{aligned}$$

This can be written in the ADMM format by using the indicator function $\delta_C(z) := g(z)$ and reformulating as:

$$\begin{aligned} \min_{(x,z)} \quad & f(x) + \delta_C(z) \\ \text{Subject to } & x - z = 0. \end{aligned}$$

The augmented Lagrangian is

$$L_\rho(x, z, \lambda) = f(x) + \delta_C(z) + \lambda^T(x - z) + \frac{\rho}{2}\|x - z\|^2.$$

We note that when $f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$ and $C = C_i \times \cdots \times C_n$ for $x := (x_1, \ldots, x_n) \in X_1 \times \cdots \times X_n$ then the augmented Lagrangian takes separable form

$$\begin{aligned}
L_\rho(x, z, \lambda) &= \sum_{i=1}^{n} \left[f_i(x_i) + \delta_{C_i}(z_i) + \lambda_i^T (x_i - z_i) + \frac{\rho_i}{2} \|x_i - z_i\|^2 \right] \\
\text{(adding 0)} \ &= \sum_{i=1}^{n} \left[f_i(x_i) + \delta_{C_i}(z_i) + \frac{\rho_i}{2} \left[\left\| \frac{\lambda_i}{\rho_i} \right\|^2 + 2 \left(\frac{\lambda_i}{\rho_i} \right)^T (x_i - z_i) + \|x_i - z_i\|^2 \right] \right] - \frac{1}{2\rho_i} \|\lambda_i\|^2 \\
&= \sum_{i=1}^{n} f_i(x_i) + \delta_{C_i}(z_i) + \frac{\rho_i}{2} \left\| \frac{\lambda_i}{\rho_i} + (x_i - z_i) \right\|^2 - \frac{1}{2\rho_i} \|\lambda_i\|^2 \text{ (completing the square)} \\
&= \sum_{i=1}^{n} f_i(x_i) + \delta_{C_i}(z_i) + \frac{\rho_i}{2} \|\mu_i + \ (x_i - z_i)\|^2 + \text{constants}, \quad \text{where } \mu_i := \frac{\lambda_i}{\rho_i}.
\end{aligned}$$

With g given as the indicator function we have

$$\begin{aligned}
x_i^{k+1} &\in \text{argmin}_{x_i} \left[f_i(x_i) + \frac{\rho_i}{2} \left\| \mu_i^k - z_i^k + x_i \right\|^2 \right] \\
z_i^{k+1} &\in \text{argmin}_{z_i \in C_i} \left\| \left(x_i^{k+1} - z_i \right) + \mu_i^k \right\| \\
&= \text{argmin}_{z_i \in C_i} \left\| x_i^k + \mu_i^k - z_i \right\| \\
&= \text{P}_{C_i}(x_i^{k+1} + \mu_i^k) \qquad \text{(the projection onto } C_i\text{)} \\
\mu_i^{k+1} &= \mu_i^k + x_i^{k+1} - z_i^{k+1} \qquad \text{for all } i = 1, \ldots, n.
\end{aligned} \tag{8.12}$$

Notice that this μ_i^{k+1} update corresponds to the multiplier update

$$\lambda_i^{k+1} = \lambda_i^k + \rho_i (x_i^{k+1} - z_i^{k+1}).$$

In this case the primal and dual residuals take the form

$$r^k := x^k - z^k \qquad \text{and} \qquad s^k = -\rho_i \left(z^k - z^{k-1} \right),$$

since $A = I$, $B = -I$, and $c = 0$.

8.1.1 Linear and Quadratic Programming

Let's look at an example that falls under the umbrella we just described. In this case we have an objective

$$\begin{aligned}
\text{Min} \quad & \frac{1}{2} x^T P x + q^T x \\
\text{Subject to } Ax = b \quad & x \geq 0.
\end{aligned}$$

In this case g is the indicator function of the set $C := \{x \mid x \geq 0\}$ and $f(x) = \frac{1}{2} x^T P x + q^T x + \delta_{\{x \mid Ax = b\}}$. Then $\text{P}_{C_i}(x_i^{k+1} + \mu_i^k) = (x_i^{k+1} + \mu_i^k)_+ := \max \left\{ (x_i^{k+1} + \mu_i^k), 0 \right\}$. Now ADMM consists of the iteration

$$\begin{aligned}
x^{k+1} &\in \text{argmin}_{x \in \{x \mid Ax = b\}} \left[f(x) + \frac{\rho}{2} \left\| \mu^k - z^k + x \right\|^2 \right] \text{ (as above)} \\
z^{k+1} &= \max \left\{ x^{k+1} + \mu^k, 0 \right\}
\end{aligned} \tag{8.13}$$

$$\mu^{k+1} = \mu^k + x^{k+1} - z^{k+1}.$$

That is the projection of z onto $C := \{x \mid x \geq 0\}$ is given by $\mathrm{P}_C(z) = \max\{z, 0\}$ where the maximum is performed component wise. To solve (8.13) we need to investigate the associated (KKT) condition to find the saddle point (x^{k+1}, ν) of the Lagrangian for the optimisation subproblem (8.13) (which is necessary and sufficient for a convex problem). The augmented Lagrangian for this sub problem is

$$\begin{aligned} L_\rho(x, \nu) &= \frac{1}{2}x^T P x + q^T x + \nu^T (Ax - b) + \frac{\rho}{2}\|\mu - z + x\|^2 \\ &= \frac{1}{2}x^T P x + x^T q + x^T A^T \nu - \nu^T b + \frac{\rho}{2}\|\mu - z\|^2 + \rho x^T(\mu - z) + \frac{\rho}{2}x^T x \\ &= \frac{1}{2}x^T (P - \rho I) x + x^T A^T \nu + x^T q - \nu^T b + \rho x^T(\mu - z) + \frac{\rho}{2}\|\mu - z\|^2 \end{aligned}$$

and so the KKT conditions say that (x^{k+1}, ν) is a saddle point if and only if

$$\begin{aligned} 0 &= \nabla_x L\left(x^{k+1}, \nu\right) = [P + \rho I] x^{k+1} + A^T \nu + q + \rho\left(\mu^k - z^k\right) \\ 0 &= \nabla_\nu L\left(x^{k+1}, \nu\right) = Ax^{k+1} - b, \end{aligned}$$

which corresponds to the system

$$\begin{bmatrix} P + \rho I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^{k+1} \\ \nu \end{bmatrix} = \begin{bmatrix} \rho\left(z^k - \mu^k\right) - q \\ b \end{bmatrix}.$$

Therefore, finding my update x^{k+1} amounts to solving the above linear system to find the saddle point (x^{k+1}, ν^{k+1}) of the Lagrangian for the optimisation problem (8.13).

The Alternating Direction Multiplier Method (ADMM) QP solver:

Initialization:

1. The initial P, q, A, b need to be supplied.
2. Initial starting point $\mathbf{x}^0$, initial $\rho^0 > 0$, values of τ^{inc}, $\tau^{\mathrm{dec}} > 1$ (usually 2) and a value of $\kappa (= 10$, say).
3. The tolerances $\epsilon^{\mathrm{abs}}, \epsilon^{\mathrm{rel}} > 0$.
4. Let $i = 0$.

While: $\mathbf{x}^i$ is unsatisfactory i.e. either $\left\|r^k\right\| > \sqrt{p}\epsilon^{\mathrm{abs}} + \epsilon^{\mathrm{rel}} \max\left\{\left\|Ax^k\right\|, \|b\|\right\}$ or $\left\|s^k\right\| > \sqrt{n}\epsilon^{\mathrm{abs}} + \epsilon^{\mathrm{rel}}\left\|A^T\nu^k\right\|$ **do**

Step 1: Solve

$$\begin{bmatrix} P + \rho I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^{k+1} \\ \nu^{k+1} \end{bmatrix} = \begin{bmatrix} \rho\left(z^k - \mu^k\right) - q \\ b \end{bmatrix}$$

$$z^{k+1} = \max\left\{x^{k+1} + \mu^k, 0\right\}$$

$$\mu^{k+1} = \mu^k + x^{k+1} - z^{k+1}.$$

Step 2: Set $r^k := x^k - z^k$ and $s^k = -\rho\left(z^k - z^{k-1}\right)$

$$\rho^{k+1} = \begin{cases} \tau^{\mathrm{inc}}\rho^k & \text{if } \left\|r^k\right\| > \kappa\left\|s^k\right\| \\ \rho^k / \tau^{\mathrm{dec}} & \text{if } \left\|s^k\right\| > \kappa\left\|r^k\right\| \\ \rho^k & \text{otherwise} \end{cases}$$

End While

8.1.2 Alternating Projections

Sometimes it is hard to obtain even a feasible point. This is when it is useful to consider iterated projections. Let us now suppose $f(x) = \delta_D(x)$ and $g(z) = \delta_C(z)$. Then our problem becomes:

$$\begin{aligned} \min_{(x,z)} \quad & \delta_D(x) + \delta_C(z) \\ \text{Subject to } & x - z = 0. \end{aligned}$$

That is we want to find $x \in D \cap C = [D_1 \times \cdots \times D_n] \cap [C_1 \times \cdots \times C_n]$. Then ADMM takes the form

$$\begin{aligned} x_i^{k+1} &\in \mathrm{P}_{D_i}(z_i^k - \mu_i^k), \quad \text{for } i = 1, \ldots, n \\ z_i^{k+1} &\in \mathrm{P}_{C_i}(x_i^{k+1} + \mu_i^k), \quad \text{for } i = 1, \ldots, n \\ \mu_i^{k+1} &= \mu_i^k + x_i^{k+1} - z_i^{k+1} \qquad \text{for all } i = 1, \ldots, n. \end{aligned}$$

This is the classic form of the Dykstra's alternating projection method for finding a feasible point. Here, the norm of the primal residual $\left\|x^k - z^k\right\|$ has a nice interpretation. Since $x^k \in C$ and $z^k \in D$, $\left\|x^k - z^k\right\|$is an upper bound on $dist(C, D)$, the Euclidean distance between C and D. If we terminate with $\left\|r^k\right\| \leq \varepsilon$, then we have found a pair of points, one in C and one in D, that are no more than ε far apart. Alternatively, the point $\frac{1}{2}\left(x^k + z^k\right)$ is no more than $\frac{\varepsilon}{2}$ from both C and D.

8.1.3 Parallel Projections

Lets now consider finding a feasible point to an intersection of N sets i.e.

$$x \in A_1 \cap A_2 \cap \cdots \cap A_N.$$

To do this we convert it to the problem of finding a feasible point to the intersection of two sets:

$$\mathcal{C} := A_1 \times A_2 \times \cdots \times A_N \quad \text{and} \quad \mathcal{D} := \{(x_1 \ldots, x_N) \mid x_1 = x_2 = \cdots = x_N\}.$$

The projection onto the second set $\mathcal{D}$ is particularly simple $\mathrm{P}_{\mathcal{D}}(z) = (\bar{z}, \bar{z}, \ldots, \bar{z})$ where $\bar{z} = \frac{1}{N}\sum_{i=1}^N z_i$. To see this consider the projection onto $\mathcal{D}$ as defined by the minimum (squared) norm problem

$$\min_{\mathbf{z}} \frac{1}{2}\sum_{i=1}^N \|z_i - z\|^2.$$

Then the minimum z is defined by

$$\begin{aligned} 0 &= \nabla_z \left[\frac{1}{2}\sum_{i=1}^N \|z_i - z\|^2\right] \\ &= \sum_{i=1}^N (z - z_i) = Nz - \sum_{i=1}^N z_i \\ \text{implying} \quad z &= \frac{1}{N}\sum_{i=1}^N z_i. \end{aligned}$$

This ADMM now looks like

$$\begin{aligned} x_i^{k+1} &\in P_{A_i}(z_i^k - \mu_i^k), \quad \text{for } i = 1, \dots, N \\ z_i^{k+1} &= \frac{1}{N} \sum_{j=1}^{N} \left(x_j^{k+1} + \mu_j^k \right), \quad \text{for } i = 1, \dots, N \\ \mu_i^{k+1} &= \mu_i^k + x_i^{k+1} - z_i^{k+1} \qquad \text{for all } i = 1, \dots, N. \end{aligned}$$

Note that the first and third steps can be carried out in parallel and so are very efficient. We can even simplify this more. Taking the average in the last equation, we get

$$\underbrace{\frac{1}{n} \sum_i \mu_i^{k+1}}_{=:\bar{\mu}^{k+1}} = \underbrace{\frac{1}{n} \sum_i \mu_i^k}_{=:\bar{\mu}^k} + \underbrace{\frac{1}{n} \sum_i x_i^{k+1}}_{=:\bar{x}^{k+1}} - \underbrace{\frac{1}{n} \sum_i z_i^{k+1}}_{=:\bar{z}^{k+1}}$$

and as the second equation is $z_i^{k+1} = \bar{x}^{k+1} + \bar{\mu}^k$ we get $\bar{\mu}^{k+1} = \left(z_i^{k+1} \right) - \bar{z}^{k+1} = 0$ (n.b. $z_i^{k+1} = \bar{z}^{k+1}$ since $z^{k+1} \in D$). Hence $z^{k+1} = (\bar{x}^{k+1}, \dots, \bar{x}^{k+1})$ and so we have

$$\begin{aligned} x_i^{k+1} &\in \mathrm{P}_{A_i}(\bar{x}^k - \mu_i^k), \\ \mu_i^{k+1} &= \mu_i^k + \left(x_i^{k+1} - \bar{x}^{k+1} \right) \qquad \text{for all } i = 1, \dots, n. \end{aligned}$$

Hence assuming $\mu^0 = 0$ then μ^k is the running sum of discrepancies $x_i^k - \bar{x}^k$.

8.1.4 The Least Absolute Deviation Problem

Suppose we wish to solve the least absolute deviation problem:

$$\begin{aligned} \min_{(x,z)} \quad & \|z\|_1 := \sum_{i=1}^{m} |z_i| \\ \text{Subject to } & Ax - z = b. \end{aligned}$$

In the case where b is in the image of A, such as when A has full rank, then the problem is trivial with $(x^*, z^*) = (x, 0)$ an optimal solution for any x that satisfies $Ax = b$. Otherwise, the problem amounts to finding the best "solution" (as measured by the one-norm) to the inconsistent system $\min_x \|Ax - b\|_1$. That is, we mimimise the sum of absolute residuals.

