Programming and Numerical Methods for Economics: Math Review

Albert Rodriguez Sala and Jacob Adenbaum

University of Edinburgh

Spring 2024

Preliminaries

Function Notation

- We often want a compact way to define a function, which clearly specifies its domain and co-domain.
- If we have a function f which takes inputs from a set A and returns outputs in a set B, we can write this compactly as

$$f : A \to B$$

- Most commonly in this course, you'll see this with variants of Euclidian space:

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

  denotes a function that takes an n dimensional vector of real numbers (in $\mathbb{R}^n$) and returns a real number

Function properties

- We say that a function $f : A \to B$ is **injective** (one to one) if for any $x, y \in A$, $f(x) = f(y)$ implies $x = y$.

  In other words, there are no distinct points in A which get mapped to the same point in B

- A function $f : A \to B$ is **surjective** (onto) if for every point $y \in B$, there exists a point $x \in A$ such that $f(x) = y$

  In other words, every point in B is "hit" by f

- A function $f : A \to B$ is a **bijection** if it is both injective and surjective
- An easy way to show that two sets are the same cardinality ("size") is to show that there exists a bijection between them.

Exercise

Prove that this is true for finite sets. I.e., show that if A and B are sets with a finite number of elements, and there exists a bijection $f : A \to B$, then $|A| = |B|$.

Linear Algebra

Scalars, vectors, and matrices

- A **scalar** a is a single number.
- A **vector** $\mathbf{a}$ is an element of Euclidean k space, written as $\mathbf{a} \in \mathbb{R}^k$.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

Scalars, vectors, and matrices

- A **matrix** is a $k \times r$ rectangular array of numbers, written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & & \ddots & \\ ak1 & a_{k2} & \dots & a_{kr} \end{bmatrix}$$

The transpose

The **transpose** of a matrix $A(k \times r)$ denoted A', A^r is obtained by flipping the matrix on its diagonal. If a matrix A is $k \times r$ then its transpose A' is $r \times k$. Properties of the transpose operator:

$$(A+B)' = A' + B'$$
$$(cA)' = cA'$$
$$(AB)' = B'A'$$

Squared, Symmetric, Diagonal and Identity matrices

- A matrix $A(k \times r)$ is **square** if $k = r$.
- A square matrix is **symmetric** if $A = A'$.
- A matrix is **diagonal** if the off-diagonal elements are all zero.
- An important diagonal matrix is the **identity matrix, or unit matrix**, $I_k(k \times k)$ which has ones on the diagonal. Let A be a $k \times r$ matrix, then:

$$AI_r = A$$
$$I_k A = A$$

Matrix Addition

if matrices A, B are of the same order:

$$\underset{(k \times r)}{A} + \underset{(k \times r)}{B} = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1r} + b_{1r} \\ \vdots & \ddots & \\ ak1 + b_{k1} & \dots & akr + b_{kr} \end{bmatrix}$$

Properties:

$$A + B = B + A (\text{Commutative law})$$
$$A + (B + C) = (A + B) + C (\text{associative law})$$

Matrix multiplication

Scalar multiplication: Let $c \in \mathbf{R}$ be a scalar and, A a matrix $(k \times r)$. The scalar product is defined st:

$$cA = Ac = (a_{ij}c) \qquad \forall ij \in k \times r.$$

Inner product (vectors): If **a** and **b** are both $k \times 1$, then their inner product is

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \ldots + a_k b_k = \sum_{j=1}^{k} a_j b_j$$

Note that $a'b = b'a$. We say that two vectors **a** and **b** are **orthogonal** if **a'b** $= 0$.

