Lecture Notes for GPED Summer Math Camp

Qiaohairuo Lin

August 11, 2024

1 Set

1.1 Basic Notation

Set is a collection of elements.

There are two common ways to write out the composition of the set. For example, let A denote a set, if it is a very simple (or, in a math jargon, "trivial") set, we can directly write down all the elements and use curly bracket to surround them, e.g.:

$$A = \{1, 2, 3, 4\}$$

If it is a rather complicated set (it could include infinite number of things!), we usually write down the general rule that determines what is inside, and put a vertical bar "|" in front of it e.g.

$$A = \{x | x \text{ is an integer and } x > 5\}$$

Let A denote a set, and x denote something. x could be either "included" in set A or not. In mathematical notations, we use $\in$ and $\notin$ to respectively denote that relationship i.e.

$$x \in (\text{or } \notin) A$$

x can be anything, even a set.

Let A and B both be sets. A could be a subset of B or not. Being a subset means that for any element included in A, it is also included in B.

$$A \subseteq B \Leftrightarrow \text{for any } x \in A, x \in B$$

Trivially, if A is a subset of B and B is a subset of A, A and B must be equal:

$$A \subseteq B \text{ and } B \subseteq A \Leftrightarrow A = B$$

1.2 Logic

Note that we use arrow connects two statements, which is quite intuitive in that $S1 \Rightarrow S2$ means that we can derive $S2$ from $S1$, or "if $S1$ holds, $S2$ also holds". Note that this not necessarily mean the reverse direction is true. For example,

$$x = 5 \Rightarrow x^2 = 25; \quad x^2 = 25 \not\Longrightarrow x = 5$$

We can also describe this as $S1$ is the **sufficient condition** for $S2$, or $S2$ is the **necessary condition** for $S1$. Both terms are very useful:

1. Sometimes, we prove $S2$ by proving a "stronger" statement $S1$.
2. Sometimes, we want to find things that hold under $S1$, but there are too many candidates, then we can firstly verify whether these candidates satisfy the statement $S1$ to decrease the number of candidates and save our efforts.

And a bi-direction arrow means the two statements are "equivalent", or $S1$ is the "sufficient and necessary condition" of $S2$ (and vice versa). We usually use "if and only if" (or iff) to denote this relationship.

1.3 Special Set

There are some sets of numbers that are frequently used and denoted with special notations as below:

1. $\mathbb{Z}$ denotes the set of all the integers, and $\mathbb{Z}^+$ or $\mathbb{N}$ denotes the set of all the positive integers.
2. $\mathbb{Q}$ denotes the set of rational numbers i.e. the numbers that could be denoted by a ratio of two integers.
3. $\mathbb{R}$ denotes the set of real numbers, and $\mathbb{R}^+$ denotes the set of all the positive integers. There are also imaginary numbers outside of $\mathbb{R}$, but those will not be studied in this course.

4. $[a, b]$ denotes the set of real numbers that are between a and b (two end points included) i.e.

$$[a, b] = \{x \in \mathbb{R} | a \leq x \leq b\}$$

we can also replace either side (or both sides) of the square bracket "[]" to round bracket "()" and replace "$\leq$" above with a strict "$<$" sign.

1.4 Operations of Set

There are some methods through which we could generate new set by manipulating some existing sets. The most two important methods are union and intersection.

1. **Union**: The union of two sets A and B, denoted A $\cup$ B, is the set of all elements that are in A, or in B, or in both. Formally:

   A $\cup$ B = $\{x \mid x \in$ A or $x \in$ B$\}$

2. **Intersection**: The intersection of two sets A and B, denoted A $\cap$ B, is the set of all elements that are common to both A and B. Formally:

   A $\cap$ B = $\{x \mid x \in$ A and $x \in$ B$\}$

Another very important operation, while seemingly meaningless so far, is to "multiply" sets. By doing so, we just obtain a new set whose element is the tuple that includes one element from each composition set. For example, we have

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x_1, x_2) | x_1 \in \mathbb{R} \text{ and } x_2 \in \mathbb{R}\}$$

Intuitively, we can call this set as two-dimensional. Similarly, we can define the $N-$dimension set $\mathbb{R}^N$ for any positive integer N, which will be extremely useful later on.

1.5 Sets with Infinite Elements

Sets can include either finite or infinite elements. When dealing with infinity, things often diverge from our common sense (and that is precisely why we need more advanced mathematical tools, and why you're here!). For finite sets, we can simply use the number of elements to denote the size of the set. But what about sets with infinite elements? First, we must ask: does size comparison still make sense for infinite sets? The answer is yes, but the method of comparison in mathematics can be counterintuitive. A famous statement illustrates this: "The set of (positive) even numbers is as large as the set of all natural numbers," despite the fact that intuitively, the former seems to contain only half the elements of the latter.

To understand this concept better, let's explore the story of Hilbert's Hotel, which provides a vivid demonstration of why these two sets are equal in terms of "size" (or in professional terms, cardinality).

Hilbert's Paradox of the Grand Hotel

Imagine a hotel with infinitely many rooms, all of which are occupied. This hotel is managed by the brilliant mathematician David Hilbert. One night, a new guest arrives and asks for a room. In a normal, finite hotel, this would be impossible. But in Hilbert's infinite hotel, here's what happens: Hilbert asks the guest in Room 1 to move to Room 2. The guest in Room 2 moves to Room 3. The guest in Room 3 moves to Room 4. This process continues indefinitely, with each guest moving to the next room. Room 1 is now vacant and can accommodate the new guest. Surprisingly, despite starting with a fully occupied infinite hotel, Hilbert found room for one more guest without evicting anyone! But the paradox doesn't stop there. What if an infinite number of new guests arrive, say, as many as there are natural numbers? Hilbert devises another clever solution: He asks the guest in Room 1 to move to Room 2; the guest in Room 2 to move to Room 4; the guest in Room 3 to move to Room 6..... In general, each guest in Room n moves to Room $2n$. Now, all the odd-numbered rooms are empty and can accommodate the infinite number of new guests!

This paradox illustrates two key points about infinite sets:

1. We can add a finite number to infinity and still have the same size of infinity.
2. We can even add infinity to infinity and still have the same size of infinity.