Soln: We use the augmented Lagrangian with $f = 0$ and $g = \|\cdot\|_1$ i.e.

$$L_\rho(x, z, \lambda) := \sum_{i=1}^{m} |z_i| + \lambda^T (Ax - z - b) + \frac{\rho}{2} \|Ax - z - b\|^2$$

and so we need to first solve $0 = \nabla_x L_\rho \left(\cdot, z^k, \lambda^k \right) \left(x^{k+1} \right)$ which is

$$\begin{aligned} 0 &= A^T \lambda^k + \rho A^T \left(Ax^{k+1} - z^k - b \right) \\ \text{or} \quad 0 &= A^T \mu^k + A^T \left(Ax^{k+1} - z^k - b \right) \quad \text{where } \mu^k := \frac{\lambda^k}{\rho} \\ \text{so} \quad \left(A^T A \right) x^{k+1} &= A^T \left(b + z^k - \mu^k \right). \end{aligned}$$

We then need to solve

$$\begin{aligned} & 0 \in \partial L_\rho\left(x^{k+1}, \cdot, \lambda^k\right)\left(z^{k+1}\right) \\ & = \partial\left[\sum_{i=1}^m |\cdot|\right]\left(z^{k+1}\right) - \lambda^k - \rho\left(Ax^{k+1} - z^{k+1} - b\right) \\ & = \times_{i=1}^m \left[\partial|\cdot|\left(z_i^{k+1}\right) + \rho z_i^{k+1}\right] - \lambda^k - \rho\left(Ax^{k+1} - b\right) \end{aligned} \tag{8.14}$$

$$\text{or} \quad 0 \in \times_{i=1}^m \left[\frac{1}{\rho}\partial|\cdot|\left(z_i^{k+1}\right) + z_i^{k+1}\right] - \left(Ax^{k+1} - b + \mu^k\right) \quad \text{where } \mu^k := \frac{\lambda^k}{\rho}$$

$$\text{or} \quad 0 \in \frac{1}{\rho}\partial|\cdot|\left(z_i^{k+1}\right) + z_i^{k+1} - \left(Ax^{k+1} - b + \mu^k\right)_i, \quad i = 1, \ldots, m. \tag{8.15}$$

To solve such problem we need to find solution to the nonsmooth (single variable) equation

$$\frac{1}{\rho}\partial|\cdot|(z) + z \ni v.$$

Now we now that

$$\partial|\cdot|(z) = \begin{cases} 1 & z > 0 \\ [-1, 1] & z = 0 \\ -1 & z < 0 \end{cases}$$

Hence

$$\frac{1}{\rho}\partial|\cdot|(z) + z = \begin{cases} \frac{1}{\rho} + z & z > 0 \\ [-\frac{1}{\rho}, \frac{1}{\rho}] & z = 0 \\ -\left(\frac{1}{\rho} - z\right) & z < 0 \end{cases}$$

So if $v \in [-\frac{1}{\rho}, \frac{1}{\rho}]$ we have $z = 0$. That is $z = 0$ if $|v| \leq \frac{1}{\rho}$. Otherwise $v = \frac{1}{\rho} + z$ (when $z > 0$) and so $z = v - \frac{1}{\rho} \geq 0$ when $v > \frac{1}{\rho}$. Similarly when $v < -\frac{1}{\rho}$ we have $z = v + \frac{1}{\rho}$. Hence

$$z = \begin{cases} v - \frac{1}{\rho} & v > \frac{1}{\rho} \\ 0 & |v| \leq \frac{1}{\rho} \\ \frac{1}{\rho} + v & v < -\frac{1}{\rho} \end{cases} := S_{\frac{1}{\rho}}(v). \tag{8.16}$$

This is call the soft thresholding function, or "shrinkage." Now to solve

$$(v_1, \ldots, v_m) \in \times_{i=1}^m \left[\frac{1}{\rho}\partial|\cdot|(z_i) + z_i\right]$$

we apply the soft thresholding function component wise i.e

$$z = (z_1, \ldots, z_m) = S_{\frac{1}{\rho}}(v) := S_{\frac{1}{\rho}}(v_1) \times \cdots \times S_{\frac{1}{\rho}}(v_m).$$

Hence the solution to (8.15) is given by

$$z^{k+1} = S_{\frac{1}{\rho}}\left(Ax^{k+1} - b + \mu^k\right)$$

and the full iteration of ADMM in this case is

$$\begin{aligned} x^{k+1} &= \text{LinearSolve}\left[\left(A^T A\right)x = A^T\left(b + z^k - \mu^k\right)\right] \text{ for } x. \\ z^{k+1} &= S_{\frac{1}{\rho}}\left(Ax^{k+1} - b + \mu^k\right) \\ \mu^{k+1} &= \mu^k + \left(Ax^{k+1} - z^{k+1} - b\right) \quad \text{where } \mu^k := \frac{\lambda^k}{\rho}. \end{aligned}$$

Problem Set 16

Write down the ADMM iteration to solve the following problems:

Problem 8.1.1 *The basis pursuit problem (used to find a sparse solution to an under determined system of linear equations):*

$$\min_{x} \quad \|x\|_1$$
$$\textit{Subj to } Ax = b.$$

[Hint: Reformulate as $\min_{(\mathbf{x},\mathbf{z})} \{f(x) + \|z\|_1 \mid x - z = 0\}$*, where* $f(x)$ *is the indicator function of the set* $C := \{x \mid Ax - b = 0\}$*.]*

Problem 8.1.2 *The Lasso problem:*

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|Ax - b\|^2 + \gamma \|x\|_1,$$

where $\|Ax - b\|^2$ *is the squared Euclidean norm.*
Hint: Reformulate as: $\min_{(x,z)} \left\{ \frac{1}{2}\|Ax - b\|^2 + \gamma \|z\|_1 \mid x - z = 0 \right\}$.

8.2 A few applications in Statistical Learning Theory

The problems addressed in this section will help illustrate why ADMM is a natural fit for machine learning and statistical problems in particular. The reason is that, unlike dual ascent or the method of multipliers, ADMM explicitly targets problems that split into two distinct parts, f and g, that can then be handled separately. Problems of this form are pervasive in machine learning, because a significant number of learning problems involve minimizing a loss function together with a regularization term or side constraints.
The regularized lasso problem:

$$\min_{x} \quad l(x) + \gamma \|x\|_1,$$

where l is any convex loss function. The ADMM form of this problem is reformulated as:

$$\min_{(x,z)} \{l(x) + g(z) \mid x - z = 0\},$$

where $g(z) = \gamma \|z\|_1$. Applying ADMM to this we need the augmented Lagrangian

$$L_\rho(x, z, \lambda) := l(x) + \gamma \sum_{i=1}^{m} |z_i| + \lambda^T (x - z) + \frac{\rho}{2}\|x - z\|^2.$$

So the first ADMM step (by completing the square) takes the form of solving

$$x^{k+1} \in \arg\min \left(l\left(x^k\right) + \frac{\rho}{2}\left\| x^k - z^k + \frac{\lambda^k}{\rho} \right\|^2 - \frac{1}{2\rho}\left\|\lambda^k\right\|^2 \right)$$
$$\text{or } x^{k+1} \in \arg\min \left(l\left(x^k\right) + \frac{\rho}{2}\left\| x^k - z^k + u^k \right\|^2 \right) \quad \text{where } u^k := \frac{\lambda^k}{\rho}.$$

The second step takes the form

$$\begin{aligned} 0 &\in \partial\left[\gamma\sum_{i=1}^{m}|\cdot|\right]\left(z^{k+1}\right)-\lambda^k-\rho\left(x^{k+1}-z^{k+1}\right) \\ \text{or } x^{k+1}-u^k &\in \left[\frac{\gamma}{\rho}\partial\left[\sum_{i=1}^{m}|\cdot|\right]\left(z^{k+1}\right)+z^{k+1}\right] \\ &= \times_{i=1}^{m}\left[\frac{\gamma}{\rho}\partial|\cdot|\left(z_i^{k+1}\right)+z_i^{k+1}\right] \\ \text{so}\quad z^{k+1} &= S_{\gamma/\rho}\left(x^{k+1}-u^k\right), \end{aligned}$$

where $S_{\gamma/\rho}$ is the component-wise soft-threshold from (8.16) (a.k.a. the shrinkage). Putting this all together we get $u^k := \frac{\lambda^k}{\rho}$ with

$$\begin{aligned} x^{k+1} &\in \arg\min\left(l\left(x^k\right)+\frac{\rho}{2}\left\|x^k-z^k+u^k\right\|^2\right) \\ z^{k+1} &= S_{\gamma/\rho}\left(x^{k+1}-u^k\right) \qquad \text{and} \\ u^{k+1} &= u^k+x^{k+1}-z^{k+1}. \end{aligned} \tag{l_1-loss}$$

The type of algorithm used to solve the first minimisation will depend on the type of function used for l. If it is differentiable, we can use a fast descent method like the one discussed in the previous sections. In general, we can interpret ADMM for regularized l_1 loss minimization as solving a sequence of l_2 (squared) regularized loss minimization problems.

Logistic Regression

Where the input is a hypothetical probability of an event (p_E), the **likelihood** function for an outcome tells us how likely we would be to observe that outcome. For example, suppose we have a two-sided coin, but do not know how fair it is. Then the probability of observing heads on a single flip is $p_H \in [0,1]$, where p_H would be 0.5 for a fair coin. Here are some likelihood function examples from the Wikipedia page[1]:

- We flip a coin once and observe the outcome heads, or $X = H$. The likelihood function for this observation is the identity map, because it takes as its input the probability of flipping heads once, and returns the likelihood that we would observe a flip of heads under this probability. In other words: $\theta \mapsto L(p_H = \theta|X = H) = \theta$.
- We flip a coin twice and observe heads twice, or $X = HH$. The likelihood function for this observation is $\theta \mapsto L(p_H = \theta|X = HH) = \theta^2$.
- We flip a coin three times, and observe $X = HHT$. The likelihood function is $\theta \mapsto L(p_H = \theta|x = HHT) = \theta^2(1-\theta)$.

The likelihood should not be confused with the associated probability density function $p_\theta(\cdot)$ that, given a fixed $P_H := \theta$, maps an observed outcome to probability of observing it: e.g. $HHT \mapsto P(X = HHT|p_H = \theta) = \theta^2(1-\theta)$. The two different functions share the property that

$$p_\theta(x) := P(X = x|p_H = \theta) = L(p_H = \theta|X = x) =: L(\theta, x).$$

[1] `https://en.wikipedia.org/wiki/Likelihood_function#Discrete_probability_distribution`

Now suppose we have made our observations $X = x$, and we want to build a model that has the best chance of describing reality as possible. Well, then we would want to find the parameter θ that maximizes the likelihood function $\theta \mapsto L(\theta, x)$. In other words, we seek the value of θ that maximizes the likelihood of us having observed what we have observed. Noting that the logarithm is strictly increasing, maximizing $L(\theta, x)$ is equivalent to maximizing $\log(L(\theta, x))$.

Maximizing the log-likelihood

The following discussion is based on [14, Section 4.4.4]. The logistic regression model arises from the need to model posterior probabilities
$P(G = 1, X = x), \ldots, P(G = K, X = x)$ of the K classes via a linear function of x, while at the same time ensuring that they sum to one and remain in $[0, 1]$. The general model has the form

$$\begin{aligned} \log \frac{P\left(G = 1|X = x\right)}{P\left(G = K|X = x\right)} &= \beta_{1,0} + \beta_1^T x \\ \log \frac{P\left(G = k|X = x\right)}{P\left(G = K|X = x\right)} &= \beta_{k,0} + \beta_k^T x \\ &\vdots \\ \log \frac{P\left(G = K-1|X = x\right)}{P\left(G = K|X = x\right)} &= \beta_{k-1,0} + \beta_{k-1}^T x \end{aligned} \tag{8.17}$$

The model is specified in terms of $K - 1$ log-odds. Taking the exponential and summing we get

$$\sum_{l=1}^{K-1} P\left(G = l|X = x\right) = \left[\sum_{l=1}^{K-1} \exp\left(\beta_{l,0} + \beta_l^T x\right)\right] P\left(G = K|X = x\right) \text{ or}$$

$$\begin{aligned} P\left(G = K|X = x\right) + \sum_{l=1}^{K-1} P\left(G = l|X = x\right) &= \left[\sum_{l=1}^{K-1} \exp\left(\beta_{l,0} + \beta_l^T x\right)\right] P\left(G = K|X = x\right) \\ &\quad + P\left(G = K|X = x\right) \\ \text{which is just} \qquad 1 &= \left[\sum_{l=1}^{K-1} \exp\left(\beta_{l,0} + \beta_l^T x\right) + 1\right] P\left(G = k|X = x\right) \\ \text{so} \qquad P\left(G = K|X = x\right) &= \frac{1}{\sum_{l=1}^{K-1} \exp\left(\beta_{l0} + \beta_l^T x\right) + 1}. \end{aligned}$$

Combining this last identity with (8.17), it then follows that

$$P\left(G = k|X = x\right) = \frac{\exp\left(\beta_{k,0} + \beta_k^T x\right)}{\sum_{l=1}^{K-1} \exp\left(\beta_{l,0} + \beta_l^T x\right) + 1}, \qquad \text{for } k = 1, 2, \ldots, K-1.$$

By the slight change of notation $\beta_k := \left(\beta_{k,0}, \beta_k\right)^T$ and $x_0 := 1$ so that $x := \left(1, x_1, \ldots\right)^T$ we can write

$$\exp(\beta_{l,0} + \beta_l^T x) = \exp(\beta_l^T x).$$

We can define the notation

$$p_k(x, \theta) := P\left(G = k|X = x; \theta\right) = \begin{cases} \frac{\exp\left(\beta_l^T x\right)}{1 + \sum_{l=1}^{K-1} \exp\left(\beta_l^T x\right)} & k = 1, \ldots, K-1 \\ \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_l^T x)} & k = K \end{cases}$$

to emphasize the dependence on the set of parameters of the model, which we denote by $\theta := \{\beta_{1,0}, \beta_1, \ldots, \beta_{K-1,0}, \beta_{K-1}\}$. In other words, $P(G = k, X = x, \theta)$ tells us the likelihood that our sample $X = x$ would belong to the observed category $G = k$ under the added assumption that θ holds. Given a the N observations $\{(g_i, x_i)\}_{i=1}^N$, we want to find the parameter θ that maximizes the likelihood that we would have made those observations together. That likelihood is the product of all the likelihoods of each individual observation:

$$\prod_{i=1}^{N} P(G = g_i | X = x_i; \theta).$$

To maximize this likelihood function is equivalent to maximizing the log of it (the "log-likelihood"), which is given by:

$$l(\theta) = \sum_{i=1}^{N} \log P(G = g_i | X = x_i; \theta)$$

When $K = 2$ the model simplifies to a single linear function. These models are binary (two classes) which frequently occur: e.g. a patient survives or dies, or a condition is present or not. In this case we code two classes: $y_i = 1$ if $g_i = 1$ (class one) or $y_i = 0$ when $g_i = 2$ (class two). Consequently, we have

$$\begin{aligned} P(G = g_i | X = x_i, \theta) &= \begin{cases} \frac{\exp\left(\beta_l^T x\right)}{1+\sum_{l=1}^{K-1} \exp\left(\beta_l^T x\right)} & \text{if } g_i = 1 \text{ (wherefore } y_i = 1) \\ \frac{1}{1+\sum_{l=1}^{K-1} \exp(\beta_l^T x)} & \text{if } g_i = 2 \text{ (wherefore } y_i = 0) \end{cases} \\ &= \begin{cases} \frac{\exp\left(y_i \beta_l^T x\right)}{1+\sum_{l=1}^{K-1} \exp\left(\beta_l^T x\right)} & \text{if } g_i = 1 \\ \frac{\exp\left(y_i \beta_l^T x\right)}{1+\sum_{l=1}^{K-1} \exp(\beta_l^T x)} & \text{if } g_i = 2 \end{cases} \end{aligned}$$

and so $\log(P(G = g_i | X = x_i, \theta)) = y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right)$.

