Mathematics for Economists: A Synopsis

Juuso Välimäki

Spring 2021

This synopsis outlines with minimal explanations what you are meant to learn in this course. It serves as a guide to orienting your studies and it is not meant to be self-contained. Additional and more detailed material can be found on the course web page organized by the week. The lectures are centered on applications of the material.

In this synopsis, I denote real valued variables by $x, y, z, ...$ In this course, we deal mostly with real vectors denoted by boldface variables $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, ...$ A real vector $\boldsymbol{x}$ with n components is then written as $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$. Vectors are interpreted as column vectors. The transpose of vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^\top$ and $\boldsymbol{x} \cdot \boldsymbol{y}$ is shorthand for the inner product $\boldsymbol{x}^\top \boldsymbol{y} := \sum_{i=1}^n x_i y_i$. The *norm* (or length) of a vector in $\mathbb{R}^n$ is denoted by $\|\boldsymbol{x}\| := \sqrt{\sum_{i=1}^n x_i^2}$.

1 Linear models

For $x \in \mathbb{R}$, and parameters $a, b \in \mathbb{R}$ a linear equation is written as:

$$ax = b.$$

If $a = 0$ and $b \neq 0$, the equation has no solution. If $a \neq 0$, it has a single solution given by:

$$x = a^{-1}b.$$

The equation has infinitely many solutions if $a = b = 0$.

For $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{A}$ an $m \times n$ matrix, $\boldsymbol{b} \in \mathbb{R}^m$, we have the linear system of equations:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

Gaussian elimination is the main methods for determining whether the system has solutions. Gaussian elimination transforms $\boldsymbol{A}$ into its *row echelon form* via *elementary row operations*. The *rank* of a matrix is the number of non-zero rows in its row echelon form. When performed on the augmented matrix $\left(\boldsymbol{A}\vdots\boldsymbol{b}\right)$ where $\boldsymbol{b}$ is added to the matrix as the last column. When the elementary row operations transforming $\boldsymbol{A}$ to the identity matrix are performed on the augmented matrix, the solution $\boldsymbol{x}$ to the system of equations can be read in the last column of the transformed augmented matrix.

The linear equation system:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

has a solution if and only if $\boldsymbol{A}$ and $\left(\boldsymbol{A}\vdots\boldsymbol{b}\right)$ have the same rank. If the rank is n, the solution is unique, if the rank is less than n, then the system has infinitely many solutions.

The most important case is when $n = m$. Then solution exists and is unique for all $\boldsymbol{b}$ if and only if $\boldsymbol{A}$ has rank n. In this case, $\boldsymbol{A}$ has an inverse matrix $\boldsymbol{A}^{-1}$ and:

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}.$$

The inverse matrix can be computed via Gaussian elimination on the augmented matrix $\left(\boldsymbol{A}\vdots\boldsymbol{I}\right)$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

The *determinant* of a square matrix is a numerical function that depends on the coefficients a_{ij} of the matrix. The most important observation is that the determinant is non-zero if and only if the matrix has full rank. *Cramer's rule* gives an alternative way for computing the solution to linear systems such as (**??**) using determinants:

$$x_i = \frac{\det(\boldsymbol{B}_i)}{\det(\boldsymbol{A})},$$

where $\boldsymbol{B}_i$ is the matrix obtained by replacing the i^{th} column of $\boldsymbol{A}$ by $\boldsymbol{b}$.

A collection of vectors $\{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n \in \mathbb{R}^m\}$ are called linearly independent if:

$$\sum_i \lambda_{i=1}^n \boldsymbol{a}_i = 0 \quad \Rightarrow \lambda_i = 0 \forall i \in \{1, ..., n\}.$$

Similarly for row vectors.

Consider the $n \times n$ matrix $\boldsymbol{A}$ that has $\{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n \in \mathbb{R}^n\}$ as its columns. The column rank of $\boldsymbol{A}$ is defined to be the maximal number of linearly independent columns. Similarly we may consider a matrix $\boldsymbol{A}$ that has $\{\boldsymbol{a}_1^\top, \boldsymbol{a}_2^\top, ..., \boldsymbol{a}_n^\top \in \mathbb{R}^n\}$ as its rows. The row rank of a matrix is the maximal number of linearly independent rows that it has. The row rank and the column rank of any matrix are the same and therefore we can speak of the rank of a matrix without ambiguity.

