Lecture Notes for GPED Summer Math Camp
Highway to KKT

Qiaohairuo Lin

August 6, 2024

Modern economics features various optimization problems: households maximizing their utility, firms minimizing their production cost, investors maximizing their expected return... More precisely, we call them **constrained optimization problems**. The intension between ”constraint” and ”optimization” points to the core of economics – allocation of resources under scarcity. This minicourse is aimed at preparing you for essential math prerequsite for handling optimization problems you may meet in future economics courses. This lecture note is a self-contained introduction to the fundamental mathematical method to address constrained optimization problems – Karush–Kuhn–Tucker (KKT)conditions. It is a concise version of the following textbook:

> Beck, A. (2014). *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*

The author of the book is an expert on operation research, and the optimization problems in the book is regularized to *minimization* problem. While in economics you will more frequently be met with *maximization* problems, although they are trivially equivalent. Thus, I follow mostly the textbook in most part of the lectures except the last part where I consider a maximization problem with a concave objective function on a convex space. The basic structure of this lecture note is as follows: firstly, we will go through some basic mathematical knowledge with rather rigorous treatment. Then we *take the short route* to prove the validity of KKT for constrained optimization problems. On the one hand, this lecture note will be a bit ”mathematical” in the sense that its focus is on *proving* the validity behind the algorithm, while some examples and pratical guides are also provided in the end of notes. On the other hand, I try to be concise in that by imposing specific condition that is usually satisfied in economic studies, the proof could be really simple and intuitive, circumventing the regular second order condition that has heavy algebra. While we begin with ”mathematical preliminaries”, I assume that the students have already taken basic calculus course in their undergraduate institution. If not, we can still cover some more fundamental concepts in the beginning of the course and supplement related materials.

1 Mathematical Preliminaries

1.1 The Space $\mathbb{R}^n$

The vector space $\mathbb{R}^n$ is the set of n-dimensional column vectors with real components endowed with the component-wise addition operator

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

and the scalar-vector product

$$\lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}$$

where in the above $x_1, x_2, ..., x_n, \lambda$ are real numbers. Throughout this mini-course we are mainly interested in problems over $\mathbb{R}^n$

Important Subset of $\mathbb{R}^n$ The *nonnegative orthant* is the subset of $\mathbb{R}^n$ consisting of all vectors in $\mathbb{R}^n$ with nonnegative components and is denoted by $\mathbb{R}^n_+$:

$$\mathbb{R}^n_+ = \left\{ (x_1, x_2, ..., x_n)^T : x_1, x_2, ..., x_n \geq 0 \right\}$$

Similarly, the *positive orthant* consists of all the vectors in $\mathbb{R}^n$ with positive components and is denoted by $\mathbb{R}^n_{++}$:

$$\mathbb{R}^n_{++} = \left\{ (x_1, x_2, ..., x_n)^T : x_1, x_2, ..., x_n > 0 \right\}$$

For a given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the *closed line segment* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is a subset of $\mathbb{R}^n$ denoted by $[\boldsymbol{x}, \boldsymbol{y}]$ and defined as

$$[\boldsymbol{x}, \boldsymbol{y}] = \{\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}) : \alpha \in [0, 1]\}$$

The *open line segment* $[\boldsymbol{x}, \boldsymbol{y}]$ is similarly defined as

$$(\boldsymbol{x}, \boldsymbol{y}) = \{\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}) : \alpha \in (0, 1)]\}$$

when $\boldsymbol{x} \neq \boldsymbol{y}$ and is the empty set when $\boldsymbol{x} = \boldsymbol{y}$. The *unit simplex*, denoted by Δ_n, is the subset of $\mathbb{R}^n$ comprising all non-negative vectors whose sum is 1.

1.2 The Space $\mathbb{R}^{m \times n}$

The set of all real-valued matrices of order $m \times n$ is denoted by $\mathbb{R}^{m \times n}$. Some special matrices that will be frequently used are the $n \times n$ identity matrix denoted by $\boldsymbol{I}_n$ and the $m \times n$ zeros matrix denoted by $0_{m \times n}$. For a special class of matrices – square and symmetric matrix on $\mathbb{R}^{n \times n}$, an important property is used frequently below.

Definition 1.1. *(positive definiteness)*

1. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called* ***positive semidefinite****, denoted by $A \succcurlyeq 0$, if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for every $\boldsymbol{x} \in \mathbb{R}^n$.*
2. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called* ***positive definite****, denoted by $A \succ 0$, if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for every $\boldsymbol{x} \in \mathbb{R}^n$.*

Similarly, the symmetric square matrix A is called **negative semidefinite/definite** if $-A$ is positive semidefinite/definite. There are multiple methods in linear algebra to test the definiteness of matrix. But in this mini-course

1.3 Inner Products and Norms

Inner Products We begin with the formal definition of an inner product.

Definition 1.2. *(inner product) An* ***inner product*** *on $\mathbb{R}^n$ is a map $\langle\cdot,\cdot\rangle : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}$ with the following properties:*

1. ***(symmetry)*** *$\langle\boldsymbol{x},\boldsymbol{y}\rangle = \langle\boldsymbol{y},\boldsymbol{x}\rangle$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$.*
2. ***(additivity)*** *$\langle\boldsymbol{x},\boldsymbol{y}+\boldsymbol{z}\rangle = \langle\boldsymbol{x},\boldsymbol{y}\rangle + \langle\boldsymbol{x},\boldsymbol{z}\rangle$ for any $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathbb{R}^n$.*
3. ***(homogeneity)*** *$\langle\lambda\boldsymbol{x},\boldsymbol{y}\rangle = \lambda\langle\boldsymbol{x},\boldsymbol{y}\rangle$ for any $\lambda \in \mathbb{R}$ and $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$.*
4. ***(positive definiteness)*** *$\langle x,x\rangle \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\langle x,x\rangle = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$.*

Example 1.3. *Perhaps the most widely used inner product is the so-called dot product defined by*

$$\langle\boldsymbol{x},\boldsymbol{y}\rangle = \boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i \quad \textit{for any } \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n.$$

Since this is in a sense the ”standard” inner product, we will by default assume–unless explicitly stated otherwise–that the underlying inner product is dot product.