Hibert's action also provides us with a way to prove that the two sets are of the same cardinality: to build a one-to-one mapping. That is, we can find a rule to "align" the elements of two sets such that for each element in one side, we can find one corresponding element in the other set, and vice versa. Manipulating this method, we could obtain the following fundamental results in real analysis:

1. The cardinality of $\mathbb{N}$ is the same as $\mathbb{Q}$. We call the set that has the same cardinality as $\mathbb{N}$ "countable", denoted by $\aleph_0$.
2. The cardinality of $\mathbb{R}$ is larger than $\mathbb{N}$. We call the set that has larger cardinality than $\mathbb{N}$ "uncountable", and we denote the cardinality of $\mathbb{R}$ by $\aleph_1$.
3. The cardinality of $\mathbb{R}^N$ is the same as $\mathbb{R}$.

1.6 A Short Introduction on Function

In the last section, we informally talk about "one-to-one mapping", which is a type of function. You probably already know something about function, which is a bridge between two sets that assign each element in one set to one element in the other. Formally, let f denote the function and A and B be the two sets, we denote it by $f : A \to B$. With our previous definition of "product" of set, it is also convenient to define a function f with two inputs as $f : A_1 \times A_2 \to B$. Essentially, the output of the function is one element in the set. If the output of one mapping takes multiple values, the mapping is called "correspondence".

1.7 Distance and Norm

For most of the time, we work on the set of real numbers $\mathbb{R}$, or vectors or matrices (which will be introduced shortly) whose entries are real numbers. But "set" only implies what elements are inside and what are not, and our common sense is far beyond that: for example, we know how to compare the two elements in one set, we know how to add or multiple two elements in the set to obtain a new element in the same set..... Advanced math course usually begins with a set of abstract rules to define these "structures" upon the set. Here we demonstrate two of the most important concepts: distance and norm.

Definition 1.1 (Distance and Norm)**.** *Let X be a set. A **distance function** or **metric** on X is a function $d : X \times X \to \mathbb{R}$ that satisfies the following properties for all $x, y, z \in X$:*

1. ***Non-negativity**: $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.*
2. ***Symmetry**: $d(x, y) = d(y, x)$.*
3. ***Triangle inequality**: $d(x, z) \leq d(x, y) + d(y, z)$.*

*A **norm** on a vector space V over $\mathbb{R}$ is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following properties for all $\boldsymbol{u}, \boldsymbol{v} \in V$ and $c \in \mathbb{R}$:*

1. ***Non-negativity**: $\|\boldsymbol{v}\| \geq 0$ and $\|\boldsymbol{v}\| = 0$ if and only if $\boldsymbol{v} = \boldsymbol{0}$.*
2. ***Scalar multiplication**: $\|c\boldsymbol{v}\| = |c|\|\boldsymbol{v}\|$.*
3. ***Triangle inequality**: $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$.*

Example 1.1. *For vectors in $\mathbb{R}^n$, the most common norm is the Euclidean norm, defined as:*

$$\|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

The corresponding distance induced by this norm, called the Euclidean distance, is:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Usually, when we use the notion of $\mathbb{R}^N$, we are referring to the set $\mathbb{R}^N$ with Euclidean metrics system, which we call "Euclidean space". For an element taken from high-dimensional Euclidean space, we usually use bold lowercase letter *e.g.*$\boldsymbol{x}$ to denote it, and write it as a "column vector" i.e. to list its elements vertically and circle them with square brackets e.g. $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$. To save space, we could also write it as $\boldsymbol{x} = [x_1, x_2, ..., x_N]^T$ where the superscript T denotes the transposition of the formation from a row to a column.

Example 1.2. *Other common norms include the L^1 norm (Manhattan norm) and L^∞ norm (maximum norm):*

- *L^1 norm:* $\|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$
- *L^∞ norm:* $\|\boldsymbol{x}\|_\infty = \max(|x_1|, |x_2|, \ldots, |x_n|)$

The concepts of distance and norm are very important. For example, almost all econometrics/statistical/machine learning models are aimed at optimizing some metrics between estimation and observed data. Moreover, only after the distance is defined, can we conveniently discuss the concept of convergence, as will be clear in the next section.

2 Linear Algebra

2.1 Vector

Element in high-dimensional space $\mathbb{R}^N$ ($N > 1$) is often referred to as "vector", while element in one-dimensional space is often called "scalar". For an element $\boldsymbol{x} \in \mathbb{R}^N$, we have the following trivial equation:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_N \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

where the right hand side (RHS) has N vectors, each only has $i-$th entry being 1 and all the other entries being 0, which we denote as $\boldsymbol{e}_i$. This trivial equation uncovers a trivial fact: we could express $\boldsymbol{x}$ as a linear combination of N elements in $\mathbb{R}^N$. Here "linear combination" means a series of operations that can only take one of the following forms: 1) addition; 2) multiplication with a scalar i.e. timing all entries by the same scalar. Clearly $\{\boldsymbol{e}_i\}_{i=1}^N$ is a

decent set of vectors as it could be used to construct anything in the space. We call such set of N vectors the "basis" of $\mathbb{R}^N$. Basis is not unique, as is indicated in the following example:

Exercise 2.1. *Prove that* $[1,2]^T$ *and* $[2,1]^T$ *is also a basis of* $\mathbb{R}^2$.

Could any N vectors form a basis? We can easily answer this question by replacing $[2,1]^T$ in the last example with $[2,4]^T$. Intuitively, we can see that the original pair of vectors could be "stretched" to match any vector on the plane. But the new pair, $[1,2]^T$, $[2,4]^T$ lies on one line. And they cannot be combined to reach anything outside of the line.

We call such case as "linear dependency". Note that if the two vectors are located on the same direction, we could find a scalar λ such that $\boldsymbol{x}_1 = \lambda \boldsymbol{x}_2$.