2 The derivative

Consider a real valued function $f : \mathbb{R}^n \to \mathbb{R}$ of n real variables and an arbitrary point $\hat{\boldsymbol{x}} \in \mathbb{R}^n$. Fix all the other coordinates $x_j = \hat{x}_j$ for $j \neq i$, but let x_i vary. Then we have the function

$$\hat{f}(\hat{x}_1, ...\hat{x}_{i-1}, x_i, \hat{x}_{i+1}, ..., x_n)$$

of the single real variable x_i. The derivative of this function at $x_i = \hat{x}_i$ is called the *partial derivative of f with respect to x_i at $\hat{\boldsymbol{x}}$* and denoted by

$$\frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_i}.$$

The derivative of f at $\hat{\boldsymbol{x}}$ is a *linear function* that approximates f well for $\boldsymbol{x}$ close to $\hat{\boldsymbol{x}}$. If the partial derivatives with respect to all x_i exist and are continuous in $\boldsymbol{x}$ at $\hat{\boldsymbol{x}}$, then a derivative at $\hat{x}$ also exists and is given by the row vector:

$$D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}) = (\frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_1}, ... \frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_n}).$$

We have:

$$f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) = D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \text{h.o.t.},$$

where h.o.t. means terms that vanish in comparison to $D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}})$ as $\boldsymbol{x}$ is close to $\hat{\boldsymbol{x}}$.

For small $\Delta \boldsymbol{x}$, we can compute the approximation:

$$f(\hat{\boldsymbol{x}} + \Delta \boldsymbol{x}) - f(\hat{\boldsymbol{x}}) \approx D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}) \Delta \boldsymbol{x}.$$

We call this the directional derivative of f in direction $\Delta \boldsymbol{x}$ at $\hat{\boldsymbol{x}}$.

The gradient $\nabla f(\hat{\boldsymbol{x}})$ is the transpose of the derivative (i.e. the column vector of partial derivatives at $\hat{\boldsymbol{x}}$.

A function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a vector of functions $f_i : \mathbb{R}^n \to \mathbb{R}$:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{pmatrix}.$$

If all the partial derivatives of all component functions exist and are continuous at $\hat{\boldsymbol{x}}$, then the derivative of such a function at $\hat{\boldsymbol{x}}$ is the $m \times n$ matrix:

$$D_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f_1(\hat{\boldsymbol{x}})}{\partial x_1} & \cdots & \frac{\partial f_1(\hat{\boldsymbol{x}})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\hat{\boldsymbol{x}})}{\partial x_1} & \cdots & \frac{\partial f_m(\hat{\boldsymbol{x}})}{\partial x_n} \end{pmatrix}.$$

Since partial derivatives can be viewed as standard derivatives in a fixed direction, the rules for computing derivatives remain valid for multivariate functions. In particular, we have the chain rule for $h(\boldsymbol{x}) := f(g(\boldsymbol{x}))$, where $f : \mathbb{R}^k \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^k$ Let $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_k) = (g_1(\hat{\boldsymbol{x}}), ..., g_1(\hat{\boldsymbol{x}}))$:

$$D_{\boldsymbol{x}} h(\boldsymbol{x})) = D_{\boldsymbol{y}} f(\hat{\boldsymbol{y}}) D_{\boldsymbol{x}} g(\hat{\boldsymbol{x}}).$$

Writing this matrix multiplication explicitly gives the ij^{th} element of $D_{\boldsymbol{x}} h(\boldsymbol{x})$ as:

$$\frac{\partial h_i(\hat{\boldsymbol{x}}}{\partial x_j} = \sum_{i=1}^{k} \frac{\partial f_i(\hat{\boldsymbol{y}})}{\partial y_k} \frac{\partial g_k(\hat{\boldsymbol{x}})}{\partial x_j}.$$

2.1 Implicit function theorem

Consider functions $\boldsymbol{f} : R^{n+m} \to \mathbb{R}^n$. Let $\boldsymbol{y} \in \mathbb{R}^n$ denote endogenous variables in an economic model and $\boldsymbol{x} \in \mathbb{R}^m$ the exogenous variables and write the model as:

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

Assume that $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})$ satisfies the model, i.e. $f(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}) = 0$. We want to know how the endogenous variables $\boldsymbol{y}$ behave when $\boldsymbol{x}$ changes a bit from $\hat{\boldsymbol{x}}$.