Example 1.4. *The dot product is not the only possible inner product on $\mathbb{R}^n$. For example, let $\boldsymbol{w} \in \mathbb{R}^n_{++}$. Then it is easy to show that the following weighted dot product is also an inner product:*

$$\langle\boldsymbol{x},\boldsymbol{y}\rangle_{\boldsymbol{w}} = \sum_{i=1}^{n} w_i x_i y_i$$

Vector Norms

Definition 1.5. *(norm) A* ***norm*** *$\|\cdot\|$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfying the following:*

1. ***(nonnegativity)*** *$\|\boldsymbol{x}\| \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$.*
2. ***(positive homogeneity)*** *$\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$ for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^n$.*
3. ***(triangle inequality)*** *$\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$.*

One natural way to generate a norm on $\mathbb{R}^n$ is to take any inner product $\langle\cdot, cdot\rangle$ on $\mathbb{R}^n$ and define the associated norm

$$\|x\| \equiv \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} \text{ for all } \boldsymbol{x} \in \mathbb{R}^n$$

which can be easily seen to be a norm. If the inner product is the dot product, then the associated norm is the so-called *Euclidean norm* or l_2 *norm*:

$$\|x\|_2 \equiv \sqrt{\sum_{i=1}^{n} x_i^2} \text{ for all } \boldsymbol{x} \in \mathbb{R}^n$$

By default, the underlying norm on $\mathbb{R}^n$ is $\|\cdot\|$, anbd the subscript 2 will be frequently omitted. The Euclidean norm belongs to the class of l_p norms (for $p \geq 1$) defined by

$$\|\boldsymbol{x}\|_p \equiv \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

1.4 Convergence and Continuity

With the definition of norms, it immediately follows to define **distance** between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$$

Then we can extend the notion of **convergence** of one-dimensional real-valued sequence to this setting.

Definition 1.6. *(convergence of sequence) The sequence $\{\boldsymbol{x}_i\} \subseteq \mathbb{R}^k$ is said to converge to $\boldsymbol{x}^* \in \mathbb{R}^k$ if*

$$\textit{for each } \varepsilon > 0, \ \textit{there is an } N \in \mathbb{N} \textit{ such that } d(\boldsymbol{x}_n - \boldsymbol{x}^*) < \varepsilon \textit{ whenever } n > N.$$

And we write it as $\boldsymbol{x}_i \to \boldsymbol{x}^*$.

With the definition of convergence, we are able to define the property of **continuity** for function f as follows:

Definition 1.7. *continuity of function Let $X \subseteq \mathbb{R}^d$, then the function $f : X \to \mathbb{R}^k$ is* ***continuous*** *on point $\boldsymbol{x}^* \in X$ if for any convergent sequence $\{\boldsymbol{x}_i\} \subseteq X$ with $\boldsymbol{x}_i \to \boldsymbol{x}^*$, we have $f(\boldsymbol{x}_i) \to f(\boldsymbol{x}^*)$ in $\mathbb{R}^k$. If f is continuous on all the points in X, we call f* ***continuous on*** *X.*

1.5 Basic Topology

We begin with the definition of a ball

Definition 1.8. *(open ball, closed ball) The* ***open ball*** *with center $\boldsymbol{c} \in \mathbb{R}^n$ and radius r is denoted by $B(\boldsymbol{c}, r)$ and defined by*

$$B(c, r) = \{x \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{c}\| < r\}.$$

The ***closed ball*** *with center $\boldsymbol{c}$ and radius r is denoted by $B[c, r]$ and defined by*

$$B[c, r] = \{x \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{c}\| \leq r\}$$

The ball $B(c, r)$ for some arbitrary $r > 0$ is also referred to as a *neighborhood* of $\boldsymbol{c}$. The first topological notion we define is that of an interior point of a set. This is a point which has a neighborhood contained in the set.

Definition 1.9. *(interior points) Given a set $U \subseteq \mathbb{R}^n$, a point $\boldsymbol{c} \in U$ is an* ***interior point*** *of U if there exists $r > 0$ for which $B(\boldsymbol{c}, r) \subseteq U$.*

The set of all interior points of a given set U is called the *interior* of the set and is denoted by $int(U)$:

$$int(U) = \{\boldsymbol{x} \in U : B(\boldsymbol{x}, r) \subseteq U \text{ for some } r > 0\}$$

Definition 1.10. *(open sets) An* ***open set*** *is a set that contains only interior points. In other words, $U \subseteq \mathbb{R}^n$ is an open set if*

$$\textit{for every } \boldsymbol{x} \in U \textit{ there exists } r > 0 \textit{ such that } B(\boldsymbol{x}, r) \subseteq U.$$

Definition 1.11. *(closed sets) A set $U \subseteq \mathbb{R}^n$ is said to be* ***closed*** *if it contains all the limits of convergent sequences of points in U: that is, U is closed if for every sequence of points $\{\boldsymbol{x}_i\} \subseteq U$ satisfying $\boldsymbol{x}_i \to \boldsymbol{x}^*$ as $i \to \infty$, it holds that $\boldsymbol{x}^* \in U$.*

Definition 1.12. *(closedness of level and contour sets of continuous functions) Let f be a continuous function defined over a closed ser $S \subseteq \mathbb{R}^n$. Then for any $\alpha \in \mathbb{R}$ the sets*

$$Lev(f, \alpha) = \{\boldsymbol{x} \in S : f(\boldsymbol{x} \leq \alpha\},$$
$$Con(f, \alpha) = \{\boldsymbol{x} \in S : f(\boldsymbol{x} = \alpha\}$$

are closed

Definition 1.13. *(boundedness and compactness)*

1. *A set $U \subseteq \mathbb{R}^n$ is called* ***bounded*** *if there exists $M > 0$ for which $U \subseteq B(\mathbf{0}, M)$.*
2. *A set $U \subseteq \mathbb{R}^n$ is called* ***compact*** *if it is closed and bounded*

Theorem 1.14. *(Weierstrass theorem) Let f be a continuous function defined over a nonempty and compact set $C \subseteq \mathbb{R}^n$. Then there exists a global minimum point of f over C and a global maximum point of f over C.*

1.6 Convex Set in $\mathbb{R}^n$

Definition 1.15. *(convex set) A set $C \subseteq \mathbb{R}^n$ is called convex if for any $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\lambda \in [0, 1]$, the point $\lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y}$ belongs to C*

The above definition is equivalent to saying that for any $\boldsymbol{x}, \boldsymbol{y} \in C$, the line segment $[\boldsymbol{x}, \boldsymbol{y}]$ is also in C. Examples of convex and nonconvex sets in $\mathbb{R}^2$ are illustrated in the following figure. We will now show some basic examples of convex sets.