Then, could any N vectors, any two of which do not lie on the same line, form a basis? The answer is "yes" when $N = 2$, but "no" when $N >= 3$. To see this, consider the following set: $\left\{[1,1,1]^T, [1,-1,1]^T, [2,0,2]^T\right\}$. As the third vector is a linear combination of the first two, it also lies on the two-dimensional plane "stretched" by the first two vectors. As they lie on the same plane, we cannot obtain any vectors outside of the plane. We extend the above result to the general case. Firstly, we define linear dependency:

Definition 2.1 (Linear Independence)**.** *A set of vectors* $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\} \subseteq \mathbb{R}^N$ *is said to be linearly independent if the only solution to the equation*

$$c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_k\boldsymbol{v}_k = \mathbf{0}$$

is $c_1 = c_2 = \cdots = c_k = 0$.

Then, we have the following intuitive result:

1. For N-dimensional space $\mathbb{R}^N$, there could be at most N vectors to be linearly independent in one set. We call any set of N linearly independent a "basis" for space $\mathbb{R}^N$.
2. The set of N vectors in $\mathbb{R}^N$ is a basis for $\mathbb{R}^N$ if and only if any element in $\mathbb{R}^N$ could be represented as a linear combination of them.

Note that for a basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$, the process of finding the proper linear combination for an element is exactly the process of solving a system of N linear equations with N unknowns, which is another great way to lead into the content of linear algebra, though not introduced here. Intuitively, whether the vector set is a basis is equivalent to whether they could be "stretched" to create the whole space. We can also quantify this operation by looking at what is the dimension the vectors could stretch, which leads to the following definitions.

Definition 2.2 (Dimension)**.** *The dimension of a vector space* V *is the number of vectors in any basis of* V*. In other words, if* $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$ *is a basis for* V*, then the dimension of* V*, denoted as* $\dim(V)$*, is* k.

We can extend this concept to a set of vectors and the linear space generated by them.

Definition 2.3 (Linear Span)**.** *Given a set of vectors* $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$ *in a vector space* V*, the linear span (or simply span) of these vectors, denoted as span* $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$*, is the set of all linear combinations of* $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$*. Formally,*

$$span\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\} = \{a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_m\boldsymbol{v}_m \mid a_1, a_2, \ldots, a_m \in \mathbb{R}\}.$$

This span is a subspace of V*.*

Definition 2.4 (Dimension of a Set of Vectors)**.** *The dimension of a set of vectors* $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$ *is the dimension of the linear span of these vectors. If the vectors are linearly independent and span a subspace* W*, then the dimension of the set of vectors is the number of vectors in the set, i.e.,* $\dim(W)$*.*

Using these definitions, we can see that the dimension of a vector space or the linear span of a set of vectors provides a measure of how many directions in the space can be independently spanned by the vectors.

2.2 Linear Mapping and Matrix

Then we consider linear mapping from one Euclidean space $\mathbb{R}^n$ to another $\mathbb{R}^k$.

Definition 2.5 (Linear Mapping)**.** *A linear mapping (or linear transformation)* f *from* $\mathbb{R}^n$ *to* $\mathbb{R}^k$ *is a function* $f : \mathbb{R}^n \to \mathbb{R}^k$ *that satisfies the following two properties for all* $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ *and all scalars* $c \in \mathbb{R}$*:*

1. *Additivity:* $f(\boldsymbol{u} + \boldsymbol{v}) = f(\boldsymbol{u}) + f(\boldsymbol{v})$
2. *Homogeneity:* $f(c\boldsymbol{u}) = cf(\boldsymbol{u})$

If we regard those elements as coordinates of the point in the space, then the linear mapping could be viewed as a transformation of the coordinate system while keeping the coordinate unchanged. To see this, let $\{\boldsymbol{v}_i\}_{i=1}^N$ denote one basis of $\mathbb{R}^n$, under which the vector $\boldsymbol{x}$ has the coordinate $(x_1, x_2, ..., x_n)$. Then we have the following result:

Exercise 2.2. *Show that* $f(\boldsymbol{x}) = \sum_{i=1}^N x_i f(\boldsymbol{v}_i)$*.*

Meanwhile, we know that $f(\boldsymbol{v}_i)$ is a vector in $\mathbb{R}^k$. Let $\{\boldsymbol{u}_j\}_{j=1}^k$ denote a basis of $\mathbb{R}^k$, then we know that $f(\boldsymbol{v}_i)$ also has a coordinate with respect to (w.r.t.) this basis. Assume that the coordinate of $f(\boldsymbol{v}_i)$ is $(a_{1i}, a_{2i}, ..., a_{ki})$. Then we have the following expression of a linear

mapping:

$$f(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}) = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n \end{bmatrix}$$

And this is exactly the result when a matrix multiply a vector. That is, if we extract all the a_{ij}'s, we could rewrite the above as

$$\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The first element on RHS lists all coefficients in n columns, k rows. We call this a matrix of $k \times n$, and let $\mathbb{R}^{k\times n}$ denote the set of all the Matrix whose entries are real numbers. Essentially, matrix could be thought as a representation of a certain linear mapping, and matrix product could then be thought as a combination of two linear mapping. For example, let matrix $A \in \mathbb{R}^{k\times n}$ denote a linear mapping $f : \mathbb{R}^n \to \mathbb{R}^k$, and matrix $B \in \mathbb{R}^{m\times k}$ denote another linear mapping $g : \mathbb{R}^{k\times m}$. Then the matrix product BA will denote the compound linear mapping (which is still a linear mapping):

$$\boldsymbol{x} \xrightarrow{f} \boldsymbol{y} \xrightarrow{g} \boldsymbol{z} \Leftrightarrow \boldsymbol{x} \xrightarrow{f} A\boldsymbol{x} \xrightarrow{g} BA\boldsymbol{x}$$

This also helps explain why we need to align the dimension of matrix when multiply them. Moreover, if we switch the dimension of the matrix, it is also easy to verify that

$$(AB)^T = B^T A^T$$

where superscript T denotes the transpose of matrix. Below is a simple example of linear transformation:

Example 2.1 (Rotating by 45 Degrees)**.** *Consider the linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ that rotates every point in the plane by 45 degrees counterclockwise. The transformation matrix for this mapping is.*

$$A = \begin{bmatrix} \cos(45°) & -\sin(45°) \\ \sin(45°) & \cos(45°) \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

For any vector $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$*, the rotated vector* $f(\boldsymbol{x})$ *is given by:*

$$f(\boldsymbol{x}) = A\boldsymbol{x} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2}(x_1 - x_2) \\ \frac{\sqrt{2}}{2}(x_1 + x_2) \end{bmatrix}$$

Thus, the coordinates (x_1, x_2) *are transformed to* $\left(\frac{\sqrt{2}}{2}(x_1 - x_2), \frac{\sqrt{2}}{2}(x_1 + x_2)\right)$ *under the rotation.*

You may notice that this rotation operation, or linear mapping, is "reversible" i.e. you can find another linear mapping to cancel out the impact of the first mapping. Or, more rigorously, for *any* vector x, we can find a special matrix M such that $MA\boldsymbol{x} = \boldsymbol{x}$. In this case, we know that MA will be equal to the following very simple matrix: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. We call such matrix that has 1 on its diagonal and 0 elsewhere **identity matrix**, as they represent the trivial linear mapping that maps one element to itself. And we call the qualified matrix M the **inverse matrix** of A, denoted by A^{-1}.