Therefore, the log-likelihood function simplifies to

$$l(\beta) = \sum_{i=1}^{N} \left(y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right)\right)$$

The l_1 regularised logistic regression problem is to maximise the log-likelihood subject to a sparsity condition:

$$\max_{\beta} \left\{ \sum_{i=1}^{N} \left\{ y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right) \right\} - \gamma \|\beta_1\|_1 \right\}$$

or converting to a convex problem:

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left\{ \log\left(1 + e^{\beta_0 + \beta_1^T x}\right) - y_i \left(\beta_0 + \beta_1^T x\right) \right\} + \gamma \|\beta_1\|_1 \right\}.$$

We don't penalise the intersect variable β_0, as β_0 as the intersect variables ensure the expected values of class 1 match those observed. The objective $f(x) = \sum_{i=1}^{N} \left\{ \log\left(1 + e^{\beta^T x_i}\right) - y_i \beta^T x_i \right\}$ is smooth, so the corresponding primal step in the ADMM can be solved using the fast gradient descent method.

Sparse Inverse Covariance Selection

In multivariate data analysis, graphical models such as Gaussian Markov Random Fields provide a way to discover meaningful interactions among variables [22]. Let $Y = \left\{y^1, \ldots, y^n\right\}$ be an n-dimensional random vector following a n-variate Gaussian distribution $N(\mu, \Sigma)$. and let $G = (V, E)$ be a graph representing the conditional dependence of the set of variables Y. That is E contains an edge (i, j) if and only if y^i is conditionally dependent on y^j, given all the remaining variables (two events A and B, are conditionally independent given C if and only if $P(A|B, C) = P(A|C)$). The lack of an edge in $(i, j) \in E$ indicates conditional independence of y^i and y^j, which corresponds to the inverse covariance matrix Σ^{-1} having a zero as its (i, j)th component. *So* we desire to estimate the sparse inverse covariance matrix. That is, we want to learn the topology of the undirected graph G. This is called the covariant selection problem. Given a sample $\left\{Y^i\right\}_{i=1}^N$ we have a sample mean $\bar{y} := \frac{1}{N}\Sigma_{i=1}^N Y^i$ and sample covariance matrix $V := \frac{1}{N}\Sigma_{i=1}^N \left(Y^i - \bar{y}\right)\left(Y^i - \bar{y}\right)^T$. The problem we want to solve is to maximize the log likelihood $\log(\det X) - \langle V, X\rangle$ where $\langle V, X\rangle = \operatorname{tr} VX$ is the Frobenious inner product, subject to a sparsity constraint. A good heuristic for this is given by the problem

$$\min_{X \in \mathcal{P}_+(n)} \quad -\log(\det X) + \langle V, X\rangle + \gamma \|X\|_1$$

where $\|X\|_1 = \sum_{i,j=1}^n |X_{ij}|$ (the component wise sum of absolute values). The $-\log(\det X)$ term makes sure we get a solution $X \succ 0$. This is a special case of the l_1 regularised loss functions problem with loss

$$l(X) = \langle V, X\rangle - \log(\det X).$$

As usual we write this as

$$\begin{aligned} \min_{X, Z \in \mathcal{P}_+(n)} \quad & l(X) + \gamma \|Z\|_1 \\ \text{Subj to } & X - Z = 0. \end{aligned}$$

Using ADMM in the form of (l_1-loss) for this function l, we get

$$\begin{aligned} X^{k+1} &= \arg\min \left(\langle V, X^k\rangle - \log\left(\det X^k\right) + \frac{\rho}{2}\left\|X^k - Z^k + U^k\right\|_F^2 \right) \\ Z^{k+1} &= \arg\min \left(\gamma \|Z\|_1 + \frac{\rho}{2}\left\|X^{k+1} - Z^k + U^k\right\|_F^2 \right) \quad \text{which is just} \\ Z_{ij}^{k+1} &= S_{\gamma/\rho}\left(X_{ij}^{k+1} + U_{ij}^k\right) \quad \text{for all } i, j = 1, \ldots, n \\ U^{k+1} &= U^k + X^{k+1} - Z^{k+1}. \end{aligned}$$

The primal X iterate turns out to have an analytical solution. As the objective is differentiable we have the first order optimality condition given by:

$$\begin{aligned} 0 &= \nabla_X \left(\langle V, X\rangle - \log(\det X) + \frac{\rho}{2}\left\|X - Z^k + U^k\right\|_F^2 \right) \\ &= V - X^{-1} + \rho\left(X - Z^k + U^k\right) \qquad \text{and } X \succ 0. \end{aligned} \tag{8.18}$$

We will construct a solution that will satisfy this necessary and sufficient optimality condition. First rewrite the equations as

$$\rho X - X^{-1} = \rho\left(Z^k - U^k\right) - V. \tag{8.19}$$

First notice that the right hand side of (8.19) lives in $S(n)$ (the covariance matrix V is always in $P(n)$) and contains data that is available at iteration k, so we can find an orthogonal eigenvalue decomposition, using the eigenvectors to construct $Q \in \mathcal{S}(n)$ for which $Q^TQ = QQ^T = I$,

$$\rho\left(Z^k - U^k\right) - V = Q\Lambda Q^T$$

where $\Lambda = (\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues. Multiplying (8.19) on the left by Q^T and on the right by Q we get

$$\rho\left(Q^TXQ\right) - Q^TX^{-1}Q = \Lambda. \tag{8.20}$$

Defining $\hat{X} = Q^TXQ$, we have

$$\hat{X}^{-1} = \left[Q^TXQ\right]^{-1} = Q^{-1}X^{-1}\left(Q^T\right)^{-1} = Q^TX^{-1}Q.$$

Hence (8.20) becomes

$$\rho\hat{X} - \hat{X}^{-1} = \Lambda.$$

Now we construct a solution using the diagonal equation $\rho\hat{X}_{ii} - \hat{X}_{ii}^{-1} = \lambda_i$. Multiplying by $\hat{X}_{ii}$ yields the quadratic $\rho\hat{X}_{ii}^2 - \lambda_i\hat{X}_{ii} - 1 = 0$. This quadratic always has one positive solution and one negative solution. Given that $X \in P(n)$, and that $X = Q\hat{X}Q^T$, we must have that $\hat{X}$ is the diagonal[2] consisting of the eigenvalues of X, and so its diagonal entries are all positive:

$$\hat{X}_{ii} = \frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho}}{2\rho} > 0, \quad \text{as } \rho > 0.$$

Now define

$$X = Q\hat{X}Q^T.$$

It is easily verified that X satisfies (8.19) and so is the unique solution of the problem. So we can perform the first step at the cost of one eigenvalue decomposition of a symmetric matrix. In summary: To solve

$$X^{k+1} = \arg\min\left(\langle V, X^k\rangle - \log\left(\det X^k\right) + \frac{\rho}{2}\left\|X^k - Z^k + U^k\right\|_F^2\right)$$

we find an orthogonal eigenvalue decomposition, using the eigenvectors to construct $Q^k \in \mathcal{S}(n)$ for which $(Q^k)^TQ^k = Q^k(Q^k)^T = I$, and

$$\rho\left(Z^k - U^k\right) - V = Q^k\Lambda^k\left(Q^k\right)^T.$$

With

$$\hat{X}_{ii} = \frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho}}{2\rho}$$

$$\text{let} \qquad X^{k+1} = Q^k \operatorname{diag}\left\{\hat{X}_{ii}\right\}\left(Q^k\right)^T.$$

When we have solved the problem, the primal optimal solution X^* approximates Σ^{-1} and we can look for negligible\zero entries.

[2]Because Q diagonalizes $\rho X - X^{-1}$, we know Q consists of the eigenvectors of $X - X^{-1}$. However, X and X^{-1} each possess the same eigenvectors as each other, and so clearly those eigenvectors are also the eigenvectors of $\rho\hat{X} - \hat{X}^{-1}$. Thus Q consists of the eigenvectors of $\hat{X}$, which are also the eigenvectors of $\hat{X}^{-1}$, whereupon Q diagonalizes both of them.

Chapter 9

Stochastic Optimisation and Consensus Problems

In this section we look at the class of problems that fall into category of seeking consensus. We have variable that we wish to agree in value. We will look at a simple two stage stochastic programming problem, which has the structure of a decision problem under uncertainty. The decision maker has to choose a set of decision variables before knowing the outcome of a scenario that will impact the objective value.

9.1 An Example

Example 9.1.1 *Consider a European farmer that has to decide how much to plant for three crops, wheat, corn and sugar beet. Let us begin by assuming that the yields per acre are known before hand. We will then consider how the problem changes when yield depend on the rain falls that are only known with to have a certain probability distribution. The Farmer has access to the following data:*

	Wheat	*Corn*	*Sugar Beet*
Yield (/acre)	*2.5*	*3*	*20*
Planting cost (/acre)	*$150*	*$230*	*$260*
Selling price (/T)	*$170*	*$150*	*$36 under 6000T* *$10 above 6000T*
Purchase Price (/T)	*$238*	*$210*	
Minimum Requirement (/T)	*200*	*240*	

The farmer has 500 acres of land to plant. The European commission imposes a quota of 6000 tons on sugar beets. The farmer can sell the sugar beets at $36 per ton within quota and only $10 per ton for that sold above the quota. The farmer requires at least 200 tons of wheat and 240 tons of corn to feed his farm animals.

The farmer has the following decision variables:

x_1 = *The acres devoted to wheat*
x_2 = *The acres devoted to corn*
x_3 = *The acres devoted to sugar beet*
y_1 = *The tons of wheat purchased*
w_1 = *The tons of wheat sold*
y_2 = *The tons of corn purchased*
w_2 = *The tons of corn sold*
w_3 = *The tons of sugar beet sold at the favourable price*

w_4 = *The tons of sugar beet sold at the unfavourable price*

We can formulate this problems as a linear programming problem which seeks to minimise costs (or maximize profit if one reverses the sign).

$$\min_{(x,y,w)} \quad 150x_1 + 230x_2 + 260x_3 + 238y_1 - 170w_1 + 210y_2 - 150w_2 - 36w_3 - 10w_4$$

Subject to

$$\begin{aligned}
&x_1 + x_2 + x_3 \leq 500\\
&2.5x_1 + y_1 - w_1 \geq 200\\
&3x_2 + y_2 - w_2 \geq 240\\
&w_3 + w_4 \leq 20x_3\\
&w_3 \leq 6000\\
&x_1, x_2, x_3, y_1, y_2, w_1, w_2, w_3, w_4 \geq 0.
\end{aligned}$$

Solving this yields the following solution at a profit of $118,600

Culture	*Wheat*	*Corn*	*Sugar beet*
Surface area (Acres)	*120*	*80*	*300*
Yields (T)	*300*	*240*	*6000*
Sales (T)	*100*	*0*	*6000*
Purchased (T)	*0*	*0*	*0*

We now consider the problem when we have the possibility of a above average, on average of below average season. These are deemed to occur with equal probability of $\frac{1}{3}$ *each. The yields will be above or below average yield by* 20% *in a good and a bad season. The decision of what to plant* (x_1, x_2, x_3) *must now be made before we know what type of season we have. To hedge against the season's outcome we will try and minimise the expected loss. The sales and purchases now depend on the yields which depend on the seasons. We have three scenarios* $s = 1, 2, 3$ *which correspond to the good, average and bad seasons. The decision variables are now* $y_{i,s}$ *for* $i = 1, 2$ *and* $s = 1, 2, 3$ *along with* $w_{j,s}$ *for* $j = 1, 2, 3, 4$ *and* $s = 1, 2, 3$*. We may now form the associate linear program:*

$$\begin{aligned}
\min_{(x,y,w)} \quad & 150x_1 + 230x_2 + 260x_3\\
& - \frac{1}{3}(-238y_{11} + 170w_{11} + 210y_{21} - 150w_{21} - 36w_{31} - 10w_{41})\\
& - \frac{1}{3}(-238y_{12} + 170w_{12} + 210y_{22} - 150w_{22} - 36w_{32} - 10w_{42})\\
& - \frac{1}{3}(-238y_{13} + 170w_{13} + 210y_{23} - 150w_{23} - 36w_{33} - 10w_{43})
\end{aligned}$$

Subject to

$$\begin{aligned}
&x_1 + x_2 + x_3 \leq 500\\
&3x_1 + y_{11} - w_{11} \geq 200\\
&3.6x_2 + y_{21} - w_{21} \geq 240\\
&w_{31} + w_{41} \leq 24x_3\\
&w_{31} \leq 6000\\
&2.5x_1 + y_{12} - w_{12} \geq 200
\end{aligned}$$

$$
\begin{aligned}
&3x_2 + y_{22} - w_{22} \geq 240\\
&w_{32} + w_{42} \leq 20x_{32}\\
&w_{32} \leq 6000\\
&2.x_1 + y_{13} - w_{13} \geq 200\\
&2.4x_2 + y_{23} - w_{23} \geq 240\\
&w_{33} + w_{43} \leq 16x_{32}\\
&w_{33} \leq 6000\\
&x_1, x_2, x_3, y_{1,s}, y_{2,s}, w_{1,s}, w_{2,s}, w_{3,s}, w_{4,s} \geq 0 \quad \textit{for } s = 1, 2, 3
\end{aligned}
$$

We see how the size of linear program explodes. One can still solve this small problem with a standard solver to obtain the following solution with an expected profit of $108,390.