Assume that the derivative of $\boldsymbol{f}$ exists at $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})$ and that the derivative with respect to endogenous variables,

$$D_{\boldsymbol{y}} f(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}) = \begin{pmatrix} \frac{\partial f_1(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})}{\partial y_1} & \cdots & \frac{\partial f_1(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})}{\partial y_1} & \cdots & \frac{\partial f_n(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})}{\partial y_n} \end{pmatrix}$$

has full rank (i.e. non-zero determinant). Then the implicit function theorem tells us that we have a differentiable implicit function $\boldsymbol{y}(\boldsymbol{x})$ with $\boldsymbol{y}(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{y}}$ defined in a neighborhood of $\hat{\boldsymbol{x}}$ such that for all $\boldsymbol{x}$ in that neighborhood,

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

Chain rule gives:

$$D_{\boldsymbol{y}} \boldsymbol{f}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}) D_{\boldsymbol{x}} \boldsymbol{y}(\boldsymbol{x}) + D_{\boldsymbol{x}} \boldsymbol{f}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}) = 0.$$

Since we have assumed that $D_{\boldsymbol{y}} f(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}})$ has full rank, it is invertible and:

$$D_{\boldsymbol{x}} \boldsymbol{y}(\hat{\boldsymbol{x}}) = -(D_{\boldsymbol{y}} \boldsymbol{f}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}))^{-1} D_{\boldsymbol{x}} \boldsymbol{f}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}).$$

The effect on y_i of a change in a particular x_k can be computed with Cramer's rule.

Illustration: indifference curves for a utility function $u(x, y)$

On an indifference curve, the utility is at a constant level:$u(x, y) = \overline{u}$. How to draw an indifference curve through point $(\hat{x}, \hat{y})$? The level of utility at that point is $u(\hat{x}, \hat{y})$. On the indifference curve through that point, the level of utility is the same:

$$u(x, y) = u(\hat{x}, \hat{y}).$$

If $\frac{\partial u(\hat{x}, \hat{y})}{\partial y} \neq 0$, then implicit function theorem tells us that for x near $\hat{x}$, we have a function $y(x)$ with $y(\hat{x}) = \hat{y}$ and

$$y'(\hat{x}) = -\frac{\frac{\partial u(\hat{x}, \hat{y})}{\partial x}}{\frac{\partial u(\hat{x}, \hat{y})}{\partial y}}.$$

3 Unconstrained optimization

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$. A point $\hat{\boldsymbol{x}}$ is a *maximum* of f if $f(\hat{\boldsymbol{x}}) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x}$. It is a *minimum* if $f(\hat{\boldsymbol{x}}) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x}$.

The point $\hat{\boldsymbol{x}}$ is called a *local maximum* (*local minimum*) if $f(\hat{\boldsymbol{x}}) \geq (\leq) f(\boldsymbol{x})$ for all $\boldsymbol{x}$ in some neighborhood of $\hat{\boldsymbol{x}}$.

A necessary condition for local (and therefore also global) minimum and maximum is that all the partial derivatives of f vanish at $\hat{\boldsymbol{x}}$ or $\nabla f(\hat{\boldsymbol{x}}) = 0$. Necessary condition means that all minima and maxima have this property.

Points satisfying the necessary condition are called critical points of f. In order to classify the critical points, we need to look at the Hessian matrix of second derivatives of f at $\hat{\boldsymbol{x}}$.

Second order Taylor approximation gives:

$$f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) = D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}) \cdot H_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}}) + h.o.t.,$$

where

$$H_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}) = \begin{pmatrix} \frac{\partial^2 f(\hat{\boldsymbol{x}})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\hat{\boldsymbol{x}})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\hat{\boldsymbol{x}})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\hat{\boldsymbol{x}})}{\partial x_n^2} \end{pmatrix},$$

and the h.o.t. vanish in comparison to the first and second order terms for $\boldsymbol{x}$ close to $\hat{\boldsymbol{x}}$.

At a critical point, $D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}) = 0$ so that the first term on the right-hand side vanishes. By Young's theorem, the Hessian matrix is a symmetric matrix and hence defines a quadratic form.