A natural question is: can we find an inverse for all the linear mappings (or matrices)? There are some cases that do not have an inverse. For example, think of a linear mapping from $\mathbb{R}^3$ to $\mathbb{R}^2$. Because for any reverse linear mapping, you cannot obtain a 3-dimensional space from a 2-dimensional plane. Similarly, if a linear mapping from $\mathbb{R}^3$ to $\mathbb{R}^3$ maps all the elements to a plane in 3-dimensional space, it will neither have a reverse linear mapping. This motivates us to think of the **rank** of a matrix, which is the dimension of the space that a linear mapping can create.

Definition 2.6 (Rank)**.** *The* ***rank*** *of a matrix A is the maximum number of linearly independent row vectors (or column vectors) in A. It is a measure of the dimension of the image of the linear transformation represented by A.*

Then, it turns out that an $n \times n$ square matrix is invertible (or non-singular) if its rank is also n (which is also called full rank). This is because an invertible matrix must span the entire n-dimensional space, meaning its rows (or columns) are linearly independent and the image of the transformation is the entire space $\mathbb{R}^n$.

In this section, we will mostly focus on other properties of square matrices.

2.3 Eigenvalues and Determinant

If a matrix is not the identity matrix, it changes the coordinates and maps one vector to another position. Vector is defined by its "direction" and "length". While the length is contingent on the metric we use (e.g. $[3, 4]^T$ and $[5, 0]^T$ are of the same length under L^2

metric, but not so under L^1 metric), direction is more "stable", as we know that the two vectors $\boldsymbol{x}, \boldsymbol{y}$ are of the same direction if we can find a scalar λ such that $\boldsymbol{x} = \lambda \boldsymbol{y}$. In the rotation example, everything changes its direction after the operation. But that is not always the case. For a square matrix, we call those special vectors whose direction does not change "eigenvectors". Their formal definition is as follows:

Definition 2.7 (eigenvectors and eigenvalues)**.** *If A is a square matrix, a non-zero vector $\mathbf{v}$ is called an eigenvector of A if it satisfies the equation*

$$A\mathbf{v} = \lambda \mathbf{v},$$

where λ is a scalar known as the eigenvalue corresponding to the eigenvector $\mathbf{v}$.

The concepts of eigenvector and eigenvalues are extremely important in linear algebra, and has a lot of application. For example, they could be used to measure the importance of each node in a connected network, which is the core of Google's algorithm to rank web pages.

Example 2.2. *Verify that $\boldsymbol{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is an eigenvector of the matrix*

$$A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$$

We need a systematic way to calculate the eigenvalues and eigenvectors for a matrix. Note that the equation $A\mathbf{v} = \lambda \mathbf{v}$ could be transformed to the system of linear equations:

$$(A - \lambda I)\boldsymbol{v} = \mathbf{0} \tag{1}$$

The non-zero $\boldsymbol{v} = [v_1, ..., v_n]^T$ gives a way to linearly combine the column vectors of matrix $(A - \lambda I)$ to be $\mathbf{0}$ i.e. to show their linear dependency. Therefore, we consider λ's that makes the matrix $(A - \lambda I)$ not full rank. A very useful tool is the determinant, which is a function that maps a square matrix to a scalar. It has a weird expression and a couple of nice properties.

Definition 2.8 (Determinant)**.** *The **determinant** of a square matrix A, denoted as* $\det(A)$*, is a scalar value that is a function of the entries of the matrix. defined explicitly as,*

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}, \tag{2}$$

where S_n is the set of all permutations of $\{1, 2, \ldots, n\}$ and $\operatorname{sgn}(\sigma)$ *is the sign of the permutation σ. The determinant has several key properties:*

- *A matrix A is invertible (or non-singular) if and only if* $\det(A) \neq 0$.
- *The determinant of a product of matrices is the product of their determinants:* $\det(AB) = \det(A)\det(B)$.
- *The determinant of a matrix is equal to the determinant of its transpose:* $\det(A) = \det(A^T)$.
- *Swapping two rows (or columns) of a matrix multiplies its determinant by* -1.
- *Multiplying a row (or column) by a scalar multiplies the determinant by that scalar.*
- *Adding a multiple of one row (or column) to another row (or column) does not change the determinant.*

The weird expression is the only function that has those nice properties. Clearly, the eigenvalues of a matrix are found by solving the characteristic equation

$$\det(A - \lambda I) = 0,$$

where det denotes the determinant and I is the identity matrix of the same dimension as A. This determinant is a polynomial of λ, and solving polynomial we get our result of λ[1], then plug in the value of λ, we can find the corresponding eigenvectors.

When calculating determinant, we usually do not directly apply the formula, as it is too complicated. Instead, we utilize its property to transform the original matrix to a triangular matrix, and then multiply the diagonal values.

2.4 Square Matrix and Quadratic Forms

Another usage of a square matrix is to summarize quadratic forms. A quadratic form in n variables is a homogeneous polynomial of degree two and can be written in matrix notation as

$$Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x},$$

where $\mathbf{x}$ is an n-dimensional column vector, $\mathbf{x}^T$ is its transpose, and A is an $n \times n$ symmetric matrix.