		Wheat	*Corn*	*Sugar beet*
First Stage	*Area (acres)*	*170*	*80*	*250*
s=1 (above)	*Yields*	*510*	*288*	*6000*
	Sales	*310*	*48*	*6000*
	Purchases	*0*	*0*	*0*
s=2 (average)	*Yields*	*425*	*240*	*5000*
	Sales	*225*	*0*	*5000*
	Purchases	*0*	*0*	*0*
s=3 (below)	*Yields*	*340*	*192*	*4000*
	Sales	*140*	*0*	*4000*
	Purchases	*0*	*48*	*0*

We can reformulate this problem in a more standard format for a stochastic optimisation problem. That is, in the form:

$$
\begin{aligned}
&\min_x \quad c^T x + E_\xi Q(x, \xi)\\
&\text{Subj to}\\
&\qquad Ax \leq b, x \geq 0
\end{aligned}
$$

where

$$
Q(x, \xi) = \min\left\{q^T y \mid Wy = h - Tx,\, y \geq 0\right\}.
$$

Here ξ is a vector made up of the data that depends on the probability distribution and effects q, h, T, and the operator E_ξ is the expected value. In our case, we have that ξ consist of the random vector $(t_1(s), t_2(s), t_3(s))$ of yields for the three crops under the three scenarios $s = 1, 2, 3$ with a probability distribution $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Then we have

$$
\begin{aligned}
&Q(x, s) = \min\left\{238y_1 - 170w_1 + 210y_2 - 150w_2 - 36w_3 - 10w_4\right\}\\
&\text{Subj to}\\
&\qquad y_1 - w_1 \geq 200 - t_1(s)\, x_1\\
&\qquad y_2 - w_2 \geq 2 - t_2(s)\, x_2\\
&\qquad w_3 + w_4 \leq t_3(s)\, x_3\\
&\qquad w_3 \leq 6000\\
&\qquad y_1, y_2, w_1, w_2, w_3 \geq 0.
\end{aligned}
$$

We can place $\mathcal{Q}(x) := E_\xi Q(x, \xi)$ is called the recourse function and so we have a problem

$$
\min_x \quad c^T x + \mathcal{Q}(x)
$$

$$\begin{aligned} &\text{Subj to} \\ &\qquad Ax \leq b, x \geq 0. \end{aligned}$$

We call x the first stage variable and y the second stage variables. The challenge with stochastic optimisation problems is that the number of variables explodes with the introduction of scenarios. The more scenarios we have, the more variables we have, and the larger the LP becomes. To make stochastic optimisation more amenable to solving, we introduce a device that allows the LP to be split into s separate LPs that can be solved separately. To do this, we introduce s copies of the first stage variables. That is, for $s = 1, 2, 3$ we have x_{1s}, x_{2s}, x_{3s}, and hence can split the larger LP into the following three problems for $s = 1, 2, 3$:

$$\begin{aligned} \min_{(x,y,w)} \quad & 150x_{2s} + 230x_{2s} + 260x_{3s} \\ & -\frac{1}{3}(-238y_{1s} + 170w_{1s} + 210y_{2s} - 150w_{2s} - 36w_{3s} - 10w_{4s}) \\ \text{Subject to} \quad & \\ & x_{1s} + x_{2s} + x_{3s} \leq 500 \\ & t_1(s)\, x_{1s} + y_{11s} - w_{1s} \geq 200 \\ & t_2(s)\, x_{2s} + y_{2s} - w_{2s} \geq 240 \\ & w_{3s} + w_{4s} \leq t_3(s)\, x_{3s} \\ & w_{3s} \leq 6000 \\ & x_{1s}, x_{2s}, x_{3s}, y_{1,s}, y_{2,s}, w_{1,s}, w_{2,s}, w_{3,s}, w_{4,s} \geq 0 \quad \text{for } s = 1, 2, 3 \end{aligned}$$

The advantage of this is that each problem is now of the same dimension as the original problem. The downside is that we really require agreement $x_{11} = x_{12} = x_{13}$, $x_{21} = x_{22} = x_{23}$ and $x_{31} = x_{32} = x_{33}$. That is

$$x_s = \begin{pmatrix} x_{1s} \\ x_{2s} \\ x_{3s} \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = z, \quad s = 1, 2, 3.$$

This is very unlikely to occur if we do not link the solutions of these separate LPs in some way.

9.1.1 The Consensus Problem

Let us consider the deterministic problem of consensus first. Consider

$$\begin{aligned} \min_{(x,z)} \quad & \sum_{i=1}^{s} f_i(x_i) \\ \text{Subj to} \quad & \\ & x_i - z = 0, \quad i = 1, \ldots, s \end{aligned}$$

We now apply ADMM. The augmented Lagrangian is given by:

$$L_\rho(x, z, \lambda) = \sum_{i=1}^{s} \left[f_i(x_i) + \lambda_i^T (x_i - z) + \frac{\rho}{2} \|x_i - z\|^2 \right]$$

The basic ADMM iteration is then

$$\begin{aligned} x^{k+1} &\in \arg\min_x L_{\rho^k}\left(x, z^k, \lambda^k\right) \\ z^{k+1} &\in \arg\min_z L_{\rho^k}\left(x^{k+1}, z, \lambda^k\right) \\ \lambda_i^{k+1} &= \lambda_i^k + \rho^k\left(x_i - z\right), \quad i = 1, \ldots, s \end{aligned}$$

If I iterate, I should be able to get consensus in the limit. This is a general framework for consensus problems. The particular case where we apply it to the stochastic optimisation problem is referred to as progressive hedging. The resulting iteration becomes separable, so we have for the first step:

$$x_i^{k+1} \in \arg\min_x \left\{ f_i\left(x_i\right) + \left(\lambda_i^k\right)^T\left(x_i - z^k\right) + \frac{\rho^k}{2}\left\|x_i - z^k\right\|^2 \right\}, \quad i = 1, \ldots, s.$$

For the second step we need to solve

$$\begin{aligned} 0 &= -\left(\sum_{i=1}^s \lambda_i + \rho^k\left(x_i^{k+1} - z^{k+1}\right)\right) \\ \text{or} \qquad z^{k+1} &= \frac{1}{s}\sum_{i=1}^s\left(x_i^{k+1} + \frac{\lambda_i}{\rho^k}\right). \end{aligned}$$

Hence we have

$$\begin{aligned} x_i^{k+1} &\in \arg\min_x \left\{ f_i\left(x_i\right) + \left(\lambda_i^k\right)^T\left(x_i - z^k\right) + \frac{\rho^k}{2}\left\|x_i - z^k\right\|^2 \right\}, \\ z^{k+1} &= \frac{1}{s}\sum_{i=1}^s\left(x_i^{k+1} + \frac{\lambda_i^k}{\rho^k}\right) \\ \lambda_i^{k+1} &= \lambda_i^k + \rho^k\left(x_i^{k+1} - z^{k+1}\right), \quad i = 1, \ldots, s. \end{aligned}$$

This can be simplified (letting the bar denote the average: $\bar{u}^k := \frac{1}{s}\sum_i u_i^k$), as the second equation states that $z^{k+1} = \bar{x}^{k+1} + \frac{1}{\rho^k}\bar{\lambda}^k$, or $\rho^k\left(z^{k+1} - \bar{x}^{k+1}\right) = \bar{\lambda}^k$, and the last equation says (on averaging) that $\bar{\lambda}^{k+1} = \bar{\lambda}^k + \rho^k\left(\bar{x}^{k+1} - z^{k+1}\right)$. Hence

$$\begin{aligned} \bar{\lambda}^{k+1} &= \bar{\lambda}^k + \rho^k\left(\bar{x}^{k+1} - z^{k+1}\right) \\ &= \rho^k\left(z^{k+1} - \bar{x}^{k+1}\right) + \rho^k\left(\bar{x}^{k+1} - z^{k+1}\right) = 0. \end{aligned}$$

That is: $z^k = \bar{x}^k$. And so we have

$$\begin{aligned} x_i^{k+1} &\in \arg\min_x \left\{ f_i\left(x_i\right) + \left(\lambda_i^k\right)^T\left(x_i - \bar{x}^k\right) + \frac{\rho^k}{2}\left\|x_i - \bar{x}^k\right\|^2 \right\} \\ \lambda_i^{k+1} &= \lambda_i^k + \rho^k\left(x_i^{k+1} - \bar{x}^{k+1}\right). \end{aligned}$$

This is the standard format in which the progressive hedging scheme is typically posed. The primal and dual residuals are

$$r^k = \left(x_1^k - \bar{x}^k, \ldots, x_s - \bar{x}^k\right) \quad \text{and} \quad s^k = -\rho^k\left(\bar{x}^k - \bar{x}^{k+1}, \ldots, \bar{x}^k - \bar{x}^{k+1}\right)$$

$$\text{and so} \quad \left\|r^k\right\|^2 = \sum_{i=1}^s\left\|x_i^k - \bar{x}^k\right\|^2 \quad \text{and} \quad \left\|s^k\right\|^2 = s\rho^k\left\|\bar{x}^k - \bar{x}^{k+1}\right\|^2.$$

9.1.2 Progressive Hedging (stochastic consensus)

We wish to now take into account the probabilistic aspect of stochastic programming and apply the same ideas that ADMM has for the solution of consensus problems. Let us consider the stochastic program in its LP format:

$$\zeta^{SP} = \min_{x,y} \left\{ c^T x + \sum_s p_s q_s^T y_s \mid (x, y_s) \in K_s,\, s \in \mathcal{S} \right\} \tag{9.1}$$

where $\{p_s\}_{s\in\mathcal{S}}$ is the probability distribution of the various scenarios that are indexed by the set $\mathcal{S}$. Also $K_s := \{(x, y_s) \mid Wy_s = h_s - T_s x,\, x \in X,\, y_s \geq 0\}$, and $X = \{x \mid Ax \leq b,\, x \geq 0\}$. We introduce copies of the first stage variables, the scenario-dependent copies x_s for each $s \in S$, and we create the following reformulation of (9.1):

$$\zeta^{SP} = \min_{x,y,z} \left\{ \textstyle\sum_{s\in\mathcal{S}} p_s (c^\top x_s + q_s^\top y_s) \;:\; (x_s, y_s) \in K_s, x_s = z,\; \forall s \in \mathcal{S}, z \in \mathbb{R}^{n_x} \right\}. \tag{9.2}$$

The constraints $x_s = z$, $s \in \mathcal{S}$, enforce *nonanticipativity* for first-stage decisions; the first-stage decisions x_s must be the same (z) for each scenario $s \in \mathcal{S}$. We need to now form the augmented Lagrangian for this problem:

$$L_\sigma(x, y, z, \lambda) = \sum_{s\in\mathcal{S}} p_s (c^\top x_s + q_s^\top y_s) + (\lambda_s)^T (x_s - z) + \frac{\sigma}{2} \|x_s - z\|^2 .$$

Now define $\omega_s = \lambda_s / p_s$ and $\rho = \sigma_s / p_s$ to get

$$\begin{aligned} L_\rho(x, y, z, \omega) &= \sum_{s\in\mathcal{S}} p_s \left[(c^\top x_s + q_s^\top y_s) + (\omega_s)^T (x_s - z) + \frac{\rho}{2} \|x_s - z\|^2 \right] \\ &= \sum_{s\in\mathcal{S}} p_s L_{\rho_s}(x_s, y_s, z, \omega_s) \end{aligned}$$

$$\text{where} \quad L_{\rho_s}(x_s, y_s, z, \omega_s) = (c^\top x_s + q_s^\top y_s) + (\omega_s)^T (x_s - z) + \frac{\rho}{2} \|x_s - z\|^2 .$$

The ADMM steps are now

$$\begin{aligned} \left(x_s^{k+1}, y_s^{k+1}\right) &\in \underset{(x_s,y_s)\in K_s}{\operatorname{argmin}} \left\{ (c^\top x_s + q_s^\top y_s) + \left(\omega_s^k\right)^T \left(x_s - z^k\right) + \frac{\rho^k}{2} \left\| x_s - z^k \right\|^2 \right\}, \quad \text{for } s \in \mathcal{S} \\ z^{k+1} &\in \arg\min_z \left\{ \sum_{s\in\mathcal{S}} p_s (c^\top x_s^{k+1} + q_s^\top y_s^{k+1}) + \left(\omega_s^k\right)^T \left(x_s^{k+1} - z\right) + \frac{\rho^k}{2} \left\| x_s^{k+1} - z \right\|^2 \right\} \\ \omega_s^{k+1} &= \omega_s^k + \rho^k \left(x_s^{k+1} - z^{k+1}\right), \quad \text{for } s \in \mathcal{S}. \end{aligned} \tag{9.3}$$

Let us now consider the second problem in more detail. Differentiating, We need

$$\begin{aligned} 0 &= \sum_{s\in\mathcal{S}} p_s \left(\rho_s^k \left(z^{k+1} - x_s^{k+1} \right) - \omega_s^k \right) \\ \text{or} \quad z^{k+1} &= \left(\sum_{s\in\mathcal{S}} p_s \right) z^{k+1} = \sum_{s\in\mathcal{S}} p_s x_s^{k+1} + \frac{1}{\rho^k} \sum_{s\in\mathcal{S}} p_s \omega_s^k. \end{aligned} \tag{9.4}$$

Now suppose that at step $k = 0$ we have the condition:

$$\sum_{s\in\mathcal{S}} p_s \omega_s^k = 0.$$

This would certainly hold true if we chose $\omega_s^k = 0$ for all s. This condition is then preserved by the update (9.3) in that

$$\begin{aligned}\sum_{s\in\mathcal{S}} p_s\omega_s^{k+1} &= \sum_{s\in\mathcal{S}} p_s\omega_s^k + \rho^k \sum_{s\in\mathcal{S}} p_s\left(x_s^{k+1} - z^{k+1}\right)\\ &= 0 + \rho^k\left(\bar{x}^{k+1} - z^{k+1}\right) = 0\end{aligned}$$

because (9.4) now enforces

$$z^{k+1} = \sum_{s\in\mathcal{S}} p_s x_s^{k+1} =: \bar{x}^{k+1}.$$

Here $\bar{x}$ is the probability-weighted mean. The fact that our zs are averages of our xs makes sense, because we have introduced the z variables to create consensus. In summary, the basic ADMM iteration is now:

$$\begin{aligned}\left(x_s^{k+1}, y_s^{k+1}\right) &\in \underset{(x_s,y_s)\in K_s}{\operatorname{argmin}} \left\{(c^\top x_s + q_s^\top y_s) + \left(\omega_s^k\right)^T\left(x_s - z^k\right) + \frac{\rho^k}{2}\left\|x_s - z^k\right\|^2\right\}, \quad \text{for } s\in\mathcal{S}\\ z^{k+1} &= \sum_{s\in\mathcal{S}} p_s x_s^{k+1} = \bar{x}^{k+1}\\ \omega_s^{k+1} &= \omega_s^k + \rho^k\left(x_s^{k+1} - z^{k+1}\right), \quad \text{for } s\in\mathcal{S}.\end{aligned}$$

Thus the task really reduces to the first step, which is just a quadratic optimisation problem with linear constraints (there are tools to solve this). The subproblem we must solve in the first step is constrained to be the same size as the original problem that had only 1 scenario.

