Lecture 4: Differentiation, IFT, Unconstrained Optimization

Originally written by Mauricio Caceres Bravo
Revised by Cole Davis
Webpage: cj-davis99.github.io

August 16, 2024

Contents

Notation

- $\forall$ translates to "for all"
- $\exists$ translates to "there exists"
- $\mathbb{N} = \{1, 2, \ldots\}$ is the set of natural numbers
- $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ is the set of integers
- $\mathbb{Q} = \{p/q : p \in \mathbb{Z} \text{ and } q \in \mathbb{Z}\backslash\{0\}\}$ is the set of rational numbers
- $\mathbb{R}$ is the set of real numbers
- If S is a set and $n \in \mathbb{N}$, then S^n is the n^{th} order Cartesian product of S. E.g., $S^2 = S \times S$
- For any $\varepsilon > 0$, $B_\varepsilon(x)$ is the Euclidean ball around x with radius ε
- Unless otherwise specified, $d(x, y)$ is a metric on the contextual set x, y belong to
- The origin is always denoted as 0 regardless of the dimension of the space considered
- If v is a vector, then both v^T and v' can represent the vector transpose. Preference is usually given to the v^T notation.

A small warning to the reader: moving forward, we will state many results without proof. We do this for a few reasons:

(a) The results themselves are often much more useful than the proof of the results (e.g., we often just take the differentiability of a function for granted in the first-year courses).

(b) The syllabus lists some textbooks that prove the results mentioned in this section if you are interested.

(c) It is expected that the majority of incoming graduate students have quite a bit of exposure to this material already through calculus, linear algebra, differential equation, and analysis courses.

Furthermore, we will be utilizing some results from linear algebra as it's rather difficult to rigorously discuss high-dimensional functions without linear algebra tools.

1 Differentiation

1.1 Single-Variable Calculus

Definition 1. Let $D \subseteq \mathbb{R}$, let $f : D \to \mathbb{R}$ be a function, and let $a \in D$ be a limit point[1] of D. We say that f is ***differentiable*** at a if

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

converges to some real number L. We write $f'(a) = \frac{df}{dx}(a) = L$. If f is differentiable $\forall a \in D$, then we say f is differentiable on D.

If D is an interval in $\mathbb{R}$ (*i.e.*, $D = (a, b), (a, b], [a, b)$, or $[a, b]$ for some $a, b \in \mathbb{R}$ with $a < b$), then every point in D is a limit point of D, and we don't need to worry about specifying that the point in which we want to evaluate a derivative is a limit point.

Intuitively, differentiation yields the slope of a function at a point:

