

# IB Optimisation

Martin von Hodenberg (mjv43@cam.ac.uk)

December 20, 2021

Optimisation concerns problems of maximising or minimising certain functions. This extends to several different types of functions and is applicable in many areas of mathematics, particularly computational mathematics.

This article constitutes my notes for the ‘IB Optimisation’ course, held in Easter 2021 at Cambridge. This course was lectured by Prof. Varun Jog.

## Contents

<b>0</b>	<b>Introduction</b>	<b>2</b>
<b>1</b>	<b>Convex optimisation</b>	<b>2</b>
1.1	Convexity . . . . .	2
1.2	Unconstrained optimisation . . . . .	3
1.2.1	First order conditions for convexity . . . . .	3
1.2.2	Second-order conditions for convexity . . . . .	4
1.3	The gradient descent algorithm . . . . .	5
1.3.1	Smoothness assumption . . . . .	5
1.3.2	Strong convexity assumption . . . . .	7

## §0 Introduction

The course is roughly divided into four parts:

1. Convex optimisation (Lectures 1-3)
2. Lagrange method (Lectures 4-5)
3. Linear programming (Lectures 6-8)
4. Applications of linear programming (Lectures 9-12)

### Definition 0.1 (Optimisation problem)

The basic structure of an optimisation problem is as follows:

minimise  $f(x)$  for  $x \in X$ , where in this course we take  $X \subset \mathbb{R}^n$ . The problem can have 'constraints'  $h(x) = b$ , where  $h : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . We will look at minimising functions WLOG.

1.  $f$  is called the objective function.
2. The components of  $x$  are called decision variables.
3.  $h(x) = b$  is called a functional constraint.
4.  $x \in X$  is called a regional constraint.
5.  $\{x : x \in X\}$  is called the feasible set  $X(b)$ .
6. Problem is feasible if  $X(b) \neq \emptyset$ , and bounded if the minimum on  $X(b)$  is bounded.
7. A point  $x^* \in X(b)$  is optimal if it minimises  $f$  over  $X(b)$ . The value  $f(x^*)$  is called the optimal set.

## §1 Convex optimisation

### §1.1 Convexity

#### Definition 1.1

A set  $S \subset \mathbb{R}^n$  is **convex** if for all  $x, y \in S$ , and all  $\lambda \in [0, 1]$ ,  $(1 - \lambda)x + \lambda y \in S$ . (The line segment joining  $x$  and  $y$  lies in  $S$ .)

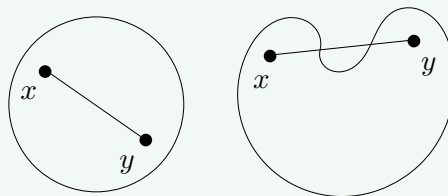


Figure 1: Two subsets of  $\mathbb{R}^2$ : the left one is convex but the right one is not.

**Definition 1.2**

A function  $f : S \rightarrow \mathbb{R}$  is **convex** if  $S$  is convex, and for all  $x, y \in S$ , and  $\lambda \in [0, 1]$ , we have

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

Intuitively, all the tangent lines to the function are below the function. We define  $f$  to be concave if  $-f$  is convex.

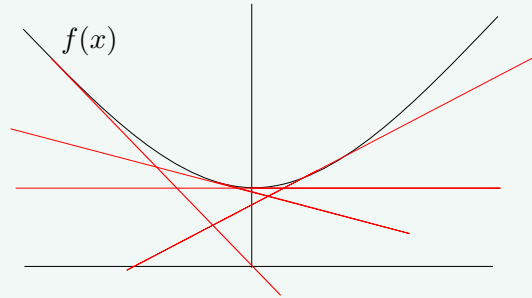


Figure 2: A convex function in  $\mathbb{R}^2$ .

**Remark.** Linear functions are always both convex and concave, since we get equality in the above equation.

**§1.2 Unconstrained optimisation**

We want to find  $\min f(x)$  where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex function. This is a particularly simple case of optimisation, since convex functions have the important property that *local information extends globally*. We will now look at a couple of ways to prove convexity, called the first- and second-order conditions.

**§1.2.1 First order conditions for convexity**

Intuitively, for a convex function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , we know that all the tangent lines to the function are below the function (see Figure 2). We can express this as

$$f(y) \geq f(x) + (y - x)f'(x).$$

We can generalise this to higher dimensions:

**Theorem 1.3 (First-order conditions)**

A differentiable  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex iff for  $x, y \in \mathbb{R}^n$ , we have

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

**Remark.** If  $\nabla f(x) = 0$ , then  $f(y) \geq f(x) \implies x$  minimises  $f$ .

*Proof.* **Convexity of  $f \implies$  F-O conditions hold:**

Let  $n = 1$ . For  $x, y \in \mathbb{R}^n$ ,  $t \in [0, 1]$  we have

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

Therefore

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

Taking  $t \rightarrow 0$ , we conclude

$$f(y) \geq f(x) + f'(x)(y - x).$$

For the general case, define a function  $g(t) = f((1 - t)x + ty)$ , i.e  $g : [0, 1] \rightarrow \mathbb{R}$ . Since  $f$  is convex,  $g$  is also convex. We can calculate

$$g'(t) = \nabla f((1 - t)x + ty)^T(y - x).$$

Since  $g$  is convex, by the above argument for  $n = 1$ , we have  $g(1) \geq g(0) + g'(0)(1 - 0)$ .