When to stop

Pseudocode for the PH algorithm is given in Algorithm 1. In Algorithm 1, $k_{max} > 0$ is the maximum number of iterations, and $\epsilon > 0$ parameterized the convergence tolerance. The initialization of Lines 3–8 provides an initial target primal value z^0 and dual values ω_s^1, $s\in\mathcal{S}$, for the main iterations $k \geq 1$. Also, an initial Lagrangian bound ϕ^0 can be computed from this initialization. For $\epsilon > 0$, the Algorithm 1 termination criterion $\sqrt{\sum_{s\in\mathcal{S}} p_s \left\|x_s^k - z^{k-1}\right\|_2^2} < \epsilon$ is motivated by the addition of the squared norms of the primal and dual residuals. In summing the squared norm primal residuals $p_s\left\|x_s^k - z^k\right\|_2^2$, $s\in\mathcal{S}$, and the squared norm dual residuals $\left\|z^k - z^{k-1}\right\|_2^2$, we have

$$\sum_{s\in\mathcal{S}} p_s\left[\left\|x_s^k - z^k\right\|_2^2 + \left\|z^k - z^{k-1}\right\|_2^2\right] = \sum_{s\in\mathcal{S}} p_s\left\|x_s^k - z^{k-1}\right\|_2^2 \tag{9.5}$$

The equality in (9.5) holds because for each $s\in\mathcal{S}$, the cross term $p_s(x_s^k - z^k)^T(z^k - z^{k-1})$ resulting from the expansion of the squared norm $\left\|(x_s^k - z^k) + (z^k - z^{k-1})\right\|_2^2$ vanishes, since we know $\sum_{s\in\mathcal{S}} p_s(x_s^k - z^k) = 0$ due to the construction of z^k. So the last term in (9.5) is the error term we monitor for our stopping criterion in progressive hedging.

Algorithm 1 Progressive Hedging Algorithm

Precondition: $\sum_{s\in\mathcal{S}} p_s\omega_s^0 = 0$
function PH(ω^0, ρ, k_{max}, ϵ)
    for $s \in \mathcal{S}$ **do**
        $(x_s^0, y_s^0) \in \arg\min_{x,y} \left\{(c+\omega_s^0)^\top x + q_s^\top y \mid (x,y) \in K_s\right\}$
    end for
    $\phi^0 \leftarrow \sum_{s\in\mathcal{S}} p_s \left[(c+\omega_s^0)^\top x_s^0 + q_s^\top y_s^0\right]$
    $z^0 \leftarrow \sum_{s\in\mathcal{S}} p_s x_s^0$
    $\omega_s^1 \leftarrow \omega_s^0 + \rho(x_s^0 - z^0)$ for all $s \in \mathcal{S}$
    for $k = 1, \ldots, k_{max}$ **do**
        for $s \in \mathcal{S}$ **do**
            $\phi_s^k \leftarrow \min_{x,y} \left\{(c+\omega_s^k)^\top x + q_s^\top y \mid (x,y) \in K_s\right\}$
            $(x_s^k, y_s^k) \in \arg\min_{x,y} \left\{L_s^\rho(x, y, z^{k-1}, \omega_s^k) \mid (x,y) \in K_s\right\}$
        end for
        $\phi^k \leftarrow \sum_{s\in\mathcal{S}} p_s \phi_s^k$
        $z^k \leftarrow \sum_{s\in\mathcal{S}} p_s x_s^k$
        if $\sqrt{\sum_{s\in\mathcal{S}} p_s \|x_s^k - z^{k-1}\|_2^2} < \epsilon$ **then**
            return $(x^k, y^k, z^k, \omega^k, \phi^k)$
        end if
        $\omega_s^{k+1} \leftarrow \omega_s^k + \rho(x_s^k - z^k)$ for all $s \in \mathcal{S}$
    end for
    return $(x^{k_{max}}, y^{k_{max}}, z^{k_{max}}, \omega^{k_{max}}, \phi^{k_{max}})$
end function

Chapter 10

Compressed Sensing and Signal Processing

10.1 Inverse Problems in Signal Processing

We will call the split feasibility problem (SFP) the problem of funding x such that

$$x \in C \quad \text{and} \quad Ax \in Q$$

where C and Q are non-empty convex sets in $\mathbf{R}^n$ and $\mathbf{R}^m$ respectively and A is an $m \times n$ matrix. The solution to the SFP can be obtained via optimisation: let

$$f(x) = \frac{1}{2} \|(I - P_Q) Ax\|^2 \quad \text{where} \quad P_Q \text{ is the projection onto } Q.$$

Then we can solve the SFP by minimizing the convex function f i.e.

$$\min_{x \in C} f(x).$$

Notice that

$$\begin{aligned} f(x) &= \frac{1}{2} \|(I - P_Q) Ax\|^2 = \frac{1}{2} ((I - P_Q) Ax)^T (I - P_Q) Ax \\ &= \frac{1}{2} x^T A^T \left(I - P_Q^T\right) (I - P_Q) Ax = \frac{1}{2} x^T A^T (I - P_Q) Ax \end{aligned}$$

noting that $P_Q^2 = P_Q P_Q = P_Q$. Clearly when the SFP has a solution we have the optimal value $f^* := \min_{x \in C} f(x) = 0$ and hence we may solve this problem via projected subgradient optimisation using the Polyak step size.

For subgradient methods at step k let $g^k = \nabla f\left(x^k\right)$ and place $x^{k+1} = P_C\left(x^k - t^k g^k\right)$ (presuming we can easily calculate the projection onto C). One can use

$$t^k = \frac{\rho_k \left(f\left(x^k\right) - f^*\right)}{\left\|g^k\right\|^2} = \frac{\rho_k f\left(x^k\right)}{\left\|\nabla f\left(x^k\right)\right\|^2} \quad \text{(Polyak step if needed)}$$

$$\text{where } \nabla f\left(x^k\right) = A^T (I - P_Q) Ax^k, \text{ (even though } P_Q \text{ may not be linear).} \tag{10.1}$$

Here P_Q and P_C are Euclidean projections, and $\rho_k > \frac{1}{2}$ (where best results occur with the choice $\rho_k = 2$).

We can improve the performance (from sublinear to superlinear convergence) by incorporating the extrapolation:

$$\begin{aligned} y_k &= x_k + \beta_k (x_k - x_{k-1}), \qquad \text{(extrapolation)} \\ x_{k+1} &= P_C (y_k - \alpha \nabla f (y_k)) \quad \text{(gradient projection step)}, \end{aligned}$$

where P_C is the projection onto C, $x_{-1} \equiv x_0$ and $\beta_k \in (0,1)$ is chosen to satisfy Nemirovski's conditions (6.33)–(6.34). As proven in Problem 6.8.3, one could use the specific choice (6.35) for the β_k given by:

$$\beta_k = \begin{cases} 0 & \text{if } k = 0 \\ \frac{k-1}{k+2} & \text{if } k = 1, 2, \ldots \end{cases} \quad \text{with} \quad \theta_k = \begin{cases} 1 & \text{if } k = -1 \\ \frac{2}{k+2} & \text{if } k = 0, 1, 2, \ldots \end{cases}. \tag{10.2}$$

In this case, an estimate on the Lipschitz constant L would be $\|A^T(I - P_Q)A\| \leq \|A^T A\|$ (since $\|I - P_Q\| = 1$). When we do not know L(and so can't use $\alpha = \frac{1}{L}$) we use the current value of $\alpha > 0$ as long as

$$\begin{aligned} f (x_{k+1}) &\leq l (x_{k+1}; y_k) + \frac{1}{2\alpha} \|x_{k+1} - y_k\|^2 \\ \text{where } l (x; y_k) &:= f (y_k) + \nabla f (y_k)^T (x - y_k). \end{aligned}$$

As soon as this condition is violated, we reduce α by some fraction, and we repeat this and repeat as many times as is necessary to get confirmation. Then hold that value of α.

SFPs naturally occur in inverse problems in signal processing. Many problems in signal processing can be formulated as an inverting the equation system

$$y = Ax + \varepsilon$$

where $x \in \mathbf{R}^N$ are the data to be recovered, and $y \in \mathbf{R}^k$ is the vector of noisy observations or measurements and represents the noise with bounded variance σ_ε^2. The matrix $A : \mathbf{R}^N \to \mathbf{R}^k$ is a bounded linear observation operator, often ill conditioned (small input change can mean big output change) because it models a process with loss of information. We will focus on the case of "compressed sensing" for which our measurements of the input signal have $k < N$. That is, we have far fewer data observations than would be required to recover non-sparse signal. Fortunately, signals are often sparse, with many zero components, making it possible to succeed in this seemingly impossible task.

A powerful approach for this problem consists of considering a solution x represented by a sparse expansion (i.e. a large number of coefficients that are zero). We attempt to find a sparse expansion by solving the unconstrained optimisation problem

$$\min_{x \in \mathbf{R}^N} \frac{1}{2} \|y - Ax\|_2^2 + \nu \|x\|_1 \tag{10.3}$$

where $\|\cdot\|_2$ is the Euclidean norm and $\|x\|_1 = \sum_{i=1}^N |x_i|$ the l^1-norm. The l^1 norm is known to produce sparse solutions (very few non-zero elements). The dual norm associated with the l^1-norm is the l^∞ norm given by $\|x^*\|_\infty := \max_{i=1,\ldots,N} |x_i^*|$.

10.2 Equivalent Formulation

Proposition 10.2.1 *Any solution to the problem*

$$\min_x \frac{1}{2} \|y - Ax\|_2^2$$

$$\text{Subject to } \|x\|_1 \leq t \tag{10.4}$$

for any fixed $t \geq 0$ is also a minimizer of (10.3) for some $\nu \geq 0$.

The problem (10.3) is now of the form of a SFP with $C = B_t^1(0) := \left\{x \in \mathbf{R}^N \mid \|x\|_1 \leq t\right\}$ and $Q = \{y\}$. The projected subgradient method (and the descent methods) would both use the follows generic projective descent step in an iteration

$$\begin{aligned} x_{k+1} &= P_{B_t^1(0)}\left[x_k - \tau_k A^T\left(I - P_{\{y\}}\right)Ax_k\right] \\ &= P_{B_t^1(0)}\left[x_k - \tau_k A^T\left(Ax_k - y\right)\right]. \end{aligned}$$

Problem Set 17

Problem 10.2.2 *1. Show that*

$$\langle x, x^* \rangle \leq \|x\|_1 \|x^*\|_\infty$$

and that equality is possible. Given any x, write down the x^* that gives the equality. Conclude (no need to prove, since it will be obvious) that any x^* has this form.

2. Use an argument similar to that used in example 4.2.1 to show that for $h(x) := \|x\|_1$ we have

$$h^*(x^*) = \delta_{B_1^\infty(0)}(x^*)$$

where $B_1^\infty(0) := \left\{x^* \in \mathbf{R}^N \mid \|x^*\|_\infty \leq 1\right\}$.

3. Deduce that $x^* \in \partial h(x)$ if and only if

$$\langle x, x^* \rangle = \|x\|_1 \|x^*\|_\infty \quad \text{and} \quad \|x^*\|_\infty \leq 1.$$

In addition, show that when $x \neq 0$, then $\|x^*\|_\infty = 1$.

4. Now consider the function $g(x) = \delta_{B_t^1(0)}(x)$ where $B_t^1(0) := \left\{x \in \mathbf{R}^N \mid \|x\|_1 \leq t\right\}$. Show that for $f(x^*) := t\|x^*\|_\infty$

$$f^*(x) = \delta_{B_t^1(0)}(x)$$

and hence deduce that $g^*(x^*) = t\|x^*\|_\infty$.

5. Deduce that $x^* \in \partial g(x)$ if and only if

$$\langle x, x^* \rangle = \|x\|_1 \|x^*\|_\infty \quad \text{and} \quad \|x\|_1 \leq t.$$

When, in addition $x^* \neq 0$, conclude that $\|x\|_1 = t$.

6. Consider the problem (10.4):

$$\begin{aligned} &\min_x \frac{1}{2}\|y - Ax\|_2^2 \\ &\text{Subject to } \|x\|_1 \leq t \end{aligned}$$

and the other problem (10.3):

$$\min_{x \in \mathbf{R}^N} \frac{1}{2}\|y - Ax\|_2^2 + \nu\|x\|_1$$

Show that a solution of (10.4) *for any* $t \geq 0$ *also solves* (10.3) *for some* $\nu \geq 0$. *To do so, first reformulate* (10.4) *as the unconstrained problem*

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \delta_{B_t^1(0)}(x).$$

Then write down the optimality condition for (10.4)*:* $0 \in \partial_x \left\{ \frac{1}{2} \|y - Ax\|_2^2 + \delta_{B_t^1(0)}(x) \right\}$. *Write down the similar optimality condition for* (10.3). *Show that a solution of the optimality condition for* (10.4) *also satisfies the optimality condition for* (10.3) *for some* ν.

We need to be able to compute the Euclidean projection onto the l^1 ball. We will now outline an efficient way to do this.

10.3 An Algorithm for the Projection Onto the l^1-Ball

By symmetry, it suffices to work in the positive orthant, so let $v \in \mathbb{R}_+^n$. We will need to consider order statistics and sorting vectors in order of decreasing size. To this end, denote $v_{(i)}$ to be the ith statistic of v where $v_{(1)} \geq v_{(2)} \geq \cdots \geq v_{(n)}$, for $v \in \mathbf{R}_+^n$. We begin by devising an algorithm to project onto the biggest ($(n-1)$-dimensional) face of the unit simplex in the positive orthant: $\sigma_t := \{w \in \mathbf{R}^n \mid \sum_{i=1}^n w_i = t,\ w_i \geq 0\}$. We need to solve

$$\min_w \quad \frac{1}{2} \|w - v\|_2^2 \quad \text{s.t.} \quad \sum_{i=1}^n w_i = t,\ w_i \geq 0. \tag{10.5}$$

The Lagrangian is

$$L(w, \theta, \lambda) = \frac{1}{2} \|w - v\|^2 + \theta \left(\sum_{i=1}^n w_i - t \right) - \langle \lambda, w \rangle$$

where $\theta \in \mathbf{R}$ and $\lambda \in \mathbf{R}_+^n$ is a vector of multipliers. The θ is not constrained, because the constraint $\sum_{i=1}^n w_i - t = 0$ is equivalent to the constraint pair $\sum_{i=1}^n w_i - t \leq 0$ and $-\sum_{i=1}^n w_i + t \leq 0$, which are together enforced by nonnegative multipliers $\theta_1, \theta_2 \geq 0$ as follows:

$$\theta_1 \left(\sum_{i=1}^n w_i - t \right) + \theta_2 \left(-\sum_{i=1}^n w_i + t \right) = (\theta_1 - \theta_2) \left(\sum_{i=1}^n w_i - t \right).$$

Here $\theta = \theta_1 - \theta_2$ is unconstrained. By complementarity, if we have $w_i > 0$, then $\lambda_i = 0$. The relevant KKT condition yields

$$\begin{aligned} \frac{dL}{dw_i} &= w_i - v_i + \theta - \lambda_i = 0 \\ \text{so} \quad w_i &= v_i - \theta + \underbrace{\lambda_i}_{=0} = v_i - \theta \quad \text{when } w_i > 0. \end{aligned} \tag{10.6}$$

Thus all nonzero elements are tied by a single variable θ.