Example 1.16. *(convex sets) Let $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ and $b \in \mathbb{R}$. The following sets in $\mathbb{R}^n$ are convex:*

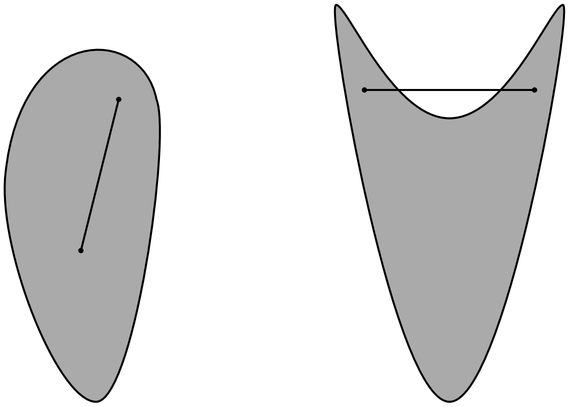

Figure 1.1: Convex and nonconvex sets

1. *a line:* $L = \{\boldsymbol{z} + t\boldsymbol{d} : t \in \mathbb{R}\}$,
2. *a hyperplane:* $H = \left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} = b\right\}$,
3. *a half-space:* $H = \left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} \leq b\right\}$,
4. *an open half-space:* $H = \left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{a}^T\boldsymbol{x} < b\right\}$.

Theorem 1.17. *(preservation of convexity under intersection) Let $C_1, C_2, ..., C_m \subseteq \mathbb{R}^n$ be convex sets, then the set $\cap_{i=1}^m C_i$ is convex.*

1.7 Differentiability

Let f be a function defined on a set $S \subseteq \mathbb{R}^n$; Let $\boldsymbol{x} \in int(S)$ and let $\mathbf{0} \neq \boldsymbol{d} \in \mathbb{R}^n$. If the limit

$$\lim_{t \to 0^+} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t}$$

exists, then it is called the *directional derivativce* of f at $\boldsymbol{x}$ along the direction $\boldsymbol{d}$ and is denoted by $f'(\boldsymbol{x}; \boldsymbol{d})$. For any $i = 1, 2, ..., n$, the directional derivative at $\boldsymbol{x}$ along the direction $\boldsymbol{e}_i$ (the ith vector in the standard basis) is called the *ith partial derivative* and is denoted by $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$:

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = \lim_{t \to 0^+} \frac{f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x})}{t}$$

If all the partial derivatives of a function f exists at a point $\boldsymbol{x} \in \mathbb{R}^n$, then the *gradient* of f at $\boldsymbol{x}$ is defined to be the column vector consisting of all the partial derivatives:

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \frac{\partial f}{\partial x_2}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}.$$

A function f is defined on an open set $U \subseteq \mathbb{R}^n$ is called *continuously differentiable* over U if all the partial derivatives exist and are continuous on U. The definition of continuous differentiability can also be extended to nonopen sets by using the convention that a function f is said to be continuously differentiable over a set C if there exists an open set U containing C on which the function is also defined and continuously differentiable. In the setting of continuous differentiablity, we have the following important formula for the directional derivative:

$$f'(\boldsymbol{x};\boldsymbol{d}) = \nabla f(\boldsymbol{x})^T\boldsymbol{d}$$

for all $\boldsymbol{x} \in U$ and $\boldsymbol{d} \in \mathbb{R}^n$. It can also be shown in this setting of continuous differentiability that the following approximation result holds.

Proposition 1.18. *Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is continuously differentiable over U. Then*

$$\lim_{\boldsymbol{d}\to 0} \frac{f(\boldsymbol{x}+\boldsymbol{d}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T\boldsymbol{d}}{\|\boldsymbol{d}\|} = 0 \textit{ for all } \boldsymbol{x} \in U$$

Another way to write the above result is as follows:

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + o(\|\boldsymbol{y}-\boldsymbol{x}\|),$$

where $o(\cdot) : \mathbb{R}^n_+ \to \mathbb{R}$ is a one-dimensional function satisfying $\frac{o(t)}{t} \to 0$ as $t \to 0^+$. A function f defined on an open set $U \subseteq \mathbb{R}^n$ is called *twice continuously differentiable* over U if all the second order partial derivatives exist and are continuous over U. Under the assumption of twice continuous differentiability, the second order partial derivatives are symmetric, meaning that for any $i \neq j$ and any $\boldsymbol{x} \in U$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{x}).$$

The *Hessian* if f ar a point $\boldsymbol{x} \in U$ is the $n \times n$ matrix

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

where all the second order partial derivatives are evaluated at $\boldsymbol{x}$. Since f is twice continuously differentiable over U, the Hessian matrix is symmetric. There are two main approximation results (linear and quadratic) which are direct consequences of Taylor's approximation theorem that will be used frequently in the mini-course and are thus recalled here.

Theorem 1.19. *(linear approximation theorem) Let $f : U \to \mathbb{R}$ be a twice continuously differentiable function over an open set $U \subseteq \mathbb{R}^n$, and let $\boldsymbol{x} \in U, r > 0$ satisfy $B(\boldsymbol{x}, r) \subseteq U$. Then for any $\boldsymbol{y} \in B(\boldsymbol{x}, r)$, there exists $\xi \in [\boldsymbol{x}, \boldsymbol{y}]$ such that*

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

Theorem 1.20. *(quadratic approximation theorem)Let $f : U \to \mathbb{R}$ be a twice continuously differentiable function over an open set $U \subseteq \mathbb{R}^n$, and let $\boldsymbol{x} \in U, r > 0$ satisfy $B(\boldsymbol{x}, r) \subseteq U$. Then for any $\boldsymbol{y} \in B(\boldsymbol{x}, r)$,*

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

2 Optimality Conditions for Unconstrained Optimization

2.1 Global and Local Optima

Although our main interest in this section is to discuss minimum and maximum points of a function over the entire space, we will nonetheless present the more general definition of global minimum and maximum points of a function over a given set.