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

**F-O conditions hold  $\implies$  convexity:**

Define  $x_t = (1 - t)x + ty$ . We have the two equations

$$f(x) \geq f(x_t) + \nabla f(x_t)^T(x - x_t) \tag{1}$$

$$f(y) \geq f(x_t) + \nabla f(x_t)^T(y - x_t) \tag{2}$$

Now we add  $(1 - t) \times$  equation (1) to  $t \times$  equation 2. Noting that  $x - x_t = tx - ty$  and  $y - x_t = (1 - t)y - (1 - t)x$ , we get

$$(1 - t)f(x) + tf(y) \geq f(x_t) + \nabla f(x_t)^T(0).$$

This concludes the proof. □

### §1.2.2 Second-order conditions for convexity

Next we look at second-order conditions. In one dimension, we would expect that if  $f'$  is increasing, i.e  $f''(x) \geq 0$  would give us convexity. We generalise this to  $n$  dimensions using the Hessian matrix, which we saw in IA Differential Equations. It has entries  $H_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ .

#### Definition 1.4 (Semidefinite matrix)

A matrix  $A$  is called **positive semidefinite** if for all  $x \in \mathbb{R}^n$  we have  $x^T A x \geq 0$ . Equivalently, all eigenvalues of  $A$  are non-negative.

**Remark.** The Taylor expansion in  $n$  dimensions is

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

#### Theorem 1.5 (Second-order condition)

A twice differentiable  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex if  $\nabla^2 f(x) \geq 0$  for all  $x \in \mathbb{R}^n$ .

*Proof.* Using the Taylor expansion of  $f$  we have

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x).$$

where  $z = (1 - \lambda)x + \lambda y$  for some  $\lambda \in [0, 1]$ . (IA Analysis).

We have  $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ . Then by the first-order condition theorem,  $f$  is convex.  $\square$

### §1.3 The gradient descent algorithm

Recall we have an unconstrained optimisation problem:

$$\text{minimise } f(x), f : \mathbb{R}^n \rightarrow \mathbb{R} \text{ is a convex function.}$$

In order to solve this, we could apply a 'greedy method' where we pick an initial point  $x_0$  and iteratively pick points  $x_1, x_2, \dots$  near this point so that the gradient of  $f$  at  $x_n$  is decreasing. In which direction should we decrease  $x_n$ ? Note that by Taylor's expansion, with small  $\epsilon > 0$ ,

$$f(x - \epsilon \nabla f(x)) \approx f(x) - \epsilon \nabla f(x)^T \nabla f(x) = f(x) - \epsilon \|\nabla f(x)\|^2 \leq f(x).$$

We call the direction  $-\nabla f(x)$  the descending direction. (Note we could also choose any other direction  $v$  so that  $f(x)^T v < 0$ .) We now formalise this process in an algorithm.

#### Definition 1.6 (Gradient descent method)

Define an algorithm:

1. Start at some point  $x_0$ . Set  $t = 0$ .
2. Repeat the following:
  - Find a descending direction  $v_t$  (e.g.  $-\nabla f(x)$ )
  - Choose a step size  $\eta_t$  (can depend on  $t$ )
  - Update  $x_{t+1} = x_t + \eta_t v_t$
3. Stop when some criterion is met (e.g.  $\nabla f(x_t) = 0$ ,  $t$  is large enough)

For the scope of this course, we need to make assumptions about our convex function.

#### §1.3.1 Smoothness assumption

##### Definition 1.7 (Smoothness assumption)

We say that a continuously differentiable function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is  $\beta$ -smooth or  $\beta$ -Lipschitz if

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|.$$

Those who have taken IB Analysis and Topology will be familiar with this concept.

**Remark.** If  $f$  is twice-differentiable, then  $\beta$ -smoothness implies that  $\nabla^2 f(x) \leq \beta I$ . Equivalently, all eigenvalues of  $\nabla^2 f(x)$  are  $\leq \beta$ .  $\beta$ -smoothness can be summed up by the fact that *the linear approximation is close to  $f$  in a small neighbourhood around  $x$ .*

**Proposition 1.8**

If  $f$  is  $\beta$ -smooth and convex, then

$$f(x) + \nabla f(x)^T(y - x) \leq f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|x - y\|^2.$$

*Proof.* The left-hand inequality follows by convexity.