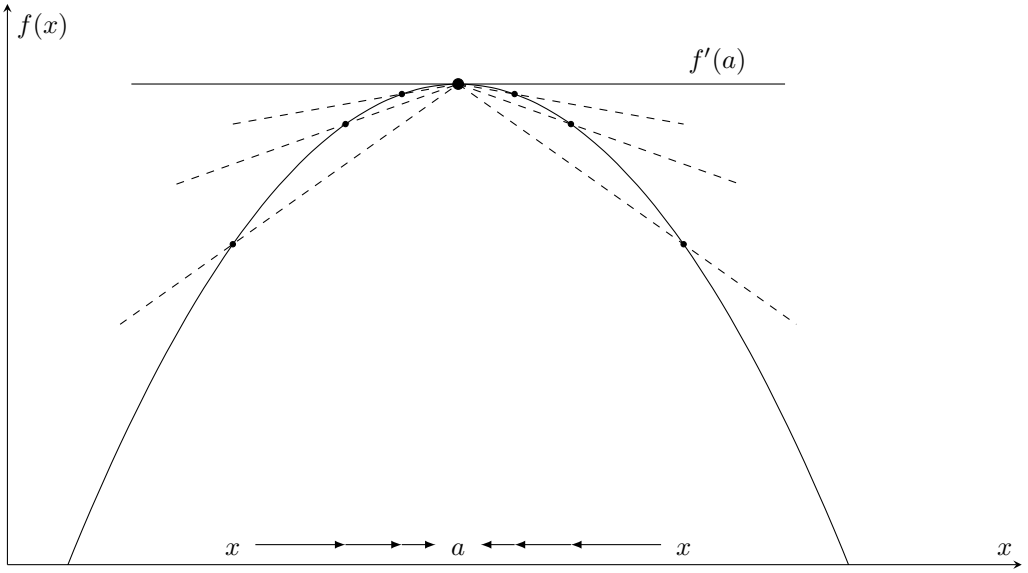


Figure 1: Graphical representation of a derivative

Theorem 1. *If $f : D \to \mathbb{R}$ is differentiable at $a \in D$ then f is also continuous at a.*

Proof. We want to show that $\forall \varepsilon > 0,\ \exists \delta > 0$ s.t.

$$|x - a| < \delta \implies |f(x) - f(a)| < \varepsilon$$

Let $\varepsilon > 0$. Since f is differentiable at a, we know we can find some $\widetilde{\delta} > 0$ s.t.

$$0 < |x - a| < \widetilde{\delta} \implies \left| \frac{f(x) - f(a)}{x - a} - L \right| < \varepsilon$$

[1] Recall that x is a ***limit point*** of a set D if $\forall \varepsilon > 0$, $(B_\varepsilon(x) \setminus \{x\}) \cap D \neq \varnothing$.

where $L = f'(a)$. That is, we know the derivative exists and it is equal to L. Thus, if $|x-a| < \widetilde{\delta}$, then

$$\frac{|f(x)-f(a)|}{\widetilde{\delta}} < \frac{|f(x)-f(a)|}{|x-a|} = \left|\frac{f(x)-f(a)}{x-a} - L + L\right| \leq \left|\frac{f(x)-f(a)}{x-a} - L\right| + |L| < \varepsilon + |L|$$

Hence whenever $|x-a| < \widetilde{\delta}$ we get

$$|f(x)-f(a)| < (\varepsilon + |L|)\,\widetilde{\delta}$$

If we can find $\delta \leq \widetilde{\delta}$ s.t. $(\varepsilon + |L|)\,\delta < \varepsilon$ then we'd be done. Take any δ such that $0 < \delta < \min\left\{\tilde{\delta}, \dfrac{\varepsilon}{\varepsilon + |L|}\right\}$. Then

$$|x-a| < \delta < \widetilde{\delta} \implies |f(x)-f(a)| < \delta(\varepsilon + |L|) < (\varepsilon + |L|)\,\frac{\varepsilon}{\varepsilon + |L|} = \varepsilon$$

□

Theorem 2. *Let $D \subseteq \mathbb{R}$, let $a \in D$ be a limit point of D, and let $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$ be differentiable at a.*

- *If $c \in \mathbb{R}$ is a constant, then $\dfrac{d[cf]}{dx}(a) = cf'(a)$.*
- $\dfrac{d[f+g]}{dx}(a) = f'(a) + g'(a)$.
- ***Product rule**: $\dfrac{d[fg]}{dx}(a) = f'(a)g(a) + f(a)g'(a)$.*
- ***Power rule**: If $f(x) = x^k$ for some $k \in \mathbb{R} \setminus \{0\}$, then $\dfrac{df}{dx}(a) = ka^{k-1}$.*
- ***Quotient rule**: If $g(a) \neq 0$, then $\dfrac{d}{dx}\dfrac{f(x)}{g(x)} = \dfrac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$.*

We state the final common derivative technique, the ***chain rule***, as a separate theorem to signify the importance of the derivative technique and also the fact that its proof (which we do not include) is not as straightforward as the proofs in Theorem 2:

Theorem 3. *Suppose $D, E \subseteq \mathbb{R}$, $f : D \to \mathbb{R}$ and $g : E \to \mathbb{R}$ are functions, $g(E) \subseteq D$, g is differentiable at a, and f is differentiable at $g(a)$. Then*

$$\frac{d[f \circ g]}{dx}(a) = \frac{d[f(g)]}{dx}(a) = f'(g(a))g'(a)$$

Some useful elementary function derivatives:

$$\begin{aligned}
\frac{d}{dx}e^x &= e^x \\
\frac{d}{dx}\ln(x) &= \frac{1}{x} \\
\frac{d}{dx}\sin(x) &= \cos(x) \\
\frac{d}{dx}\cos(x) &= -\sin(x)
\end{aligned}$$

Definition 2. A function $f : D \to \mathbb{R}$ is ***continuously differentiable*** on D if f' is continuous on D. It's not uncommon to write $f \in C^1(D)$.

Example 1. $f(x) = x^2$ is continuously differentiable since $f'(x) = 2x$ is continuous. However,

$$f(x) = \begin{cases} x^2 \sin(1/x) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is *not* continuously differentiable. In particular, for $x \neq 0$,

$$f'(x) = 2x\cos(1/x) - x^2 \frac{1}{x^2}\cos(1/x) = 2x\cos(1/x) - \cos(1/x)$$

and for $x = 0$ the derivative is 0:

$$\lim_{x\to 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x\to 0} \frac{x^2 \sin(1/x)}{x} = \lim_{x\to 0} x\sin(1/x)$$

With $\sin(1/x)$ bounded, $x \to 0 \implies x\sin(x) \to 0$. However, the derivative itself is not continuous at 0. Note that while $2x\cos(1/x) \xrightarrow{x\to 0} 0$, $-\cos(1/x)$ does not have a limit, so the derivative does not have a limit as $x \to 0$ either, meaning it cannot be continuous.

Theorem 4 (L'Hôpital's rule)**.** *Let f and g be continuous functions on $[a, b]$ and differentiable on (a, b). Suppose that $c \in [a, b]$ and that $f(c) = g(c) = 0$. Suppose also that for some $\varepsilon > 0$, $g'(x) \neq 0$ for all $x \in ([a, b] \cap B_\varepsilon(c)) \setminus \{c\}$. If*

$$\lim_{x\to c} \frac{f'(x)}{g'(x)} = L$$

then

$$\lim_{x\to c} \frac{f(x)}{g(x)} = L$$

If f, g are differentiable at c (which is not actually required by the Theorem 4) then the intuition can be made plain, as this implies $f(c) = g(c) = 0$, and

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \lim_{x\to c} \frac{f(x) - 0}{g(x) - 0} \cdot \frac{x - c}{x - c} = \lim_{x\to c} \frac{(f(x) - f(c))/(x - c)}{(g(x) - g(c))/(x - c)} = \frac{f'(c)}{g'(c)}$$

1.1.1 Mean Value Theorem (MVT)

We now move onto what is one of the most crucial results in single variable calculus: the Mean Value Theorem.

Theorem 5 (Mean Value Theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Then $\exists c \in (a, b)$ s.t.*

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

We will prove an equivalent, albeit perhaps conceptually easier, version of this:

Theorem 6 (Rolle's Theorem)**.** *Let $g : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b) with $g(a) = g(b)$. Then $\exists c \in (a, b)$ s.t.*