Definition 2.1. *(global and minimum and maximum) Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$. Then*

1. *$\boldsymbol{x}^* \in S$ is called a **global minimum point** of f if $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \in S$*
2. *$\boldsymbol{x}^* \in S$ is called a **strict global minimum point** of f if $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \neq \boldsymbol{x}^* \in S$*
3. *$\boldsymbol{x}^* \in S$ is called a **global maximum point** of f if $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \in S$*
4. *$\boldsymbol{x}^* \in S$ is called a **strict global maximum point** of f if $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$ for any $\boldsymbol{x} \neq \boldsymbol{x}^* \in S$*

The set S on which the optimization off is performed is also called the *feasible set*, and any point $\boldsymbol{x} \in S$ is called a *feasible solution.* We will frequently omit the adjective "global" and just use the terminology "minimum point" and "maximum point." It is also customary to refer to a global minimum point as a *minimizer* or a *global minimizer* and to a global maximum point as a *maximizer* or a *global maximizer.* A vector $\boldsymbol{x}^* \in S$ is called a *global optimum* of f over S if it is either a global minimum or a global maximum. The *maximal value* of f over S is defined as the supremum off over S:

$$max\,\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\} = sup\,\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$

Similarly the *minimal value* of f over S is the infimum of f over S,

$$min\,\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\} = inf\,\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$

and is equal to $f(\boldsymbol{x}^*$ when $\boldsymbol{x}^*$ is a global minimum of f over S. Note that the maximum or minimum may not be actually attained. As opposed to global maximum and minimum points, minimal and maximal values are always unique. There could be several global minimum points, but there could be only one minimal value. The set of all global minimizers o f f over S is denoted by

$$argmin\,\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$

and the set of all global maximizers of f over S is denoted by

$$argmax\,\{f(\boldsymbol{x}) : \boldsymbol{x} \in S\}$$

Example 2.2. *Consider the two-dimensional function*

$$f(x,y) = \frac{x+y}{x^2+y^2+1}$$

defined over the entire space $\mathbb{R}^2$*. The surface plot of the function are given in the following figure. The function has two optima points: a global maximizer* $(x,y) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ *and a global minimizer* $(x,y) = (-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$*. The maximal value of the function is* $\frac{1}{\sqrt{2}}$ *and the minimal value is* $-\frac{1}{\sqrt{2}}$*.*

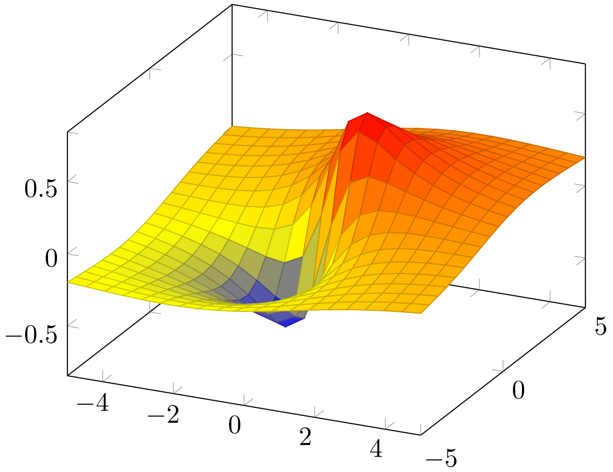

Figure 2.1: Surface plots of $f(x,y) = \frac{x+y}{x^2+y^2+1}$

Our main task will usually be to find and study global minimum or maximum points; however, most of the theoretical results only characterize local minima and maxima which are optimal points with respect to a neighborhood of the point of interest. The exact definitions follow.

Definition 2.3. *(local minima and maxima) Let* $f : S \to \mathbb{R}$ *be defined on a set* $S \subseteq \mathbb{R}^n$*. Then*

1. $\boldsymbol{x}^* \in S$ *is called a* ***local minimum point*** *of* f *over* S *if there exists* $r > 0$ *for which* $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ *for any* $\boldsymbol{x} \in S \cap B(\boldsymbol{x}^*, r)$*,*

2. $\boldsymbol{x}^* \in S$ *is called a* ***strict local minimum point*** *of* f *over* S *if there exists* $r > 0$ *for which* $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$ *for any* $\boldsymbol{x} \neq \boldsymbol{x}^* \in S \cap B(\boldsymbol{x}^*, r)$*,*

3. $\boldsymbol{x}^* \in S$ *is called a* ***local maximum point*** *of* f *over* S *if there exists* $r > 0$ *for which* $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$ *for any* $\boldsymbol{x} \in S \cap B(\boldsymbol{x}^*, r)$*,*

4. $\boldsymbol{x}^* \in S$ *is called a* ***strict local maximum point*** *of* f *over* S *if there exists* $r > 0$ *for which* $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$ *for any* $\boldsymbol{x} \neq \boldsymbol{x}^* \in S \cap B(\boldsymbol{x}^*, r)$*.*

Of course, a global minimum (maximum) point is also a local minimum (maximum) point. As with global minimum and maximum points, we will also use the terminology *local minimizer* and *local maximizer* for local minimum and maximum points, respectively.

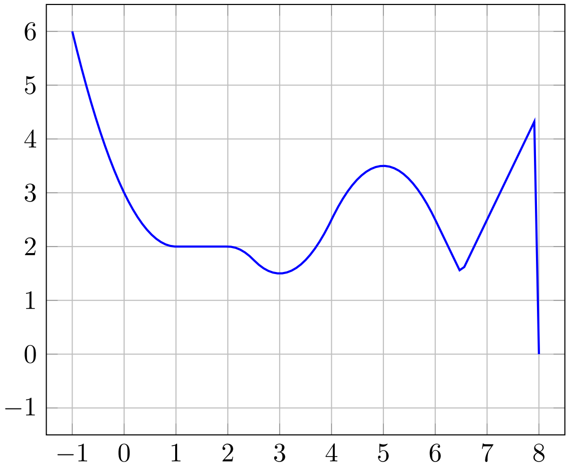

Figure 2.2: Local and global optimum points of a one-dimensional function

2.2 First Order Optimality Condition

A well-known result is that for a one-dimensional function f defined and differentiable over an interval (a, b), if a point $x^* \in (a, b)$ is a local maximum or minimum, then$f'(x^*) = 0$. This is also known as Fermat's theorem. The multidimensional extension of this result states that the gradient is zero at local optimum points. We refer to such an optimality condition as a *first order optimality condition*, as it is expressed in terms of the first order derivatives. In what follows, we will also discuss second order optimality conditions that use in addition information on the second order (partial) derivatives.

Theorem 2.4. *(first order optimality condition for local optima points) Let $f : U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $\boldsymbol{X}^* \in int(U)$ is a local optimum point and that all the partial derivatives of f exist at $\boldsymbol{x}^*$. Then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.*

Proof. Let $i \in \{1, 2, ..., n\}$ and consider the one-dimensional function $g(t) = f(\boldsymbol{x}^* + t\boldsymbol{e}_i)$. Note that g is differentiable at $t = 0$ and that $g'(0) = \frac{\partial f}{\partial x_i}(\boldsymbol{x}^*)$. Since $\boldsymbol{x}^*$ is a local optimum point of f, it follows that $t = 0$ is a local optimum of g, which immediately implies that $g'(0) = 0$. The latter is exactly the same as $\frac{\partial f}{\partial x_i}(\boldsymbol{x}^*) = 0$. Since this is true for any $i \in \{1, 2, ..., n\}$, the result $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ follows ■

Note that the proof of the first order optimality conditions for multivariate functions strongly relies on the first order optimality conditions for one-dimensional functions. The theorem presents a *necessary* optimality condition: the gradient vanishes at all local optimum points, which are interior points of the domain of the function; however, the re- verse claim is not true-there could be points which are not local optimum points, whose gradient is zero. For example, the derivative of the one-dimensional function $f(x) = x^3$ is zero at $x = 0$, but this point is neither a local minimum nor a local maximum. Since points in which the gradient vanishes are the only candidates for local optima among the points in the interior of the domain of the function, they deserve an explicit definition.