Matrix multiplication

Matrix product: If the number of columns in A is equal to the number of rows in B, then A and B are **conformable**. In this event the matrix product AB is defined as:

$$\underset{(k\times r)}{A}\underset{(r\times s)}{B} = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_k \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \dots & \mathbf{b}_k \end{bmatrix} = \begin{bmatrix} \mathbf{a_1}'\mathbf{b_1} & \dots & \mathbf{a_1}'\mathbf{b_s} \\ \vdots & \ddots & \\ \mathbf{a_k}'\mathbf{b_k} & \dots & \mathbf{a_k}'\mathbf{b_s} \end{bmatrix}$$

Matrix multiplication

Properties:

- Matrix multiplication is not commutative in general:

$$AB \neq BA$$

- Matrix multiplication is associative and distributive:

$$A(BC) = (AB)C$$
$$A(B + C) = AB + AC$$

the $(k \times r)$ Matrix A, $r \leq k$, is called orthonormal if $A'A = I_r$:

The trace and the rank

The **trace** of a square matrix A is the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{k} a_{ii}$$

- The **rank** of the matrix $A(k \times r)$ with $r \leq k$ is the number of linearly independent columns $\mathbf{a}_j$ and is written as rank(A).
- We say A has full rank if rank(A)=r.
- A square matrix A is said to be **nonsingular** if it has full rank.
- This means that there is no vector $\mathbf{c}$ $(k \times 1)$, with $\mathbf{c} \neq 0$ such that $A\mathbf{c} = 0$.

Nonsingular matrices and the inverse

If a square matrix ($k \times k$) A is nonsingular then there exists a unique matrix ($k \times k$) A^{-1} called the **inverse** of A which satisfies:

$$AA^{-1} = A^{-1}A = I_k$$

Properties for non-singular matrices A and B include:

- $AA^{-1} = A^{-1}A = I_k$
- $(A^{-1})^{-1} = A$
- $(A^{-1})' = (A')^{-1}$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A + B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}$
- $det(A^{-1}) = det(A)^{-1}$

Also if A is an **orthonormal** matrix then $A^{-1} = A'$.

The determinant

The **determinant** is a measure of the volume of a matrix $(k \times k)$. Let A be a (2×2) matrix, then its determinant is:

$$det(A) = |A| = a_{11}a_{22} - a_{12}a_{21}$$

Properties of the determinant:

- $det(A) = det(A')$
- $det(cA) = c^k det(A)$
- $det(AB) = det(A)det(B)$
- $det(A^{-1} = det(A)^{-1}$
- $det(A) \neq 0 \iff A$ is non-singular.

The characteristic equation and the eigenvalues of a matrix

The **characteristic equation** of a $(k \times k)$ matrix is:

$$det(A - \lambda I_k) = 0$$

It has k roots not necessarily distinct and real or complex. the **eigenvalues** or **latent roots** or **characteristic roots**; λ, of A are the roots of the characteristic function.

The eigenvector

If λ_i is an eigenvalue of A, then $A - \lambda_i I_k$ is singular so that there exists a non-zero vector $\mathbf{h_i}$ such that

$$(A - \lambda_i I_k)\mathbf{h_i} = 0$$

the vector h_i is called a **latent vector** or **characteristic vector** or **eigenvector** of A corresponding to λ_i. Let Λ be a diagonal matrix with the eigenvalues in the diagonal and let $H = [\mathbf{h_1} \ldots \mathbf{h_k}]$. Some useful properties are:

Properties of eigenvalues and eigenvectors

- $det(A) = \prod_{i=1}^{k} \lambda_i$
- $tr(A) = \sum_{i=1}^{k} \lambda_i$
- A is non-singular $\iff \lambda_i \neq 0 \forall i = 1 \ldots k$
- if A has distinct eigenvalues, then there exists a nonsingular matrix P such that $A = P^{-1} \Lambda P$ and $PAP^{-1} = \Lambda$.
- If A is symmetric, then $A = H \Lambda H'$, which is called the **spectral decomposition** of A, and $H'AH = \Lambda$. In that case the eigenvalues are all real.
- the eigenvalues of A^{-1} are $\lambda_1^{-1}, \ldots, \lambda_k^{-1}$,
- The matrix H has the orthonormal properties: $H'H = I$, $HH' = I$, or alternatively, $H^{-1} = H'$ and $(H')^{-1} = H$

Note that if an eigenvalue has multiplicity 2, it can have 1 or 2 latent vectors linearly independent. If it has multiplicity 3, it can gave 1, 2, 3 latent vectors linearly independent.