$$g'(c) = 0$$

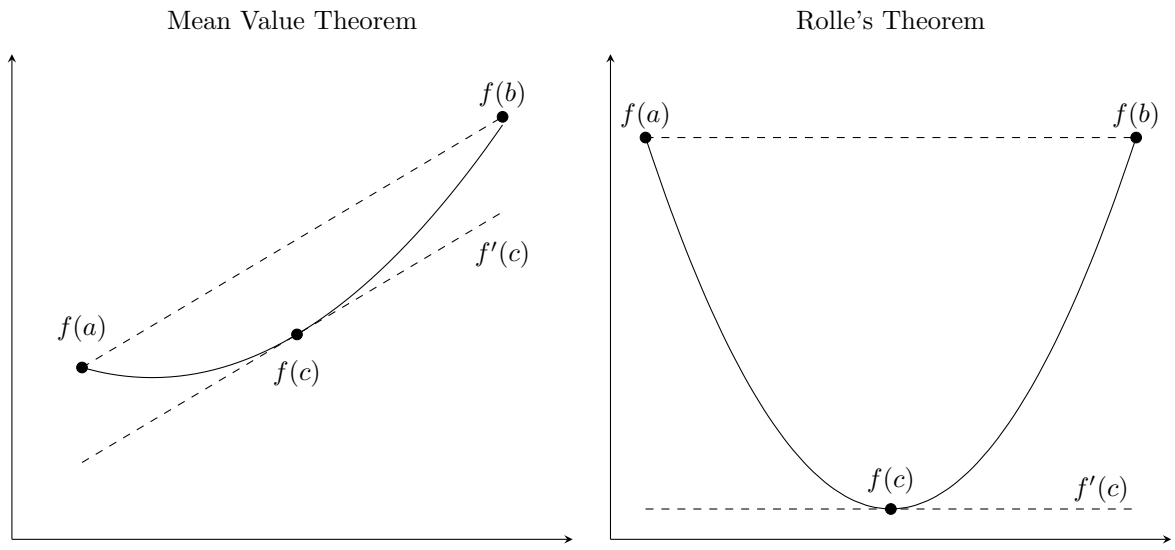


Figure 2: Graphical depiction of MVT

Claim 1. *Rolle's Theorem iff Mean Value Theorem.*

Proof. The mean value theorem implies Rolle's theorem by definition. Simply note that by the MVT there is some c s.t.

$$g'(c) = \frac{g(b) - g(a)}{b - a} = 0$$

Now to show Rolle's theorem implies MVT, we only need a simple transformation:

$$g(x) = (f(x) - f(a)) - \frac{f(b) - f(a)}{b - a}(x - a)$$

(This is subtracting the line with slope $\dfrac{f(b) - f(a)}{b - a}$ that goes through a from the function f, which is what we need to get it to "flatten.") Note that $g(a) = g(b) = 0$. Hence $\exists c$ s.t.

$$g'(c) = 0$$

But we can see that

$$g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} = 0 \implies f'(c) = \frac{f(b) - f(a)}{b - a}$$

□

Now we show Rolle's Theorem.

Proof. For Rolle's theorem we will use the extreme value theorem to assert that g is bounded and attains its sup and its inf in $[a, b]$. If the sup and the inf are *both* at $\{a, b\}$, then because by assumption $g(a) = g(b)$, $g(x) = 0$ everywhere on $[a, b]$. Take any $c \in (a, b)$, and

$$g'(c) = \lim_{x \to c} \frac{g(x) - g(c)}{x - c} = \lim_{x \to c} \frac{0}{x - c} = \lim_{x \to c} 0 = 0$$

Suppose then that either the sup or the inf do not occur at $x = a$ nor at $x = b$. Take the sup (WLOG; the inf is analogous or you can perform the analysis to $-g$). By assumption g is differentiable, so $g'(c)$ exists. That is, we know that

$$g'(c) = \lim_{x \to c} \frac{g(x) - g(c)}{x - c} = L$$

Consider $x \to c^-$, that is x approaching c from the left. Since the sup is attained at c, $g(c) \geq g(x)$ for all $x < c$, which in turn gives

$$\frac{g(x) - g(c)}{x - c} \geq 0 \qquad \forall x < c$$

Now take $x \to c^+$, that is x approaching c from the right. Again, since the sup is attained at c, $g(c) \geq g(x)$ for all $x > c$, which in turn gives

$$\frac{g(x) - g(c)}{x - c} \leq 0 \qquad \forall x > c$$

Which means that

$$L_- = \lim_{x \to c^-} \frac{g(x) - g(c)}{x - c} \geq 0 \quad \text{and} \quad L_+ = \lim_{x \to c^+} \frac{g(x) - g(c)}{x - c} \leq 0$$