Definition 2.5. *(stationary points) Let $f : U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $\boldsymbol{x}^* \in int(U)$ and that f is differentiable over some neighborhood of $\boldsymbol{x}^*$. Then $\boldsymbol{x}^*$ is called a **stationary point** of f if $\nabla f(\boldsymbol{x}^*) = 0$.*

Thus, local optimum points are necessarily stationary points.

2.3 Second Order Optimality Conditions

Recall the criterion of local optimum for one-dimensional twice continuous differentiable function $f(x)$:

1. if $f'(x^*) = 0$ and $f''(x) > 0$, then x^* is a local minimizer.
2. if $f'(x^*) = 0$ and $f''(x) < 0$, then x^* is a local minimizer.

This motivates us to consider the exntension of the second order derivative characterization of optimum criterion. Essentially we have the following theorem.

Theorem 2.6. *Let $f : U \to \mathbb{R}$ be a function defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over $U$and that $\boldsymbol{x}^*$ is a stationary point. Then the following hold:*

1. *If $\boldsymbol{x}*$ is a local minimum point of f over U, then $\nabla^2 f(\boldsymbol{x^*}) \succcurlyeq 0$,*
2. *If $\boldsymbol{x}*$ is a local maximum point of f over U, then $\nabla^2 f(\boldsymbol{x^*}) \preccurlyeq 0$,*
3. *If $\nabla^2 f(\boldsymbol{x^*}) \succ 0$, then $\boldsymbol{x}*$ is a local minimum point of f over U,*
4. *If $\nabla^2 f(\boldsymbol{x^*}) \prec 0$, then $\boldsymbol{x}*$ is a local minimum point of f over U,*

Intuitively, to be a local minimum, there should not be any descending direction when starting from the minimizer around a neighborhood. The subtle difference between $\succcurlyeq$ and $\succ$ emerges when one applies the second order approximation to prove the theorem. Meanwhile we have another way to guarantee the sufficiency of optimum with a stronger condition:

Theorem 2.7. *Let f be a twice continuously differentiable function defined over $\mathbb{R}^n$. Suppose that $\nabla^2 f(\boldsymbol{x}) \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$. Let $\boldsymbol{x}^*$ Rn be a stationary point of f. Then $\boldsymbol{x}^*$ is a global minimum point of f.*

3 Convex Function

3.1 Definition and Examples

Definition 3.1. *(convex functions) A function $f : C \to \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^n$ is called **convex** (or **convex over** C) if*

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in C, \lambda \in [0,1] \tag{1}$$

The fundamental inequality 1 is illustrated in the following figure.

In case when no domain is specified, then we naturally assume that f is defined over the entire space $\mathbb{R}^n$. If we do not allow equality in 1 when $\boldsymbol{x} \neq \boldsymbol{y}$ and $\lambda \in (0,1)$, the function is called *strictly convex.*

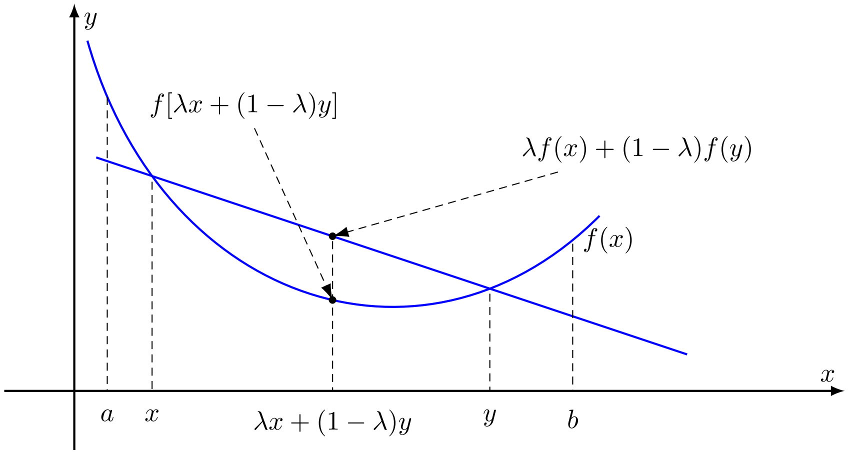
y
$f[\lambda x + (1-\lambda)y]$
$\lambda f(x) + (1-\lambda)f(y)$
$f(x)$
x
a
x
$\lambda x + (1-\lambda)y$
y
b

Figure 3.1: Illustration of inequality $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \le \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y})$

Definition 3.2. *(strictly convex functions) A function $f : C \to \mathbb{R}$ defined on a convext set $C \subseteq \mathbb{R}^n$ is called **strictly convex** if*

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) \text{ for any } \boldsymbol{x} \ne \boldsymbol{y} \in C, \lambda \in (0,1)$$

Another important concept is concavity. A function is called concave if $-f$ is convex. Similarly, f is called strictly concave if $-f$ is strictly convex. We can of course write a more direct definition of concavity based on the definition of convexity. A function f is concave if and only if for any $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\lambda \in [0,1]$ we have

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \ge \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y})$$

Equipped only with the definition of convexity, we can give some elementary examples of convex functions. We begin by showing the convexity of **affine functions**, which are functions of the form $f(x) = \boldsymbol{a}^T\boldsymbol{x} + b$, where $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. (If $b = 0$, then f is also called linear.)

Example 3.3. *(convexity of affine functions) Let $f(\boldsymbol{x} = \boldsymbol{a}^T\boldsymbol{x} + b$, where $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. To show that f is convex, take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Then*

$$\begin{aligned} f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) &= \boldsymbol{a}^T(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) + b \\ &= \lambda(\boldsymbol{a}^T\boldsymbol{x}) + (1-\lambda)(\boldsymbol{a}^T\boldsymbol{y}) + \lambda b + (1-\lambda)b \\ &= \lambda(\boldsymbol{a}^T\boldsymbol{x} + b) + (1-\lambda)(\boldsymbol{a}^T\boldsymbol{y} + b) \\ &= \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) \end{aligned}$$

and thus in particular $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \le \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y})$, and convexity follows. Meanwhile, it is also trivial that affine functions are both convex and concave. ■

The basic property characterizing a convex function is that the function value of a convex combination of two points **x** and **y** is smaller than or equal to the corresponding convex combination of the function values $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$. An interesting result is that convexity implies that this property can be generalized to convex combinations of any number of vectors. This is the so-called Jensen's inequality.

