Mathematics for Economists

Juuso Välimäki

Spring 2020

Contents

1 Economics **3**
1.1 Optimization problem 3
1.2 Endogenous and exogenous variables 4
1.3 Mathematical formulation 4
1.4 Mathematical structure 6
1.5 Language and vocabulary of mathematics: 7

2 Preliminary material on functions **10**
2.1 Functions and linear functions 10
2.2 Non-linear functions 13
2.2.1 Solving non-linear models 15
2.3 The derivative 15

3 First look at multivariate calculus **20**
3.1 Partial derivative 20
3.2 The derivative of a multivariate real-valued function 21
3.3 The derivative of a vector-valued function 23
3.4 Applications 24
3.5 Computing the derivative 27

4 Comparative statics and the implicit function theorem **31**
4.1 Implicit function theorem for $n = m = 1$ 32
4.2 Linear implicit function theorem for $n > 1$ 34
4.3 Implicit function theorem for $n > 1$ 36
4.4 Applications 38

4.4.1 Drawing indifference curves 38
4.4.2 Drawing an isoquant 40
4.4.3 Comparative statics of consumer choice 41

5 Unconstrained optimization **43**
5.1 Local optima for $n = 1$ 43
5.1.1 First-order condition 44
5.1.2 Second-order conditions 44
5.2 Taylor's theorem 45
5.3 Quadratic functions 46
5.3.1 Extrema of quadratic functions 48
5.3.2 Application: Sum of least squares 48
5.4 Classifying extrema of quadratic functions 51
5.4.1 Quadratic forms and the definiteness of matrices . . 52
5.4.2 Definiteness with linear constraints 54
5.4.3 Examples 55
5.5 Definiteness and comparative statics 59

6 Convex and concave functions **61**
6.1 Basic definitions 61
6.2 Convexity and concavity of differentiable functions 64
6.3 Second derivatives and convexity 66
6.4 Quasiconvex and quasiconcave functions 66
6.5 Quasiconcavity and differentiability 69

1 Economics

- studies the allocation of scarce resources amongst competing ends
  - what are the ways to allocate?
  - how to evaluate the results?
  - what do we mean by scarcity?
  - how can we formalize such questions?
- individualistic approach: economic agents are autonomous decision makers
- they act in pursuit of individual objectives or goals
  - agents do not make systematic mistakes in their choices
  - they act within constraints
  - they react to changes on their environment
- equilibrium analysis to guarantee the consistency of individual decisions
  - in competitive markets: equilibrium brought about by price mechanism
  - in games: equilibrium from consistency of expectations and realized behavior

This course introduces mathematical methods that allow us to present and analyze the problem of an individual economic agent and basic tools for equilibrium analysis.

1.1 Optimization problem

- Economic agents (also called decision makers) have objectives summarized in their objective functions
  - utility function of a consumer
  - profit function of a firm

- total surplus for an economic planner

- Autonomous decisions:
  - each agent chooses her own actions
- Decisions in line with the objectives
  - each agent maximizes her objective function
- Economic choices constrained by scarcity

1.2 Endogenous and exogenous variables

- Economic models are chosen by the modeler
- Idea is to pick the most important features of an economic situation and ignore the rest
- Every model has variables that are determined within the model
  - endogenous variables
- Interesting models have variables not determined within the model
  - exogenous variables
- Exogenous variables and parameters of the model are similar in nature

1.3 Mathematical formulation

- In order to formulate the problem, we need the following ingredients:
  - Choices x from the set of choice variables: X
  - Evaluation of choices: objective function $f : X \to \mathbb{R}$
  - Scarcity in the form of feasible set: $F = \{x|g(x) \leq 0\}$
  - Possible parameters and other (exogenous) variables, $\{\alpha, \beta, ...\}$ to include in f, g

- Exogenous variables are not determined in the model

- For concreteness, let's consider some economic problems from Principles of Economics I

- Consumer choice between food and leisure
  - x_1 food consumption, x_2 leisure. $X = \{(x_1, x_2) | x_i \in \mathbb{R}, \text{ for } i \in \{1, 2\}, x_1 \geq 0, 0 \leq x_2 \leq 24\}$.
  - Utility from x: $f(x; \alpha) = f(x_1, x_2; \alpha)$, where α is a preference parameter
  - Feasible set: $p_1 x_1 \leq w(24 - x_2)$
  - Price of food p_1 and wages w are exogenous variables
  - Exercise: Write the constraint in form $g(x_1, x_2) \leq 0$.

- Best responses of player 1 in two player games:
  - Own action $x_1 \in X_1$ (row in the matrix)
  - Payoff from own action: $f(x_1; x_2)$, where x_2 is the exogenous variable (for best responses of 1, we just compute the payoff for all possible choices $x_2 \in X_2$, i.e. for all rows in the matrix)
  - In this context, no further feasibility constraint
  - Of course when solving the game, x_2 becomes also an endogenous variable.

- In general, we have the problem

$$\max_{x \in X} f(x; \alpha)$$
$$\text{subject to } g(x; \beta) \leq 0.$$

- What is a solution to the problem? An x^* such that
  - i)
  $$g(x^*; \beta) \leq 0.$$
  i.e. the solution is feasible.

- ii)
$$f(x^*;\alpha) \geq f(y;\alpha),$$
for all y such that:
$$g(y;\beta) \leq 0.$$

- In other words, optimal choice attains the highest value of the objective function within the feasible set

1.4 Mathematical structure

- What kinds of variables are x, α, β?
  - most often real numbers, real vectors or sometimes discrete choices (such as choosing the row in a matrix or choosing between a red and a blue car)
- When does a solution exist?
  - Weierstrass theorem (you will see this soon) or other existence results (to be just hinted at)
- How to find a solution?
  - one of the main questions for this course
  - usually with the help of calculus
  - calculus is not of much help for discrete problems, in more advanced courses tools for handling this to some extent
- Is the solution unique?
  - concavity and convexity of the objective function key for this
- How do endogenous variables react to changes in exogenous variables?
  - comparative statics
  - implicit function theorem is the key tool for this and one of our first goals in this course

1.5 Language and vocabulary of mathematics:

- A set is a collection of elements
- Sets are denoted by capital letters: $X, Y, Z, ...$
  - Sets can be defined extensively by enumerating all its elements (difficult for large sets such as real numbers):

$$X = \{x_1, x_2, ..., x_K\}.$$

  - Or based on a property:

$$X = \{x \in \mathbf{R} \,|0 \leq x < 1\}$$

  - Much used sets: real numbers, $\mathbb{R}$, integers $\mathbb{Z}$, rational numbers $\mathbb{Q}$, natural numbers $\mathbb{N}$ etc.
- Vectors:
  - Cartesian product of X and Y is a new set denoted by: $X \times Y$. It is defined as:

$$X \times Y = \{(x, y) \,|x \in X, y \in Y\}.$$

  - If $X = Y$, we write $X \times Y = X^2$.
  - For example real plane vectors are elements in $\mathbb{R}^2$:

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

  - Similarly n -dimensional real vectors, $x \in \mathbb{R}^n$:

$$\mathbb{R}^n = \{(x_1, ..., x_n) \,|x_1 \in \mathbb{R}, ..., x_n \in \mathbb{R}\}.$$

- Vector inequalities:
  - Consider $x, y \in \mathbb{R}^n$.

- $x = y$ if $x_i = y_i$ for all $i \in \{1, ..., n\}$.

  $x \geq y$ if $x_i \geq y_i$ for all $i \in \{1, ..., n\}$ and for some $i, x_i > y_i$.

  $x \gg y$ if $x_i > y_i$ for all $i \in \{1, ..., n\}$.

- Note that unlike for real numbers, we do not have for all x, y

  $$(x \geq y) \text{ or } (y \leq x) \text{ or both.}$$

- Consider an economy with k consumers and n goods.
  Vector $x^1 = (x_1^1, ..., x_n^1)$ is the consumption vector of consumer 1, vector $x^2 = (x_1^2, ..., x_n^2)$ is the consumption vector of 2 etc..
  Consumer j receives utility $u^j(x^j)$ from consumption x^j.
  Let $\overline{x}_i$ be the total quantity of good i available in the economy.
  An allocation for the economy is a list $x = \left(x^1, x^2, ..., x^k\right)$ of all consumption vectors.
  An allocation is feasible if for all goods $i \in \{1, ..., n\}$, we have:

  $$\Sigma_{j=i}^k x_i^j \leq \overline{x}_i.$$

  In words, an allocation is feasible if total consumption of each goods is no larger than total quantity available.
  Allocation $y = \left(y^1, ..., y^k\right)$ *Pareto-dominates* allocation $x = \left(x^1, x^2, ..., x^k\right)$, if

  $$u^j\left(y^j\right) \geq u^j\left(x^j\right).$$

  Allocation x is *Pareto-efficient* if x is feasible and there exists no feasible allocation y, Pareto-dominating x.

- Functions:

  - A function is a rule assigning for each element $x \in X$ of its domain X an element $y \in Y$ in its codomain Y
  - We write:

    $$f : X \to Y,$$

    and

    $$y = f(x).$$

- Functions $f : X \to Y$ and $g : Y \to Z$ can be composed to for the composite function $h : X \to Z$ as follows:

$$h(x) = g(f(x)).$$

- A function f is *injective* or *one-to-one* if

$$f(x_1) = f(x_2) \Rightarrow x_1 = x_2.$$

- A function f on *surjective* or *onto* if for all $y \in Y$ there exists $x \in X$ such that

$$y = f(x).$$

- A function is bijective if it is injective and surjective.
- A bijective function

$$f : X \to Y$$

has an inverse function

$$f^{-1} : Y \to X,$$

such that for all $x \in X$ and for all $y \in Y$:

$$x = f^{-1}(f(x)) \text{ and } y = f\left(f^{-1}(y)\right).$$

- Quantification: propositions with variables should be quantified.
- A proposition is a statement that can be assigned a truth value, i.e. it is either true or false.
- Statements involving free variables are not propositions
  - for example $x^3 \geq 0$ does not have a truth value because there are some x for which it is true and others for which it is false
  - all variables in any statement must be quantified
  - quantified statements are propositions
- Examples: For all x, we have $x^2 \geq 0$:

$$\forall x,\ x^2 \geq 0.$$

- There exists an x such that $x^2 \leq 0$:

$$\exists x,\ x^2 \leq 0.$$

- Denote any statement involving variable x by the $q(x)$. The previous examples are propositions of the form:

$$\forall x,\ q(x) \text{ and } \exists x,\ q'(x).$$

- Logical notation We write $p \Rightarrow q$ to denote 'if p is true, then q is true'. We write $p \Leftrightarrow q$ to denote 'p is true if and only if q is true' or '$p \Rightarrow q$ and $q \Rightarrow p$'.

- Let p be a proposition. We define the negation $\neg p$ by the rule:

$$\neg p \text{ is true } \Leftrightarrow p \text{ is false.}$$

- Examples of propositions involving quantified statements:

$$\neg (\forall x,\ q(x)).$$

  This is the same as:

$$(\exists x,\ \neg q(x)).$$

- Finally, we the notation $A := B$ to denote 'A is defined by B, where A, B can be sets, functions etc.

- In the materials for this course, you can find an additional handout on logic and the language of mathematics It is not required reading but may help you to get used to more formal mathematical notation.

2 Preliminary material on functions

2.1 Functions and linear functions

Recall that a function is a relation defined on the cartesian product of two sets X and Y. We call X the domain of a function f and we call Y its co-domain. A function associates for each element $x \in X$ a single element $y \in Y$. The common notation for this is that $y = f(x) \in Y$. We write also

$$f : X \to Y.$$

Functions are sometimes called mappings, maps or transformations. The sets X, Y can be very general and the function can take many forms. Here are some examples:

- X is the set of all humans, dead and alive and $Y = X$. For each x, $f(x)$ is the mother of x.
- X is the population of Finland, $Y = \{0, 1\}$. $f(x) = 1$ if x has Covid -antibodies in her blood and $f(x) = 0$ otherwise.
- X is the set of feasible portfolios, $Y = \mathbb{R}$. $f(x)$ is the expected return on portfolio x.
- X is the set of all humans, dead and alive and $Y = X$. The following relation $y = f(x)$ if y is a child of x is not a function since some x have many children and some have none.

Our goal in this part of the course is to gain an understanding of real-valued and vector-valued functions of multiple variables, i.e. functions

$$f : \mathbb{R}^m \to \mathbb{R}^n.$$

Such functions cover most of the cases that are relevant for mathematical models in economics. Examples that are particularly important include the following:

1. Preferences of a consumer expressed through a utility function defined on consumption vectors
2. Profit of a firm as a function of the vector of outputs and the inputs chosen
3. Simultaneous equilibrium in many markets as the intersection of demand and supply functions
4. Likelihood function of observed data as a function of the parameters of the distribution of error terms

When are functions easy to understand? Intuitively, one would require that this is the case when the function can be understood based on a few representative cases and then extrapolated to the entire population.

This is exactly the reason why linear functions are so nice to work with. In order to get simplicity of the type desired above, we need first of all an easy way to describe the domain X in terms of a few typical cases. We get this by assuming that X is a vector space. This is a big word for a simple idea. It just says that scalar multiples of elements of X are also elements of X. More formally, we require that if $a \in \mathbb{R}$ and $x \in X$, then $ax \in X$. Verify that the real line and k- dimensional vectors as defined at the beginning of these notes satisfy this requirement. The other requirement for a vector space is that whenever both x and x' are elements of X, we can define an operation $x + x'$ so that $x + x' \in X$. Again verify that the coordinate-wise addition for vectors (and real numbers) satisfies this.

For vectors of real numbers $X = \mathbb{R}^n$, we can make the useful observation based on the definitions above that with our definition of addition and scalar multiplication, we can write

$$x = (x_1, ..., x_n) = x_1\mathbf{e_1} + x_2\mathbf{e_2} + ... + x_n\mathbf{e_n}$$

$$= \sum_{i=1}^{n} x_i\mathbf{e_i},$$

where $\mathbf{e_i}$ is the i^{th} unit coordinate vector.

A function $f : X \to Y$ between two vector spaces X and Y is called linear if for all $x, x' \in X$ and all $a \in \mathbb{R}$, i) $f(ax) = af(x)$ and ii) $f(x + x') = f(x) + f(x')$.