A necessary condition for a local maximum (minimum) at a critical point $\hat{\boldsymbol{x}}$ is that for all $\boldsymbol{x}$,

$$(\boldsymbol{x} - \hat{\boldsymbol{x}}) \cdot H_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}}) \leq (\geq) 0.$$

A sufficient condition for local maximum or at $\hat{\boldsymbol{x}}$ is that the inequality above is strict for all $\boldsymbol{x}$.

A symmetric matrix is said to be positive (negative) definite if $\boldsymbol{x} \cdot \boldsymbol{A}\boldsymbol{x} > (<) 0$ for all $\boldsymbol{x}$. It is positive (negative) semi-definite if the $\boldsymbol{x} \cdot \boldsymbol{A}\boldsymbol{x} \geq (\leq) 0$ for all x.

If $\boldsymbol{x} \cdot \boldsymbol{A}\boldsymbol{x}$ is strictly positive for some $\boldsymbol{x}$ and strictly negative for some $\boldsymbol{x}$, $\boldsymbol{A}$ is said to be indefinite. The criteria for checking definiteness are given in the textbook and in the lecture notes and slides.

Notice that for functions of a single real variable, the Hessian is just the second derivative of the function.

3.1 Convex and concave functions

A set $X \in \mathbb{R}^n$ is convex if the line connecting any two points $\boldsymbol{x}, \boldsymbol{y}$ in the set also belongs to X. A function f on a convex set is said to be *concave* if its graph on any line connecting $\boldsymbol{x}, \boldsymbol{y} \in X$ lies above the line joining $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$, i.e. for all $\boldsymbol{x}, \boldsymbol{y} \in X$ and all $\lambda \in [0, 1]$,

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

A function is *convex* if for all $\boldsymbol{x}, \boldsymbol{y} \in X$ and $\lambda \in [0, 1]$,

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

Examples and criteria for convex sets and concave and convex functions are given in the lecture notes.

Every critical point of a concave (convex) function f is a global maximum (minimum) of f.

A twice differentiable function f is concave (convex) if and only if the Hessian at $\boldsymbol{x}$ is negative (positive) semi-definite for all $\boldsymbol{x}$.

If $g(\boldsymbol{x}, \boldsymbol{a})$ are affine (linear plus a constant) functions of $\boldsymbol{a}$, then $f(\boldsymbol{a}) = \max_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{a})$ is a convex function of $\boldsymbol{a}$. Similarly, the minimum of affine functions is concave.

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

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

It is *quasiconvex* if

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

An alternative characterization is that function f is quasiconcave (quasiconvex) if and only if its upper (lower) contour sets $\{\boldsymbol{x} : f(\boldsymbol{x}) \geq (\leq) f(\hat{\boldsymbol{x}})\}$ are convex for all $\hat{\boldsymbol{x}} \in X$.

Any critical point of a quasiconcave (quasiconvex) function with non-vanishing derivative is a global maximum (minimum).

4 Optimization with equality constraints

Consider the maximization problem:

$$\max f(x_1, ..., x_n)$$

subject to

$$\begin{aligned} g_1(x_1, ..., x_n) &= 0, \\ &\vdots \\ g_k(x_1, ..., x_n) &= 0. \end{aligned}$$

If f is continuous and the feasible set is compact, then Weierstrass' theorem guarantees that a maximum exists.

Constraint qualification at $\boldsymbol{x}$ requires that the matrix of the derivatives of the constraints $D_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})$ has full rank at $\boldsymbol{x}$. For a single constraint, this just requires a non-zero gradient.

With a single constraint, consider the following intuition. The feasible set cannot intersect the level curve of the objective function at optimum. If there was such an intersection, part of the feasible set would give a strictly higher value to the objective function than the intersection point. Therefore an intersection point cannot be optimum and we get the tangency of the level curve and the feasible set. Since the feasible set is the level curve of the constraint function, the gradients of the objective function and the constraint must be collinear.

A necessary condition for a maximum at a point $\hat{\boldsymbol{x}}$ where constraint qualification holds is that $\hat{\boldsymbol{x}}, \hat{\boldsymbol{\mu}}$ be a critical point of the following Lagrangean function

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{j=1}^{k} \mu_j g_j(\boldsymbol{x}).$$

In other words, we must look for $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\mu}})$ such that

$$\frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_i} - \sum_{j=1}^{k} \hat{\mu}_j \frac{\partial g_j(\hat{\boldsymbol{x}})}{\partial x_i} = 0 \text{ for all } i \in \{1, ..., n\},$$

$$g_j(\hat{x}_1, ..., \hat{x}_n) = 0 \text{ for all } j \in \{1, ..., k\}.$$

For a linear constraint, the feasible set is convex and therefore by the results in the previous section, critical points are global maxima if the objective function is concave (or quasiconcave with a non-zero gradient).