Theorem 3.4. *(Jensen's inequality) Let $f : C \to \mathbb{R}$ be a convex function where $C \subseteq \mathbb{R}^n$ is a convex set. Then for any $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k \in C$ and $\lambda \in \Delta_k$, the following inequality holds:*

$$f(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i) \leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i). \tag{2}$$

Proof. We will prove the inequality by induction on k. For $k = 1$ the result is obvious (it amounts to $f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_1)$ for any $\boldsymbol{x}_1 \in C$). The induction hypothesis is that for any k vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k \in C$ and any $\lambda \in \Delta_k$, the inequality 2 holds. We will now prove the theorem for $k + 1$ vectors. Suppose that $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{k+1} \in C$ and that $\lambda \in \Delta_{k+1}$. We will show that $f(\boldsymbol{z}) \leq \sum_{i=1}^{k+1} \lambda_i f(\boldsymbol{x}_i)$, where $\boldsymbol{z} = \sum_{i=1}^{k+1} \lambda_i \boldsymbol{x}_i$. If $\lambda_{k+1} = 1$, then $\boldsymbol{z} = \boldsymbol{x}_{k+1}$ and 2 is obvious. If $\lambda_{k+1} < 1$, then

$$\begin{aligned} f(\boldsymbol{z}) &= f(\sum_{i=1}^{k+1} \lambda_i \boldsymbol{x}_i) \\ &= f(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i + \lambda_{k+1} \boldsymbol{x}_{k+1}) \\ &= f((1 - \lambda_{k+1}) \underbrace{\sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} \boldsymbol{x}_i}_{\boldsymbol{v}} + \lambda_{k+1} \boldsymbol{x}_{k+1}) \\ &\leq (1 - \lambda_{k+1} f(\boldsymbol{v}) + \lambda_{k+1} f(\boldsymbol{x}_{k+1}). \end{aligned}$$

Since $\sum_{i=1}^{k} = \frac{1-\lambda_{k+1}}{1-\lambda_{k+1}=1}$, it follows that $\boldsymbol{v}$ is a convex combination of k points from C, and hence by the induction hypothesis we have that $f(\boldsymbol{v}) \leq \sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} f(\boldsymbol{x}_i)$, which combined with the ineuqality above yields

$$f(\boldsymbol{z}) \leq \sum_{i=1}^{k+1} \lambda_i f_i(\boldsymbol{x}_i)$$

■

3.2 First Order Characterization of Convex Functions

Convex functions are not necessarily differentiable, but in case they are, we can replace the Jensen's inequality definition with other characterizations which utilize the gradient of the function. An important characterizing inequality is the *gradient inequality*, which essentially states that the tangent hyperplanes of convex functions are always underestimates of the function.

Theorem 3.5. *(the gradient inequality) Let $f : C \to \mathbb{R}$ b e a continuously differentiable function defined on a convex set $C \subseteq \mathbb{R}^n$. Then f is comvex over C if and only if*

$$f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \leq f(\boldsymbol{y}) \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in C. \tag{3}$$

Proof. Suppose that f is convex. Let $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\lambda \in (0, 1]$. If $\boldsymbol{x} = \boldsymbol{y}$, then 3 trivially holds. We will therefore assume that $\boldsymbol{x} \neq \boldsymbol{y}$. Then

$$f(\lambda \boldsymbol{y} + (1 - \lambda)\boldsymbol{x}) \leq \lambda f(\boldsymbol{y}) + (1 - \lambda) f(\boldsymbol{x}),$$

and hence

$$\frac{f(\boldsymbol{x}+\lambda(\boldsymbol{y}-\boldsymbol{x}))-f(\boldsymbol{x})}{\lambda} \leq f(\boldsymbol{y})-f(\boldsymbol{x}).$$

Taking $\lambda \to 0^+$, the left-hand side converges to the directional derivative of f at $\boldsymbol{x}$ in the direction $\boldsymbol{y}-\boldsymbol{x}$, so that

$$f'(\boldsymbol{x};\boldsymbol{y}-\boldsymbol{x}) \leq f(\boldsymbol{y})-f(\boldsymbol{x})$$

Since f is continuously differentiable, it follows that $f'(\boldsymbol{x},\boldsymbol{y}-\boldsymbol{x}) = \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})$, and hence 2 follows. To prove the reverse direction, assume that the gradient inequality holds. Let $\boldsymbol{z},\boldsymbol{w} \in C$, and let $\lambda \in (0,1)$. We will show that $f(\lambda\boldsymbol{z}+(1-\lambda)\boldsymbol{w}) \leq \lambda f(\boldsymbol{z})+(1-\lambda)f(\boldsymbol{w})$. Let $\boldsymbol{u} = \lambda\boldsymbol{z}+(1-\lambda)\boldsymbol{w} \in C$. Then

$$\boldsymbol{z}-\boldsymbol{u} = \frac{\boldsymbol{u}-(1-\lambda)\boldsymbol{w}}{\lambda}-\boldsymbol{u} = -\frac{1-\lambda}{\lambda}(\boldsymbol{w}-\boldsymbol{u}).$$

Invoking the gradient inequality on the pairs $\boldsymbol{z},\boldsymbol{u}$ and $\boldsymbol{w},\boldsymbol{u}$, we obtain

$$\begin{aligned} f(\boldsymbol{u})+\nabla f(\boldsymbol{u})^T(\boldsymbol{z}-\boldsymbol{u}) \leq& f(\boldsymbol{z}), \\ f(\boldsymbol{u})-\frac{\lambda}{1-\lambda}\nabla f(\boldsymbol{u})^T(\boldsymbol{z}-\boldsymbol{u}) \leq& f(\boldsymbol{w}). \end{aligned}$$

Multiplying the first inequality by $\frac{\lambda}{1-\lambda}$ and adding it to the second one, we obtain

$$\frac{1}{1-\lambda}f(\boldsymbol{u}) \leq \frac{\lambda}{\lambda}f(\boldsymbol{z})+f(\boldsymbol{w}),$$

which after multiplication by $1-\lambda$ amounts to the desired inequality

$$f(\boldsymbol{u}) \leq \lambda f(\boldsymbol{z})+(1-\lambda)f(\boldsymbol{w}).$$

■

Geometrically, the gradient inequality essentially states that for convex functions, the tangent hyperplane is below the surface of the function. A two-dimensional illustration is given in the following figure. A direct result of the gradient inequality is that the first order optimality condition $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is sufficient for global optimality.