Here is the great thing about linear functions. If we know $f(\mathbf{e_i})$ for $i \in \{1, ..., n\}$, then we know $f(x)$ for all $x \in X$.

$$f(x) = f(\sum_{i=1}^{n} x_i\mathbf{e_i}) = \sum_{i=1}^{n} x_i f(\mathbf{e_i}).$$

Suppose now that $Y = \mathbb{R}^m$. Define $\mathbf{a_i} = f(\mathbf{e_i}) \in \mathbb{R}^m$. Then we can represent the linear function f by an $(m \times n)$ -matrix A:

$$A = [\mathbf{a_1}, ..., \mathbf{a_n}].$$

and

$$f(x) = Ax.$$

Similarly multiplication by an arbitrary $(m \times n)$ -matrix A gives rise to a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$.

We say that a function $f : X \to Y$ is surjective or onto if for all $y \in Y$, there is (at least one) $x \in X$ such that $y = f(x)$. f is called injective or one-to-one if for each $yinY$, there is at most one $x \in X$ such that $y = f(x)$. Finally, it is called bijective if it is both injective and surjective.

Note that for a bijective function, we can define an inverse function $f^{-1} : Y \to X$ so that we have for all $x \in X$ and all $y \in Y$:

$$x = f^{-1}(f(x)), \quad y = f(f^{-1}(y)).$$

Hence for a linear function f represented by an ($m \times n$ -matrix A, we can use our results from the systems of equations to classify f. f is surjective if for all $y \in Y$, there is an $x \in X$ such that:

$$Ax = y.$$

Recall that this is true if $rank(A) = m$.

It is injective if $rank(A) = n$. It is bijective if $m = rank(A) = n$. In this case, the inverse function f^{-1} is represented by the inverse matrix A^{-1}.

Exercise: A real valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *homogenous of degree 1* or *linearly homogenous* if for all $x \in \mathbb{R}^n$ and all real λ, $f(\lambda x) = \lambda f(x)$. Give an example of a function that is linearly homogenous but not linear. Exercise (harder): A real valued function $f : \mathbb{R}^k \to \mathbb{R}$ is said to be additive if for all $x, y \in \mathbb{R}^k$, $f(x + y) = f(x) + f(y)$. Show that for additive functions on $\mathbb{R}$, we have $f(qx) = qf(x)$ for all rational q. This shows that for functions defined on Q, linearity coincides with additivity. Do not try the following two parts unless you really like abstract mathematics. (Still harder): With this you can show that for continuous functions on the real line, additivity and linearity coincide. (Really hard). It can be shown that without continuity, real-valued functions on the real that are additive but not linear exist. Luckily for us, continuity is not a bad property to assume.

2.2 Non-linear functions

Unfortunately linear functions do not fit well in all economic situations. A partial list of issues that limit the usefulness of linear models is:

i) Diminishing marginal returns are not captured by linear models

ii) Only linear indifference curves are consistent with linear utility functions

iii) There is no reason why prices and quantities should have linear dependencies

Because of these and other problems, we often have to resort to models where functions are not linear, i.e. they are non-linear.

Examples

1. Utility function
   - A consumer considers consuming k different goods
   - A consumption plan is a positive vector $x \in \mathbb{R}^k_+$.
   - A consumer has rational preferences if i) for all consumption plans x, y, she either prefers x to y or $y$to x or both, ii) for all consumption plans x, y, z, if she prefers x to y and y to z, then she prefers x to z.
   - In a later course in microeconomics, you will see that (continuous) rational preferences can be represented by a utility function $u : \mathbb{R}^k_+ \to \mathbb{R}$
   - This means that $u(x) = u(x_1, ...x_k) \geq u(y_1, ...y_k) = u(y)$ if and only if x is preferred to y
   - In other words, the utility function is just a convenient summary of the preferences
   - Non-linearity of u reflects non-constant marginal rates of substitution between goods, diminishing marginal utility, in models of uncertainty the risk attitudes etc.
2. Production function
   - A firm produces output from two inputs, labor $L \in \mathbb{R}_+$ and capital $K \in \mathbb{R}_+$
   - the output is given by $Y = f(K, L) \in \mathbb{R}_+$
   - Non-linearity arises from diminishing marginal returns, diminishing diminishing or increasing returns to scale

2.2.1 Solving non-linear models

The main problem in analyzing non-linear functions is that their shape varies sometimes greatly as x moves around in the domain X. Consider a linear function $f : \mathbb{R}^2 \to \mathbb{R}$. For all $\hat{x}_0 = (\hat{x}_1, \hat{x}_2)$, we have

$$f(\hat{x} + y) - f(\hat{x}) = f(\hat{x}_1 + y_1, \hat{x}_2 + y_2) - f(\hat{x}_1, \hat{x}_2)$$

$$= f(\hat{x}) + f(y_1, y_2) - f(\hat{x}) = f(y_1, y_2) = f(y).$$

In other words, moving in the direction y induces a change in the value of the function that is independent of the starting point of the movement $\hat{x}$. This is not true of non-linear functions and the analysis is unfortunately a lot more difficult.

Options:

i) Numerical methods. Graph the function using Matlab, R or some other language and find solutions to the model using numerical algorithms. Not pursued in this course.

ii) Local methods Idea: A well behaved function is well approximated by a (different) linear function near any point in its domain. Use the linear approximation to make inferences about the true non-linear function. Analogy, the surface of the earth is curved, i.e. non-linear but for most everyday uses, a two-dimensional approximation (a map) is good enough.

Issue: What does 'near enough' mean?→ definitions for distance and limits come useful here

The main tool for local analysis: differential calculus

2.3 The derivative

Recall the definition of a function $f : \mathbb{R} \to \mathbb{R}$ at point $x_0 \in \mathbb{R}$ for a real-valued function of a single variable. Form the difference quotient

$$\frac{f(x) - f(x_0)}{x - x_0}$$

and consider the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

If the limit exists, we denote it by

$$Df\left(x_0\right)$$

or sometimes by

$$f'\left(x_0\right)$$

and call it the derivative of f at x_0.

Remark 1 (Approximation) *When a derivative exists, we have from the definition of limits:*

$$\forall \varepsilon > 0, \exists \delta > 0 \textit{ such that whenever } |x - x_0| < \delta,$$

$$\textit{we have: } |\frac{f\left(x\right) - f(x_0)}{x - x_0} - Df(x_0)| < \varepsilon.$$

In other words,

$$f\left(x\right) - f\left(x_0\right) = Df\left(x_0\right)\left(x - x_0\right) + r(x),$$

where $\frac{r(x)}{|x-x_0|} \to 0$ *when* $x \to x_0$*. In this case, we say that* $r(x)$ *consists of higher order terms in* $(x - x_0)$*. Higher order terms vanish more quickly than the first-order term* $Df\left(x_0\right)\left(x - x_0\right)$ *as* $x \to x_0$*.*

This says simply that near x_0*, the changes in* $f(x)$ *are well approximated by the linear function* $Df(x_0)(x - x_0)$ *of changes in* x*. This view of the derivative generalizes to functions of many variables and also to vector-valued functions.*

From now on, we simply say for differentiable functions that for x near x_0 or for small $|x - x_0|$,

$$f(x) = f(x_0) + Df(x_0)(x - x_0).$$

Let's make some preliminary observations. Recall that a function is continuous at x_0 if $\lim_{x \to x_0} f(x) = f(x_0)$. Hence we see immediately that if a derivative exists at x_0, then f must be continuous at x_0.

You can easily compute the derivative of the linear function $f(x) = ax$ at x_0 on the real line to be $Df(x_0) = a$. Notice that the derivative is the same at all points. For a non-linear function on the real line, e.g. $g(x) = x^2$, you get by computing the limit that $Df(x_0) = 2x_0$. Notice that the approximating linear function is different for all points x_0 in the domain.

Exercise: Show that the continuous function with $f(x) = |x|$ does not have a derivative at x_0.

If f has a derivative at x_0, we say that it is *differentiable at* x_0. If it differentiable at all points $x_0 \in \mathbb{R}$, we just say that it is differentiable. We use this same terminology also for multivariate and vector-valued functions.

We can also write the derivative $Df(x)$ as a function depending on the point x where it is evaluated.

Rules for derivatives for $f : \mathbb{R} \to \mathbb{R}$

1. If $f(x) = a$ for all x, then $Df(x_0) = 0$ for all x_0.
2. If $f(x) = x$ then $Df(x_0) = 1$ for all x_0.

   (Linear homogeneity) If $g(x) = af(x)$, then $Dg(x_0) = aDf(x_0)$.
3. (Additivity) Let $h(x) = f(x) + g(x)$. Then $Dh(x_0) = Df(x_0) + Dg(x_0)$.
4. (Product rule) Let
$$\phi(x) = f(x) g(x).$$
   Then
$$\begin{aligned} D\phi(x_0) &= Df(x) g(x) = \lim_{h\to 0} \frac{\phi(x_0+h) - \phi(x_0)}{h} \\ &= \lim_{h\to 0} \frac{f(x_0+h) g(x_0+h) - f(x_0) g(x_0)}{h} \\ &= \lim_{h\to 0} \frac{(f(x_0+h) - f(x_0)) g(x_0+h) - f(x_0)(g(x_0) - g(x_0+h))}{h} \\ &= f'(x_0) g(x_0) + f(x_0) g'(x_0). \end{aligned}$$
5. (Chain rule) Let
$$\zeta(x) = g(f(x)).$$

Then

$$\begin{aligned} D\zeta(x_0) &= Dg(f(x)) = \lim_{h\to 0} \frac{\zeta(x_0+h)-\zeta(x_0)}{h} \\ &= \lim_{h\to 0} \frac{g(f(x_0+h))-g(f(x_0))}{h} \\ &= \lim_{h\to 0} \frac{g(f(x_0)+f'(x_0)h)-g(f(x_0))}{h} \\ &= \lim_{h\to 0} \frac{f'(x_0)(g(f(x_0+f'(x_0)h))-g(f(x_0)))}{f'(x_0)h} \\ &= g'(f(x_0))f'(x_0). \end{aligned}$$

6. With these formulas, we can compute most derivatives that we need. For example, the derivative $D\phi(x_0)$ at x_0 for the function

$$\phi(x) = x^2$$

is obtained from the product rule

$$f(x) = g(x) = x.$$

We get

$$D\phi(x_0) = x_0 + x_0 = 2x_0.$$

By 'mathematical induction', we can see that for

$$f(x) = x^a,$$

$$Df(x_0) = ax_0^{a-1}.$$

By additivity and linear homogeneity, we can extend this to get the derivatives of all polynomial functions.

7. The rule for derivatives of quotients follows from the product rule. For $g(x) \neq 0$,

$$h(x) = \frac{f(x)}{g(x)}$$

can be written as:

$$h(x)g(x) = f(x).$$

Therefore

$$h'(x_0)\,g(x_0) = f'(x_0) - h(x_0)\,g'(x_0)$$

and therefore

$$h'(x_0) = \frac{f'(x_0)}{g(x_0)} - \frac{f(x_0)\,g'(x_0)}{(g(x_0))^2}.$$

8. The inverse function rule is a consequence of the chain rule: For all x, we have:

$$f^{-1}(f(x)) = x$$

Taking derivatives on both sides and denoting $y_0 = f(x_0)$:

$$Df^{-1}(y_0)\,Df(x_0) = 1$$

and therefore:

$$Df^{-1}(y_0) = \frac{1}{Df(x_0)}.$$

9. A case that is not covered by the previous ones is the exponential function:

$$f(x) = e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}...$$

We have not proved this, but for convergent power series as the one above, we may differentiate element by element to get:

$$Df(x_0) = 0 + 1 + x_0 + \frac{x_0^2}{2} = \sum_{n=1}^{\infty} \frac{x_0^{n-1}}{(n-1)!} = e^{x_0}.$$

10. The logarithmic function denoted by $\ln(y)$ is the inverse function of the exponential function (defined for strictly positive y):

$$g(y) = \ln y,$$

$$f(x) = e^x,$$

$$g(f(x)) = x.$$

By chain rule:

$$Dg(y_0)\,Df'(x_0) = 1, \text{ for all } x_0 \text{ and } y_0 = e^{x_0}.$$

Therefore

$$Dg\left(y_0\right) = \frac{1}{Df\left(x_0\right)} = \frac{1}{f\left(x_0\right)} = \frac{1}{y_0}.$$

So we have:

$$D \ln y_0 = \frac{1}{y_0}.$$

11. Trigonometric functions etc. can be differentiated using their representations as complex power series or via direct limit arguments using basic identities from trigonometry.

Let's pause to take stock of what we have learned so far. The derivative gives a linear approximation to differentiable functions of a real variable. What does it mean if we know that $f'(x_0) > 0$? For x near x_0, the function is strictly increasing. This means that for x, x' near x_0, $f(x) > f(x')$ whenever $x > x'$. Similarly, if $f'(x_0) < 0$, then the function is strictly decreasing near x_0. If the sign of the derivative is always strictly positive (strictly negative), then the function is strictly increasing (strictly decreasing) everywhere.

To see that it is harder to say if a function is increasing, decreasing or neither around a point x_0 where $f'(x_0) = 0$, consider $f(x) = x^2, g(x) = -x^2, h(x) = x^3$ near point $x_0 = 0$.

3 First look at multivariate calculus

3.1 Partial derivative

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, where we write

$$y = f(x) = f\left(x_1, x_2, ..., x_n\right).$$

For the moment, consider f as a function of x_i :n only and fix the other variables x_j for $j \neq i$ at $x_j = \widehat{x}_j$. We can then define the derivative of f with respect to x_i exactly as before around the point $\widehat{x} = (\widehat{x}_1, ..., \widehat{x}_n)$:

$$\frac{\partial f\left(\widehat{x}\right)}{\partial x_i} = \lim_{x \to \hat{x}_i} \frac{f\left(\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_{i-1}, x_i, \widehat{x}_{i+1}, ..., \widehat{x}_n\right) - f\left(\widehat{x}\right)}{x_i - \hat{x}_i}.$$

We call this limit the partial derivative of f with respect to x_i at $\hat{x}$. All the rules for computing derivatives remain valid for computing partial

derivatives. As the name suggests, partial derivatives capture the effect on the function from changing a single coordinate from a fixed initial point $\hat{x}$ in the domain. Since all the other coordinates are fixed,

$$f\left(\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_{i-1}, x_i, \widehat{x}_{i+1}, ..., \widehat{x}_n\right)$$

is really a function of the single real variable x_i.

Similar to the univariate case, we see that if $\frac{\partial f(\widehat{x})}{\partial x_i} > 0$, then f is strictly increasing in variable x_i near point $\hat{x}$.

We can also write:

$$\frac{\partial f\left(\widehat{x}\right)}{\partial x_i} = \lim_{h \to 0} \frac{f\left(\widehat{x} + h\mathbf{e}_i\right) - f\left(\widehat{x}\right)}{h},$$

where $\mathbf{e}_i$ is the i^{th} unit vector and $h = x_i - \hat{x}_i$.