Example

Exercise

Compute the eigenvalues and eigenvectors of the following matrices. Indicate clearly the multiplicity of each vector.

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Positive definiteness

A symmetric matrix is **positive (semi-)definite**, $A \underset{(\geq)}{>} 0$, if and only if $\forall \mathbf{c} \neq 0$:

$$\mathbf{c}' A \mathbf{c} \underset{(\geq)}{>} 0$$

A matrix A is larger than B if $\mathbf{c}'(A - B)\mathbf{c} > 0$.

Properties:

- $A > 0 \iff \lambda_i \geq 0 \forall i$
- $A > 0 \longrightarrow$ A is non-singular and $A^{-1} \exists$ and $A^{-1} > 0$
- $A = G'G \longrightarrow A \geq 0$ and if G has full rank then, $A > 0$
- $A > 0 \longrightarrow \exists B$ st$A = BB'$ where B is **the matrix square root**.

Differentiation

The derivative

The **derivative** of a function $f : \mathbb{R} \to \mathbb{R}$ at a point $x \in \mathbb{R}$ is a real number that describes the instantaneous change of the value f as x changes:

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

Local minimum/maximum

A point $x^\star$ is said to be **local maximum** or **maximizer** if there exists $\delta > 0$ such that $f(x^\star) \geq f(x)$ for all $x \in N_\delta(x^*)$, where $N_\delta(x^\star)$ denotes the ball of radius δ about $x^\star$.

I.e, $N_\delta(x^\star) = (x^\star - \delta, x^\star + \delta)$.

A **local minimum** or **minimizer** is defined in an analogous way: $x^\star$ is a minimizer if there exists some $\delta > 0$ where $f(x^\star) \leq f(x)$ for all $x \in N_\delta(x^\star)$.

Proposition: Maximum/minimum in open sets

Proposition: Let O be an open subset of $\mathbb{R}$, and let $f : O \to \mathbb{R}$ a function with derivative at x, and let x^* be a local maximum (minimum). Then, $f'(x^*) = 0$

Proposition: Maximum/minimum in open sets

Proof.

Suppose x^* is a local maximum (for a local minimum is analogous). Then, $f(x^* + h) - f(x^*) \forall h > 0$, small enough, $(|h| < \delta$. Taking limits as h goes fo 0 from the right, it must be that

$$\lim_{h \to 0^+} \frac{f(x^* + h) - f(x^*)}{h} \leq 0 \quad \forall h \in (0, \delta)$$

Moreover, taking limits from the left

$$\lim_{h \to 0^-} \frac{f(x^* + h) - f(x^*)}{h} \geq 0 \quad \forall h \in (-\delta, 0)$$

Thus, since $f(x)$ has a derivative at x, then, it must be that:

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

$\square$

Properties of derivatives

- If $f(x) = c$ for all x. Then, $f'(x) = 0$
- If $f(x) = x$ then, $f'(x) = 1$
- $[f(x) \underset{(-)}{+} g(x)]' = f'(x) \underset{(-)}{+} g'(x)$ (Sum rule)
- $[kf(x)]' = kf'(x)$
- $[f(x) \cdot g(x)]' = f'(x) \cdot g(x) + g'(x) \cdot f(x)$ (Product rule)
- $\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)g(x) - g'(x)f(x)}{[g(x)]^2}$ (quotient rule)
- $(f(g(x))' = f'(g(x)) \cdot g'(x)$ Equivalently, $\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$ (Chain rule)
- $\left[\frac{1}{f(x)}\right]' = -\frac{f'(x)}{[]f(x)]^2}$

Mean value theorem

Theorem 1

(Mean value theorem) *Let* $f : [a, b] \to \mathbb{R}$ *be a continuous function and differentiable on* (a, b)*. Then there exists a point* $c \in [a, b]$ *such that* $f'(c) = \frac{f(b)-f(a)}{b-a}$

Taylor Polynomials

When we approximate a function by its tangent line, we suffer an error. We can reduce such approximation error by approximating the function by a more sophisticated object than the tangent line: a polynomial of order higher than one.
If $f(a), f'(a), \dots f^{(n)}(a)$ all exist, the **nth Taylor polynomial** of the function f at the point a is

$$T_n(x) = \sum_{k=0}^{n} \frac{f^k(a)}{k!}(x-a)^k$$

$$= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \dots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Taylor series

If the function f has derivatives of all orders a, the **Taylor series** is a series expansion of a function about a point a given by

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} T_n(x) \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \\ &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots \end{aligned}$$

Then, we say we do a **Taylor approximation of order nth** when we approximate $f(x)$ by computing a Taylor expansion till its *nth* polynomial.