Since the limit exists (f is differentiable at c), we know $L = L_- = L_+$. Thus $0 \leq L \leq 0 \implies L = 0$. □

Corollary 1. *If f is continuous on $[a, b]$ and differentiable on (a, b) and obtains a local minimum or maximum at c, then $f'(c) = 0$.*

Corollary 2. *If f is continuous on $[a, b]$ and differentiable on (a, b) and $f'(x) > 0$ for every $x \in (a, b)$ then f is increasing on (a, b). Conversely, if $f'(x) < 0$ for every $x \in (a, b)$ then f is decreasing on (a, b).*

Theorem 7 (Cauchy's MVT)**.** *Let $f, g : [a, b] \to \mathbb{R}$ be continuous functions on $[a, b]$ and differentiable on (a, b). Then $\exists c \in (a, b)$ s.t.*

$$g'(c)\,(f(b) - f(a)) = f'(c)\,(g(b) - g(a))$$

The MVT is a case where $g(x) = x$, the identity function.

1.1.2 Taylor's Theorem

Note that Theorem 7 is a generalization of the Mean Value Theorem (i.e., Theorem 5); however, it's maybe not the most natural generalization of the MVT that one could formulate. Moreover, the scenarios where Theorem 7 is useful and Theorem 5 is not useful are scarce.

A primary reason the MVT is so useful is that it provides us with an approximation of a differentiable function f. To see this, suppose $f : [a, b] \to \mathbb{R}$ meets all the requirements of Theorem 5 and we know the value of $f(a)$ and $f'(a)$, but we don't know the value of $f(b)$ (this is referred to as an interpolation problem in numerical analysis). By the MVT, we know there exists a $c \in (a, b)$ such that

$$f(b) = f(a) + f'(c)(b - a)$$

We don't know c (or $f'(c)$ for that matter), but if b is sufficiently close to a, then it stands to reason that

$$f(b) \approx f(a) + f'(a)(b - a)$$

More generally, it allows us to write

$$f(x) \approx f(a) + f'(a)(x - a)$$

whenever $x \in [a, b]$.

A few natural questions are raised through this thought process:

- How accurate is this approximation for a general $x \in (a, b)$?
- Can we generate a more accurate approximation for a general $x \in (a, b)$?

The following theorem provides a useful generalization of the MVT that addresses the questions above:

Theorem 8 (Taylor's theorem)**.** *Suppose $f : [a, b] \to \mathbb{R}$ is continuously differentiable $n + 1$ times on (a, b) and $x_0 \in (a, b)$. Then for each $x \in [a, b]$, there exists a c between x and x_0 such that*

$$f(x) = \underbrace{\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k}_{T_n(x)} + \underbrace{\frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}}_{R_n(x)}$$

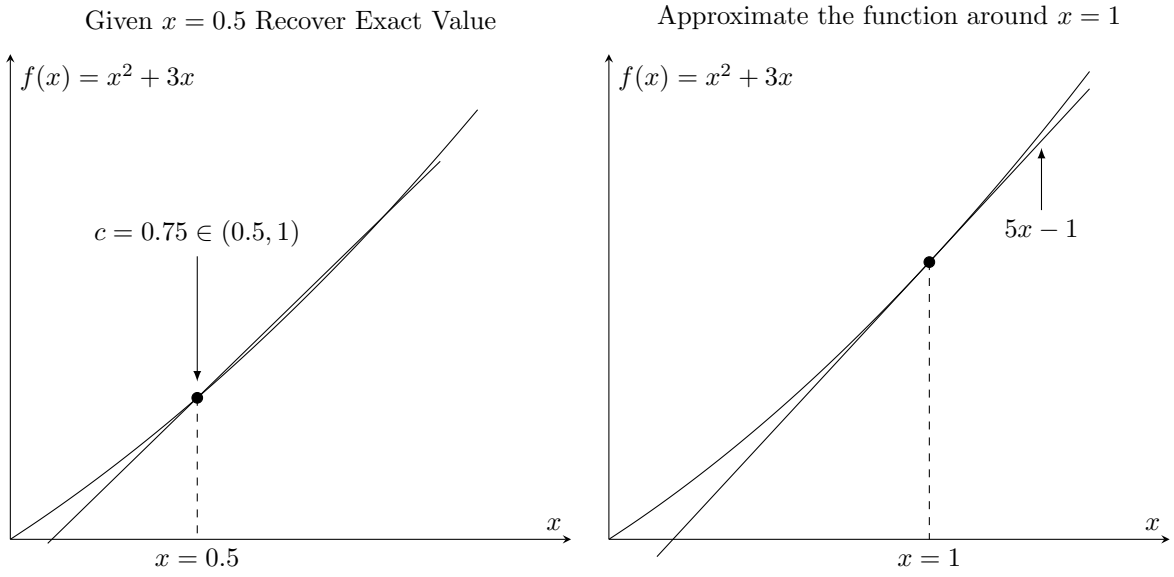
*where $T_n(x)$ is called the n^{th} **order Taylor polynomial** and $R_n(x)$ is called the remainder or error.*