Example 2.3. *Consider the simplest quadratic form: $x_1^2 + x_2^2 + 2x_1x_2$. It can be represented as*

$$Q(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 2x_1x_2 + x_2^2.$$

[1]There could be solutions with imaginary numbers, but here we focus on the case with only real numbers.

Here, the matrix A is

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Quadratic forms can exhibit different properties depending on the nature of the matrix A. One important property is positive definiteness, which ensures that the quadratic form is always positive except at the origin.

Definition 2.9 (Positive Definiteness)**.** *A symmetric matrix A is called **positive definite** if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$,*

$$\mathbf{x}^T A\mathbf{x} > 0.$$

This property ensures that the quadratic form $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is always positive except at the origin, where it is zero.

Similarly, we can define positive semidefiniteness (by changing "$>$" to "$\geq$") and negative (semi)definiteness. There are multiple methods in linear algebra to test the definiteness of matrix. Below is some most important ones.

Theorem 2.1 (Leading Principal Minors Test)**.** *A symmetric matrix A is positive definite if and only if all leading principal minors of A are positive. Specifically, for an $n \times n$ matrix A,*

$$A_k > 0 \quad \textit{for} \quad k = 1, 2, \ldots, n,$$

where A_k denotes the determinant of the k-th leading principal submatrix of A.

Theorem 2.2 (Eigenvalue Test)**.** *A symmetric matrix A is:*

- *positive definite if and only if all its eigenvalues are positive,*
- *positive semidefinite if and only if all its eigenvalues are non-negative,*
- *negative definite if and only if all its eigenvalues are negative,*
- *negative semidefinite if and only if all its eigenvalues are non-positive.*

3 Function

We already introduced function above, in this section we lay out more properties an definitions related to function.

3.1 Limit of Sequence

Sequence is a useful tool to simplify the notion of other concepts. We consider a sequence $\{x_i\} = \{x_1, x_2, ...\}$ that has infinite (but clearly, countable) elements from Euclidean space R^k.

Definition 3.1. *(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 $\boldsymbol{x}_i \to \boldsymbol{x}^*$.

This "$N - \epsilon$" language makes convergence and limit questions tractable. For a sequence with a written expression, the question of whether or not it is convergent is transformed to the question of finding a proper expression of N as a function of ϵ. Take the following exercise as an example.

Exercise 3.1. *Prove that $x_n = \frac{1}{n^2+1}$ converges to 0.*

3.2 Open Set and Closed Set

With the concepts of convergent sequence, we are able to give a formal definition for the openness/closedness of the set.

Definition 3.2. *(Open Set) A set $U \subseteq \mathbb{R}^k$ is said to be open if for every sequence $\{\boldsymbol{x}_i\} \subseteq \mathbb{R}^k$ that converges to a point $\boldsymbol{x} \in U$, there exists a subsequence $\{\boldsymbol{x}_{i_k}\}$ such that $\boldsymbol{x}_{i_k} \in U$ for all $k \in \mathbb{N}$.*

Definition 3.3. *(Closed Set) A set $F \subseteq \mathbb{R}^k$ is said to be closed if it contains all its limit points. Equivalently, F is closed if whenever a sequence $\{\boldsymbol{x}_i\} \subseteq F$ converges to a limit $\boldsymbol{x} \in \mathbb{R}^k$, then $\boldsymbol{x} \in F$.*

Note that while the statement is now concise and tidy, it may not be that straightforward to think through. In Euclidean space, however, there is a type of set whose openness/closeness is easy to determine:

1. $[a, b] = \{x \geq a \text{ and } x \leq b | x \in \mathbb{R}\}$ is closed.
2. $(a, b) = \{x > a \text{ and } x < b | x \in \mathbb{R}\}$ is open.

We can (though not rigorously) verify that they agree with our definition above. Moreover, it is intuitive to see that for high dimensional space, we have similar result:

1. $\{x_i \geq a_i \text{ and } x_i \leq b_i \text{ for every } i = 1, 2, ..., k | (x_1, ..., x_k) \in \mathbb{R}^k\}$ is closed.
2. $\{x_i > a_i \text{ and } x_i < b_i \text{ for every } i = 1, 2, ..., k | (x_1, ..., x_k) \in \mathbb{R}^k\}$ is open.

3.3 Continuity

Convergent sequences can also be used to define the continuity of functions:

Definition 3.4 (continuity of function)**.** *Let $X \subseteq \mathbb{R}^d$. A function $f : X \to \mathbb{R}^k$ is* ***continuous*** *at a 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 at all points in X, we call f* ***continuous on*** *X.*

Another way to define the continuity of a function is to use the famous "$\epsilon - \delta$" language:

Definition 3.5 (epsilon-delta continuity)**.** *Let $X \subseteq \mathbb{R}^d$. A function $f : X \to \mathbb{R}^k$ is* ***continuous*** *at a point $\boldsymbol{x}^* \in X$ if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $\boldsymbol{x} \in X$, if $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta$, then $\|f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\| < \varepsilon$. If f is continuous at all points in X, we call f* ***continuous on*** *X.*

A very important peroperty of continuous function is that it is guaranteed to have optimum values on a bounded and closed domain, as is stated below.

Definition 3.6. *(boundedness and compactness)*

1. *A set $U \subseteq \mathbb{R}^n$ is called* ***bounded*** *if there exists $M > 0$ such that for any $\boldsymbol{x} \in U$, $d(\boldsymbol{x}, \mathbf{0}) < M$.*
2. *A set $U \subseteq \mathbb{R}^n$ is called* ***compact*** *if it is closed and bounded*

Theorem 3.1. *(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.*

3.4 Derivative

The concept of the derivative is a fundamental tool in calculus. It provides a measure of how a function changes as its input changes.

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 derivative* of f at $\boldsymbol{x}$ along the direction $\boldsymbol{d}$ and is denoted by $f'(\boldsymbol{x}; \boldsymbol{d})$. A special sets of directional derivative is those along the simplest directions: 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:

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

Differentiability is a stronger requirement than continuity. Every differentiable function is continuous, but not every continuous function is differentiable. Fortunately, most of the functions we work with everyday are elementary functions, on which we could confidently take derivatives.