Taylor's Theorem

Theorem 2

Let $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at the point $a \in \mathbb{R}$. Then

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + o((x-a)^k)$$

where $o((x-a)^k)$ denotes a function which converges to zero faster than $(x-a)^k$.

- This tells us that in a local neighborhood of the point a, a Taylor series is a "good" approximation of the function f.
- Adding more terms will make it "better" the sense that the error term will converge to zero faster
- Note: we can do Taylor approximations for functions from $\mathbb{R}^n$ to $\mathbb{R}$ as well, we just have to be careful about the proper matrix notation.

Derivatives of real multivariable functions

Now suppose the case of of a real value multivariate function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Then,

- The **directional derivative** of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at point x in the direction of $\mathbf{u}$ is defined by

$$D_u f(x) = \lim_{\alpha \rightarrow 0} \frac{f(x + \alpha \mathbf{u}) - f(x)}{\alpha}$$

  where $\mathbf{u}$ is a vector $\mathbf{u} = (u_1, u_2, \ldots, u_n)$, that gives the direction (set of coordinates) to where to move.

Derivatives of real multivariable functions: the partial derivative

- The **Partial derivative** of f with respect to the $i-th$ argument, x_i, is defined by

$$\frac{\partial f(x)}{\partial x_i} = f_i(x) = D_{x_i} f(x) = Df(x, \mathbf{u_i}) \quad \text{with } \mathbf{u_i} = (\underset{1}{0}, \ldots, 0, \underset{i}{1}, 0, \ldots, \underset{n}{0})$$

Derivatives of real multivariable functions: The Jacobian

- The **Jacobian** of f is defined as the vector $(1 \times n)$ formed by all the partial derivatives of the function
$$\mathcal{J}_x(f) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Derivatives of real multivariable functions: The Hessian

- the **Hessian** of f is defined as the matrix $(n \times n)$ formed by all the second-order partial derivatives of the function

$$\mathcal{H}_x(f) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \\ \frac{\partial f}{\partial x_m \partial x_1} & \cdots & \frac{\partial f}{\partial^2 x_m} \end{bmatrix}$$

The Hessian allows us to determine the convexity/concavity of the function f

1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave if $\mathcal{H}_x(f)$ is negative semidefinite.
2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathcal{H}_x(f)$ is positive semidefinite.
3. A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly concave if $\mathcal{H}_x(f)$ is negative definite.
4. A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if $\mathcal{H}_x(f)$ is positive definite.

Evaluating concavity/convexity with the Hessian

Then, the following equivalences are true

$$\begin{aligned}
f \text{ concave} \iff \text{Hessian semidefinite negative} &\iff \text{its latent roots are } \lambda_i \leq 0 \\
&\longleftrightarrow \text{principal minors order } k\text{, have sign } -1^k \\
f \text{ convex} \iff \text{Hessian semidefinite positive} &\iff \text{its latent roots are } \lambda_i \geq 0 \\
&\longleftrightarrow \text{principal minors are} \geq 0
\end{aligned}$$

Hessian matrix

A Hessian matrix will be symmetric if the second partial derivatives are continuous (Schwarz's theorem).

Differentiability

On the one hand, the derivative of a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ at a point $x \in \mathbb{R}$ is defined as a real number that describes the instantaneous change of the value of f as x changes. On the other hand, the notion of derivative is tightly linked to the line that best approximates f near x.