Proposition 3.6. *(sufficiency of stationary) Let f be a continuously differentiable function which is convex over a convex set $C \subset \mathbb{R}^n$. Suppose that $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ for some $\boldsymbol{x}^* \in C$. Then $\boldsymbol{x}^*$ is a global minimizer of f over C.*

Proof. Let $\boldsymbol{z} \in C$. Plugging $\boldsymbol{x} = \boldsymbol{x}^*$ and $\boldsymbol{y} = \boldsymbol{z}$ in the gradient inequality 3, we obtain that

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}^*)+\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{z}-\boldsymbol{x}^*),$$

which by the fact that $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ implies that $f(\boldsymbol{z}) \geq f(\boldsymbol{x}^*)$, thus establishing that $\boldsymbol{x}^*$ is the global minimizer of f over C. ■

We note that the above proposition establishes only the sufficiency of the stationarity condition $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ for guaranteeing that $\boldsymbol{x}^*$ is a global optimal solution. There could be some cases that the global minimizer does not satisfy the assumption (e.g. corner solution in a closed set). When C is not the entire space, this condition is not necessary. However, on most occasions of our interest (e.g. $C = \mathbb{R}^n$) this is not the case. **Analogously, the same logic applies to the sufficiency of stationarity for guaranteeing a global maximizer when the function is concave.**

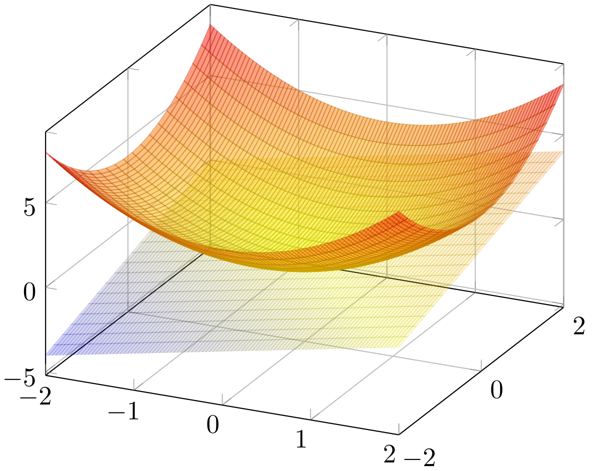

Figure 3.2: The function $f(x, y) = x^2 + y^2$ and its tangent hyperplane at $(1, 1)$, which is a lower bound of the function's surface.

3.3 Second Order Characterization of Convex Functions

When the function is twice continuously differentiable, convexity can be characterized by the positive semidefiniteness of the Hessian matrix.

Theorem 3.7. *(second order characterization of convexity) Let f be a twice continuously differentiable function over an open convex set $C \subseteq \mathbb{R}^n$. Then f is convex if and only if $\nabla^2 f(\boldsymbol{x}) \succcurlyeq 0$ for any $\boldsymbol{x} \in C$.*

Proof. Suppose that $\nabla^2 f(\boldsymbol{x} \succcurlyeq 0$ for all $\boldsymbol{x} \in C$. We will prove the gradient inequality, which by Theorem 3.5 is enough in order to establish convexity. Let $\boldsymbol{x}, \boldsymbol{y} \in C$. Then by the linear approximation theorem we have that there exists $\boldsymbol{z} \in [\boldsymbol{x}, \boldsymbol{y}]$ (and hence $\boldsymbol{z} \in C$) for which

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{z})^T(\boldsymbol{y} - \boldsymbol{x}) \tag{4}$$

Since $\nabla f(\boldsymbol{z}) \succcurlyeq 0$, it follows that $(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{z})^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0$, and hence by 4, the inequality $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$ holds.

To prove the opposite direction, assume that f is convex over C. Let $\boldsymbol{x} \in C$ and let $y \in \mathbb{R}^n$. Since C is open, it follows that $\boldsymbol{x} + \lambda\boldsymbol{y} \in C$ for $0 < \lambda < \varepsilon$, where ε is a small enough positive number. Invoking the gradient inequality we have

$$f(\boldsymbol{x} + \lambda\boldsymbol{y}_{=} f(\boldsymbol{x}) + \lambda\nabla f(\boldsymbol{x})^T\boldsymbol{y}$$

In addition, by the quadratic approximation theorem we have that

$$f(\boldsymbol{x} + \lambda\boldsymbol{y}) = f(\boldsymbol{x}) + \lambda\nabla f(\boldsymbol{x})^T\boldsymbol{y} + \frac{\lambda^2}{2}\boldsymbol{y}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{y} + o(\lambda^2\|\boldsymbol{y}\|^2),$$

Combine the two inequalities above we will have

$$\frac{\lambda^2}{2}\boldsymbol{y}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{y} + o(\lambda^2\|\boldsymbol{y}\|^2) \geq 0$$

for any $\lambda \in (0, \varepsilon)$. Dividing the latter inequality by λ^2 and taking $\lambda \to 0^+$, we conclude that

$$\boldsymbol{y}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{y} \geq 0$$

for any $y \in \mathbb{R}^n$, implying that $\nabla^2 f(\boldsymbol{x}) \succcurlyeq 0$ for any $\boldsymbol{x} \in C$. ■

3.4 Operations Preserving Convexity

There are several important operations that preserve the convexity property. First, the sum of convex functions is a convex function and a multiplication of a convex function by a nonnegative number results with a convex function.

Theorem 3.8. *(preservation of convexity under summation and multiplication by nonnegative scalars)*

1. *Let f be a convex function defined over a convex set $C \subseteq \mathbb{R}^n$ and let $\alpha \geq 0$. Then αf is a convex function over C.*
2. *Let $f_1, f_2, ..., f_p$ be convex functions over a convex set $C \subseteq \mathbb{R}^n$. Then the sum function $f_1 + f_2 + \cdots + f_p$ is convex over C*

Theorem 3.9. *(preservation of convexity under composition with a nondecreasing convex function) Ler $f : C \to \mathbb{R}$ be a convex function over the convex set $C \subseteq \mathbb{R}^n$. Let $g : I \to \mathbb{R}$ be a one-dimensional nondecreasing convex function over the interval $I \subseteq \mathbb{R}$. Assume that the image of C under f is contained in I: $f(C) \subseteq I$. Then the composition of g with f defined by*

$$h(\boldsymbol{x}) \equiv g(f(\boldsymbol{x})), \quad \boldsymbol{x} \in C$$

is a convex function over C.