Definition 3.7 (elementary function)**.** *Elementary functions are functions built from basic functions such as polynomials, exponentials, logarithms, trigonometric functions, and their inverses using a finite number of arithmetic operations, compositions, and solutions of algebraic equations.*

The gradient is easy to calculate. As the term of limit is essentially a univariate function of t, it turns out that we can apply similar methods of univeriate calculus to calculate it.

Note that all of the definitions of derivatives in this section is done on the general $\mathbb{R}^N$ space, because unlike the univariate case where you only have two directions to approach a point, in high-dimensional space you have infinitely many (uncountable) directions. For elementary functions, the gradient provides us with all the information we need to calculate any directional derivative, as it could be calculated as:

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

for all $\boldsymbol{x} \in U$ and $\boldsymbol{d} \in \mathbb{R}^n$.

3.5 Approximation

Derivative could be used not only to decide the "momentum" of function, but also to approximate the function. It can also be shown in this setting of continuous differentiability that the following approximation result holds.

Proposition 3.1. *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*

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

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 $t \to 0^+$ or $t \to \infty$. Another similar notation is the big O "O", which denotes that $\frac{O(t)}{t}$ approaches to some constant as $t \to 0$ or $t \to \infty$. With the notation of big and small o, we are able to analyze the complicated functions with a system of polynomial series.

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 3.2. *(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 3.3. *(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 $\boldsymbol{y} \in B(\boldsymbol{x}, r)$,*

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

3.6 Composite Function and Implicit Function

Functions could be combined to generate new composite functions. In the univariate case, we have the chain rule for finding the derivative of a composite function. Let $F(x) = f(g(x))$, then we have

$$F'(x) = f'(g(x))g'(x)$$

We now extend this chain rule to a general multi-dimensional scenario:

Suppose we have a scalar-valued function $z = f(\mathbf{x})$, where $\mathbf{x}$ is a vector of n variables, $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, and each x_i is a function of another vector $\mathbf{t}$ with m variables, $\mathbf{t} = [t_1, t_2, \ldots, t_m]^T$. Then, the composite function $z = f(\mathbf{x}(\mathbf{t}))$ depends on the variables $\mathbf{t}$ through $\mathbf{x}$, and we can compute the gradient of z with respect to $\mathbf{t}$ using the chain rule in the multivariate case.

Formally, the gradient of z with respect to $\mathbf{t}$ is given by:

$$\frac{\partial z}{\partial \mathbf{t}} = \nabla_{\mathbf{x}} f(\mathbf{x}) \cdot \mathbf{J}_{\mathbf{x}}(\mathbf{t}),$$

where:

- $\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]$ is the gradient of f with respect to $\mathbf{x}$.
- $\mathbf{J}_{\mathbf{x}}(\mathbf{t})$ is the Jacobian matrix of $\mathbf{x}$ with respect to $\mathbf{t}$, given by:

$$\mathbf{J}_{\mathbf{x}}(\mathbf{t}) = \begin{bmatrix} \frac{\partial x_1}{\partial t_1} & \frac{\partial x_1}{\partial t_2} & \cdots & \frac{\partial x_1}{\partial t_m} \\ \frac{\partial x_2}{\partial t_1} & \frac{\partial x_2}{\partial t_2} & \cdots & \frac{\partial x_2}{\partial t_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial t_1} & \frac{\partial x_n}{\partial t_2} & \cdots & \frac{\partial x_n}{\partial t_m} \end{bmatrix}.$$

This formulation is particularly useful when dealing with complex systems where variables are interdependent. It allows us to compute the sensitivity of the function z with respect to the underlying variables $\mathbf{t}$ without explicitly solving for $\mathbf{x}$.

In addition to composite functions, we also encounter cases where we implicitly define one variable in terms of others. For example, if we have a function $G(\mathbf{x}, y) = 0$ that implicitly defines y as a function of $\mathbf{x}$, the implicit function theorem provides a way to differentiate y with respect to $\mathbf{x}$.

If $G(\mathbf{x}, y) = 0$ and $\frac{\partial G}{\partial y} \neq 0$, the implicit function theorem states that there exists a function

$y = h(\mathbf{x})$ such that $G(\mathbf{x}, h(\mathbf{x})) = 0$, and the derivative of y with respect to $\mathbf{x}$ is given by:

$$\frac{\partial y}{\partial \mathbf{x}} = -\left(\frac{\partial G}{\partial y}\right)^{-1} \cdot \nabla_{\mathbf{x}} G(\mathbf{x}, y).$$

This result is crucial in many areas of mathematics and economics, where it allows us to determine how an implicitly defined variable changes in response to changes in other variables.

4 Optimality Conditions for Unconstrained Optimization

4.1 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 4.1. *(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 4.1. *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* $-\frac{1}{\sqrt{2}}$.

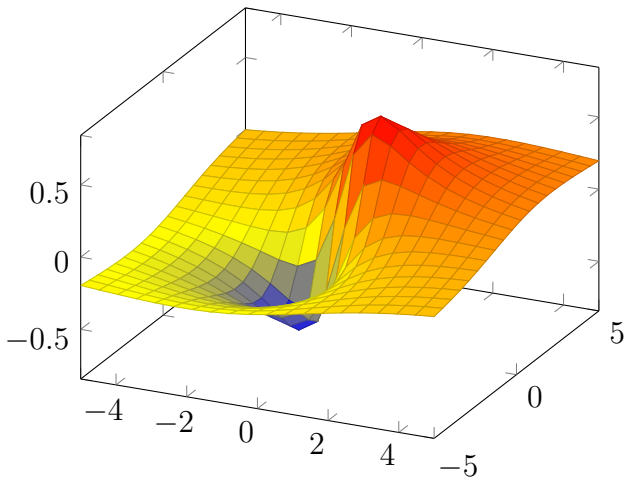

Figure 4.1: Surface plots of $f(x, y) = \frac{x+y}{x^2+y^2+1}$

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

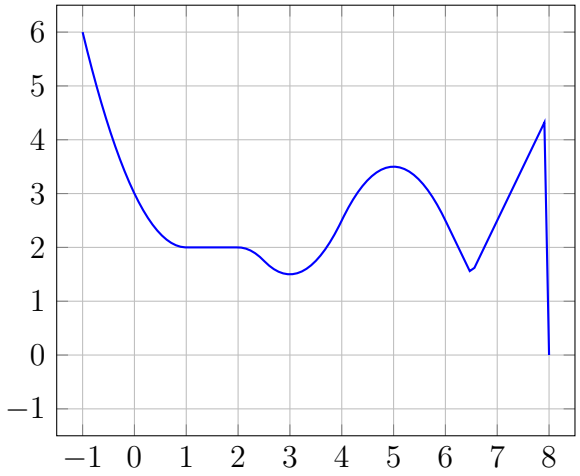

Figure 4.2: Local and global optimum points of a one-dimensional function

optimal points with respect to a neighborhood of the point of interest. The exact definitions follow.