Let f be a function from a vector space $V \subset \mathbb{R}^n$ to another vector space $W \subset \mathbb{R}$. We say that f is a linear map if for any two vectors $\mathbf{x}, \mathbf{y} \in V$ and any scalar $\alpha \in \mathbb{R}$, two conditions are satisfied:

1. $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ (additivity)
2. $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$ (homogeneity of degree 1)

Therefore, we can only consider the map L s.t. $L(t) = f(x) - f'(x)(t - x)$

Differentiability

A function f is said to be **differentiable** at x is there exists such a linear map. For functions on the reals this is equivalent to the existence of a derivative: the linear map will be the one intersecting $f(x)$ with slope $f'(x)$. For functions defined on other spaces the definition of differentiability is more general.

A function $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$ is said **differentiable** at $x \in O$ if there exists a matrix $\mathcal{J}_x(n \times 1)$, the Jacobian, such that

$$\lim_{\|h\| \to 0} \frac{|f(x+h) - f(x) - J_x h|}{\|h\|} = 0$$

Differentiability

Notice the following:

1. The derivative $Df : O \to \mathbb{R}^{\ltimes}$ assigns a Jacobian to each x.
2. the differential $df_x : \mathbb{R}^n \to \mathbb{R}$ is a mapping that assigns to each x the linear operator $df_x(h) = \mathcal{J}_x h$

Differentiable and continuously differentiable functions

Theorem 3

Consider the function $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$. if f is differentiable at $x \in O$ then f is continuous at x. Suppose f has a well-defined partial derivatives and those are continuous. Then f is differentiable.

A function $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$ is said **continuously differentiable** if it is differentiable and Df is a continuous function. Therefore, a function is continuous differentiable if and only if the partial derivatives of the function f exists and are continuous.

Inverse function theorem

Theorem 4

Inverse Function Theorem Let $f : O \subseteq \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and let $x^0 \in O$. If the determinant of the Jacobian of f is not equal to zero at x^0, then there exists an open neighborhood of x^0, U, such that:

1. *f is a one-to-one correspondance in U and f^{-1} is well defined.*
2. *$V = f(U)$ is an open set containing $f(x^0)$.*
3. *f^{-1} is continuously differentiable with $D[f^{-1}(x^0)] = [Df(x^0)]^{-1}$*

This theorem is very useful when we cannot invert $f(x, y)$. We can find the Jacobian of the inverse, by taking the jacobian and then invert it.

Derivatives of matrices

Let $\mathbf{x} = [x_1 \dots x_k]$ be a $(k \times 1)$ vector and $g(\mathbf{x}) : \mathbf{R}^k \to \mathbf{R}$ a vector function. Then, the **vector derivative** is defined as:

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial g(\mathbf{x})}{\partial x_k} \end{bmatrix} \qquad \text{and} \qquad \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g(\mathbf{x})}{\partial x_k} \end{bmatrix}$$

Properties

Some properties:

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}'\mathbf{a}}{\partial \mathbf{x}} = a$$

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = A$$

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (A + A')\mathbf{x}$$

$$\frac{\partial^2 \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}\partial \mathbf{x}'} = A + A'$$

Example: Deriving the OLS estimator in matrix notation

Exercise

Show that the OLS estimator in matrix notation is $\hat{\beta} = (X'X)^{-1}X'y$

Solution Take the linear model $y = X\beta + u$ where y is a $(n \times 1)$ vector of observations; X is a $(n \times k)$ matrix of regressors; β is a $(k \times 1)$ vector of coefficients; and u is a $(n \times 1)$ vector of errors. Then, the OLS estimator $\hat{\beta}$ is the vector of coefficients that minimize squared residuals:

$$\underset{\beta}{min}\{u'u\} \qquad \{u'u\} = (y - X\beta)'(y - X\beta)$$

$$= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta$$

$$= y'y - 2\beta'X'y + \beta'X'X\beta$$

Example: Deriving the OLS estimator in matrix notation

note that the second and third term in the second line are scalars [1] and the fact that the transpose of a scalar is the scalar i.e. $y'X\beta = (y'X\beta)' = \beta'X'y$.