Lemma 10.3.1 ([9], Lemma 1) *Let* w *be the optimal solution to the minimization problem (10.5). Let* s *and* j *be two indicies such that* $v_s \geq v_j$. *If* $w_s = 0$ *then* $w_j = 0$ *as well.*

Proof. First we show the case when $v_s > v_j$ strictly. Assume by contradiction that $w_s = 0$ yet $w_j > 0$. Let $\tilde{w} \in \mathbf{R}^k$ be a vector whose elements are equal to the elements of w except for $\tilde{w}_s$ and $\tilde{w}_j$ which are interchanged, that is, $\tilde{w}_s = w_j$, and $\tilde{w}_j = w_s$ and for every other $r \notin \{s, j\}$ we have $\tilde{w}_r = w_r$. It is immediate to verify that the constraints of Eq. (10.5) still hold. In addition we have that,

$$\begin{aligned}\|w - v\|^2 - \|\tilde{w} - v\|^2 &= (\underbrace{0}_{(=w_s)} - v_s)^2 + (w_j - v_j)^2 - (\underbrace{w_j}_{(=\tilde{w}_s)} - v_s)^2 - (\underbrace{0}_{(=\tilde{w}_j)} - v_j)^2 \\ &= 2w_j \left(v_s - v_j\right)^2 > 0\end{aligned}$$

Therefore, we obtain that $\|w - v\|^2 > \|\tilde{w} - v\|^2$, which contradicts the fact that w is the optimal solution. This concludes the case $v_s > v_j$.

Now we deal with the case $v_s = v_j$. If we proceed as before, we get

$$\|w - v\|^2 - \|\tilde{w} - v\|^2 = 2w_j(v_s - v_j)^2 = 0,$$

which shows that $\tilde{w} \in P_{B_t^1}(v)$. However, since B_t^1 is Chebyshev, it holds that $P_{B_t^1}(v)$ is a singleton, and so we must have $w = \tilde{w}$, a contradiction.

You can also show the result as follows, without using the Chebyshev property. Take a sequence v^n such that $v_n \to v$, and $v_s^n > v_s$, and $v_j^n = v_j$. The triangle inequality yields

$$|w_j - 0| \leq |w_j - w_j^n| + |w_j^n|.$$

Moreover, because the projection operator is continuous and $v^n \to v$, we must have that $|w_j - w_j^n| \to 0$. It remains only to show that $w_j^n \to 0$.

Case: Suppose there exists an infinite subsequence where $w_s^n = 0$. Then, since $v_s^n > v_s = v_j = v_j^n$, we have $v_s^n > v_j^n$ and so we may apply our result from the first part to obtain that we must have $w_j^n = 0$. Since it is true for all n, we are done.

Case: When the subsequence satisfying $w_s^n = 0$ is of finite length, then we can consider instead the subsequence that satisfies $w_s^n > 0$. Whenever $w_j^n = 0$, there is obviously nothing more to show, so consider the case when $w_j^n > 0$. Then we have that

$$\begin{aligned}0 < w_j^n &= v_j^n - \theta \quad \text{(by (10.6))} \\ &= v_j - \theta \quad (\text{since } v_j = v_j^n) \\ &= v_s - \theta \quad \text{since } v_j = v_s \\ &< v_s^n - \theta \quad \text{since } v_s^n > v_s \\ &= w_s^n \quad \text{by (10.6) since } w_s^n > 0 \\ &\to w_s = 0,\end{aligned}$$

which is what we needed to show. □

Denote by $I := \left\{i \in \{1, \ldots, n\} \mid w_{(i)} > 0\right\}$ and so $I = \{1, \ldots, \rho\}$ for some $0 \leq \rho \leq n$. If we had ρ then by)

$$t = \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_{(i)} = \sum_{i=1}^{\rho} w_{(i)} + \underbrace{\sum_{i=\rho+1}^{n} w_{(i)}}_{=0} \overset{(10.6)}{=} \sum_{i=1}^{\rho} \left(v_{(i)} - \theta\right).$$

and so

$$\theta = \frac{1}{\rho}\left(\sum_{i=1}^{\rho} v_i - t\right).$$

Given θ we have the optimal w given by

$$w_i = \max\{v_i - \theta, 0\}.$$

To find ρ we have the following.

Lemma 10.3.2 ([21], Lemma 3) *Let w be the optimal solution to (10.5). Let μ be the order statistic of v: that is the vector obtained from v by sorting v in descending order. Then the number of strictly positive elements in w is*

$$\rho = \rho(t, \mu) := \max\left\{ j \in \{1, \ldots, n\} \mid \mu_j - \frac{1}{j}\left(\sum_{r=1}^{j} \mu_r - t\right) > 0 \right\}.$$

We now have the algorithm.

Projection onto the simplex: Initially we start with $v \in \mathbf{R}^n_+$ and $t > 0$.

Sort v into descending order $\mu : \mu_1 \geq \mu_2 \geq \cdots \geq \mu_\rho$.
Find $\rho = \max\left\{ j \in \{1, \ldots, n\} \mid \mu_j - \frac{1}{j}\left(\sum_{r=1}^{j} \mu_r - t\right) > 0 \right\}$.
Define $\theta = \frac{1}{\rho}\left(\sum_{i=1}^{\rho} \mu_i - t\right)$.
Output w given by $w_i = \max\{v_i - \theta, 0\}$.

We may now derive an efficient algorithm to project onto the l^1 ball. We need to now solve the problem

$$\min_w \quad \|w - v\|_2^2 \quad \text{s.t.} \quad \|w\|_1 \leq t. \tag{10.7}$$

We will reduce this to the problem of projecting onto the positive simplex. If $\|v\|_1 \leq t$ then clearly $v = w$ is the solution. Otherwise $\|v\|_1 > t$ and so the optimal solution w must satisfy $\|w\|_1 = t$ so we now study

$$\min_w \quad \|w - v\|_2^2 \quad \text{s.t.} \quad \|w\|_1 = t.$$

Lemma 10.3.3 *Let w be the optimal solution to (10.7). Then for all i we have $w_i v_i \geq 0$ (i.e. v_i and w_i take the same sign, using symmetry is valid).*

Proof. Assume, by contradiction, that the claim does not hold. Thus, there exists i for which $w_i v_i < 0$. Let $\hat{w}$ be the vector such that $\hat{w}_i = 0$ and $\hat{w}_j = w_j$ for $j \neq i$. Then $\|\hat{w}\|_1 = \|w\|_1 - |w_i| \leq t$ and hence $\hat{w}$ is feasible. In addition

$$\begin{aligned} \|w - v\|_2^2 - \|\hat{w} - v\|_2^2 &= (w_i - v_i)^2 - (0 - v_i)^2 \\ &= w_i^2 - 2w_i v_i > w_i^2 > 0. \end{aligned}$$

Thus we see that the feasible solution $\hat{w}$ attains a smaller objective value than w, the optimal solution. This is a contradiction. □

Based on this Lemma and the symmetry of the objective, we are ready to present our reduction. Let u be a vector containing the absolute values of v i.e. $u_i = |v_i|$. Then we replace (10.7) by

$$\min_\beta \quad \frac{1}{2}\|\beta - u\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 = \sum_i \beta_i \leq t \quad \text{and } \beta \geq 0.$$

Once we obtain the solution to this problem we obtain that for (10.7) via $w_i = \text{sign}(v_i)\,\beta_i$.

Problem Set 18

1. Devise pseudocode for the solution of the Euclidean norm projection onto the l^1 ball of radius t. For writing it, you can use, for example, the algorithm2e Latex package (or similar).

2. Turn this code into functioning computer code for this problem. For presenting the code in your homework, you can include screenshots of the relevant portions (preferred), or use the verbatim package. Remember to attach your code to your submission.

3. Take some points in $\mathbb{R}^2$, and test your code using these example points. Provide some images, or a list of input and output points, to check the results.

   [Hint: If $v \in N_{B_t^1(0)}(w)$, where $B_t^1(0) := \{x \in \mathbf{R}^n \mid \|x\|_1 \leq t\}$, then w will be the solution to (10.7) for the given v. Note also that $N_{B_t^1(0)}(w) = \partial\delta_{B_t^1(0)}(w)$]

10.4 Compressed Sensing

In this section we will use the formulation (10.4) of the problem (10.3) to solve the signal recovery problem as a SFP problem. We use $C = B_t^1(x) := \left\{x \in \mathbf{R}^N \mid \|x\|_1 \leq t\right\}$ and $Q = \{y\}$. We have an N sample signal x with exactly m nonzero components to recover from $m \ll k < N$ measurements. That is, the number of measurements k is smaller than the number of signal samples N and at the same time much larger than the sparsity m of x. Likewise the measurements are required to be incoherent, that is the information is spread out over the entire domain. Since $k < N$ the problem of recovering x from k measurements is ill conditioned, because we encounter an under-determined system of linear equations. But, by using sparsity as a prior, we can recover x from y as long as the number of nonzero elements is small enough.

Problem Set 19

We use subgradient optimisation to solve the problem (10.4), using the fast projective descent algorithm of section 6.7.

1. Construct some random examples of signals with $N = 2^{14} = 16384$ elements, only 50 of which are nonzero. We will reconstruct using $k = 2^{13} = 8192$ measurements. We want $m = 50$ spikes (height ± 1) located at random within the interval $[0, 16384]$ (choose the $\pm$ randomly as well). The observations y are generated via $y = Ax + \varepsilon$. Here A is a $k \times N$ matrix randomly obtained with independent sample from a normalised standard Gaussian distribution (i.e. the mean paramter of A is 0 and the covariance matrix for A is the $N \times N$ identity) and $\varepsilon \backsim N\left(0, \sigma_\varepsilon^2 = 10^{-4}\right)$ (i.e. ε is a randomly generated noise vector of size k whose entries have mean parameter zero and standard deviation parameter $\sqrt{10^{-4}}$). Write code to generate examples of this kind.

2. Write code for the fast convex projective gradient descent algorithm to solve (10.4). You will need to use your code for the projection onto $B_t^1(0)$.

3. Apply your code to the signal recovery problem for an example generated in 1. Use $t = 50$ (this parameter is usually estimated on the fly and should be close to the sparsity of x) and a termination criteria of

$$\|x_n - x_{n-1}\| < 10^{-15}.$$

4. Produce plots of the original signal x and the signal recovered from y. Either position the plots with one above the other, or superimpose them with different colours, so that they can be compared.

5. Similarly, try an experiment where you use $t = 30$ and $t = 60$, and report what happens.

Problem Set 20: ADMM and Positive Definite Programming Problems

Problem 10.4.1 *We studied Linear and Quadratic programming in the context of ADMM earlier on. This approach translates to the context of semi-definite programming with just the change of the projection step onto the positive order cone $\mathcal{P}(n)$. In this case we have the objective*

$$Min \quad \langle C, X\rangle \tag{10.8}$$

$$Subject\ to\ \langle A_i, X\rangle = b_i,\ i = 1, \ldots, m,\ X \succeq 0. \tag{10.9}$$

Referring back to section 7.1.1, we set $g := \delta_{\mathcal{P}(n)=\{x|X\succeq 0\}}$ and $f(X) = \langle C, X\rangle + \delta_{\{X|\langle A_i,X\rangle=b_i,\ i=1,\ldots,m\}}$. Then the ADMM iteration for this linear program consists of:

$$X^{k+1} \in \operatorname{argmin}_{X\in\mathcal{S}(n)} \left\{ \langle C, X\rangle + \frac{\rho}{2} \left\| U^k - Z^k + X \right\|_F^2 \mid \langle A_i, X\rangle = b_i,\ i = 1, \ldots, m \right\} \tag{10.10}$$

$$Z^{k+1} = P_{\mathcal{P}(n)}(X^{k+1} + U^k) \tag{10.11}$$

$$U^{k+1} = U^k + X^{k+1} - Z^{k+1}.$$

Part 1. *To solve (10.10) we need to investigate the associated KKT conditions (which are necessary and sufficient for a convex problems). So:*

1. *Write down the augmented Lagrangian $L_\rho(c, \nu)$ for solving the subproblem (10.10);*

2. *Next, write down the KKT conditions $0 = \nabla_X L(\cdot, \nu^*)(X^{k+1})$ and $0 = \nabla_\nu L(X^{k+1}, \cdot)(\nu^{k+1})$ where (X^{k+1}, ν^{k+1}) are the optimal primal-dual pair;*

3. *Use the KKT conditions to show that*

$$X^{k+1} = -\left(U^k - Z^k\right) - \frac{1}{\rho}C - \sum_{i=1}^{m} \frac{\nu_i^{k+1}}{\rho} A_i, \quad \text{where } \nu^{k+1} \text{ solves the system} \tag{10.12}$$

$$\sum_{j=1}^{m} \frac{\nu_j^{k+1}}{\rho} \langle A_i, A_j\rangle = -b_i + \langle A_i, \left(-U^k + Z^k - \frac{1}{\rho}C\right)\rangle, \quad i = 1, \ldots, m. \tag{10.13}$$

You may use the fact that the Frobenius norm and trace inner product differentiate in the way we are used to: i.e. $\nabla_X \|X\|_F^2 = 2X$ and $\nabla_X \langle X, B\rangle = B$.

Part 2. *Once we have solved the system (10.13) for ν_i for $i = 1, \ldots, m$ we have the solution for X^{k+1} from (10.12). This leaves the projection step (10.11) to discuss. Before doing so, we note that the primal and dual residuals are given by:*

$$R^k := X^k - Z^k \quad \text{and} \quad S^k = -\rho\left(Z^k - Z^{k-1}\right).$$

We terminate when both $\left\|R^{k+1}\right\|_F$ *and* $\left\|S^k\right\|_F$ *are small. We can also update the* ρ *as before:*

$$\rho^{k+1} = \begin{cases} \tau^{inc}\rho^k & \text{if } \left\|R^k\right\|_F > \kappa \left\|S^k\right\|_F \\ \rho^k/\tau^{dec} & \text{if } \left\|S^k\right\|_F > \kappa \left\|R^k\right\|_F \\ \rho^k & \text{otherwise} \end{cases}.$$

Now let's consider the projection. To project onto the positive semi-definite cone, consider the projection problem:

$$\min_{Y \in \mathcal{P}(n)} \frac{1}{2} \|Y - X\|_F^2 . \tag{10.14}$$

$$\text{or} \quad \min_{Y} \frac{1}{2} \|Y - X\|_F^2 \quad \text{s.t. } Y \succeq 0 \tag{10.15}$$

We can formulate this using a Lagrangian:

$$L(Y, \Lambda) := \frac{1}{2} \|Y - X\|_F^2 + \langle \Lambda, -Y \rangle, \quad \text{with} \quad \Lambda \succeq 0. \tag{10.16}$$

The complementarity condition associated with the constraint $Y \succeq 0$ *would be* $\langle \Lambda, Y \rangle = 0$*. Next, we'll show that this implies* $Y\Lambda = \Lambda Y = 0$*. Do it the following way:*

1. *Expand* Λ *via the spectral decomposition* $\Lambda = \sum_i \lambda_i u_i u_i^T$*, substitute the spectral decomposition form into* $\langle \Lambda, Y \rangle = 0$*, and use this to show that whenever* $\lambda_i \neq 0$*, we have* $u_i^T Y u_i = 0$ *(Hint: remember that* $(Y \in P(n)) \implies u_i^T Y u_i \geq 0$ *and that* $(\Lambda \in \mathcal{P}(n)) \implies \lambda_i \geq 0$*)*
2. *Expand* Y *via the spectral decomposition* $Y = \sum_j \mu_j v_j v_j^T$*, and combine this with what you just showed to prove that whenever* $\lambda_i, \mu_j \neq 0$*, we have* $u_j \perp v_i$*.*
3. *Use this final identity to show that* $Y\Lambda = \Lambda Y = 0$*.*

Part 3.