Otherwise, one must check whether the critical points are maxima or minima using the bordered Hessian at the critical point as explained in the textbook.

5 Optimization with inequality constraints

The problem is now to

$$\max f(x_1, ..., x_n)$$

subject to

$$\begin{gathered} g_1(x_1, ..., x_n) \leq 0, \\ \vdots \\ g_k(x_1, ..., x_n) \leq 0. \end{gathered}$$

The difficulty is now that we do not know which of the constraints are binding. The necessary conditions can be formulated using a Lagrangean function, but now we need also complementary slackness conditions for the constraints. If the constraint binds, its multiplier is positive, if it does not bind, the multiplier is zero.

The Lagrangean of the problem is:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{j=1}^{k} \lambda_j g_j(\boldsymbol{x}).$$

The necessary first-order Kuhn-Tucker conditions for the problem at a point where the constraint qualification holds (i.e. the derivatives of the binding constraints are linearly independent) are given by:

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} - \sum_{j=1}^{k} \lambda_j \frac{\partial g_j(\boldsymbol{x})}{\partial x_i} = 0 \text{ for all } i \in \{1, ..., n\},$$

$$\lambda_j g_j(\boldsymbol{x}) = 0 \text{ for all } j \in \{1, ..., k\},$$

$$g_j(\boldsymbol{x}) \leq 0 \text{ for all } j \in \{1, ..., k\},$$

$$\lambda_j \geq 0 \text{ for all } j \in \{1, ..., k\}.$$

If the objective function is quasiconcave and has a nonzero gradient and if the feasible set is convex, then any point satisfying the K-T conditions is a global maximum.

Of course, it is a nightmare to solve a system such as this one in general. Therefore it is advisable to look carefully at the problem and argue which of the constraints cannot bind at optimum (this can often be done for non-negativity constraints) and argue which constraints are binding (typically budget constraints). The art of solving optimization problems is often the art of simplifying the constraints.

5.1 Utility maximization: an example

Consider the problem of maximizing

$$u(x, y) = (a_x x^\rho + a_y y^\rho)^{\frac{1}{\rho}},$$

subject to

$$p_x x + p_y y \leq w, x, y \geq 0,$$

where $\rho < 1, \quad \rho \neq 0$ and $a_x, a_y > 0$.

The constraint set is compact and the utility function is convex so therefore an optimum exists by Weierstrass' theorem. The utility function is quasiconcave with non-vanishing gradient and the feasible set is convex and thus K-T conditions are also sufficient. The binding constraints are also linearly independent (since the non-negativity constraints are orthogonal and the budget constraint cannot bind simultaneously with all the other constraints).

Lagrangean for the problem:

$$\mathcal{L}(x, y, \lambda, \lambda_x, \lambda_y) = (a_x x^\rho + a_y y^\rho)^{\frac{1}{\rho}} - \lambda(p_x x + p_y y - w) + \lambda_x x + \lambda_y y.$$

Kuhn-Tucker conditions:

$$\frac{\partial u(x,y)}{\partial x} - \lambda p_x + \lambda_x = \rho a_x x^{\rho-1} \frac{1}{\rho}(a_x x^\rho + a_y y^\rho)^{\frac{1}{\rho}-1} - \lambda p_x + \lambda_x = 0,$$

$$\frac{\partial u(x,y)}{\partial y} - \lambda p_y + \lambda_y = \rho a_y y^{\rho-1} \frac{1}{\rho}(a_x x^\rho + a_y y^\rho)^{\frac{1}{\rho}-1} - \lambda p_y + \lambda_y = 0,$$

$$\lambda(p_x x + p_y y - w) = 0,$$

$$\lambda_x x = \lambda_y y = 0,$$

$$\lambda, \lambda_x, \lambda_y, x, y \geq 0.$$

Since $\rho < 1$, $\frac{\partial u(0,y)}{\partial x}$ and $\frac{\partial u(x,0)}{\partial y}$ are not defined and we see that the first and second lines in K-T conditions cannot be satisfied at $x = 0$ or $y = 0$. Therefore we know that $\lambda_x = \lambda_y = 0$. Budget constraint must bind since:

$$\frac{\partial u(x,y)}{\partial x} > 0 \text{ and } \frac{\partial u(x,y)}{\partial y} > 0 \text{ for all } (x,y) > 0.$$