The FOC is:

$$\frac{\partial u'u}{\partial \beta} = \frac{\partial y'y - 2\beta'X'y + \beta'X'X\beta}{\partial \beta} = 0$$

[1] the size of the second term is $(1 \times n)(n \times k)(k \times 1) = (1 \times 1)$ while the size of the third is $(1 \times k)(k \times n)(n \times 1) = (1 \times 1)$

Example: Deriving the OLS estimator in matrix notation

Using matrix calculus properties (1) and (3):

$$\frac{\partial 2\beta' X' y}{\partial \beta} = 2X'y \qquad \frac{\partial \beta' X' X \beta}{\partial \beta} = 2X'X\beta$$

so that the FOC becomes:

$$\frac{\partial u'u}{\partial \beta} = -2X'y + 2X'X\beta = 0$$

Isolating β:

$$(X'X)\beta = X'y$$
$$(X'X)^{-1}(X'X)\beta = (X'X)^{-1}X'y$$

Hence, The OLS estimator is:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Static optimization

3 cases of static optimization

1. Convex constraint set (unconstraint optimization).
2. Lagrange problem.
3. Kuhn-Tucker problem.

Convex constraint set

Let S be a convex set in $\mathbb{R}^n, f : S \to \mathbb{R}$ and consider the problem

$$\max_{x \in S} \; f(x)$$

Suppose S is an open set. Then, a necessary condition for the optimum is to be a **stationary point**:

$$Df(x^*) = 0$$

That is all partial derivatives must be zero. In the more general case, this condition is not necessary,for example, if S was closed; nor sufficient, x^* could be a saddle point if f is not differenciable.

Convex constraint set: Sufficient conditions for global optimums

Sufficient conditions for global optimums: Suppose $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$ is a differentiable function defined on a convex set S and let x^* be a starionary point int the interior of S. Then:

1. If f is strictly concave in $S \longrightarrow x^*$ is a global maximum.
2. If f is strictly convex in $S \longrightarrow x^*$ is a global minimum.
3. The set of optimizers $\sigma(\theta) :=$ argmax/argmin$\{f(x) : x \in X\}$ is either empty or convex. That is, there cannot be multiple isolated points as maximizers (minimizers).

convexity and concavity of f with the Hessian

Thus, to determine whether a stationary point x^* is a maximum or a minimum we need to find the convexity and concavity of the function f.

$$f \text{ is concave} \iff \text{Hessian at } x^* \text{ semidefinite negative} \iff \text{its latent roots are } \lambda_i \leq 0$$
$$f \text{ is convex} \iff \text{Hessian at } x^* \text{ semidefinite positive} \iff \text{its latent roots are } \lambda_i \geq 0$$

Lagrange problem

Consider the problem of maximize or minimize a function $f : S \subseteq \mathbb{R}^n$ knowing that the variables must satisfy a system of J equations $g_j(x) = 0$. To do so, we use the **Lagrange method** defined by the **lagrange function**

$$L(x, \lambda) = f(x) - \sum_{j=1}^{n} \lambda_j g_j(x)$$

Where note that we impose a penalization λ_j from deviating from each constraint g_j.

FOC Lagrange problem

For a given λ_j we consider the following *first order conditions*:

$$\frac{\partial L}{\partial x} = \frac{\partial f}{\partial x} - \sum_{j=1}^{n} \lambda_j \frac{\partial g_j}{\partial x} = 0$$

$$\frac{\partial L}{\partial \lambda_j} = g_j(x) = 0 \ \ \forall j = 1, \ldots, J$$

Any stationary point x^* of the Lagrangian function $L(x, \lambda)$ is a candidate to global maximum or minimum if the gradient vector of the functions g_i in the stationary point are linearly independent.