3.2 The derivative of a multivariate real-valued function

How to define the derivative of a multivariate function t at $\widehat{x}$? We could try with something along the lines:

$$Df\left(\widehat{x}\right) = \lim_{x \to \hat{x}} \frac{f\left(x\right) - f\left(\widehat{x}\right)}{x - \hat{x}}.$$

But this is nonsense since on the right hand side, we have a real number divided by a vector and this is not defined.

If we stick with our goal of finding a linear approximation around point $\hat{x}$ to the non-linear function f, we could say that f has a derivative at $\hat{x}$ if there exists a linear function $Df(\hat{x})$ such that for small $\Delta x = (\Delta x_1, ..., \Delta x_n)$:

$$f\left(\widehat{x} + \Delta x\right) - f(\hat{x}) = Df\left(\widehat{x}\right)\Delta x$$

Remark 2 (Approximation for multivariate functions) *How to make the above statement mathematically precise? We say that a vector $y \in \mathbb{R}^n$ is small if $\|y\| := \sqrt{\sum_i^n y_i^2}$ is small. In particular, if $\|y\| < \varepsilon$, then $y_i < \varepsilon$ for all $i \in \{1, ..., n\}$. Notice that by the Pythagorean theorem, $\|y\|$ measures the distance of y form the origin. This is also called the norm of vector y.*

The exact mathematical formulation for the statement above is thus the following: $\forall \varepsilon > 0, \exists \delta > 0$ such that

$$\frac{|f(\widehat{x}+\Delta x)-f(\hat{x})-Df(\widehat{x})\,\Delta x|}{\|\Delta x\|}<\varepsilon \text{ whenever } \|\Delta x\|<\delta.$$

From now on, this is the formal mathematical meaning for our statements that assert equality of two objects whenever some quantity is small.

We have already seen that linear functions from $\mathbb{R}^n$ to $\mathbb{R}$ are given by an inner product with a vector $a \in \mathbb{R}^n$. In other words, $Df(\widehat{x})$ is a row vector. Since $Df(\hat{x})$ is defined as a linear function, it is completely determined by its values for unit vectors. This means that we can write the derivative as:

$$Df(\widehat{x})=\left(\frac{\partial f(\widehat{x})}{\partial x_1},...,\frac{\partial f(\widehat{x})}{\partial x_n}\right).$$

One can also show that if all partial derivatives of f exist and if they are continuous in $\hat{x}$ at $\hat{x}$, then f has a derivative at $\hat{x}$. It is good to know that for most models in economics, partial derivatives do exist and they are continuous. Hence the derivatives exist in almost all cases.

The following alternative way of denoting the partial derivative of f at $\widehat{x}$ is often used:

$$\frac{\partial f(\widehat{x})}{\partial x_i}=f_{x_i}(\widehat{x})$$

Sometimes one needs the column vector of partial derivatives at $\hat{x}$. It is called the gradient of f at $\hat{x}$ and denoted by:

$$\nabla f(\widehat{x})=\begin{pmatrix}\frac{\partial f(\widehat{x})}{\partial x_1}\\ \vdots \\ \frac{\partial f(\widehat{x})}{\partial x_n}\end{pmatrix}=\begin{pmatrix}f_{x_1}(\widehat{x})\\ \vdots \\ f_{x_n}(\widehat{x})\end{pmatrix}=Df(\hat{x})^{\top}.$$

The gradient can be given a geometric interpretation as the direction of fastest growth of f at $\widehat{x}$. We can ask for the direction Δx, that achieves the biggest change in the value of f.

$$f(\widehat{x}+\Delta x)-f(\widehat{x})=Df(\widehat{x})\cdot\Delta x,$$

The quantity $Df(\widehat{x})\cdot\Delta x$ is sometimes called the directional derivative of f at $\hat{x}$ in direction Δx. By the geometric interpretation of dot product, you

will recall from high school that for fixed length vectors, the dot product is maximized when the vectors are parallel:

$$\Delta x = \nabla f\left(\widehat{x}\right).$$

Because of this, the gradient at $\hat{x}$ gives the direction of steepest increase for the value of f at $\hat{x}$.

3.3 The derivative of a vector-valued function

A vector valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ can be written as a vector of real valued multivariate functions:

$$y = f\left(x\right) = \begin{pmatrix} f_1\left(x\right) \\ \vdots \\ f_m\left(x\right) \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix},$$

where for each i,

$$f_i\left(x\right) : \mathbb{R}^n \to \mathbb{R}.$$

Recall that f has a derivative at $\widehat{x}$ if there is a linear function $Df(\widehat{x})$ such that for small Δx,

$$f\left(\widehat{x} + \Delta x\right) - f\left(\widehat{x}\right) = Df(\widehat{x}) \cdot \Delta x.$$

Remark 3 (Approximating vector-valued functions) *The exact mathematical formulation for the statement above is almost identical to the formulation in the previous subsection:* $\forall \varepsilon > 0, \exists \delta > 0$ *such that*

$$\frac{\left\| f\left(\widehat{x} + \Delta x\right) - f(\hat{x}) - Df\left(\widehat{x}\right)\Delta x \right\|}{\|\Delta x\|} < \varepsilon \text{ whever } \|\Delta x\| < \delta.$$

The only change is that we use the norm of the change in value since f *is vector-valued.*

If f has a derivative at $\widehat{x}$, the results in the previous section imply that:

$$f_i\left(\widehat{x} + \Delta x\right) - f_i\left(\widehat{x}\right) = Df_i(\widehat{x}) \cdot \Delta x.$$

This means that we can write the derivative of f at $\hat{x}$ as:

$$Df(\widehat{x}) = \begin{pmatrix} Df_1(\widehat{x}) \\ \vdots \\ Df_m(\widehat{x}) \end{pmatrix}.$$

In other words,

$$Df(\widehat{x}) = \begin{pmatrix} \frac{\partial f_1(\widehat{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\widehat{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\widehat{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\widehat{x})}{\partial x_n} \end{pmatrix}.$$

Again, it can be shown that when all the partial derivatives $\frac{\partial f_m(\widehat{x})}{\partial x_n}$ exist and are continuous in $\hat{x}$, then the derivative $Df(\hat{x})$ exists.

We note finally that since all rules for computing derivatives are valid for computing partial derivatives, many rules have multidimensional generalizations. In particular, the chain rule remains valid. If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^k$, are both differentiable at $\hat{x}$ and $f(\hat{x})$ respectively, then the function $h(x) := g(f(x))$ is differentiable at $\hat{x}$ and

$$Dh(\hat{x}) = Dg(f(\hat{x}))Df(\hat{x}).$$

Write out the sum in terms of partial derivatives to see the total effect of a change in x_i on a $z_j = g_j(f_1(x_1, ..., x_n), ..., f_m(x_1, ..., x_n))$.

3.4 Applications

- Unconstrained optimization in $\mathbb{R}^n$
  - A real-valued function $f(x)$ has a maximum at $\widehat{x}$ if

    $$f(y) \leq f(\widehat{x}) \text{ kaikille } y \in \mathbb{R}^n.$$

    Since

    $$y = \widehat{x} + (y - \widehat{x}),$$

    we can write:

    $$f(\widehat{x} + (y - \widehat{x})) - f(\widehat{x}) \leq 0.$$

for $|y - \hat{x}|$ small, we have

$$f(\widehat{x} + (y - \widehat{x})) - f(\widehat{x}) = Df(\widehat{x}) \cdot (y - \widehat{x}).$$

If $\hat{x}$ is a maximum, then

$$Df(\widehat{x}) \cdot (y - \widehat{x}) \leq 0.$$

Since y is arbitrary, we get a necessary condition for the optimum:

$$Df(\widehat{x}) = 0,$$

eli

$$\frac{\partial f(\widehat{x})}{\partial x_i} = 0 \text{ kaikille } i.$$

- We will soon see examples of unconstrained optimization.

- Comparative statics:

  - Assume that the endogenous variables $y \in \mathbb{R}^n$ and exogenous variables $x \in \mathbb{R}^k$ satisfy the equations

$$\begin{array}{rcl} f_1(y, x) & = & 0, \\ \vdots & & \\ f_n(y, x) & = & 0, \end{array} \tag{1}$$

    at point $(\widehat{y}, \widehat{x})$.

  - How do small changes in the exogenous variables affect the endogenous variables?

  - Tool: Implicit function theorem. This is the main topic in the next few lectures. We are interested in two questions: i) Do solutions $(y(x), x)$ to 1 exist for x near $\hat{x}$ (i.e. $\|x - \hat{x}\|$ small)? ii) How can we characterize

$$\begin{array}{c} y_1(x_1, ..., x_k), \\ \vdots \\ y_n(x_1, ..., x_k). \end{array}$$

To answer both of these questions, we must consider the derivative of f.

- Constrained optimization:
  - Let B be the feasible set for an optimization problem.
  - The problem has an optimal solution at $\widehat{x}$ if

$$f(y) \leq f(\widehat{x}) \text{ for all } y \in B.$$

  - Notice that now the y cannot be chosen freely. They must satisfy $y \in B$.
  - Therefore not all

$$y = \widehat{x} + (y - \widehat{x})$$

    are feasible even when $\|y - \widehat{x}\|$ is small.
  - We have to find the directions Δx, for which

$$\widehat{x} + \Delta x \in B$$

    when Δx is small. When B is given by a constraints $g : \mathbb{R}^n \to \mathbb{R}^k$ so that $x \in B$ if

$$\begin{aligned} g_1(x) &\leq 0, \\ &\vdots \\ g_k(x) &\leq 0, \end{aligned} \tag{2}$$

    then we can characterize the feasible directions Δx the derivative $Dg(\hat{x})$.
  - If $\hat{x}$ is an optimum, must have

$$Df(\widehat{x}) \cdot \Delta x \leq 0,$$

    for all feasible directions.
  - In a few weeks, we'll see in detail, how this is done.
- Linearizing non-linear dynamical systems. Maybe we'll have time for a few words on this at the end of the course.

3.5 Computing the derivative

- Compute at $(x_1, x_2, x_3) = (1, 2, 1)$ the derivative of the following function:

$$f(x_1, x_2, x_3) = x_1 \ln x_2 + \sqrt{x_2 x_3}.$$

Since we have a real-valued function f, its derivative is the row vector of its partial derivatives at an arbitrary point $x = (x_1, x_2, x_3)$:

$$Df(x) = \left(\frac{\partial f(x_1, x_2, x_3)}{\partial x_1}, \frac{\partial f(x_1, x_2, x_3)}{\partial x_2}, \frac{\partial f(x_1, x_2, x_3)}{\partial x_3} \right)$$

$$= \left(\ln x_2, \frac{x_1}{x_2} + \frac{1}{2} x_2^{-\frac{1}{2}} x_3^{\frac{1}{2}}, \frac{1}{2} x_2^{\frac{1}{2}} x_3^{-\frac{1}{2}} \right).$$

Evaluating at $(1, 2, 1)$

$$Df(1, 2, 1) = \left(\ln 2, \frac{1}{2} + \frac{1}{2\sqrt{2}}, \frac{\sqrt{2}}{2} \right).$$

- Consider the function

$$f : \mathbb{R}^3 \to \mathbb{R}^2,$$

where

$$f(x, y, z) = \begin{pmatrix} f_1(x, y, z) \\ f_2(x, y, z) \end{pmatrix} = \begin{pmatrix} x^2 y + yz \\ \frac{y - z}{x} \end{pmatrix}.$$

Our task is to compute the derivative of f at $(x, y, z) = (1, 1, 1)$.

Form the derivative of f as a martix of partial derivatives at an arbitrary $x \in \mathbb{R}^3$:

$$\begin{pmatrix} \frac{\partial f_i(x,y,z)}{\partial x} & \frac{\partial f_i(x,y,z)}{\partial y} & \frac{\partial f_i(x,y,z)}{\partial z} \\ \frac{\partial f_2(x,y,z)}{\partial x} & \frac{\partial f_2(x,y,z)}{\partial y} & \frac{\partial f_2(x,y,z)}{\partial z} \end{pmatrix} = \begin{pmatrix} 2xy & x^2 + z & y \\ -\frac{y-z}{x^2} & \frac{1}{x} & -\frac{1}{x} \end{pmatrix}.$$

Evaluate the derivative at $(1, 1, 1)$:

$$Df(1, 1, 1) = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

- The CES utility function

$$u\left(x_1, x_2, x_3\right) = \left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}}.$$

Compute the partial derivatives at $x = (x_1, x_2, x_3)$.

$$\frac{\partial u}{\partial x_i}\left(x_1, x_2, x_3\right) = \frac{1}{\rho}\left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}-1} \rho x_i^{\rho-1} = x_i^{\rho-1}\left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}-1}.$$

The gradient at x is then:

$$\nabla u\left(x_1, x_2, x_3\right) = \begin{pmatrix} x_1^{\rho-1}\left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}-1} \\ x_2^{\rho-1}\left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}-1} \\ x_3^{\rho-1}\left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}-1} \end{pmatrix}.$$

Remark 4 (Extra material on utility functions) *If a consumer prefers consumption vector $x \in \mathbb{R}^n$ to $y \in \mathbb{R}^n$, we write $x \succsim y$. A utility function*

$$u : \mathbb{R}^n \to \mathbb{R}$$

is said to represents preferences $\succsim$ if

$$u(x) \geq u(y) \iff x \succsim y.$$

As discussed before, a utility representation for preferences exists if i) for all x, y, either $x \succsim y$ or $y \succsim x$ or both, ii)

$$x \succsim y \text{ and } y \succsim z \Rightarrow x \succsim z,$$

and iii) the preference relation is continuous (more on this later).

The utility representation is not unique. To make this concrete, consider the following argument. A function

$$v : \mathbb{R} \to \mathbb{R}$$

is said to be strictly increasing if

$$x > y \Rightarrow v\left(x\right) > v\left(y\right).$$

If u represents $\succsim$ then also the compound function

$$v\left(u\left(x\right)\right)$$

is a representation of the same preferences.

Exercise: Show that $u(x)$ and $v(u(x))$ represent the same preferences if v is a strictly increasing function. Show also that this is not true if v is not strictly increasing.

In light of this result, the previous utility function

$$u\left(x_1, x_2, x_3\right) = \left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}}$$

represents the same preferences as

$$u\left(x_1, x_2, x_3\right) = x_1^\rho + x_2^\rho + x_3^\rho$$

if $\rho > 0$.

From the economics point of view, the preferences are the primitive notion. A utility function is used only because functions are much easier to manipulate than preferences.