1. *Write down the KKT conditions* $0 = \nabla_Y L(\cdot, \Lambda)(Y)$ *associated with Lagrangian (10.16).*
2. *Take what you just showed, and use the fact that the optimal* Y *satisfies* $\Lambda Y = Y\Lambda = 0$ *to show that* $XY = YX$*.*

Part 4. *The following fact is well known:* $A, B \in \mathcal{S}(n)$ *commute if and only if* A *and* B *can be simultaneously diagonalised i.e.* $AB = BA$ *if and only if there exists an orthogonal* $Q \in \mathcal{S}(n)$ *such that*

$$\begin{aligned} Q^{-1}AQ &= \text{diag}[\lambda_1, \ldots, \lambda_m] \\ \text{and} \quad Q^{-1}BQ &= \text{diag}[\mu_1, \ldots, \mu_m]. \end{aligned}$$

[Note: we will use this without proving it.]

1. *Show that a matrix* $A \in \mathcal{P}(n)$ *satisfies* $\|A\|_F = \sqrt{\sum_{i=1}^n a_i^2}$ *where the* a_i *are the eigenvalues of* A*. (Hint: use Lemma 2.1.3 and the nonnegativity of the eigenvalues of* A*).*
2. *If* Q *diagonalizes* $A \in \mathcal{P}(n)$*, why does it hold that* $\|Q^T A Q\|_F^2 = \|A\|_F^2$*?*

3. *Show that where μ_i are the eigenvalues of Y and ξ_i are the eigenvalues of X, then*

$$\|Y - X\|_F^2 = \|(\mu_1 - \xi_1, \ldots, \mu_n - \xi_n)\|_F^2$$

(Hint: $Y - X \in \mathcal{P}(n)$, and, since X and Y commute, they possess a simultaneous diagonalization via a matrix Q)

4. *Use what you have shown to explain why $Y = Q\,\mathrm{diag}[\max\{\xi_i, 0\}]Q^T$ is the solution to problem (10.15). Interpret this as a projection onto a cone.*

Part 5. *Summarize the ADMM algorithm associated with the problem (10.9).*

Appendix A

Proof of Theorem 4.1.7

First we need the following lemma.

Lemma A.0.1 *Suppose $p : \mathbf{X} \to \overline{\mathbf{R}}$ is sublinear and the point $\bar{x} \in \operatorname{core}(\operatorname{dom} p)$. Then $q(\cdot) = p'(\bar{x}, \cdot)$ is convex and satisfies*

1. *$q(\lambda\bar{x}) = \lambda p(\bar{x})$ for all real λ.*
2. *$q \leq p$, and*
3. *$\operatorname{lin} q \supseteq \operatorname{lin} p + \operatorname{span}\{\bar{x}\}$*

Proof. Clearly $p'(\bar{x}, \cdot)$ is finite (by Proposition 4.1.6). We know that sublinearity implies p is positively homogeneous and so

$$\begin{aligned} q(\lambda\bar{x}) &= \lim_{\delta \downarrow 0} \frac{1}{\delta}\left(p(\bar{x} + \delta(\lambda\bar{x})) - p(\bar{x})\right) \\ &= \lim_{\delta \downarrow 0} \frac{1}{\delta}\left(p(\bar{x}(1 + \delta\lambda)) - p(\bar{x})\right) \quad (\text{and } 1 + \delta\lambda > 0 \text{ for } \delta \text{ small}) \\ &= \lim_{\delta \downarrow 0} \frac{1}{\delta}\left((1 + \delta\lambda)p(\bar{x}) - p(\bar{x})\right) = \lim_{\delta \downarrow 0} \frac{1}{\delta}(\delta\lambda)p(\bar{x}) = \lambda p(\bar{x}). \end{aligned}$$

As p is sublinear it is convex and so by Lemma 4.1.1 $d \mapsto q(d)$ is convex and

$$\begin{aligned} q(d) := p'(\bar{x}, d) &= \inf_{\delta > 0} \frac{1}{\delta}\left(p(\bar{x} + \delta d) - p(\bar{x})\right) \\ &\leq \inf_{\delta > 0} \frac{1}{\delta}\left(p(\bar{x}) + \delta p(d) - p(\bar{x})\right) = p(d). \end{aligned}$$

Finally we show $\operatorname{lin} q \supseteq \operatorname{lin} p + \operatorname{span}\{\bar{x}\}$. First note that $\operatorname{lin} q$ is a subspace since if d_1, $d_2 \in \operatorname{lin} q$ then for any α_1, $\alpha_2 \in \mathbf{R}$ we have $-q(-d_i) = q(d_i)$ implying $-\alpha_i q(-d_i) = \alpha_i q(d_i)$. If $\alpha_i \geq 0$ then $-q(-\alpha_i d_i) = q(\alpha_i d_i)$ by positive homogeneity (or $0q(d_i) = q(0d_i) = q(0) = 0$ when $\alpha_i = 0$). When $\alpha_i < 0$ then $\alpha_i = -|\alpha_i|$ and $-\alpha_i q(-d_i) = \alpha_i q(d_i)$ implies $q(-|\alpha_i| d_i) = -q(|\alpha_i| d_i)$ or $q(-\alpha_i d_i) = -q(-\alpha_i d_i)$ and again $\alpha_i d_i \in \operatorname{lin} q$. From (4.3) we know that $-q(-d) \leq q(d)$ always for any d. Thus to show $d \in \operatorname{lin} q := \{d \mid -q(-d) = q(d)\}$ we only need to show that $-q(-d) \geq q(d)$. Now consider d_1, $d_2 \in \operatorname{lin} q$ and α_1, $\alpha_2 \in \mathbf{R}$ then $\alpha_1 d_1$, $\alpha_2 d_2 \in \operatorname{lin} q$ implying

$$q(\alpha_1 d_1 + \alpha_2 d_2) \leq q(\alpha_1 d_1) + q(\alpha_2 d_2)$$

$$\begin{aligned}
&\leq -q(-\alpha_1 d_1) - q(-\alpha_2 d_2) \quad \text{since } \alpha_1 d_1,\, \alpha_2 d_2 \in \lin q \\
&= -\left(q(-\alpha_1 d_1) + q(-\alpha_2 d_2)\right) \\
&\leq -\left(q\left((-\alpha_1 d_1) + (-\alpha_2 d_2)\right)\right) \quad \text{by subadditivity} \\
&= -q\left(-\left(\alpha_1 d_1 + \alpha_2 d_2\right)\right)
\end{aligned}$$

implying $\alpha_1 d_1 + \alpha_2 d_2 \in \lin q$. Next we show that if $d \in \lin p$ then $d \in \lin q$. Indeed

$$-p\left(-d\right) = p(d) \geq q(d) \quad \text{by part 2.}$$

Also since 2. implies $p(-d) \geq q(-d)$ we have $-p(-d) \leq -q(-d)$ and so

$$-q(-d) \geq q(d) \quad \Longrightarrow \quad d \in \lin q.$$

Finally we note that by 1.

$$-q(-\lambda \bar{x}) = -\left(-1\right) q(\lambda \bar{x}) = q(\lambda \bar{x})$$

implying $\operatorname{span}\{\bar{x}\} \subseteq \lin q$ and so by linearity $\lin q$ we have 3. holding. □

The following is an important fact (see [4] Theorem 3.1.8 (finite dimensions) or [20] page 23 (infinite dimensions)). The power of the idea of a subgradient comes from the fact that the subdifferential if often nonempty. We now prove the Theorem 4.1.7.

Theorem A.0.2 *If $f : \operatorname{dom}(f) \subseteq \mathbf{X} \to \overline{\mathbf{R}}$ is proper, convex then at any point $\bar{x} \in \operatorname{icr}\operatorname{dom}(f)$ we have $\partial f(\bar{x}) \neq \emptyset$ and*

$$f'(\bar{x}, d) = \max\left\{\langle \lambda, d\rangle \mid \lambda \in \partial f(\bar{x})\right\}.$$

Proof. Restrict attention to $\mathbf{X} := \operatorname{affine}\operatorname{dom}(f)$ and then we may assume $\bar{x} \in \operatorname{core}\operatorname{dom}(f)$. In view of Lemma 4.1.2 we only need to show that for all d there exists $\phi \in \partial f(\bar{x})$ such that $\langle \phi, d\rangle = f'(\bar{x}, d)$. Choose a basis $\{e_1, \ldots, e_n\}$ for $\mathbf{X}$ with $e_1 = d$ if d is nonzero. Now define a sequence p_0, $p_1, \ldots, p_n$ recursively by $p_0 = f'\left(\bar{x}, \cdot\right)$, and $p_k = p'_{k-1}(e_k, \cdot)$ for $k = 1, 2, \ldots, n$. We essentially show $p_n(\cdot)$ is the required subgradient. First note that Proposition 4.1.6 and the fact that $\bar{x} \in \operatorname{core}\operatorname{dom}(f)$ implies each p_k is finite and sublinear. Consequently

$$\lin p_k \supseteq \lin p_{k-1} + \operatorname{span}\{e_k\} \quad \text{for } k = 1, \ldots, n$$

implying $\lin p_n = \mathbf{X}$ and so p_n is linear. Thus there is a $\phi \in X'$ such that $\langle \phi, \cdot\rangle = p_n(\cdot)$. Now apply part 2. of Lemma A.0.1 to obtain

$$\begin{aligned}
&p_n \leq p_{n-1} \leq \cdots \leq p_0 = f'(\bar{x}, \cdot) \\
\text{implying} \quad &\langle \phi, x - \bar{x}\rangle = p_n(x - \bar{x}) \leq p_0(x - \bar{x}) = f'(\bar{x}, x - \bar{x}) \leq f(x) - f(\bar{x})
\end{aligned}$$

where the last inequality follows from the observation that

$$\begin{aligned}
f'(\bar{x}, x - \bar{x}) &= \inf_{\delta > 0} \Delta f(\bar{x}, \delta, x - \bar{x}) \leq \Delta f(\bar{x}, 1, x - \bar{x}) \\
&= \frac{1}{1}\left(f(\bar{x} + 1\left(x - \bar{x}\right)) - f(\bar{x})\right) = f(x) - f(\bar{x}).
\end{aligned}$$

By Lemma 4.1.2 we have $\phi \in \partial f(\bar{x}) \neq \emptyset$. If $\phi = 0$ then $p_n(0) = 0 = f'(\bar{x}, 0)$ without further work. When $\phi \neq 0$ from part 1. of A.0.1 we have

$$\begin{aligned}
p_n(d) &\leq p_0(d) = p_0(e_1) = -p'_0(e_1, -e_1) \\
&= -p_1(-e_1) = -p_1(-d) \leq -p_n(-d) = p_n(d),
\end{aligned}$$

(the last inequality follows from the linearity of p_n) whence $p_n(d) = p_0(d) = f'(\bar{x}, d)$. □

Remark A.0.1 *It can also be shown that when* $\nabla f(\bar{x})$ *exists then* $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$ *is unique and so in finite dimension* $f'(\bar{x}, d) = \nabla f(\bar{x})^T d$ *which is the classical characterization of the directional derivative and* $\lin f'(\bar{x}, \cdot) = \mathbf{R}^n$. *Indeed when* $\nabla f(\bar{x})$ *does not exist* $\partial f(\bar{x})$ *may contain many values prompting it to be referred to as a multi-function (i.e. one to many mapping).*

Appendix B

Proof of Theorem 4.2.2

Corollary B.0.1 *Suppose $f : X \to \overline{\mathbf{R}}$ and $g : X \to \overline{\mathbf{R}}$ and $A : X \to Y$ be a linear mapping. At any point $x \in X$ the calculus rule*

$$\partial\left(f + g \circ A\right)(x) \supseteq \partial f(x) + A^* \partial g\left(Ax\right)$$

with equality holding when $0 \in \operatorname{core}\left(\operatorname{dom} g - A \operatorname{dom} f\right)$.

Proof. Suppose $x^* \in \partial f(x)$ and $y^* \in \partial g(Ax)$. Then

$$\begin{aligned} g^*(y^*) + g\left(Ax\right) &= \langle y^*, Ax\rangle = \langle A^* y^*, x\rangle \\ \text{and} \quad f^*(x^*) + f(x) &= \langle x, x^*\rangle. \end{aligned}$$

Adding these we find

$$f(x) + g\left(Ax\right) + f^*(x^*) + g^*(y^*) = \langle A^* y^* + x^*, x\rangle. \tag{B.1}$$

Placing $z^* := A^* y^* + x^*$ we have $x^* = z^* - A^* y^*$ and note that

$$\begin{aligned} f^*(x^*) + g^*(y^*) &= f^*(z^* - A^* y^*) + g^*(y^*) \\ &\geq \inf_{y^*} \left\{f^*(z^* - A^* y^*) + g^*(y^*)\right\} := f^* \square g^*\left(z^*\right). \end{aligned} \tag{B.2}$$

Next note that $f^{**} = f$ and $g^{**} = g$ (since f and g are closed proper convex functions)

$$\begin{aligned} \left[f^* \square g^*\right]^*\left(z\right) &= \sup_{z^*} \left\{\langle z^*, z\rangle - \inf_{y^*} \left\{f^*(z^* - A^* y^*) + g^*(y^*)\right\}\right\} \\ &= \sup_{(z^*, y^*)} \left\{\langle z^* - A^* y^*, z\rangle - f^*(z^* - A^* y^*) + \langle A^* y^*, z\rangle - g^*(y^*)\right\} \\ &= \sup_{(z^*, y^*)} \left\{\langle z^* - A^* y^*, z\rangle - f^*(z^* - A^* y^*) + \langle y^*, Az\rangle - g^*(y^*)\right\} \\ &\leq \sup_{(z^*, y^*)} \left\{\langle z^* - A^* y^*, z\rangle - f^*(z^* - A^* y^*)\right\} + \sup_{y^*} \left\{\langle y^*, Az\rangle - g^*(y^*)\right\} \\ &\leq \sup_{k^*} \left\{\langle k^*, z\rangle - f^*(k^*)\right\} + \sup_{y^*} \left\{\langle y^*, Az\rangle - g^*(y^*)\right\} = f^{**}(z) + (Az) \\ &= f(z) + g(Az). \end{aligned}$$