Definition 4.2. *(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.

4.2 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 4.1. *(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 4.3. *(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.

4.3 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 4.2. *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 4.3. *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*.*

5 Convex Function

5.1 Definition and Examples

Definition 5.1. *(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*

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in C, \lambda \in [0,1] \tag{3}$$

The fundamental inequality **??** is illustrated in the following figure.

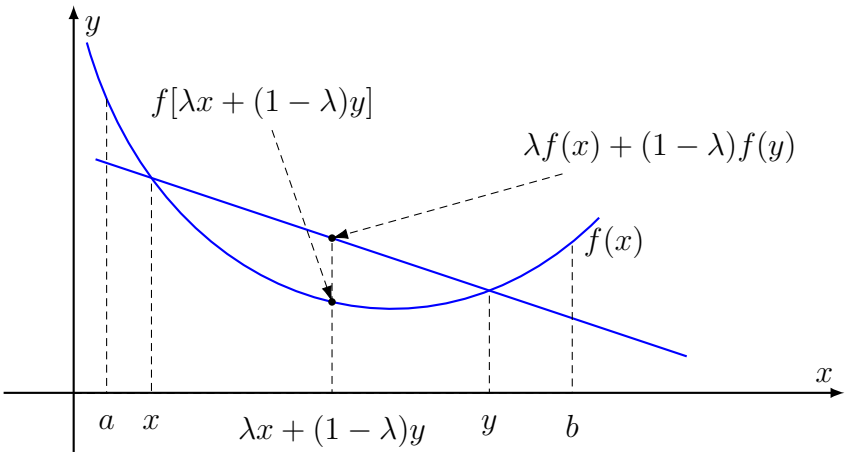

Figure 5.1: Illustration of inequality $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$

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 **??** when $\boldsymbol{x} \neq \boldsymbol{y}$ and $\lambda \in (0,1)$, the function is called *strictly convex*.

Definition 5.2. *(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*

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

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

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

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 5.1. *(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 $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$, 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 $\mathbf{x}$ and $\mathbf{y}$ 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 5.1. *(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:*

$$f(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i) \leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i). \tag{4}$$

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 **??** 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 **??** 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

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

■

5.2 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 5.2. *(the gradient inequality) Let $f : C \to \mathbb{R}$ be a continuously differentiable function defined on a convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if*

$$f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \leq f(\boldsymbol{y}) \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in C. \tag{5}$$

Proof. Suppose that f is convex. Let $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\lambda \in (0, 1]$. If $\boldsymbol{x} = \boldsymbol{y}$, then **??** 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 **??** 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

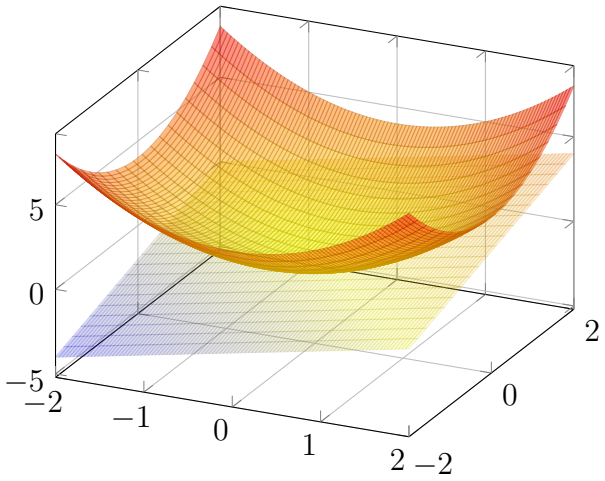

Figure 5.2: The function $f(x, y) = x^2 + y^2$ and its tangent hyperplane at $(1, 1)$, which is a lower bound of the function's surface.

optimality condition $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is sufficient for global optimality.

Proposition 5.1. *(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 **??**, 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.**

5.3 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 5.3. *(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

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{z})^T(\boldsymbol{y} - \boldsymbol{x}) \tag{6}$$

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 **??**, 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 $\boldsymbol{x} \in C$. ■

5.4 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 5.4. *(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 5.5. *(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.

5.5 Relationship between Concavity and Optimization

Note that by utilizing theorem 5.2 we immediately know that for a concave (convex) function, stationary point is also a global minimum (maximum). Therefore, if the objective function is concave (convex), we could save our efforts from verifying the sufficient conditions.

Dive a bit deeper, we can see that the second order Hessian matrix of a concave function is guaranteed to be negative semi-definite everywhere on the domain. And this is stronger than the sufficient condition which only requires the matrix being semi-definite locally around the stationary point. Given a set of normal concave function and operations that preserve the concavity, we are able to conviniently skip the complicated calculation of Hessian matrix when figuring out optimum.

6 Constrained Optimization

Constrained optimization refers to the case of finding maximum/minimum of a function on a non-conventional domain. A typical constrained optimization takes the following form:

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

s.t. means either "such that" or "subject to", and the following equations and inequalities define the domain of the function.