Note that

$$u\left(x_1, x_2, x_3\right) = \left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}}$$

is a linearly homogenous function:

$$\begin{aligned} u\left(\lambda x_1, \lambda x_2, \lambda x_3\right) &= \left(\left(\lambda x_1\right)^\rho + \left(\lambda x_2\right)^\rho + \left(\lambda x_3\right)^\rho\right)^{\frac{1}{\rho}} \\ &= \left(\lambda^\rho \left(x_1^\rho + x_2^\rho + x_3^\rho\right)\right)^{\frac{1}{\rho}} \\ &= \lambda \left(x_1^\rho + x_2^\rho + x_3^\rho\right)^{\frac{1}{\rho}}. \end{aligned}$$

Preferences that have a linearly homogenous representation occupy are called homothetic and they have a particularly prominent place in consumer theory. We will see that for homothetic preferences, the relative consmuption shares are unaffected by the wealth of the consumer.

On the other hand, the non-homothetic form is much easier to differentiate etc. The bottom line is that you can use whichever representation is more convenient since they describe the same consumer preferences.

- Profit maximizing firm The profit of a firm can be computed as:

$$g(k,l) = pf(k,l) - rk - wl,$$

where p is the price of the output f is the production function of the firm, k is the amount of capital used l the labor used r is the rental cost of capital and w is the wage cost of labor.

Compute the gradient:

$$\nabla g(k,l) = \begin{pmatrix} p\frac{\partial f(k,l)}{\partial k} - r \\ p\frac{\partial f(k,l)}{\partial l} - w \end{pmatrix}.$$

The partial derivative of the production function with respect to k:n is called the marginal product of capital $MP_k(k,l)$ and the partial derivative w.r.t. l is called marginal product of labor $MP_l(k,l)$.

The gradient can therefore be written as:

$$\nabla g(k,l) = \begin{pmatrix} pMP_k - r \\ pMP_l - w \end{pmatrix}.$$

In the previous section, we saw that for unconstrained optimization,

$$\nabla g\left(\hat{k},\hat{l}\right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In other words,

$$MP_k(\hat{k},\hat{l}) = \frac{r}{p}, MP_l(\hat{k},\hat{l}) = \frac{w}{p}.$$

At optimum, the marginal product of each input is equals its cost normalized by the output price.

Specify further:

$$f(k,l) = k^{\alpha}l^{\beta}.$$

Then

$$MP_k(k,l) = \alpha k^{\alpha-1}l^{\beta}, MP_l(k,l) = \beta k^{\alpha}l^{\beta-1}.$$

Now at optimum $(\hat{k}, \hat{l})$,

$$\frac{MP_k(\hat{k}, \hat{l})}{MP_l(\hat{k}, \hat{l})} = \frac{\alpha \hat{l}}{\beta \hat{k}} = \frac{r}{w},$$

so that

$$\hat{k} = \frac{\alpha w}{\beta r} \hat{l}.$$

Plug into

$$MP_l = \beta k^{\alpha} l^{\beta - 1} = \frac{w}{p}$$

to get

$$\beta \left(\frac{\alpha w}{\beta r} \right)^{\alpha} \hat{l}^{\alpha + \beta - 1} = \frac{w}{p}.$$

This gives:

$$\hat{l} = \left(\beta \left(\frac{\alpha w}{\beta r} \right)^{\alpha} \frac{p}{w} \right)^{\frac{1}{1-\alpha-\beta}} = \left(\alpha^{\alpha} \beta^{1-\alpha} w^{\alpha - 1} r^{-\alpha} p \right)^{\frac{1}{1-\alpha-\beta}}.$$

Similarly for capital:

$$\begin{aligned} \hat{k} &= \frac{\alpha w}{\beta r} l = \frac{\alpha w}{\beta r} \left(\alpha^{\alpha} \beta^{1-\alpha} w^{\alpha - 1} r^{-\alpha} p \right)^{\frac{1}{1-\alpha-\beta}} \\ &= \left(\alpha^{1-\beta} \beta^{\beta} w^{-\beta} r^{\beta - 1} p \right)^{\frac{1}{1-\alpha-\beta}}. \end{aligned}$$

Is this the solution to the problem? If the problem has a solution, then we have found it. In Section 5, we'll see if the problem has a solution.

4 Comparative statics and the implicit function theorem

In the notes on basic linear algebra, we saw that linear economic models as described by systems of equations have solutions whenever the rank of the matrix representing the system is equal to the number of variables in the model. Let's consider a system of equations with$n$ endogenous variables

$(y_1, ..., y_n)$ and m exogenous variables $(x_1, ..., x_m)$ The simplest case is of course when $n = m = 1$.

$$f(y; x) = ay + bx = 0.$$

How does a change in x affect the solution of the model i.e. how does y behave as a function of x?

Let's solve:

$$y = -\frac{b}{a}x,$$

whenever $a \neq 0$. This tells us that a change of Δx in x causes a change of $\Delta y = -\frac{b}{a}\Delta x$ in y.

$$\frac{\Delta y}{\Delta x} = -\frac{b}{a}.$$

For Δx on small, denote it by dx.

We get:

$$\frac{dy}{dx} = -\frac{b}{a} = -\frac{\partial f(y; x)}{\partial y}^{-1} \frac{\partial f(y; x)}{\partial x}.$$

We want to find an analogous formula for linear systems of equations and then non-linear systems of equations. In the linear case under full rank, we can solve for an explicit solution vector y for each value of x. With non-linear equations, there is no hope for this in general.

4.1 Implicit function theorem for $n = m = 1$

We start this section with an example of a univariate function.

Example 1

$$f(y, x) = xy + \ln(xy + x) = 0. \tag{3}$$

Note that $(\widehat{y}, \widehat{x}) = (0, 1)$ satisfies equation 3. What is the impact of a small change dx in $\hat{x}x$ on the value of y satisfying the equation. We are interested in all points (y, x) near $(0, 1)$ satisfying equation 3. Let's assume that such a $y(x)$ exists for all x near $\hat{x}$. Assume also that $y(x)$ has a derivative at $\hat{x}$. We can then write:

$$g(x) = f(y(x), x) = xy(x) + \ln(xy(x) + x) = 0$$

for all x near $\hat{x} = 1$.

We see that the original equation has been reduced to an equation in a single variable x. Since the composite function is constant in x (=0), the composite function g must have a zero derivative in x near $\hat{x} = 1$.

By the chain rule:

$$g'(x) = \frac{\partial f(y;x)}{\partial y} y'(x) + \frac{\partial f(y;x)}{\partial x}$$

$$= \left(x + \frac{x}{xy + x} \right) y'(x) + y + \frac{y+1}{xy+x}.$$

By requiring $g'(1) = 0$, we get:

$$y'(1) = -\frac{\frac{\partial f(0,1)}{\partial x}}{\frac{\partial f(0,1)}{\partial y}} = -\frac{1}{2}.$$

Notice that this is a valid computation only if $\frac{\partial f(0,1)}{\partial y} \neq 0$.

The theorem below generalizes the message of this example. We say that a function is continuously differentiable at point $(\hat{y}, \hat{x})$ if its partial derivatives w.r.t. x and y exists and are continuous at $(\hat{y}, \hat{x})$. An ε -neighborhood $B^{\varepsilon}(\hat{x})$ of $\hat{x} \in \mathbb{R}^k$ is the open set of all points at distance less than ε from $\hat{x}$:

$$B^{\varepsilon}(\hat{x}) := \{x \in \mathbb{R}^k : \|x - \hat{x}\| < \varepsilon\}.$$

Theorem 1 *Let $f(y, x)$ be a continuously differentiable at $(\widehat{y}, \widehat{x})$ in an ε -neighborhood $B^{\varepsilon}(\widehat{y}, \widehat{x})$, for some $\varepsilon > 0$ and also that*

$$f(\widehat{y}, \widehat{x}) = 0.$$

If $\frac{\partial f(\widehat{y},\widehat{x})}{\partial y} \neq 0$, then there exists a $\delta > 0$ and a continuously differentiable function $y(x)$ in some δ -neighborhood of $\widehat{x}$ $B^{\delta}(\widehat{x})$, such that:

1. *$f(y(x), x) = 0$ for all $x \in B^{\delta}(\widehat{x})$,*
2. *$y(\widehat{x}) = \widehat{y}$,*
3. *The derivative of y at $\hat{x}$ satisfies:*

$$y'(\widehat{x}) = -\frac{\frac{\partial f(\widehat{y},\widehat{x})}{\partial x}}{\frac{\partial f(\widehat{y},\widehat{x})}{\partial y}}$$

I have given the theorem in its full mathematical generality. The important point to remember from all this is that the key to the theorem is that i) f is continuously differentiable at the solution $(\hat{y}, \hat{x})$, ii) $\frac{\partial f(\hat{y},\hat{x})}{\partial y} \neq 0$.

Let's write this a bit differently:

$$Df\left(0,1\right) = \left(\frac{\partial f\left(0,1\right)}{\partial y}, \frac{\partial f\left(0,1\right)}{\partial x}\right).$$

For small (dy, dx), we have by the definition of the derivative:

$$f\left(0+dy, 1+dx\right) - f\left(0,1\right) = \left(\frac{\partial f\left(0,1\right)}{\partial y}, \frac{\partial f\left(0,1\right)}{\partial x}\right)\left(\begin{array}{c} dy \\ dx \end{array}\right).$$

For the equation to remain valid at $(0+dy, 1+dx)$,the change in f has to be zero:

$$\frac{\partial f\left(0,1\right)}{\partial y}dy + \frac{\partial f\left(0,1\right)}{\partial x}dx = 0.$$

By solving for dy as a function of dx, we get:

$$dy = -\frac{\frac{\partial f(0,1)}{\partial x}}{\frac{\partial f(0,1)}{\partial y}}dx.$$

This approach is the easiest to generalize to get the full implicit function theorem. Notice how nicely this plays on the linearity of the derivative for small changes.

4.2 Linear implicit function theorem for $n > 1$

Consider the system of equations:

$$\begin{aligned} a_{11}y_1 + ... + a_{1n}y_n + b_{11}x_1 + ... + b_{1m}x_m &= 0, \\ &\vdots \\ a_{n1}y_1 + ... + a_{nn}y_n + b_{n1}x_1 + ... + b_{nm}x_m &= 0. \end{aligned}$$

In matrix form:

$$Ay + Bx = 0,$$

where A on $n \times n$ matrix and B on $n \times m$ matrix, $y = \left(y_1, ..., y_n\right)$, $x = \left(x_1, ..., x_m\right)$.

Write this as:

$$f(y;x)=0.$$

Assume that

$$f(\widehat{y};\widehat{x})=0 \text{ or } A\widehat{y}+B\widehat{x}=0,$$

and consider the effect of a small change $(dy;dx)=(dy_1,...,dy_n;dx_1,...,dx_m)$ on the value of f :

$$f(\widehat{y}+dy,\widehat{x}+dx)-f(\widehat{y},\widehat{x})=Ady+Bdx$$

$$=D_yf(\widehat{y};\widehat{x})\,dy+D_xf(\widehat{y};\widehat{x})\,dx,$$

where $D_yf(\hat{y},\hat{x})$ consists of the partial derivatives of f w.r.t. the endogenous variables y and $D_xf(\hat{y},\hat{x})$ w.r.t. the exogenous variables x.

For

$$f(y;x)=0.$$

to hold at $(y,x)=(\widehat{y}+dy,\widehat{x}+dx)$, the change must be zero:

$$D_yf(\widehat{y};\widehat{x})dy+D_xf(\widehat{y};\widehat{x})dx=0.$$

In other words

$$dy=-D_yf(\widehat{y};\widehat{x})^{-1}D_xf(\widehat{y};\widehat{x})dx=A^{-1}Bdx.$$

If a single exogenous variable changes, then Bdx is a row vector and dy can be solved using Cramer's rule. This equation has a solution for all dx only if A^{-1} exists, i.e. if $A=D_yf(\widehat{y};\widehat{x})$ has full rank. This result can be generalized for the non-linear case in a neighborhood of $(\widehat{y};\widehat{x})$ and it is the implicit function theorem.

Example

$$\begin{aligned}2y_1+y_2+3x &= 0,\\ y_1-y_2-x &= 0.\end{aligned}$$

In matrix form:

$$\begin{pmatrix}2&1\\1&-1\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}+\begin{pmatrix}3\\-1\end{pmatrix}x=\begin{pmatrix}0\\0\end{pmatrix},$$

or

$$\begin{pmatrix}2&1\\1&-1\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}-3\\1\end{pmatrix}x.$$

By Cramer's rule:

$$y_1 = \frac{\det \begin{pmatrix} -3 & 1 \\ 1 & -1 \end{pmatrix} x}{\det \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}} = \frac{-2}{3}x, y_2 = \frac{\det \begin{pmatrix} 2 & -3 \\ 1 & 1 \end{pmatrix} x}{\det \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}} = \frac{-5}{3}x.$$

In other words, if dx is the change in the exogenous variable, then

$$dy_1 = \frac{-2}{3}dx, \quad dy_2 = \frac{-5}{3}dx.$$

4.3 Implicit function theorem for $n > 1$

Consider now a continuously differentiable non-linear function

$$f : \mathbb{R}^{n+m} \to \mathbb{R}^n$$

in a neighborhood of the point $(\hat{y}, \hat{x}) \in \mathbb{R}^n + m$, where

$$f\left(\widehat{y}, \widehat{x}\right) = 0.$$

Use the derivative of $Df(\hat{x}, \hat{x})$ to approximate f at $(\hat{y} + dy, \hat{x} + dx)$:

$$f\left(\widehat{y} + dy, \widehat{x} + dx\right) - f\left(\widehat{y}, \widehat{x}\right) = Df\left(\widehat{y}, \widehat{x}\right) = D_y f\left(\widehat{y}; \widehat{x}\right) dy + D_x f\left(\widehat{y}, \widehat{x}\right) dx,$$

and require that:

$$D_y f\left(\widehat{y}, \widehat{x}\right) dy + D_x f\left(\widehat{y}, \widehat{x}\right) dx = 0.$$

Since $D_y f\left(\widehat{y}; \widehat{x}\right)$ ja $D_x f\left(\widehat{y}; \widehat{x}\right)$ are matrices, we continue here exactly as in the linear case. With differential calculus, we have reduced the really complicated non-linear problem to the much simpler linear case locally, i.e. in a neighborhood of the solution point $(\hat{y}, \hat{x})$.

Before stating the full implicit function theorem, we consider an example.