Consequently

$$\left(f(\cdot) + g(A\cdot)\right)^*\left(z^*\right) \leq \left[f^* \square g^*\right]^{**}\left(z^*\right) \leq f^* \square g^*\left(z^*\right)$$

and (B.1) and (B.2) imply

$$\begin{aligned} &f(x) + g(Ax) + (f(\cdot) + g(A\cdot))^*(z^*) \\ &\leq f(x) + g(Ax) + f^*(x^*) + g^*(y^*) = \langle z^*, x\rangle. \end{aligned}$$

Since we always have $f(x) + g(Ax) + (f(\cdot) + g(A\cdot))^*(z^*) \geq \langle z^*, x\rangle$ it follows that $z^* = A^*y^* + x^* \in \partial(f + g \circ A)(x)$ and so $\partial(f + g \circ A)(x) \supseteq \partial f(x) + A^*\partial g(Ax)$. Now suppose $x^* \in \partial(f + g \circ A)(x)$ then x is the optimal solution of the problem

$$p := \inf_{x \in X} \{f(x) - \langle x^*, x\rangle + g(Ax)\},$$

that is $0 \in \partial(f + g \circ A) - x^*$. Now use the calculus in Proposition 4.1.17 to obtain

$$h^*(z^*) := (f - \langle x^*, \cdot\rangle)^*(z^*) = f^*(x^* + z^*).$$

Applying the Fenchel duality we obtain $d := \sup_{y^* \in Y^*} \{-h^*(A^*y^*) - g^*(-y^*)\}$. When (4.14) holds and $p = d$ we find that if y^* attains the maximum then it follows that $A^*y^* \in \partial h(x) = \partial f(x) - x^*$ and $-y^* \in \partial g(Ax)$. That is

$$\begin{aligned} x^* = A^*y^* + x^* - A^*y^* \in &\partial f(x) + A^*\partial g(Ax) \\ \text{implying} \quad &\partial(f + g \circ A)(x) \subseteq \partial f(x) + A^*\partial g(Ax). \end{aligned}$$

□

Note: Suppose $A^* = I$ the identity operator. The operation $f^* \square g^*$ that takes two convex functions f^* and g^* and produces another convex function $f^* \square g^*$ is referred to as the infimal convolution. Under the qualification assumption $0 \in \text{core}(\text{dom}\, g - \text{dom}\, f)$ the above proof shows that

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*).$$

In finite dimensions (or a Hilbert space) we may reverse the roles of X and X^* and replace f by f^* and g by g^* to obtain

$$(f^* + g^*)^*(x) = (f^{**} \square g^{**})(x) = (f \square g)(x)$$

and so

$$(f \square g)(x) = (f^* + g^*)^*(x).$$

The so called infimal convolution smoothing results on using $g(x) = \frac{\|x\|^2}{\epsilon}$ for a small parameter ϵ.

Appendix C

Closedness of the Infimal Convolution

We prove that the infimal convolution of two closed proper convex functions f and g, denoted by

$$(f^*\square h^*)(x^*) := \inf_{y^*}\{f(y^*) + h(x^* - y^*)\}$$

under the condition

$$0 \in \operatorname{core}(\operatorname{dom} f - \operatorname{dom} h) \tag{C.1}$$

is closed and the infimum is exact i.e. there exists y^* such that

$$(f^*\square h^*)(x^*) := \ f(y^*) + h(x^* - y^*).$$

Proof. We need to use the following observation that a set $H \subseteq \mathbf{R}^{2n}$ is bounded iff its support function is uniformly bounded i.e.

$$\begin{aligned} &H \text{ is bounded} \\ \iff \quad &\exists C(x,y) \text{ so that } S(H,(x,y)) \le C(x,y) < +\infty, \quad \text{for all } (x,y) \in \mathbf{R}^n \times \mathbf{R}^n. \end{aligned}$$

The implication $\Rightarrow$ is obvious (continuous functions on closed bounded sets attain their supremum) and so we only show the $\Longleftarrow$ implication. Suppose the contrary that there exists a norm unbounded $(x_n^*, y_n^*) \in H$. Then

$$\begin{aligned} \|x_n^*\| + \|y_n^*\| &= \sup_{\|x\|=1,\|y\|=1}\{\langle x_n^*, x\rangle + \langle y_n^*, y\rangle\} \\ &\le \sup_{\|x\|=1,\|y\|=1} S(H,(x,y)) \le \sup_{\|x\|=1,\|y\|=1} C(x,y) = C < +\infty \end{aligned}$$

where $\|x_n^*\| + \|y_n^*\| \to \infty$, a contradiction. Now let

$$H(K,r) := \{(x_1^*, x_2^*) \mid f^*(x_1^*) + h^*(x_2^*) \le K, \ \|x_1^* + x_2^*\| \le r\}.$$

We show this is bounded under the assumption of (C.1). Let $(x^*, y^*) \in H(K,r)$ (so that $f^*(x^*) + h^*(y^*) \le K$ and $\|x^* + y^*\| \le r$) and $(x,y) \in \mathbf{R}^{2n}$ then since
$0 \in \operatorname{core}(\operatorname{dom} f - \operatorname{dom} h)$ there exists $\lambda > 0$ such that

$$\frac{1}{\lambda}(x - y) \in \operatorname{dom} f - \operatorname{dom} h$$

and so $u \in \operatorname{dom} f$ and $v \in \operatorname{dom} h$ such that $(x-y) = \lambda(u-v)$. We will also use the Fenchel inequalities

$$\begin{aligned} \langle x^*, u\rangle &\leq f^*(x^*) + f(u) \quad \text{and} \\ \langle y^*, v\rangle &\leq h^*(y^*) + h(v). \end{aligned}$$

Then for all $(x,y) \in \mathbf{R}^{2n}$ (noting λ, u and v are determined by (x,y))

$$\begin{aligned} \langle x^*, x\rangle + \langle y^*, y\rangle &= \lambda\langle x^*, u\rangle + \lambda\langle y^*, v\rangle + \langle x^*+y^*, y-\lambda v\rangle \\ &\leq \lambda(f^*(x^*) + f(u)) + \lambda(h^*(y^*) + h(v)) + \|x^*+y^*\|\,\|y-\lambda v\| \\ &\leq \lambda[K + f(u) \ + h(v)] + r\|y-\lambda v\| := C(x,y). \end{aligned}$$

Thus $H(K,r)$ is bounded. Now we show that $x^* \mapsto (f^*\square h^*)(x^*)$ is lower semi–continuous. Let $x_n^* \to x^*$, $K > 0$, $\varepsilon_n \downarrow 0$ and y_n^* such that

$$f(y_n^*) + h(x_n^* - y_n^*) \leq (f^*\square h^*)(x_n^*) + \varepsilon_n \leq K,$$

and hence

$$(y_n^*, x_n^* - y_n^*) \in H(K, 2\|x^*\|), \quad \text{for } n \text{ large.}$$

Thus $(y_n^*, x_n^* - y_n^*)$ is bounded and so there exists a convergent subsequence $y_{n_m}^* \to y^*$ and consequently

$$\left(y_{n_m}^*, x_{n_m}^* - y_{n_m}^*\right) \to (y^*, x^* - y^*)$$

implying

$$\begin{aligned} (f^*\square h^*)(x^*) &\leq f(y^*) + h(x^*-y^*) \leq \left[\liminf_n f(y_n^*) + \liminf_n h(x_n^* - y_n^*)\right] \\ &\leq \liminf_n [f(y_n^*) + h(x_n^* - y_n^*)] \\ &\leq \liminf_n [(f^*\square h^*)(x_n^*) + \varepsilon_n] = \liminf_n (f^*\square h^*)(x_n^*) \end{aligned}$$

implying $x^* \mapsto (f^*\square h^*)(x^*)$ is lower semi–continuous and $(f^*\square h^*)(x^*) = f(y^*) + h(x^*-y^*)$. □

Appendix D

Proof of Theorem 8.0.4.

Proof. We use the inequalities (8.10a) and (8.10b), adding these, regrouping terms and multiplying by 2 i.e.

$$2(\lambda^{k+1}-\lambda^*)^T r^{k+1}-2\rho(B(z^{k+1}-z^k))^T r^{k+1}+2\rho(B(z^{k+1}-z^k))^T(B(z^{k+1}-z^*)) \leq 0. \quad \text{(D.1)}$$

Substituting $\lambda^{k+1} = \lambda^k \rho r^{k+1}$ into the first term gives

$$2(\lambda^k - \lambda^*)^T r^{k+1} + \rho\|r^{k+1}\|^2 + \rho\|r^{k+1}\|^2, \quad \text{(D.2)}$$

and substituting $r^{k+1} - \frac{1}{\rho}(\lambda^{k+1} - \lambda^k)$ in the first two terms of (D.2) we get

$$\frac{2}{\rho}(\lambda^k - \lambda^*)^T(\lambda^{k+1} - \lambda^k) + \frac{1}{\rho}\|\lambda^{k+1} - \lambda^k\|^2 + \rho\|r^{k+1}\|^2.$$

As $\lambda^{k+1} - \lambda^k = (\lambda^{k+1} - \lambda^*) - (\lambda^k - \lambda^*)$ this can be written as:

$$\frac{1}{\rho}\left(\|\lambda^{k+1} - \lambda^*\|^2\right) - \left(\|\lambda^k - \lambda^*\|^2\right) + \rho\|r^{k+1}\|^2. \quad \text{(D.3)}$$

Taking the last term of (D.3) and the remaining terms of (D.1) i.e.

$$\rho\|r^{k+1}\|^2 - 2\rho(B(z^{k+1} - z^k))^T r^{k+1} + 2\rho(B(z^{k+1} - z^k))^T(B(z^{k+1} - z^*))$$

we substitute

$$z^{k+1} - z^k = (z^{k+1} - z^k) - (z^k - z^*)$$

in the last term gives (on completing the square) that

$$\rho\|r^{k+1} - B(z^{k+1} - z^k)\|^2 + \rho\|B(z^{k+1} - z^k)\|^2 + 2\rho(B(z^{k+1} - z^k))^T(B(z^k - z^*)),$$

and using $z^{k+1} - z^k = (z^{k+1} - z^*) - (z^k - z^*)$ in the last two terms gives

$$\rho\|r^{k+1} - B(z^{k+1} - z^k)\|^2 + \rho\left(\|B(z^{k+1} - z^*)\|^2 - \|B(z^k - z^*)\|^2\right).$$

Combining all of these we can rewrite (D.1) as

$$\begin{aligned} V^k - V^{k+1} &\geq \rho\|r^{k+1} - B(z^{k+1} - z^k)\|^2 \\ &= \rho\|r^{k+1}\|^2 - 2\rho B(z^{k+1} - z^k)^T r^{k+1} + \rho\|B(z^{k+1} - z^k)\|^2. \end{aligned} \quad \text{(D.4)}$$

We are finished once we have shown that $\rho B(z^{k+1} - z^k)^T r^{k+1} \leq 0$ and we can then drop this form the right hand side of (D.4). To this end note that z^{k+1} minimizes

$g(z) + \lambda^{k+1^T} Bz$ and z^k minimizes $g(z) + \lambda^k Bz$ so we can add the following two inequalities

$$g(z^{k+1}) + (\lambda^{k+1})^T Bz^{k+1} \leq g(z^k) + (\lambda^{k+1})^T Bz^k$$
$$g(z^k) + (\lambda^k)^T Bz^k \leq g(z^{k+1}) + (\lambda^k)^T Bz^{k+1}.$$

to get the desired inequality

$$(\lambda^{k+1} - \lambda^k)^T B(z^{k+1} - z^k) \leq 0$$

□

Bibliography

[1] M. S. Bazaraa, C. M. Shetty, *Nonlinear Programming*, Theory and Algorithms, second edition, John Wiley and Sons, 1979.

[2] H.H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* 2nd edition, Springer 2017.

[3] D. Bertsekas, *Nonlinear Programming*, first edition, Athena Scientific, 1999.

[4] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization; Theory and Examples,* Springer-Verlag, 2000.

[5] S. Boyd and L. vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] Brézis, Haim, *Functional analysis, Sobolev spaces and partial differential equations*, Vol. 2. No. 3. New York: Springer, 2011.

[7] P. Camerini, L. Fratta and F. Maffioli, On improving relaxation method by modifying gradient techniques. Math. Programming Study, vol. 3, pp. 26-34, ,1975.

[8] F.H.Clarke, Yu.S Ledyaev, R.J. Stern and P.R. Wolenski, (1998) *Anonsmooth Analysis and Control Theorey,* Graduate texts in Mathematics, Springer.ond edition, John Wiley and Sons, 1979.

[9] J. Duchi, Shai Shalev-Shwartz and Y.S. Tushar Chandra, (2008), Efficient Projections onto the l^1-Ball for Learning in High Dimensions, Proceedings of the 25 th International Conference on Machine Learning, Helsinki, Finland, 2008

[10] Jonathan Eckstein and Wang Yao, Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. Pac. J. Optim., 11(4):619–644, 2015.

[11] R. Fletcher, *Practical Methods of Optimization*, Vol. 1 and 2, John Wiley and Sons, 1985.

[12] D. Gabay. Applications of the method of multipliers to variational inequalities. In Studies in mathematics and its applications, volume 15, chapter ix, pages 299–331. Elsevier, 1983.

[13] P. Gill, W. Murray and M. Wright, *Practical Optimization*, Academic Press 1981.

[14] T. Hastie, R TYibshirani and J. Friedman. *The Elements of Statistical Learning Theory,* Springer Series in Statistics, second editions 2008.

[15] J-B Hiriart-Urrity and C. Lemarechal, *Convex analysis and Minimization Algorithms I*, Springer–Verlag A series of comprehensive studies in mathematics 305, 1991.

[16] C.T. Kelly, *Iterative Methods for Optimization*, Frontiers in Applied Mathematics, Philadelphia, SIAM, 1999.

[17] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research, 1999.

[18] *Handbook of Applied Optimization*, Eds P. Pardalos and M. resende, Oxford Univ. Press, 2002.

[19] B. R. Hundt, R. L. Lipsman and J. M. Rosenberg, *A guide to MATLAB, for beginners and experienced users*, second ed., Cambridge Univ. Press, 2006.

[20] R. Holmes *Geometric Functional Analysis,* Springer-Verlag, Graduate Text in Maths 24, New York Heildelberg Berlin, 1975.

[21] S. Shalev-Shwartz and Y. Singer, (2006, Efficient Learning of Label Ranking by Soft Projections onto Polyhedra, Journal of Machine Learning Research 7 (2006) 1567–1599.

[22] K. Scheinberg, S. Ma, and D. Goldfarb, "Sparse inverse covariance selection via alternating linearization methods," in Advances in Neural Information Processing Systems, 2010.

[23] G. Strang, *Linear algebra and its applications*. Belmont, CA: Thomson, Brooks/Cole, 2006.

[24] S. J. Wright, Primal–Dual Interior Point Methods, SIAM, 1997.