For the right hand inequality, by Taylor's theorem,

$$\begin{aligned} f(y) &= \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \\ &\leq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T(\beta I)(y - x) \\ &= f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|x - y\|^2 \end{aligned}$$

□

**Corollary 1.9**

At any point  $x$ ,

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

*Proof.* Let's look at

$$f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|x - y\|^2,$$

and try to minimise it over  $y$  for a fixed  $x$ .

$$\nabla_y (f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|x - y\|^2) = \nabla f(x) - \beta(x - y) = 0.$$

$$\implies \nabla \frac{\nabla f(x)}{\beta} = x - y \implies y = x - \frac{1}{\beta} \nabla f(x).$$

Plug this value of  $y$  into the above claim:

$$\begin{aligned} f(x - \frac{1}{\beta} \nabla f(x)) &\leq f(x) + \nabla f(x)^T(\frac{-1}{\beta} \nabla f(x)) + \frac{\beta}{2} \|\frac{1}{\beta} \nabla f(x)\|^2 \\ &= f(x) - \frac{1}{\beta} \|\nabla f(x)\|^2 + \frac{1}{2\beta} \|\nabla f(x)\|^2 \\ &= f(x) - \frac{1}{2\beta} \|\nabla f(x)\|^2 \end{aligned}$$

□

**Corollary 1.10** (Improved first-order condition)

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

*Proof.* For any  $z$ , we have

$$f(x) + \nabla f(x)^T(z - x) \leq f(z) \leq f(y) + \nabla f(y)^T(z - y) + \frac{\beta}{2}\|z - y\|^2.$$

This implies

$$f(x) - f(y) \leq \nabla f(x)^T(x - z) + \nabla f(y)^T(z - y) + \frac{\beta}{2}\|z - y\|^2.$$

To minimise the RHS, set  $\nabla_z = 0$ . We get  $-\nabla f(x) + \nabla f(y) + \beta(z - y) = 0$ , which implies

$$z = \frac{f(x) - f(y)}{\beta} + y.$$

Subbing into the RHS, we get

$$f(x) - f(y) \leq \nabla f(x)^T(x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2.$$

□

### §1.3.2 Strong convexity assumption

We now introduce a new type of convexity we assume. In essence, it tells us that *if the gradient is small, we are close to the optimum*.

#### Definition 1.11 (Strong convexity assumption)

A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\alpha$ -strongly convex if

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\alpha}{2}\|x - y\|^2.$$

If  $f$  is twice differentiable, then

$$\nabla^2 f(x) \geq \alpha I \quad \forall x.$$

#### Proposition 1.12

Let  $f$  be  $\alpha$ -strongly convex. Let  $p^*$  be the optimal cost. Then for any  $x$  we have

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

**Remark.** If  $\|\nabla f(x)\| \leq \sqrt{2\alpha\epsilon}$  then  $p^* \geq f(x) - \epsilon$ , i.e

$$p^* \leq f(x) \leq p^* + \epsilon.$$

*Proof.* The  $\alpha$ -strong convexity assumption gives us

$$\min_y f(y) \geq \min_y f(x) + \nabla f(x)^T(y - x) + \frac{\alpha}{2}\|x - y\|^2.$$

Setting  $\nabla_y$  of the RHS=0, we get  $y = x - \frac{\nabla f(x)}{\alpha}$ .

Plugging this value of  $y$  in the RHS, we get

$$f(x) + \nabla f(x)^T \left( \frac{-\nabla f(x)}{\alpha} \right) + \frac{\alpha}{2} \left\| \frac{\nabla f(x)}{\alpha} \right\|^2 = f(x) - \frac{1}{2\alpha} \|\nabla f(x)\|^2.$$

□

We might also want to ask: how close is the true optimiser  $x^*$  to our current point  $x$ ? It turns out  $\alpha$ -strong convexity gives us an answer to this too.

### Proposition 1.13

Let  $x \in S$ , and  $x^*$  be the true optimiser. Then

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

*Proof.*

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)(x^* - x) + \frac{\alpha}{2} \|x - x^*\|^2 \\ &\geq f(x) - \|\nabla f(x)\| \|x^* - x\| + \frac{\alpha}{2} \|x - x^*\|^2 \text{ by Cauchy-Schwartz} \end{aligned}$$

We already know that  $f(x^*) \leq f(x)$ . Therefore

$$0 \geq f(x^*) - f(x) \geq -\|\nabla f(x)\| \|x^* - x\| + \frac{\alpha}{2} \|x - x^*\|^2.$$

So  $\|\nabla f(x)\| \|x - x^*\| \leq \frac{2}{\alpha} \|\nabla f(x)\|^2$ , and thus

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

□