Exercise 3.10. *Write the analogous theorem of 3.9 for concave functions.*

4 The KKT Conditions

In this section, we finally arrive at Karush–Kuhn–Tucker (KKT) conditions that underlie the common solution to constrained optimization problem. The preassumptions of applying the conditions vary, and in this note we pay attention to a very special case that is frequently met in economics studies and whose validity is easily proved. After stating and proving the theorem in a rather comprehensive setup, some intuition is provided and one example of practice follows.

4.1 Constrained optimization problems

Theorem 4.1. *(sufficiency of the KKT conditions for concave optimization problems) Let $\boldsymbol{x}^*$ be a feasible solution of the problem*

$$\begin{aligned} \max \quad & f(\boldsymbol{x}) \\ s.t. \quad & g_i(\boldsymbol{x}) \geq 0, \quad i = 1, 2, ..., , m, \\ & h_j(\boldsymbol{x}) = 0, \quad j = 1, 2, ..., p, \end{aligned}$$

where $f, g_1, ..., g_m$ are continuously differentiable concave functions over $\mathbb{R}^n$ and $h_1, h_2, ..., h_p$ are afffine functions. Suppose that there exist multipliers $\lambda_1, \lambda_2, ..., \lambda_m \geq 0$ and $\mu_1, \mu_2, ..., \mu_p \in \mathbb{R}$ such that

$$\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{n} \mu_j \nabla h_i(\boldsymbol{x}^*) = 0,$$

$$\lambda_i g_i(\boldsymbol{x}^*) = 0, i = 1, 2, ..., m$$

Then $\boldsymbol{x}^*$ is an optimal solution of the problem.

Proof. Let $\boldsymbol{x}$ be a feasible solution of the problem. We will show that $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$. Note that the function

$$s(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x})$$

is concave, and since $\nabla s(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(\boldsymbol{x}^*) = \mathbf{0}$, it follows by Proposition 3.6 that $\boldsymbol{x}^*$ is a maximizer of $s(\cdot)$ over $\mathbb{R}^n$, and in particular $s(\boldsymbol{x}^*) \geq s(\boldsymbol{x})$. We can thus conclude that

$$\begin{aligned} f(\boldsymbol{x}^*) &= f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}^*) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x}^*) \\ &= s(\boldsymbol{x}^*) \\ &\geq s(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \mu_j h_j(\boldsymbol{x}) \\ &\geq f(\boldsymbol{x}) \end{aligned}$$

■

4.2 Example: GDP and price index

The concept of GDP is usually the content of the first class in macroeconomics. It is the summation of value added across all industries. As apples and bananas cannot be directly added up, we firstly transform them into monetary terms and then calculate the summation of these numbers. By doing so we obtain the *nominal GDP* of the economy, which is usually accompanied with a *price index* to help tease out the impact of purely price change. The following practice help establish a link between the math we have learned and this daily economic concepts. We restrict our attention to a specific topic i.e. we simplify the real world economy, to consider only consumption over a range of different goods in a representative agent world. Let $x_1, x_2, ..., x_N$ denote the consumption amount of N various goods with prices being respectively p_i. The total income of the consumer is M. Then the agent allocates its consumption by solving the following constrained optimization problem (termed *utility maximization problem* in economics):

$$\begin{aligned} &\max U(x_1, x_2, ..., x_N) \\ \text{such that} \quad &\sum_{i=1}^{N} p_i x_i = M \end{aligned}$$

$U(x_1, x_2, ..., x_N)$, a concave function (to get rid of the sufficiency of optimum), is called *utility function* in microeconomics, while in macroeconomics it is also sometimes called *aggregator*, as it aggregates consumption over all goods to generate utility, and, with a bit of craziness at a first glance, we can directly treat it as GDP! To see this, we firstly impose a reasonable assumption on the aggregator function.

Definition 4.2. *(homogeneous function) A function $f : C \to \mathbb{R}$ defined on a convex hull $C \subseteq \mathbb{R}^n$ is called **homogeneous of degree** k ($k \in \mathbb{N}$) if for any $\lambda > 0$ we have*

$$f(\lambda \boldsymbol{x}) = \lambda^k f(\boldsymbol{x})$$

We assume that the aggregator function is *homogeneous of degree 1*, which is also usually called *constant return to scale* in economics. And we will also utilize the following properties of homogenous function:

Theorem 4.3. *(Euler's homogeneous function theorem) Let $f : C \to \mathbb{R}$ where $C \subseteq \mathbb{R}^n$ be a continuously differentiable function homogeneous of degree k, then we have*

$$kf(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} x_i \frac{\partial f}{\partial x_i}$$

Now let's follow the regular procedures of constrained optimization. Firstly, we establish a Lagrangian function with λ being the Lagrangian multiplier for the only constriant:

$$L(\lambda, \boldsymbol{x}) = U(\boldsymbol{x}) + \lambda(M - \sum_{i=1}^{N} p_i x_i)$$

Then, we derive the first order condition (FOC) for the function:

$$\frac{\partial U}{\partial x_i} = \lambda p_i \quad \text{for } i = 1, 2, ..., N$$

We have N such conditions, and we combine them with the constraint to form a system of equations, from which we could solve out exactly $N + 1$ variables: x_i's and λ. Even before plugging in the specific functional form, we could treat the FOC with some tricks: multiplying each side by x_i, and sum all the N FOCs up we have

$$\sum_{i=1}^{N} x_i \frac{\partial U}{\partial x_i} = \lambda \sum_{i=1}^{N} p_i x_i$$

For the left hand side (LHS) of equation, we apply Euler's homogeneous function theorem. For the right hand side (RHS) of equation, we apply the constraint. Combine them together we will have

$$U = \lambda M$$

Note that M is the total income. We can imagine a simplified case where the economy has only one homogeneous good and consumers spend all income to consume that good. In such a world, it is easy to calculate GDP and inflation, and this imagination is represented mathematically by the equation above, if we treat U as the consumption amount of the "final good", and λ the inverse of the price of the good. In state-of-the-art economics researches on multiple sector economy, this is exactly the case. λ is exactly the inverse of **price index** to dictate the change in prices. This will be more clear if we have a specific functional form and derive λ as a function of all prices.

Example 4.4. *Calculate the constrained optimization result above with following specific functional forms of U:*

1. $U(x_1, x_2) = x_1^\alpha x_2^{1-\alpha}$
2. $U(x_1, x_2) = (x_1^\alpha + x_2^\alpha)^{\frac{1}{\alpha}}$