Example 2

$$f\left(y; x\right) = \begin{pmatrix} f_1\left(y_1, y_2; x_1, x_2\right) \\ f_2\left(y_1, y_2; x_1, x_2\right) \end{pmatrix}.$$

$$\begin{aligned} f_1(y_1, y_2; x_1, x_2) &= y_1 y_2^2 - x_1 x_2 + x_2 + 2 = 0, \\ f_2(y_1, y_2; x_1, x_2) &= y_1 - \frac{x_1}{y_2} + x_2 - 1 = 0. \end{aligned}$$

Consider the system of equations in a neighborhood of the point

$$(\widehat{y}_1, \widehat{y}_2; \widehat{x}_1, \widehat{x}_2) = (1, 1, 2, 2).$$

Check first that the equation is satisfied at $(1, 1, 2, 2)$ *and form the appropriate matrices of partial derivatives:*

$$D_y f(\widehat{y}; \widehat{x}) = \begin{pmatrix} \frac{\partial f_1((\widehat{y};\widehat{x}))}{\partial y_1} & \frac{\partial f_1((\widehat{y};\widehat{x}))}{\partial y_2} \\ \frac{\partial f_2((\widehat{y};\widehat{x}))}{\partial y_1} & \frac{\partial f_2((\widehat{y};\widehat{x}))}{\partial y_2} \end{pmatrix} = \begin{pmatrix} \widehat{y}_2^2 & 2\widehat{y}_1\widehat{y}_2 \\ 1 & \frac{-\widehat{x}_1}{\widehat{y}_2^2} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix},$$

$$D_x f(\widehat{y}; \widehat{x}) = \begin{pmatrix} \frac{\partial f_1((\widehat{y};\widehat{x}))}{\partial x_1} & \frac{\partial f_1((\widehat{y};\widehat{x}))}{\partial x_2} \\ \frac{\partial f_2((\widehat{y};\widehat{x}))}{\partial x_1} & \frac{\partial f_2((\widehat{y};\widehat{x}))}{\partial x_2} \end{pmatrix} = \begin{pmatrix} -\widehat{x}_2 & 1 - \widehat{x}_2 \\ \frac{-1}{\widehat{y}_2} & 1 \end{pmatrix} = \begin{pmatrix} -2 & -1 \\ -1 & 1 \end{pmatrix}.$$

We see that $\det(D_y f(\widehat{y}; \widehat{x})) \neq 0$*, and therefore the matrix* $D_y f(\hat{y}, \hat{x})$ *has full rank and an inverse matrix* $[D_y f(\hat{y}, \hat{x})]^{-1}$

Exercise: Show that

$$[D_y f(\hat{y}, \hat{x})]^{-1} = \frac{-1}{4} \begin{pmatrix} -2 & -2 \\ -1 & 1 \end{pmatrix},$$

and therefore:

$$dy = \frac{-1}{4} \begin{pmatrix} -2 & -2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -2 & -1 \\ -1 & 1 \end{pmatrix} dx.$$

We could single out e.g. the effect of a change in x_1 *on the endogenous variables near* $(\hat{y}_1, \hat{y}_2, \hat{x}_1\hat{x}_2) = (1, 1, 2, 2)$*:*

$$\begin{pmatrix} \frac{\partial f_1((\widehat{y},\widehat{x}))}{\partial y_1} & \frac{\partial f_1((\widehat{y},\widehat{x}))}{\partial y_2} \\ \frac{\partial f_2((\widehat{y},\widehat{x}))}{\partial y_1} & \frac{\partial f_2((\widehat{y},\widehat{x}))}{\partial y_2} \end{pmatrix} \begin{pmatrix} dy_1 \\ dy_2 \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1((\widehat{y},\widehat{x}))}{\partial x_1} \\ \frac{\partial f_2((\widehat{y},\widehat{x}))}{\partial x_1} \end{pmatrix} dx_1 = 0$$

Plugging in $(1, 1, 2, 2)$*, we get:*

$$\begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} dy_1 \\ dy_2 \end{pmatrix} + \begin{pmatrix} -2 \\ -1 \end{pmatrix} dx_1 = 0$$

Solving by Cramer's rule gives:

$$dy_1 = \frac{\det \begin{pmatrix} 2 & 2 \\ 1 & -2 \end{pmatrix} dx_1}{\det \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}} = \frac{1}{2} dx_1, dy_2 = \frac{\det \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}}{\det \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}} = \frac{1}{4} dx_1.$$

We are now ready for the main theorem in this section.

Theorem 2 *Let $f : \mathbb{R}^{n+m} \to \mathbb{R}^n$ be continuously differentiable in a neighborhood $B^{\varepsilon}(\widehat{y}, \widehat{x})$, of $(\hat{y}, \hat{x})$ for some $\varepsilon > 0$ and*

$$f(\widehat{y}, \widehat{x}) = 0.$$

If $\det(D_y f(\widehat{y}; \widehat{x})) \neq 0$, then there exists a $\delta > 0$ and a continuously differentiable function $y(x)$ in a neighborhood $B^{\delta}(\widehat{x})$ of $\hat{x}$ such that :

1. *$f(y(x), x) = 0$ for all $x \in B^{\delta}(\widehat{x})$,*
2. *$y(\widehat{x}) = \widehat{y}$,*
3. *The derivative of the function y satisfies:*

$$Dy(\widehat{x}) = -(D_y f(\widehat{y}; \widehat{x}))^{-1} D_x f(\widehat{y}; \widehat{x})$$

Proving this theorem is beyond the scope of this course. Let me just make some comments. Assuming points 1. and 2. above, point 3. is an application of the chain rule in the vector-valued multivariate case. It is nothing more than a local version of the linear implicit function theorem. Parts 1. and 2. require some more sophisticated mathematics. Proving the existence of the implicit function $y(x)$ near $\hat{x}$ requires the use of a fixed point theorem (similar to the case of showing the existence of local solutions to differential equations). This is beyond the scope of this course.

4.4 Applications

4.4.1 Drawing indifference curves

Consider the consumer's problem with two consumption goods x_1, x_2:

$$\max_{x_1, x_2 \geq 0} u(x_1, x_2)$$

such that

$$p_1 x_1 + p_2 x_2 \leq w.$$

We shall return to this problem a number of times during this course. For now, we are not interested in its solution but just understanding some parts of the underlying structure.

An indifference curve through point $(\hat{x}_1, \hat{x}_2)$ traces in the positive part of the plane $x_1, x_2 \geq 0$ (also called the positive orthant) all the points that are considered equally good as $(\hat{x}_1, \hat{x}_2)$. Using the utility representation, this just means all the points (x_1, x_2) such that $u(x_1, x_2) = u(\hat{x}_1, \hat{x}_2)$. Denote the value of the utility function on this indifference curve by $\hat{u} = u(\hat{x}_1, \hat{x}_2)$.

How should x_2 change if x_1 is changed to $\hat{x}_1 + dx_1$ (recall that whenever we write dx, we mean small changes)? The implicit function theorem tells us when we can answer this question definitively. The only condition we need is that the partial derivative $\frac{\partial u}{\partial x_2}(\hat{x}_1, \hat{x}_2)$ be non-zero.

Write the linear approximation:

$$u(\hat{x}_1 + dx_1, \hat{x}_2 + dx_2) - u(\hat{x}_1, \hat{x}_2) = \frac{\partial u}{\partial x_1}(\hat{x}_1, \hat{x}_2)dx_1 + \frac{\partial u}{\partial x_2}(\hat{x}_1, \hat{x}_2)dx_2.$$

In order to have

$$u(\hat{x}_1 + dx_1, \hat{x}_2 + dx_2) = \hat{u},$$

we get:

$$\frac{\partial u}{\partial x_1}(\hat{x}_1, \hat{x}_2)dx_1 + \frac{\partial u}{\partial x_2}(\hat{x}_1, \hat{x}_2)dx_2 = 0,$$

and therefore:

$$\frac{dx_2}{dx_1} = -\frac{\frac{\partial u}{\partial x_1}(\hat{x}_1, \hat{x}_2)}{\frac{\partial u}{\partial x_2}(\hat{x}_1, \hat{x}_2)}$$

whenever $\frac{\partial u}{\partial x_2}(\hat{x}_1, \hat{x}_2) \neq 0$.

In almost all consumer models, we assume that consumers have monotonic preferences i.e. higher consumption in any of the goods (keeping the others fixed) results in an increase in the utility function. This is guaranteed by the assumption that the marginal utility from good $i \in \{1, 2\}$ is strictly positive, i.e. $\frac{\partial u}{\partial x_i}(x_1, x_2) > 0$ for $i \in \{1, 2\}$ and for all (x_1, x_2). The indifference curve is therefore downwards sloping and the absolute value of its slope, i.e. the marginal rate of substitution MRS_{x_1,x_2} is given by the ratio of the marginal utilities.

If we have k goods, the analysis is almost identical. Consider $\hat{x} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_k)$ and $\hat{u} = u(\hat{x})$. We can find the indifference hypersurface by looking for $\{x | u(x) = \hat{u}\}$. Using the linear approximation around $\hat{x}$, we are looking for $dx = (dx_1, ..., dx_k)$ such that:

$$\sum_{i=1}^{k} \frac{\partial u}{\partial x_i}(\hat{x})dx_i = 0.$$

Fixing $\hat{x}_i$ for $i \neq k, l$, we can define the marginal rate of substitution between x_k and x_l as before:

$$MRS_{x_k,x_l}(\hat{x}) = \frac{\frac{\partial u}{\partial x_k}(\hat{x})}{\frac{\partial u}{\partial x_k}(\hat{x})}.$$

Exercise: Show that for $u(x_1, x_2) = (a_1 x_1^\rho + a_2 x_2^\rho)^{\frac{1}{\rho}}$,

$$MRS_{x_1,x_2} = \frac{a_1 x_1^{\rho-1}}{a_2 x_2^{\rho-1}}.$$

Show also that this result generalizes for k goods.

4.4.2 Drawing an isoquant

In the theory of the firm, an analogous problem involves inputs (k, l) giving rise to the same level of output. We call such combinations the isoquants of the firm. Using the production function $f(k, l)$, we can look for (k, l) such that for a fixed level of output $f(\hat{k}, \hat{l} = \hat{y}$:

$$f(k, l) = \hat{y}.$$

Again, implicit function theorem gives us the right conditions for finding say the level of capital needed to adjust for changes in labor along an isoquant. The only requirement is that the marginal product of capital be non-zero at $(\hat{k}, \hat{l})$, i.e.

$$\frac{\partial f}{\partial k}(\hat{k}, \hat{l}) \neq 0.$$

But it is safe to assume that at any reasonable level of operation, additional capital gives additional output so that this partial derivative is strictly positive.

By repeating the steps from the consumer's problem, we see that in a neighborhood of $(\hat{k}, \hat{l})$,

$$\frac{dk}{dl} = -\frac{\frac{\partial f}{\partial l}(\hat{k}, \hat{l})}{\frac{\partial f}{\partial k}(\hat{k}, \hat{l})}.$$

This means that isoquants are downward sloping and the marginal rate of technical substitution (the absolute value of the slope of the isoquant) is

given by:

$$MRTS_{l,k} = \frac{\frac{\partial f}{\partial l}}{\frac{\partial f}{\partial k}} = \frac{MP_l}{MP_k}.$$

4.4.3 Comparative statics of consumer choice

This subsection anticipates some of the results from the section on constrained optimization and we will return to these issues. As argued in Principles of Economics I, a key insight in constrained optimization is that at optimum, the marginal rate of substitution equals the marginal rate of transformation. For consumer choice, the marginal rate of transformation between two goods 1 and 2 is given by their price ratio. If I give up one unit of good 1, I can buy $\frac{p_1}{p_2}$ more units of good 2.

Let $w > 0$ denote the wealth available to the consumer so that her budget constraint is

$$\sum_{i=1}^{k} p_i x_i = w.$$

Specialize now to the case of two goods. The solution (x_1, x_2) to the consumer's optimization problem is then given by the solution to the two equations:

$$\begin{array}{rcl} \frac{\frac{\partial u}{\partial x_1}(x_1,x_2)}{\frac{\partial u}{\partial x_2}(x_1,x_2)} & = & \frac{p_1}{p_2}, \\ p_1 x_1 + p_2 x_2 & = & w. \end{array} \tag{4}$$

Cross multiplying the first row, we get:

$$\begin{array}{rcl} p_2 \frac{\partial u}{\partial x_1}(x_1, x_2) - p_1 \frac{\partial u}{\partial x_2}(x_1, x_2) & = & 0, \\ p_1 x_1 + p_2 x_2 & = & w. \end{array} \tag{5}$$

Suppose now that we have a solution to the system at $(\hat{x}_1, \hat{x}_2, \hat{p}_1, \hat{p}_2, \hat{w})$. We are now treating (p_1, p_2, w) as exogenous variables in the problem and (x_1, x_2) as endogenous. When can we solve for the changes in optimal consumptions resulting from small changes in the prices and wealth around $(\hat{p}_1, \hat{p}_2, \hat{w})$?

If we assume that the partial derivatives of u are themselves continuously differentiable, then the function h defined as:

$$h(x_1, x_2) := p_2 \frac{\partial u}{\partial x_1}(x_1, x_2) - p_1 \frac{\partial u}{\partial x_2}(x_1, x_2)$$

is continuously differentiable.

Write the equation system (4) as:

$$\begin{array}{rcl} h(x_1, x_2) & = & 0, \\ p_1 x_1 + p_2 x_2 - w & = & 0. \end{array} \tag{6}$$

We can now write the matrices of the endogenous and exogenous variables as follows:

$$\begin{pmatrix} \frac{\partial h}{\partial x_1}(x_1, x_2) & \frac{\partial h}{\partial x_2}(x_1, x_2) \\ p_1 & p_2 \end{pmatrix}, \begin{pmatrix} -\frac{\partial u}{\partial x_2}(x_1, x_2) & \frac{\partial u}{\partial x_1}(x_1, x_2) & 0 \\ x_1 & x_2 & 1 \end{pmatrix}.$$

The implicit function theorem tells us that near any solution

$$(\hat{x}_1, \hat{x}_2, \hat{p}_1, \hat{p}_2, \hat{w}),$$

we can find a neighborhood of $\hat{p}_1, \hat{p}_2, \hat{w})$ and $x_1(p_1, p_2, w)$ and $x_2(p_1, p_2, w)$ such that (6) is satisfied if the determinant of

$$\begin{pmatrix} \frac{\partial h}{\partial x_1}(x_1, x_2) & \frac{\partial h}{\partial x_2}(x_1, x_2) \\ p_1 & p_2 \end{pmatrix}$$

is non-zero.