Solving for λ from the first two lines and cross-multiplying gives:

$$\frac{\frac{\partial u(x,y)}{\partial x}}{\frac{\partial u(x,y)}{\partial y}} = \frac{p_x}{p_y}.$$

Also, the budget constraint holds with equality:

$$p_x x + p_y y = w.$$

Plugging in the marginal utilities gives:

$$\frac{a_x x^{\rho-1}}{a_y y^{\rho-1}} = \frac{p_x}{p_y},$$

or

$$\frac{x}{y} = (\frac{a_y p_x}{a_x p_y})^{\frac{1}{\rho-1}},$$

or

$$y = x(\frac{a_y p_x}{a_x p_y})^{\frac{1}{1-\rho}}. \tag{2}$$

Substituting into the budget constraint, we get:

$$p_x x + p_y x(\frac{a_y p_x}{a_x p_y})^{\frac{1}{1-\rho}} = w.$$

We can solve for x_1 to get

$$x = \frac{w}{p_x + p_y(\frac{a_y p_x}{a_x p_y})^{\frac{1}{1-\rho}}}.$$

Substituting this into (**??**) lets us solve for y:

$$y = \frac{w}{p_y + p_x(\frac{a_x p_y}{a_y p_x})^{\frac{1}{1-\rho}}}.$$

5.2 Value function and envelope theorem

Consider an unconstrained maximization problem of a function of a single real variable x, where the objective function depends on a parameter $\alpha \in \mathbb{R}$.

$$\max_{x \in \mathbb{R}} f(x, \alpha).$$

Let $x(\alpha)$ be the solution to this problem. Consider the maximum value of the objective function that is achievable at the parameter $\hat{\alpha}$

$$V(\alpha) := f(x(\alpha), \alpha).$$

At the (unconstrained) optimum $x(\hat{\alpha})$, by the first-order condition:

$$\frac{\partial f(x(\hat{\alpha}), \hat{\alpha})}{\partial x} = 0.$$

Compute the change in V from a change in the parameter:

$$V'(\hat{\alpha}) = \frac{\partial f(x(\hat{\alpha}), \hat{\alpha})}{\partial x} x'(\hat{\alpha}) + \frac{\partial f(x(\hat{\alpha}), \hat{\alpha})}{\partial \alpha} = \frac{\partial f(x(\hat{\alpha}), \hat{\alpha})}{\partial \alpha},$$

since $\frac{\partial f(x(\hat{\alpha}), \hat{\alpha})}{\partial x} = 0$ by the first order condition for maximizing with respect to x.

This observation is called the envelope theorem. For $\boldsymbol{x} \in \mathbb{R}^n$, the message is exactly the same. The first order-condition is now:

$$\frac{\partial f(\boldsymbol{x}(\hat{\alpha}), \hat{\alpha})}{\partial x_i} = 0 \text{ for all } i \in \{1, ..., n\}.$$

Assuming the conditions for implicit function theorem, we have by chain rule:

$$V'(\hat{\alpha}) = \sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}(\hat{\alpha}), \hat{\alpha})}{\partial x_i} x_i'(\hat{\alpha}) + \frac{\partial f(\boldsymbol{x}(\hat{\alpha}), \hat{\alpha})}{\partial \alpha}.$$

Again, the first term vanishes by first-order condition and we are left with

$$V'(\hat{\alpha}) = \frac{\partial f(\boldsymbol{x}(\hat{\alpha}), \hat{\alpha})}{\partial \alpha}.$$

Suppose that we have an equality constrained parametric maximization problem for $x \in \mathbb{R}^n$:

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}, \alpha)$$
$$\text{subject to } g(\boldsymbol{x}, \alpha) = 0.$$

The value function is still defined as before:

$$V(\alpha) = f(\boldsymbol{x}(\alpha), \alpha).$$

Begin the analysis by forming the Lagrangean:

$$\mathcal{L}(\boldsymbol{x}, \mu; \alpha) = f(\boldsymbol{x}, \alpha) - \mu g(\boldsymbol{x}, \alpha).$$

The envelope theorem relates the derivative of the value function with respect to the parameter to the partial derivatives of the Lagrangean.