6.1 Intuition

Constrained optimization usually corresponds to the real world case of allocating a scarce resources. Therefore, the objective function is usually unbounded on its natural domain, and the optimum is contingent on the constraint we have. In the univariate case, we know that if we want to restrict the domain, we usually result in really simple one direction or two direction inequalities e.g. $x < 1, 1 \leq x \leq 5$. If the optimum is contingent on the constraint, it must lie on the "edge" of constraint, and in this case one of two end points. In multivariate case, however, the edges usually consist of not two, but infinitely many points. The problem is then turned to selecting a point on the edge. Consider the simplest case of one constraint:

$$\begin{aligned} \max \quad & f(\boldsymbol{x}) \\ s.t. \quad & g(\boldsymbol{x}) = \mathbf{0}. \end{aligned}$$

We could transform this to an unconstrained optimization problem by replacing one variable with the others using the constraint. Specifically, let $\boldsymbol{x}_{-n}$ denote the collection of the first 1 to $n-1$ variables, and assume that we could derive from constraint an implicit function: $x_n = G(\boldsymbol{x}_{-n})$. From implicit function theorem, we have

$$\nabla G(\boldsymbol{x}_{-n}) = -(\frac{\partial g}{\partial x_n})^{-1}(\nabla g)_{-n}$$

where $(\nabla g)_{-n}$ denote the first $n-1$ terms of the gradient of g. With this replacement we can transform the problem to

$$\max f(\boldsymbol{x}_{-n}, G(\boldsymbol{x}_{-n}))$$

Consider the first order condition, we have

$$(\nabla f)_{-n} + \frac{\partial f}{\partial x_n}\nabla G(\boldsymbol{x}_{-n}) = \mathbf{0}$$
$$\Rightarrow (\nabla f)_{-n} = \frac{\partial f}{\partial x_n}(\frac{\partial g}{\partial x_n})^{-1}(\nabla g)_{-n}$$

The above equation means: for any i of the first $(n-1)$ variable, the ratio between $\frac{\partial f}{\partial x_i}$ and $\frac{\partial g}{\partial x_i}$ is the same. This conclusion also holds when we consider the $n-$th variable, as

$$\frac{\partial f}{\partial x_n} = \frac{\partial f}{\partial x_n}(\frac{\partial g}{\partial x_n})^{-1}\frac{\partial g}{\partial x_n}$$

To summarize, the first order condition indicates that the optimum should be some point on which the gradient of objective function and the gradient of constraint function are of the same direction i.e. we are able to find some scalar λ such that

$$\nabla f(\boldsymbol{x}_0) = \lambda \nabla g(\boldsymbol{x}_0)$$

Geometrically, for a function f defined on $\mathbb{R}^n$, the equation $f(\boldsymbol{x}) = m$ defines an isoquant contour of the function, and the gradient evaluated at the point, $\nabla f(\boldsymbol{x}_0)$ is the "direction" of the line/plane/hyperplane that is tangent to the contour at the point $\boldsymbol{x}_0$. Take a two-dimensional function as an example: let $f(x, y) = x^2 + y^2$. We know that for any positive real number m, $x^2 + y^2 = m$ defines an isoquant contour which is a circle. And at any given point (x_0, y_0), the gradient $\begin{bmatrix} 2x_0 \\ 2y_0 \end{bmatrix}$ defines the line that is tangent to the contour at the point. Specifically, we know that the line $\begin{bmatrix} 2x_0 \\ 2y_0 \end{bmatrix}[x, y] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is tangent to the circle $x^2 + y^2 = x_0^y + y_0^2$ at the point (x_0, y_0). Therefore, we conclude that if the optimum lies on the edge defined by $g(\boldsymbol{x}) = 0$, it must be the point where the tangent hyperplane of both the objective function and the constraint are the same.

We can extend this argument in two dimensions. Firstly, if we consider more than one edge and attempt to find optimum on the intersection of several edges. Then the optimum must be the point where the tangent hyperplane of isoquant contour overlaps with the hyperplane of the intersection set i.e. there exists a set of scalars $\{\lambda_i\}$ such that

$$\nabla f(\boldsymbol{x}_0) = \sum_i \lambda_i \nabla g_i(\boldsymbol{x}_0)$$

Secondly, there may also be some constraints which the optimum do not lie on. We usually call such constraints "slack" as they are not binding at the optimum. The above formula also

holds for these constraints by simply setting the corresponding λ to 0.

To streamline the above procedures, we usually construct the **Lagrangian function**, denoted by $\mathcal{L}(\boldsymbol{x};\boldsymbol{\lambda})$, as follows:

$$\mathcal{L} = f(\boldsymbol{x}) - \sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{x})$$

And state the necessary first-order conditions as follows:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}} = \mathbf{0}$$
$$\lambda_i g_i(\boldsymbol{x}) = 0, i = 1, ..., p.$$

6.2 The KKT Conditions

After providing the intuition and practical steps in finding constrained optimum, we formally state the necessary conditions, called "Karush–Kuhn–Tucker (KKT) conditions", of constrained optimum. 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.

Theorem 6.1. *(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 linear 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 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

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

■

6.3 Envelope Theorem

Now suppose we address a well-defined constrained optimization, and obtain the result $\boldsymbol{x}^*$. We may then be interested in the properties of the solution. That is, if we have several other coefficients in the constraints and objective function, the optimum value $y^* = f(\boldsymbol{x}^*)$ will be a function of these coefficients. For example, if you want to maximize your production under a given budget of purchasing inputs, the increase in price or change in technology will impact your maximum output. If we have explicit functional forms, we could directly write out the closed-form solution, and the thing will go easy. While we also have some general properties that do not depend on specific functional forms, which we usually call "envelope theorem". Formally, consider an objective function $f(\boldsymbol{x}, \theta)$ that we wish to maximize subject to constraints $g(\boldsymbol{x}, \theta) = 0$, where θ is a parameter of interest. Let $\boldsymbol{x}^*(\theta)$ denote the optimal solution and $y^*(\theta) = f(\boldsymbol{x}^*(\theta), \theta)$ denote the optimal value. The envelope theorem states that the derivative of the optimal value with respect to the parameter θ is given by:

$$\frac{dy^*(\theta)}{d\theta} = \frac{\partial \mathcal{L}(\boldsymbol{x}^*, \lambda^*, \theta)}{\partial \theta}$$

where $\mathcal{L}(\boldsymbol{x}, \lambda, \theta)$ is the Lagrangian of the problem, and λ^* is the vector of Lagrange multipliers at the optimum. This result allows us to easily assess the sensitivity of the optimal value to changes in the parameters.

6.4 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 6.1. *(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 6.2. *(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 6.1. *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. $U(x_1, x_2) = (x_1^{\alpha} + x_2^{\alpha})^{\frac{1}{\alpha}}$