At this point, it is probably a good idea to specialize a little bit to get more explicit solutions. Consider the case where

$$u(x_1, x_2) = a_1 u(x_1) + a_2 u(x_2).$$

In this case,

$$MRS_{x_1, x_2} = \frac{a_1 u'(x_1)}{a_2 u'(x_2)}.$$

With this utility function, we can now write our system for the optimum by cross multiplying:

$$\begin{array}{rcl} p_2 a_1 u'(x_1) - p_1 a_2 u'(x_2) & = & 0, \\ p_1 x_1 + p_2 x_2 - w & = & 0. \end{array} \tag{7}$$

The matrix of partial derivatives with respect to endogenous variables is now:

$$\begin{pmatrix} p_2 u''(x_1) & -p_1 u''(x_2) \\ p_1 & p_2 \end{pmatrix},$$

where we have written $u''(x) = Du'(x)$. This is called the second derivative of u and we will have a lot more to say about such higher order derivatives very soon.

Since the prices are positive and the two elements on the first row of the matrix have opposite signs, we see immediately that the determinant is non-zero if the second derivative u'' has constant sign. We will return to the meaning of the sign of second derivatives soon.

The matrix of the partial derivatives with respect to the exogenous variables is unchanged from before. Hence we can apply the implicit function theorem to find the changes in the endogenous variables resulting from small changes in the prices and wealth.

5 Unconstrained optimization

Consider first the case of a single variable:

$$f : \mathbb{R} \to \mathbb{R}.$$

Function f has a maximum at point x_0 if fo all $y \in \mathbb{R}$,

$$f(y) \leq f(x_0).$$

Function f has a minumum at x_0 if for all $y \in \mathbb{R}$,

$$f(y) \geq f(x_0)$$

5.1 Local optima for $n = 1$

The function $f$has a local maximum at x_0 if there exists an $\varepsilon > 0$ such that for all $y \in B^{\varepsilon}(x_0)$, we have:

$$f(y) \leq f(x_0).$$

A local minimum is defined analogously.

- How do we know whether $f$has a maximum or a minimum at x_0?
- How to find minima and maxima?
- Local or global minima and maxima?

5.1.1 First-order condition

Assume that f is differentiable. We know from the definition of the derivative that if f has a minimum or maximum at x_0, then

$$Df(x_0) = f'(x_0) = 0.$$

We call this the *first-order necessary condition*. It must be satisfied at any optimum, but the fact that $Df(x_0) = 0$ does not tell us that we have a minimum or a maximum at x_0 Just think about $f(x) = x^3$ at $x = 0$.

5.1.2 Second-order conditions

Assume conversely that $f'(x_0) = 0$. If f has a maximum at x_0, then f is increasing for $x < x_0$ and decreasing for $x > x_0$. If an increasing (decreasing) function has a derivative, it is positive (negative).

In other words:

$$f'(x_0)$$

when viewed as a function of x_0. Therefore if $f'(x_0)$has a derivative, we know that at a maximum,

$$Df'(x_0) \leq 0$$

This means that we need to consider the derivative of the derivative. We denote:

$$f''(x_0) = \lim_{h \to 0} \frac{f'(x_0 + h) - f(x_0)}{h}$$

and we call $f''(x_0)$ the second derivative of f at x_0.

The third derivative is the derivative of the second derivative etc. The definitions are as before. If a function f has a k^{th} derivative at point x_0, we say that f is k times differentiable at x_0. We denote the k^{th} derivative at x_0 by $f^{[k]}(x_0)$. We say that it is k times continuously differentiable if the k^{th} derivative is continuous in x_0.

5.2 Taylor's theorem

Consider a function $f : \mathbb{R} \to \mathbb{R}$, and assume that it is $k+1$ times continuously differentiable at x_0. Taylor's theorem asserts that then

$$f(x_0+h) = f(x_0) + f'(x_0)\,h + \frac{1}{2}f''(x_0)\,h^2 + \ldots + \frac{1}{k!}f^{[k]}(x_0) + \frac{1}{(k+1)!}f^{[k+1]}(x)\,h^{k+1},$$

for some x with $x_0 < x < x_0 + h$.

Remark 5 (Idea of proof) *To get the idea behind this result, consider the linear approximation:*

$$f(x_0+h) = f(x_0) + f'(x_0)h + \textit{ higher order terms } r(h).$$

We need to show that for some $x_0 < x < x_0 + h$, we have

$$r(h) = \frac{1}{2}f''(x)h^2.$$

Let

$$g(x) = f(x) - f(x_0) - f'(x_0)h - \frac{1}{h^2}[f(x_0+h) - f(x_0) - f'(x_0)h](x-x_0)^2.$$

We see that $g(x_0) = g(x_0+h) = 0$. Since g is continuously differentiable, Rolle's theorem guarantees that $g'(x') = 0$ for some $x_0 < x' < x_0 + h$. Since g is twice continuously differentiable, a second application of Rolle's theorem guarantees that there is an $x_0 < x < x' < x_0 + h$ such that $g''(x) = 0$. Since $g''(x) = f''(x) - 2\frac{1}{h^2}[f(x_0+h) - f(x_0) - f'(x_0)h]$, the claim follows. The idea for the proof of the general case is based on the same idea, but involves more terms and $k+1$ uses of Rolle's theorem.

For local analysis, i.e. for h arbitrarily small, we need to look for the first term with a non-zero coefficient in the Taylor approximation. The other terms vanish much more quickly when $h \to 0$ (since they involve the multiplier h^k for $k > 1$. For twice (or more times) continuously differentiable functions, Taylor's theorem gives a precise reason why we called the remainder term as higher-order terms in the first-order approximation by derivatives.

With the help of Taylor's theorem, we can classify all points with $f'(x_0) = 0$:

1. If the first l for which $f^{[l]}(x_0) \neq 0$, is odd, then f does not have an extremum (i.e. minimum or maximum) at x_0.
2. If the first l for which $f^{[l]}(x_0) \neq 0$, is even and $f^{[l]}(x_0) < 0$, then f has a local maximum at x_0.
3. If the first l for which $f^{[l]}(x_0) \neq 0$, is even and $f^{[l]}(x_0) > 0$, then f has a local minimum at x_0.

Make sure that you understand why this classification holds. For l defined as above, divide the Taylor approximation by $h^{l-1}$and let $h \to 0$.

The case $f'(x_0) = 0 and f''(x_0) < 0$ is called the *second-order sufficient condition* for maximum at x_0.

One more point should be kept in mind. The function f may have several local maxima and not all of them are maxima. We will have more to say about global extrema when we discuss convex and concave functions.

5.3 Quadratic functions

Quadratic functions of a real variable take the form:

$$f(x) = ax^2 + bx + c.$$

Taylor's series around x_0 gives:

$$f(x_0 + h) = f(x_0) + (2ax_0 + b)h + \frac{1}{2}2ah^2.$$

At $x_0 = -\frac{b}{2a}$, $f'(x_0) = 0$ and f has a minumum at x_0, if $a > 0$ and a maximum if $a < 0$.

Consider next multivariate polynomial functions of second degree. These consist of a constant term, $c \in \mathbb{R}$, a first-order linear term $b \cdot x$, where

$$b = (b_1, b_2, ..., b_n), x = (x_1, x_2, ..., x_n)$$

$$b \cdot x = b^\top x = \Sigma_{i=1}^n b_i x_i$$

and a second-order term

$$\begin{aligned} &a_{11}x_1^2 + a_{12}x_1x_2 + ... + a_{1n}x_1x_n \\ &+a_{21}x_2x_1 + ... + a_{2n}x_2x_n \\ &+a_{n1}x_nx_1 + ... + a_{nn}x_n^2. \end{aligned}$$

We can express the second order term via an $n \times n$ matrix A:

$$x \cdot Ax = x^\top Ax =$$
$$(x_1, x_2, ..., x_n) A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Since $x_i x_j = x_j x_i$ we can write A as a symmetric matrix A' by taking

$$A' = \frac{1}{2} A + A^\top.$$

What is the long form of:

$$(x_1, x_2, x_3) \begin{pmatrix} 1 & 3 & 2 \\ 3 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}?$$

By multiplying the second product, we get:

$$(x_1, x_2, x_3) \begin{pmatrix} x_1 + 3x_2 + 2x_3 \\ 3x_1 + 2x_2 + x_3 \\ 2x_1 + x_2 + 2x_3 \end{pmatrix}$$

$$\begin{aligned} = \; & x_1^2 + 3x_1x_2 + 2x_1x_3 + 3x_1x_2 \\ & +2x_2^2 + x_2x_3 + 2x_1x_3 + x_2x_3 + 2x_3^2 \\ = \; & x_1^2 + 2x_2^2 + 2x_3^2 + 6x_1x_2 + 4x_1x_3 + 2x_2x_3. \end{aligned}$$

Here is an example of a general quadratic function of two variables $x = (x_1, x_2)$:

$$\begin{aligned} f(x) &= 6 + 7x_1 + 3x_2 + 2x_1^2 + 5x_1x_2 + 4x_2^2 = \\ c &= 6, b = (7, 3), A = \begin{pmatrix} 2 & \frac{5}{2} \\ \frac{5}{2} & 4 \end{pmatrix}. \end{aligned}$$

5.3.1 Extrema of quadratic functions

To find the local extrema of a quadratic f, compute the gradient:

$$\nabla f(\widehat{x}) = 0.$$

$$\frac{\partial f(\widehat{x})}{\partial x_i} = b_i + 2a_{ii}\widehat{x}_i + \Sigma_{j \neq i}(a_{ij} + a_{ji})\widehat{x}_j = b_i + 2\mathbf{a}_i \cdot \widehat{x},$$

where $\mathbf{a}_i$ is the i^{th} row of matrix A.

Since

$$\nabla f(\widehat{x}) = \begin{pmatrix} \frac{\partial f(\widehat{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\widehat{x})}{\partial x_n} \end{pmatrix},$$

we get

$$\nabla f(\widehat{x}) = b + 2A\widehat{x}.$$

Hence we see that

$$\nabla f(\widehat{x}) = 0 \Leftrightarrow \widehat{x} = -\frac{1}{2}A^{-1}b.$$

For this to make sense, A must have full rank so that A^{-1} exists.

5.3.2 Application: Sum of least squares

Consider a statistical sample consisting of on N pairs of observations

$$(y_1, x_1), ..., (y_N, x_N).$$

Suppose that we want to find a linear relation between x and y. We would like to find a coefficient β that rationalizes the observations as

$$y_i = \beta x_i.$$

If we have many observations, this will not be satisfied in general. To account for errors in the linear relationship, specify the following statistical model

$$y_i = \beta x_i + \varepsilon_i,$$

where ε_i is an identically and independently distributed error term for all i.

Our task is to infer β from the sample. One way of doing this is based on minimizing the sum of squared error terms $(\varepsilon_i)^2$, i.e. to

$$\min_{\beta} f(\beta) = \Sigma_{i=1}^{N} (y_i - \beta x_i)^2 .$$

Compute $f'(\beta)$ and consider $\widehat{\beta}$ such that:

$$f'\left(\widehat{\beta}\right) = 0.$$

By taking the derivative, we get:

$$f'\left(\widehat{\beta}\right) = \sum_{i=1}^{N} -2x_i \left(y_i - \widehat{\beta} x_i\right).$$

As a result, $f'\left(\widehat{\beta}\right) = 0$ if

$$\widehat{\beta} = \frac{\Sigma_{i=1}^{N} x_i y_i}{\Sigma_{i=1}^{N} x_i^2}.$$

Since $f''(\hat{\beta}) = \sum_{i=1}^{N} x_i^2 > 0$, we have found the minimum.

If we want to include a constant term α, we get:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

function of (α, β):

$$f(\alpha, \beta) = \Sigma_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 .$$

To find $\left(\widehat{\alpha}, \widehat{\beta}\right)$ such that

$$\frac{\partial f\left(\widehat{\alpha}, \widehat{\beta}\right)}{\partial \alpha} = \frac{\partial f\left(\widehat{\alpha}, \widehat{\beta}\right)}{\partial \beta} = 0,$$

we get:

$$\begin{aligned} \Sigma_{i=1}^{N} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i\right) &= 0, \\ \Sigma_{i=1}^{N} - 2x_i \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i\right) &= 0. \end{aligned}$$

Solving for α from the first equation gives:

$$\widehat{\alpha} = \frac{\Sigma_{i=1}^{N} y_i - \widehat{\beta}\Sigma_{i=1}^{N} x_i}{N} := \overline{y} - \widehat{\beta}\overline{x}.$$

Using the first equation we also see that:

$$\Sigma_{i=1}^{N} \overline{x} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right) = 0.$$

By substituting into the second, we get:

$$\widehat{\beta} = \frac{\Sigma_{i=1}^{N} \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right)}{\Sigma_{i=1}^{N} \left(x_i - \overline{x} \right)^2} = \frac{Cov\left(y, x \right)}{Var\left(x \right)}.$$

More generally, we can consider samples with more explanatory variables: $\left(y_1, x_{11}, x_{21}, ..., x_{K1} \right), ... \left(y_N, x_{1N}, ..., x_{KN} \right)$ and a linear model

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \begin{pmatrix} = \\ \vdots \\ = \end{pmatrix} \begin{pmatrix} \beta_1 x_{11} + & \cdots & \beta_K x_{K1} \\ \vdots & & \vdots \\ \beta_1 x_{1N} & \cdots & \beta_K x_{KN} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

or in matrix form:

$$y = X\beta + \varepsilon.$$

We can compute the sum of squares now as:

$$\begin{aligned} f\left(\beta \right) &= \varepsilon \cdot \varepsilon = \left(y - X\beta \right)^{\top} \left(y - X\beta \right) \\ &= y \cdot y - \left(X\beta \right)^{\top} y - y^{\top} X\beta + \beta^{\top} X^{\top} X\beta \\ &= y \cdot y - 2y^{\top} X\beta + \beta^{\top} X^{\top} X\beta. \end{aligned}$$

We can now use the general formula found earlier for the quadratic functions:

$$\nabla f\left(\widehat{\beta} \right) = -2X^{\top} y + 2X^{\top} X\widehat{\beta}.$$

Therefore we can find a candidate for the extremum by setting

$$\nabla f\left(\widehat{\beta} \right) = 0.$$

Solving for β, we get:

$$\widehat{\beta} = \left(X^{\top} X \right)^{-1} X^{\top} y.$$

Exercise: Is there a need to add a constant term to get a more general formula?