Theorem 1 (Envelope theorem)**.** In an optimization problem subject to an equality constraint, we have:

$$V'(\alpha) = \frac{\partial \mathcal{L}(\boldsymbol{x}, \mu; \alpha)}{\partial \alpha}.$$

The envelope theorem gives us a nice way of understanding the Lagrange multipliers in utility maximization problems. The Lagrangean for the UMP with a single binding equality constraint is:

$$\mathcal{L}(\boldsymbol{x}, \lambda) = u(\boldsymbol{x}) - \mu \left[\sum_{i=1}^{n} p_i x_i - w \right].$$

The maximum value function

$$v(\boldsymbol{p}, w) = \max u(\boldsymbol{x}) \text{ subject to } \boldsymbol{p} \cdot \boldsymbol{x} = w,$$

is called the indirect utility function. It computes the optimal utility level for all combinations of prices $\boldsymbol{p} \in \mathbb{R}^n_{++}$ and income $w > 0$.

Envelope theorem tells us that:

$$\frac{\partial v(\boldsymbol{p}, w)}{\partial w} = \mu.$$

The multiplier tells that if income goes up (down) by dw, then utility goes up (down) by λdw. Because of this the multiplier is called the shadow price of income.

6 Difference equations

Difference equations relate the future state $\boldsymbol{x}_{n+1} \in \mathbb{R}^m$ of a dynamical system to its current state $\boldsymbol{x}_n \in \mathbb{R}^m$:

$$\boldsymbol{x}_{n+1} = \boldsymbol{f}(\boldsymbol{x}_n) \text{ for all } n \in \mathbb{Z}_+.$$

The function $\boldsymbol{f}$ is called the *system equation*. The unknown in the problem is the sequence of vectors $\{\boldsymbol{x}_n\}_{n=0}^{\infty}$ satisfying the system equation. Think back to the linear exchange model of Lecture 2 for a concrete example.

The simplest form of difference equations are linear difference equations with constant coefficients. These can be written as:

$$\boldsymbol{x}_{n+1} = \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}_n,$$

where $\boldsymbol{b}_n$ is a given sequence. If $\boldsymbol{b}_n = 0$ for all n, we have a homogenous equation. We start with the simplest homogenous equations where $x_n \in \mathbb{R}$ and $\boldsymbol{A} = a \in \mathbb{R}$.

Solving the homogenous equation is very easy. If $x_{n+1} = ax_n$ for all n, then $x_{n+k} = a^k x_n$. Hence any sequence of the form $x_n = ca^n$ solves the difference equation. If we are given the initial value x_0, the solution is $x_n = x_0 a^n$. In other words, the initial value pins down the coefficient c of the general solution.

Consider next an inhomogenous equation,

$$x_{n+1} = ax_n + b,$$

where $b_n = b$ for all n. Clearly the constant solution $x_n = \frac{b}{1-a}$ for all n solves the equation. I claim that also $x_n = ca^n + \frac{b}{1-a}$ solves the equation. But this follows immediately from the fact that $ca^{n+1} = aca^n$.

This principle holds more generally. If you have a particular solution x_n^P to the inhomogenous equation and the general solution of the homogenous equation x_n^H, then the general solution to the problem is $x_n^P + x_n^H$. This is called the principle of superposition and it arises from the linearity of the equations in x_{n+1}, x_n. It is valid also for the case with $\boldsymbol{x}_n \in \mathbb{R}^m$.

Consider next linear systems with constant coefficients. Let $\boldsymbol{x}_n \in \mathbb{R}^m$ for all n and let A be an $m \times m$ matrix of real numbers. A linear homogenous system is then given by:

$$\boldsymbol{x}_{n+1} = \boldsymbol{A}\boldsymbol{x}_n.$$

As before, we can 'solve' this by repeated substitution to get

$$\boldsymbol{x}_{n+k} = \boldsymbol{A}^k \boldsymbol{x}_n.$$

Hence I could write the general solution as $\boldsymbol{x}_n = \boldsymbol{A}^n \boldsymbol{c}$ for some vector $\boldsymbol{c} = (c_1, ..., c_k)$. I do not consider this a real solution since it is almost impossible to see what $\boldsymbol{A}^n$ is except in some very special cases. If A is a diagonal matrix with diagonal elements $a_1, ..., a_k$, then the solution becomes:

$$x_{i,n} = c_i a_i^n \text{ for } i \in \{1, ..., m\}.$$

Here we have essentially independent variables and the difference equation for each can be solved separately.