Kuhn-Tucker problem

Consider the problem of maximize (minimize) a function $f : S \subseteq \mathbb{R}^n$ subject to J constraints $g_j(x) \leq 0$, where both f and g are continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$.

$$\max_x f(x)$$
$$\text{s.t. } g(x) \leq 0$$

The constraints can be

1. **binding** or **active** at a feasible point x^0 if they hold with equality, in which case we are in the Lagrangian case. That is there exists $\lambda_j \neq 0$
2. **Inactive** otherwise, in which case they will not have effects on the local properties of the solution. That is $\forall \lambda_j, \lambda_j = 0$

Kuhn-Tucker conditions

To solve this problem, we define the function

$$L(x, \lambda) = f(x) - \sum_{j=1}^{n} \lambda_j g_j(x)$$

Then, the **Kuhn-Tucker conditions** for the problem are

1. $D_x L(x, \lambda) = Df(x) - \sum_{j=1}^{n} \lambda_j Dg_j(x) = 0$
2. $\lambda_j \underset{(\leq)}{\geq} 0, g_j(x) \leq 0$ and $\lambda_j [g_j(x)] = 0$ for $j = 1, \ldots, J$

   Where

   - If x^* is a maximum $\Longrightarrow$ all $\lambda_i \geq 0$
   - If x^* is a minimum $\Longrightarrow$ all $\lambda_i \leq 0$

Example: Kuhn-Tucker problem

Example (Optimization in a non-closed region). Find the maximum and minimum of the function

$$f(x,y) = \frac{x}{2} - y \quad s.t. \begin{cases} x + e^{-x} \leq y \\ x \geq 0 \end{cases}$$

Solution: First we define the Lagrangian function

$$L(x,y,\lambda) = x/2 - y - \lambda_1(x + e^{-x} - y) - \lambda_2(-x)$$

the first order conditions are

$$\begin{aligned} D_x L(x,y,\lambda) = 1/2 - \lambda_1(1 - e^{-x}) + \lambda_2 = 0 \\ D_y L(x, y\lambda) = -1 + \lambda_1 = 0 \end{aligned}$$

Example: Kuhn-Tucker problem

From which we get $\lambda_1 = 1, e^{-x} + 0 = 1/2$. Then, we have the following 4 cases to study:

- 1. $[\lambda_1 = 0 \text{ and } \lambda_2 \neq 0]$
  we do not have any candidate.
- 2. $[\lambda_1 \neq 0 \text{ and } \lambda_2 = 0]$
  We have a candidate given by the conditions $e^{-x} + \lambda_2 = 1/2$ and $x + e^{-x} = y$.
  So that,
  $$e^{-x} = \frac{1}{2} \qquad x = ln\frac{1}{2}^{-1} \qquad x = ln2$$
  and
  $$y = ln2 + e^{-ln2} = ln2 + 1/2$$
  so that we find the point $(ln2, 1/2 + ln2)$ which is a candidate to a maximum $(\lambda = 1 \geq 0)$.

Example: Kuhn-Tucker problem

- 3. [$\lambda_1 = 0$ and $\lambda_2 = 0$]
  We do not have any candidate since $\lambda_1 = 1$
- 4. [$\lambda \neq 0$ and $\lambda \neq= 0$]
  In this case we have that $x = 0$ and $y = x + e^{-x} = y$ so that we find the point $(0, 1)$ which is neither a maximum or a minimum because $\lambda_1 = 1 \geq 0$ and $\lambda_2 = -1/2$

Example: Kuhn-Tucker problem

We found $(ln2, 1/2 + ln2)$ as a candidate to maximum with $\lambda_1 = 1, \lambda_2 = 0$, so that the Lagrangian function becomes $L(x, y, \lambda) = x/2 - y - (x + e^{-x} - y)$ and its Hessian is:

$$\mathcal{H}(f) = \begin{bmatrix} -e^{-x} & 0 \\ 0 & 0 \end{bmatrix}$$

And $det(\mathcal{H}(f)|_{(ln2,1/2+ln2)} = 0$ so that we cannot determine the convexity/concavity with the Hessian. However, we can do so computing the eigenvalues of $\mathcal{H}(f)$. Solving the characteristic equation we find $\lambda_1 = -1/2$ and $\lambda_2 = 0$. Because $\lambda \leq 0$, the Hessian is semidefinite negative and thus f is concave. The point **(ln2, 1/2 +ln2) is a maximum**. Also note that in this exercise there is no minimum (the region is not bounded!) :

$$\lim_{x=0, y\to\infty} f(x, y) = \lim_{x=0, y\to\infty} \frac{0}{2} - y = -\infty$$