5.4 Classifying extrema of quadratic functions

How can we determine if a quadratic function has a minimum or a maximum at $\widehat{x}$ such that $f'(\hat{x}) = 0$? Denote by $D^2 f(\widehat{x})$ the second derivative of f at $\widehat{x}$. Assume that Taylor's theorem is true for multivariate functions as well (as it is). Then we would have locally:

$$f(\widehat{x} + h) = f(\widehat{x}) + \nabla f(\widehat{x}) \cdot h + \frac{1}{2} h \cdot D^2 f(\widehat{x})\, h$$

Consider the gradient as a function of $\widehat{x}$.

$$\nabla f(\widehat{x}) : \mathbb{R}^n \to \mathbb{R}^n,$$

Then the derivative of the gradient is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^n$ and we define:

$$D^2 f(\widehat{x}) := D(\nabla f(\widehat{x})) = \begin{pmatrix} \frac{\partial f(\widehat{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial f(\widehat{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\widehat{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial f(\widehat{x})}{\partial x_n \partial x_n} \end{pmatrix}.$$

We call this matrix of second order partial derivatives the Hessian matrix of f. Taylor's theorem tells us that for $\hat{x}$ such that

$$\nabla f(\widehat{x}) = 0,$$

we have:

$$f(\widehat{x} + h) - f(\widehat{x}) = \frac{1}{2} h \cdot D^2 f(\widehat{x})\, h.$$

Whether f has a minimum, a maximum or neither at $\widehat{x}$ depends on whether

$$h \cdot D^2 f(\widehat{x})\, h \gtreqless 0$$

for all h.

We conclude this subsection with the important Young's theorem.

Theorem 3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. Then for all $i, j \in \{1, ..., n\}$ and all x, we have*

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

As a result, the Hessian matrix of f at x is a symmetric matrix for all x.

5.4.1 Quadratic forms and the definiteness of matrices

The following few subsection are long and at times cumbersome. Do not mistake its length to be a sign that it is of overwhelming importance. I discuss definiteness in some detail in the notes as it is not covered so much in the lectures. For reference, let me list the most important points:

1. The definitions of definiteness and semidefiniteness
2. Second-order Taylor approximation of multivariate functions f and the Hessian matrix of f
3. Connecting definiteness of the Hessian to convexity and concavity discussed in the next section

A quadratic form is a homogenous second-degree polynomial whose terms are all of second order. They can be written as:

$$x \cdot Ax$$

for some symmetric matrix A.

A quadratic form is *positive definite* if for all $x \neq 0$, $x \cdot Ax > 0$. It is *positive semidefinite* if for all x, $x \cdot Ax \geq 0$.

A quadratic form is *negative definite* if for all $x \neq 0$, $x \cdot Ax < 0$. It is *negative semidefinite* if for all x, $x \cdot Ax \leq 0$. In all other cases, we say that the quadratic form is indefinite.

By Taylor's theorem, we can use definiteness to classify the local extrema of quadratic functions. Let $\nabla f(\widehat{x}) = 0$. Then if $D^2 f(\widehat{x})$ is positive definite, then $\widehat{x}$ is a local minimum. If $D^2 f(\widehat{x})$ is negative definite, then $\widehat{x}$ is a local maximum.

When is A positive definite? The easiest case is when A is a diagonal matrix. In this case, it is positive definite if and only if all of its diagonal elements are strictly positive. Moe generally the any positive definite matrix has strictly positive diagonal elements.

Another easy case is when A is a 2×2 matrix:

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

so that the quadratic form is:

$$ax_1^2 + 2bx_1x_2 + cx_2^2.$$

View this as a second degree function in x_2. If $c > 0$, this function has a minimum at

$$x_2 = -\frac{bx_1}{c}.$$

Plugging into the quadratic form:

$$ax_1^2 - 2\frac{b^2x_1^2}{c} + \frac{b^2x_1^2}{c} = \left(a - \frac{b^2}{c}\right)x_1^2.$$

This is strictly positive

$$\begin{aligned}\left(a - \frac{b^2}{c}\right) &> 0 \text{ or} \\ ac &> b^2.\end{aligned}$$

In other words, the quadratic form is positive definite if i) $a > 0$ ja ii) $\det A > 0$.

For negative definiteness, assume that $a, c < 0$. Solving for the maximal x_2 for eachx_1 gives:

$$x_2 = -\frac{bx_1}{c}$$

and plugging into the quadratic form and require that:

$$ax_1^2 - 2\frac{b^2x_1^2}{c} + \frac{b^2x_1^2}{c} = \left(a - \frac{b^2}{c}\right)x_1^2 < 0.$$

We get:

$$a < \frac{b^2}{c} \text{ or } ac > b^2.$$

In other words,

$$\det A > 0.$$

Unfortunately, the general case is tedious. I give it here for completeness, but it is not particularly illuminating. We need to consider the leading principal minors of A:

$$\begin{aligned}M_1 &= \det a_{11}, M_2 = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}, \\ M_3 &= \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}, ...\end{aligned}$$

A quadratic form

$$x \cdot Ax$$

is positive definite if $M_i > 0$ for all i. It is negative definite if $M_i (-1)^i > 0$ for all i, i.e. M_i is negative for odd i and positive for even i.

To analyze semidefiniteness of A, more is needed. Define for all $1 \leq i_1 < i_2 < ... < i_n \leq n$

$$A^n_{\{i_1,i_2,...,i_n\}} = \begin{pmatrix} a_{i_1 i_1} & a_{i_1 i_2} \cdot \cdot & a_{i_1 i_n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{i_n i_1} & a_{i_n i_2} ... & a_{i_n i_n} \end{pmatrix}.$$

and

$$M^n_{\{i_1,i_2,...,i_n\}} = \det \left(A^n_{\{i_1,i_2,...,i_n\}} \right).$$

The matrix A is positive semidefinite if

$$M^n_{\{i_1,i_2,...,i_n\}} \geq 0 \text{ kaikille } n \text{ ja kaikille } \{i_1, i_2, ..., i_n\}.$$

for all n and $\{i_1, i_2, ..., i_n\}$.

It is negative semidefinite if

$$M^n_{\{i_1,i_2,...,i_n\}} \leq 0 \text{ for all odd } n \text{ and for all } \{i_1, i_2, ..., i_n\},$$

$$M^n_{\{i_1,i_2,...,i_n\}} \geq 0 \text{ odd } n \text{ and for all } \{i_1, i_2, ..., i_n\}.$$

At the end of Part II of these lectures, we will discuss the eigenvalues of a matrix. It turns out that for symmetric matrices, A, there is a simple connection between definiteness and the sign of the eigenvalues. First of all, all eigenvalues of a symmetric matrix are real. If they are all positive (negative), then A is positive (negative) semidefinite. If they are all strictly positive (strictly negative), then it is positive (negative) semidefinite. A is indefinite only if it has a strictly positive and a strictly negative eigenvalue.

5.4.2 Definiteness with linear constraints

The definiteness of the quadratic form

$$x \cdot Ax$$

can also be considered under linear constraints. In other words, we require that

$$b \cdot x = 0.$$

Let x be a column vector so that $b \cdot x = 0$ restricts the set of vectors that we consider. We can ask whether A is definite in these directions

Consider the matrix

$$H = \begin{pmatrix} 0 & b_1 & \cdots & b_n \\ b_1 & a_{11} & & a_{1n} \\ \vdots & & & \\ b_n & a_{n1} & & a_{nn} \end{pmatrix},$$

and assume that $b_1 \neq 0$.

A is positive definite in directions $\{x \,|b \cdot x = 0\}$ if all the leading principal minors of H except for the first one are negative. It is negative definite in directions $\{x \,|b \cdot x = 0\}$ if all the leading principal minors of H except for the first one alternate in sign.

5.4.3 Examples

1. Consider the definiteness of

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

(a) $M^1 = \det(a_{11}) = 2.$

(b) $M^2 = \det \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 3.$

(c) $M^3 = \det \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 1 \end{pmatrix} = (-1)^{3+3} \det \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} + (-1)^{3+2} \det \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} +$
$(-1)^{3+1} \det \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} = 3 + 1 - 1 = 3.$

Therefore A is positive definite.

2. Consider matrix
$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & -1 & -1 \\ 0 & 1 & 1 \end{pmatrix}.$$
This is easily seen to be indefinite (why?).

3. Consider
$$A = \begin{pmatrix} -1 & -4 & -1 \\ -4 & 0 & 1 \\ -1 & 1 & -1 \end{pmatrix}.$$
(a) $M_1^1 = -1, M_2^1 = 0, M_3^1 = -1.$
(b) $M_{\{1,2\}}^2 = -16, M_{\{1,3\}}^2 = 0, M_{\{2,3\}}^2 = -1.$
We see already that A is indefinite.

4. Consider the function
$$f(x_1, x_2, x_3) = x_1^2 - x_2^3 + x_1 x_3$$
around $(x_1, x_2, x_3) = (0, 0, 0)$. The gradient is
$$\nabla f(x_1, x_2, x_3) = \begin{pmatrix} 2x_1 + x_3 \\ -3x_2^2 \\ x_1 \end{pmatrix}$$
Compute
$$\nabla f(0, 0, 0) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$
The Hessian matrix is given by:
$$D^2 f(x_1, x_2, x_3) = \begin{pmatrix} 2 & 0 & 1 \\ 0 & -6x_2 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$
Evaluate at $(0, 0, 0)$:
$$D^2 f(0, 0, 0) = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

This matrix is indefinite since $M_1^1 = 2 > 0$ and $M_{\{1,3\}}^2 = \det \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} = -1$.

5. $$f(x_1, x_2) = x_1^\rho + x_2^\rho.$$

   Form the gradient

   $$\nabla f(x_1, x_2) = \begin{pmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} \\ \frac{\partial f(x_1,x_2)}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \rho x_1^{\rho-1} \\ \rho x_2^{\rho-1} \end{pmatrix}.$$

   Form the Hessian matrix by taking the derivative of the gradient:

   $$D^2 f(x_1, x_2) = \begin{pmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_2} \end{pmatrix}.$$

   We get:

   $$D^2 f(x_1, x_2) = \begin{pmatrix} \rho(\rho-1) x_1^{\rho-2} & 0 \\ 0 & \rho(\rho-1) x_2^{\rho-2} \end{pmatrix}.$$

   $D^2 f(x_1, x_2)$ is thus negative definite when $x_i \neq 0$ ja $0 < \rho < 1$.

6. Consider the function

   $$f(x_1, x_2) = (x_1^\rho + x_2^\rho)^{\frac{1}{\rho}}.$$

   Form the gradient:

   $$\nabla f(x_1, x_2) = \begin{pmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} \\ \frac{\partial f(x_1,x_2)}{\partial x_2} \end{pmatrix} = \begin{pmatrix} (x_1^\rho + x_2^\rho)^{\frac{1}{\rho}-1} x_1^{\rho-1} \\ (x_1^\rho + x_2^\rho)^{\frac{1}{\rho}-1} x_2^{\rho-1} \end{pmatrix}.$$

   The the Hessian matrix is:

   $$D^2 f(x_1, x_2) = \begin{pmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_2} \end{pmatrix}.$$

   By the product rule:

   $$\begin{aligned} \frac{\partial^2 f(x_1, x_2)}{\partial x_1 \partial x_1} &= (\rho-1) x_1^{\rho-2} (x_1^\rho + x_2^\rho)^{\frac{1}{\rho}-1} \\ &\quad + \left(\frac{1}{\rho} - 1\right) (x_1^\rho + x_2^\rho)^{\frac{1}{\rho}-2} \rho x_1^{2\rho-2}, \end{aligned}$$

$$\frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} = \left(\frac{1}{\rho}-1\right)\left(x_1^\rho + x_2^\rho\right)^{\frac{1}{\rho}-2} \rho x_2^{\rho-1} x_1^{\rho-1},$$

$$\begin{aligned}\frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_2} &= (\rho-1)\, x_2^{\rho-2}\left(x_1^\rho + x_2^\rho\right)^{\frac{1}{\rho}-1} \\ &\quad + \left(\frac{1}{\rho}-1\right)\left(x_1^\rho + x_2^\rho\right)^{\frac{1}{\rho}-2} \rho x_2^{2\rho-2}.\end{aligned}$$

By collecting the common terms, we get:

$$\begin{aligned}D^2 f(x_1,x_2) &= \begin{pmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_2} \end{pmatrix} \\ &= \left(x_1^\rho + x_2^\rho\right)^{\frac{1}{\rho}-2} \begin{pmatrix} (\rho-1)\, x_1^{\rho-2} x_2^\rho & (1-\rho)\, x_2^{\rho-1} x_1^{\rho-1} \\ (1-\rho)\, x_2^{\rho-1} x_1^{\rho-1} & (\rho-1)\, x_2^{\rho-2} x_1^\rho \end{pmatrix}.\end{aligned}$$

When computing the determinant, we can separate the common factor:

$$\det\left(D^2 f(x_1,x_2)\right) =$$

$$\left(x_1^\rho + x_2^\rho\right)^{\frac{1}{\rho}-2} x_1^{2\rho-2} x_2^{2\rho-2} \det\begin{pmatrix} (\rho-1) & (1-\rho) \\ (1-\rho) & (\rho-1) \end{pmatrix} = 0.$$

$D^2 f(x_1,x_2)$ is therefore negative semidefinite if $\rho < 1$ and positive semidefinite if$\rho > 1$.

7. Consider the definiteness of the matrix

$$A = \begin{pmatrix} 2 & 5 \\ -5 & 2 \end{pmatrix}$$

with the linear constraint

$$x_1 + x_2 = 0.$$

Form the bordered matrix

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 5 \\ 1 & 5 & 2 \end{pmatrix}$$

and analyze its two last principal minors:

(a) $\det\begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix} = -1$

(b) $\det\begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 5 \\ 1 & 1 & 2 \end{pmatrix} = (-1)^{2+1}\det\begin{pmatrix} 1 & 1 \\ 5 & 2 \end{pmatrix} + (-1)^{3+1}\det\begin{pmatrix} 1 & 1 \\ 2 & 5 \end{pmatrix} = 8.$

Therefore we conclude that $A = \begin{pmatrix} 2 & 5 \\ 5 & 2 \end{pmatrix}$ is negative definite under the constraint $x_1 + x_2 = 0$.

5.5 Definiteness and comparative statics

Consider the unconstrained optimization of

$$f(y; x).$$

Take y as the endogenous variable and x as exogenous. Write the problem of maximizing $y : n$ as follows:

$$\max_y f(y; x)$$

The first order condition for optimum:

$$\frac{\partial f}{\partial y}(\widehat{y}; \widehat{x}) = 0.$$

A sufficient condition for local maximum is obtained from Taylor's theorem:

$$f(\widehat{y} + dy; \widehat{x}) - f(\widehat{y}; \widehat{x}) = \frac{\partial f}{\partial y}(\widehat{y}; \widehat{x})\, dy + \frac{1}{2}\frac{\partial^2 f}{\partial y \partial y}(\widehat{y}; \widehat{x})\,(dy)^2 + ...$$

If

$$\frac{\partial^2 f}{\partial y \partial y}(\widehat{y}; \widehat{x}) < 0,$$

then $f$has a local maximum at$(\widehat{y}; \widehat{x})$.