To deal with the general case, we want to change the basis in $\mathbb{R}^m$ so that $\boldsymbol{A}$ is diagonal in that basis. This involves the eigenvectors and eigenvalues of $\boldsymbol{A}$. You can visualize the effect of matrix multiplication on vectors as consisting of two operations: a rotation and a stretching or shrinking.

Eigenvectors of $\boldsymbol{A}$ are those vectors that are not rotated, i.e. if $\boldsymbol{x} \neq 0$ is an eigenvector of $\boldsymbol{A}$, then for some $\lambda \in \mathbb{R}$,

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}.$$

We may write this more compactly as

$$(\boldsymbol{A} - \lambda I)\boldsymbol{x} = 0,$$

where I is the $m \times m$ identity matrix. But from basic linear algebra, we know that a homogenous linear equation can have a non-zero solution only if the matrix does not have full rank, i.e. if $\det(\boldsymbol{A} - \lambda I) = 0$. The values of λ for which this determinant is zero are called the eigenvalues of A.

The determinant of $(\boldsymbol{A} - \lambda I)$ is called the characteristic polynomial of $\boldsymbol{A}$ so the eigenvalues are the roots of the characteristic polynomial. If $\boldsymbol{A}$ has n distinct eigenvelues $\lambda_1, ..., \lambda_m$, then it has also n linearly independent eigenvectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_m$ so that

$$\boldsymbol{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i.$$

In this case, we can express any $\boldsymbol{x} \in \mathbb{R}^n$ given in the usual coordinate system in the new coordinate system spanned by the eigenvectors by simple matrix multiplication. Let $\boldsymbol{P} = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \dots \boldsymbol{v}_n]$ be the matrix formed by the eigenvectors. Then for any vector $\boldsymbol{y}$ expressed in the coordinate system of the eigenvectors, we can translate it to the standard system by $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}$. Similarly any $\boldsymbol{x}$ in the standard system is $\boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{x}$ in the system of the eigenvectors.

$$\boldsymbol{y}_{n+1} = \boldsymbol{P}^{-1}\boldsymbol{x}_{n+1} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{x}_n = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}\boldsymbol{y}_n.$$

Now we want to show that $\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. But this is the same claim (as can be seen by pre-multiplying by $\boldsymbol{P}$) as:

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda}.$$

But this follows immediately from the fact that P consists of the eigenvectors of $\boldsymbol{A}$.

Hence we have: $\boldsymbol{y}_n = (y_{1,n}, ..., y_{k,n}) = (c_1\lambda_1^n, ..., c_k\lambda_k^n)$. Since $\boldsymbol{x}_n = \boldsymbol{P}\boldsymbol{y}_n$, we have the general solution:

$$\boldsymbol{x}_n = c_1\lambda_1^n\boldsymbol{v}_1 + ... + c_k\lambda_k^n\boldsymbol{v}_k.$$

The values of the c_i are determined from the initial condition for $\boldsymbol{x}_0$.

Note that $\boldsymbol{A}^k = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$. Therefore we could have also concluded that

$$\boldsymbol{x}_n = \boldsymbol{P}\boldsymbol{\Lambda}^k\boldsymbol{P}^{-1}\boldsymbol{x}_0.$$

The two methods give the same results since $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{x}_0$ or $\boldsymbol{c} = \boldsymbol{P}^{-1}\boldsymbol{x}_0$.

The general solution to a non-homogenous equation is the sum of a particular solution and the general solution of the homogenous problem.

Higher orders difference equations such as

$$x_n = a_1x_{n-1} + a_2x_{n-2},$$

can be transformed into a system of first order difference equations.

The analysis of linear equations with non-constant coefficients such as

$$x_{n+1} = a_nx_n,$$

for a nonconstant sequence a_n of coefficients is a lot more tricky, but problems of this type arise frequently in e.g. combinatorics (where n has no connection to time but refers to the size of the problem).

The analysis of non-linear difference equations such as

$$x_{n+1} = \mu x_n(1 - x_n),$$

is often very hard and only qualitative characterizations relying on *phase diagrams* are possible.