Note that then also the function

$$\frac{\partial f}{\partial y}(\widehat{y}; \widehat{x})$$

has a non-zero derivative w.r.t. the endogenous variable at $(\widehat{y};\widehat{x})$ and we can apply the implicit function theorem $y$to get the optimal y as a function of x.

Since

$$\frac{\partial f}{\partial y}\left(y\left(x\right);x\right)=0.$$

for all x near $\widehat{x}$, we get:

$$\frac{\partial^2 f\left(\widehat{y};\widehat{x}\right)}{\partial y\partial y}dy+\frac{\partial^2 f\left(\widehat{y};\widehat{x}\right)}{\partial y\partial x}dx=0$$

or

$$\frac{dy}{dx}=-\frac{\frac{\partial^2 f(\widehat{y};\widehat{x})}{\partial y\partial x}}{\frac{\partial^2 f(\widehat{y};\widehat{x})}{\partial y\partial y}}.$$

Since $\frac{\partial^2 f(\widehat{y};\widehat{x})}{\partial y\partial y}<0$ by second-order condition for optimum, we see that $\frac{dy}{dx}$ has the same sign as $\frac{\partial^2 f(\widehat{y};\widehat{x})}{\partial y\partial x}$.

Example 3 (Optimal monopoly production) *Let x be the output by the monopolist. $P\left(q\right)=\alpha-b\left(q\right)$ is the inverse demand function and cq^2 is the cost function of the monopolist. The monopolist's maximization problem is then*

$$\max_q \pi\left(q;\alpha,c\right)=q\left(\alpha-b\left(q\right)\right)-cq^2.$$

First-order condition for optimality:

$$D\pi\left(q;\alpha,c\right)=\alpha-b\left(q\right)-qb'\left(q\right)-2cq=0.$$

Second-order condition:

$$D^2\pi\left(q\right)<0.$$

How does the optimal output change when α or c changes?

By the previous result, the sign of the change in the endogenous variable depends on the signs of

$$\frac{\partial^2\pi\left(q;\alpha,c\right)}{\partial q\partial\alpha}$$

and

$$\frac{\partial^2\pi\left(q;\alpha,c\right)}{\partial q\partial c}.$$

6 Convex and concave functions

In this last section of Part I of the course, we take a first look at the extremely important question of convexity and concavity of functions. These notions revolve around quite general geometric notions in $\mathbb{R}^n$ and you will see applications in many different areas of economic theory (in particular under the title of 'duality theory'. For us now, the most immediate questions relate to the curvature of non-linear functions and their extrema. The really useful observation for optimization is that for concave functions, the first-order decessary conditions are also sufficient. In other words if f is concave and $Df(\hat{x}) = 0$, then $\hat{x}$ is a maximum of f.

6.1 Basic definitions

We start with a definition of convex sets.

Definition 1 *A set X* is convex *if for all $x, y \in X$ and for all $\lambda \in [0, 1]$, we have:*

$$\lambda x + (1 - \lambda) y \in X.$$

We call $\lambda x + (1 - \lambda) y$ a convex combination *of x and y.*

On the real line, convex sets are intervals $a \leq x \leq b$ for some $-\infty \leq a \leq b \leq \infty$. In $\mathbb{R}^n$, convex sets are sets X with the property that when you connect linearly two points in X, the entire connecting line is also in X. Hence a disk in the plane is convex and a cube in the three dimensional space are convex, but the circle in the plane is not, a disk with the center removed is not, a doughnut in three dimensions is not etc.

Consider a real-valued function $f : X \to \mathbb{R}$, where X is a convex set.

Definition 2 *The function f* is convex *if for all $x, y \in X$ and for all $\lambda \in [0, 1]$, we have:*

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

f is concave *if*

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

Observations:

- If $f(x)$ is convex, then $g(x) = -f(x)$ is concave.
- If $f(x)$ is convex, then $af(x)$ is convex if $a > 0$.
- If $f(x)$ and $g(x)$ are convex, then $h(x) = f(x) + g(x)$ is convex.
- If $f(x)$ and $g(x)$ are convex, then $h(x) = f(x)g(x)$ is not necessarily convex. (Give an example for both cases, i.e. where the product of convex functions is convex and where it is not).
- Exercise: Assume that $f : X \to \mathbb{R}$ is convex and $g : \mathbb{R} \to \mathbb{R}$ is also convex. Is $g(f(x))$ convex? What if g is increasing and convex?
- (Optional Exercise): Assume that $f : X \to \mathbb{R}$ is a convex function. Show that the set

$$\{(x, y) \in \mathbb{R}^{n+1} \quad | \quad x \in X, y \geq f(x)\}$$

  is a convex set.
- If $f(x)$ and $g(x)$ are convex, then $h(x) = \max\{f(x), g(x)\}$ is convex.

  Proof: Since by assumption, f and g are convex, we have:

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

  and

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

  By definition,

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

  The first inequality follows from the convexity of f and g.The second follows by choosing the larger of $f\cdot, g(\cdot)$ for x, y. The last equality is just the definition of h.

- The same result is true for an arbitrary set of convex functions. Let $f(x;\alpha)$ be convex in $x$for all α. Then

$$g(x) = \max_{\alpha} f(x;\alpha)$$

  is convex. The proof is identical to the one above.

- Since linear functions are convex, this result holds for any set of linear functions.

- Since

$$\max\{f(x), g(x)\} = -\min\{-f(x), -g(x)\},$$

  and since $-f$ is concave when f is convex, we get:

$$g(x) = \min_{\alpha} f(x;\alpha)$$

  is concave if $f(x;\alpha)$ is concave in x for all α.

Example 4 (Profit function of a firm) *A competitive firm sells output y at price p_0 and buys inputs $x = (x_1, ..., x_n)$ at input prices $(p_1, ..., p_n)$.*

Its profit is

$$p_0 y - \Sigma_{i=1}^{n} p_i x_i.$$

The maximization problem is then

$$\max_{y,x\in F} p_0 y - \Sigma_{i=1}^{n} p_i x_i,$$

where F is the feasible set determined by technological possibilities.

The profit function of the firm gives the maximum achievable profit amongst the feasible set at input and output prices p.

$$\pi(p) = \pi(p_0, p_1, ..., p_n) = \max_{y,x\in F} p_0 y - \Sigma_{i=1}^{n} p_i x_i$$

Since the profit from a fixed feasible production is a linear function of the prices p, the profit function is the maximum over linear functions and therefore convex in p.

Example 5 (Expenditure minimization) *Let X be the feasible set for inputs $x = (x_1, ..., x_n)$ and $p = (p_1, ..., p_n)$ be the input prices. The expenditure function*

$$e(p; X) = \min_{x\in X} p.x = \min_{x\in X} \Sigma_{i=1}^{n} p_i x_i$$

is a concave function by the same argument as above.

These two examples show that convexity and concavity play a real role in economic applications. We shall see more applications when we discuss constrained optimization and value functions of optimization problems. Is there an economic intuition for the maximum of linear functions being convex? We'll return to this after some further characterizations of convex functions.

6.2 Convexity and concavity of differentiable functions

When $f : \mathbb{R} \to \mathbb{R}$, and f is convex and differentable, it it is easy to see by drawing a picture that for all x, y we have:

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

This just says that the graph $(x, f/x))$ of a convex function f is above all of its tangent lines.

Proposition 1 *A differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if*

$$f(y) - f(x) \geq f'(x)(y - x) \text{ for all } x, y.$$

Proof i) Let f be convex. Then for all x, y:

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

or

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

or

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

Letting $\lambda \to 1$, we get:

$$(y - x) f'(x) \leq f(y) - f(x).$$

ii) Assume that

$$f(y) - f(x) \geq f'(x)(y - x) \text{ for all } x, y.$$

Then

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

and

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

Multiply the first inequality by λ and the second by $(1-\lambda)$and sum together to get the definition of convex functions. □

Consider next $f : X \to \mathbb{R}$, where X is a convex subset of $\mathbb{R}^n$.

Proposition 2 *A differentiable function $f : X \to \mathbb{R}$ is convex if and only if*

$$f(y) - f(x) \geq Df(x)(y-x).$$

Proof We start with a preliminary result: f is convex if and only if $g_{x,y}(\lambda) := f((1-\lambda)x + \lambda y)$ is convex for all x, y. In other words, convexity is equivalent to convexity along convex combinations. The proof of this is left as a relatively easy exercise.

Using this result and the chain rule,

$$\begin{aligned} g'_{x,y}(\lambda) &= \Sigma_{i=1}^{n} \frac{\partial f(x + \lambda(y-x))}{\partial x_i}(y_i - x_i) \\ &= Df(x + \lambda(y-x))(y-x). \end{aligned}$$

By the previous theorem, $g_{x,y}(\lambda)$ is convex if and only if

$$g_{x,y}(1) - g_{x,y}(0) \geq g'_{x,y}(0).$$

In other wirds if and only if

$$f(y) - f(x) \geq Df(x)(y-x).$$

□

This is the multidimensional generalization to the geometric notion that the graphs of convex functions lie above their tangent lines. Can you formulate this condition in terms of level curves and gradients? What is the corresponding result to concave functions?

Exercise: Using this condition, show that if f is convex (concave) on the convex set X and $Df(\hat{x}) = 0$, then $\hat{x}$ is a global minumum (maximum) of f on X

6.3 Second derivatives and convexity

Start again with functions of a single variable. By Taylor's theorem without te remainder term,

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2}f''(x)(y-x)^2 + \frac{1}{6}f'''(x)(y-x)^3 + ...$$

In order to have

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

for $|y-x|$ small, we must have

$$f''(x) \geq 0.$$

In other words, convex functions have a positive second derivative.

Taylor's theorem with a remainder term of second degree:

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

for some $z \in [x, y]$. If f'' is everywhere non-negative, we get:

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

for all y, x and f is therefore convex.

Let's generalize now to $f : X \to \mathbb{R}$, where X is a convex subset of $\mathbb{R}^n$ n.

We use again the function

$$g_{x,y}(\lambda) = f((1-\lambda)x + \lambda y) = f(x + \lambda(y-x))$$

and consider the second derivatives of g.

Convexity corresponds to positive semidefiniteness of the Hessian matrix. Concavity corresponds to negative semidefiniteness of the Hessian matrix. Hence we see an immediate connection between convexity and the second order conditions for optimality.

6.4 Quasiconvex and quasiconcave functions

Even though the name suggests something extremely technical and tedious, quasiconcavity is actually one of the most important notions for functions in economic theory. We begin with the definitions and properties of quasiconcave functions, but at the end of this section, I will discuss why this is such a useful definitions for economic modeling.

Definition 3

A function f on a convex set X is quasiconcave *if for all $x, y \in X$ and for all $\lambda \in [0,1]$*

$$f(\lambda x + (1-\lambda) y) \geq \min\{f(x), f(y)\}.$$

f is quasiconvex *is for all $x, y \in X$ and for all $\lambda \in [0,1]$*

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

Exercise: f is quasiconcave, then $-f$ is quasiconvex.
We can make some observations:

- If f is quasiconcave, then af is quasiconcave if $a > 0$.
- If f and g are quasiconcave $f + g$ is not necessarily quasiconcave.
- All monotone (i.e. all increasing and all decreasing) functions of a single variable are both quasiconcave and quasiconvex.
- All concave functions are quasiconcave. Show this as an exercise.
- Not all quasiconcave functions are concave.
- If f is a quasiconcave function and g is a strictly increasing function, then $h(x) = g(f(x))$ is a quasiconcave function.

  Proof:

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

  Exercise: where was the increasing property of g used in the proof?

An upper contour set of function f for value α is denoted by $U(f;\alpha)$ and defined as:

$$U(f;\alpha) := \{x \in X | f(x) \geq \alpha\}.$$

Interpretation: if f is a utility function, $U(f;\alpha)$ is the better side of the indifference curve giving utility level α.

Proposition 3 *A function f is quasiconcave if and only if $U(f;\alpha)$ is a convex set for all α.*

Proof i) Assume that f is quasiconcave and $x, y \in U(f;\alpha)$. Then $f(x) \geq \alpha$, $f(y) \geq \alpha$ and by quasiconcavity of f,

$$f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\} \geq \alpha.$$

In other words

$$\lambda x + (1-\lambda)y \in U(f;\alpha),$$

and therefore $U(f;\alpha)$ is convex.

ii) Assume that $U(f;\alpha)$ is a convex set for all α. Then

$$x, y \in U(f, \min\{f(x), f(y)\}),$$

and

$$\lambda x + (1-\lambda)y \in U(f, \min\{f(x), f(y)\}).$$

But then by the definition of $U(f;\alpha)$:

$$f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}.$$

□

Let me make a methodological point here. For economic modeling, the exact mathematical form of the utility function is unimportant since many different functions represent the same preferences as discussed before. The meaningful properties for an economic model relate to the preferences, i.e. to the indifference curves. As a result, an assumption on the shape of these curves or their upper contour sets are meaningful. The convexity of upper contour sets of utility functions is a meaningful property and it is often assumed in models of consumer choice. Notice that the shapes of the upper contour sets remain unchanged when going from $u(x)$ to $v(u(x))$ for a strictly increasing v. This follows from the observation that

$$U(v(u); v(\alpha)) = U(u;\alpha) \text{ for all } \alpha.$$

We end this subsection with a useful special case of quasiconcave functions:

Definition 4 *A function f on a convex set X is* strictly quasiconcave *if for all $x, y \in X$ and for all $\lambda \in (0,1)$*

$$f(\lambda x + (1-\lambda) y) > \min\{f(x), f(y)\}.$$

The following exercise shows why strict quasiconcavity is very useful for optimization problems.

Exercise: show that if a strictly quasiconcave function has a maximum, then the maximum is unique.

6.5 Quasiconcavity and differentiability

A differentiable function f on a convex set X is quasiconcave if and only if:

$$f(y) \geq f(x) \Rightarrow Df(x)(y-x) \geq 0.$$

Exercise: Compare this to the definition of concavity for differentiable functions and relate this condition to the geometry of upper contour sets and tangent planes to the upper contour sets.

The second order conditions for quasiconcavity based on bordered Hessian matrices are extremely complex and contain little economic intuition. The textbook on pages 527-531 gives an introduction to this.

With these preliminaries, we are ready for constrained optimization in Part II of this course.