Example 2. • A high enough order Taylor expansion of any polynomial is the polynomial itself. Consider $f(x) = x^2 + 3x$

$$\begin{aligned} f(x) &= \frac{f(1)}{0!}(x-1)^0 + \frac{f'(1)}{1!}(x-1)^1 + \frac{f''(c)}{2!}(x-1)^2 \\ &= 4 + 5(x-1) + \frac{2}{2}(x-1)^2 \\ &= 4 + 5x - 5 + x^2 - 2x + 1 = x^2 + 3x \end{aligned}$$

However, for lower-order Taylor expansions the theorem still applies. Visually:

Figure 3: Visualizing Taylor's Theorem



The left and right figures correspond to:

$$\begin{aligned} f(x) &= \frac{f(1)}{0!}(x-1)^0 + \frac{f'(c)}{1!}(x-1)^1 \\ &= 4 + (2c+3)(x-1) \\ f(x) &\approx \frac{f(1)}{0!}(x-1)^0 + \frac{f'(1)}{1!}(x-1)^1 \\ &= 4 + 5(x-1) \end{aligned}$$

We can see on the left that there is indeed some number c s.t. Taylor's theorem holds. For our sample value of $x = 0.5$ we find $c = 0.75 \in (0.5, 1)$. On the right figure, on the other hand, we plot the *approximation*. In this case, the function and the approximation are exactly equal at 1, since the approximation at that point simplifies to $f(1)$. We can also see that around $x = 1$ the approximation is fairly good! However, farther away the error increases, as we would expect.

- One common Taylor approximation is for the logarithm. In particular, the first order Taylor expansion around 1:

$$\log(x) \approx \log(1) + \frac{1}{1}(x-1) = x - 1$$

Another version of this approximation is for $\log(1+x)$ around 0:

$$\log(1+x) \approx \log(1) + \frac{1}{1+0}(x-0) = x$$

Take the relation:

$$\log Y = \alpha \log K + (1-\alpha) \log L$$

Using the approximation above, we can say that if K increases by 10%, then Y increases by $\alpha \log(1.1) \approx \alpha \cdot 0.1$, that is, $10\alpha\%$.

This second approximation is often used when dealing with percentage changes. You will hear the term ***log-linearization*** thrown around, and this is what that's in reference to: Logarithms can be approximated as a percentage for small values.

1.2 Partial Derivatives

To generalize our notion of the derivative into multiple dimensions, we first need to adopt some more notation. Let $S \subseteq \mathbb{R}^N$, $f : S \to \mathbb{R}^M$, $\mathcal{E} = \{e_1, \ldots, e_N\}$ a ***standard normal basis*** of $\mathbb{R}^N$, and $\mathcal{U} = \{u_1, \ldots, u_M\}$ a standard normal basis for $\mathbb{R}^M$; that is, $e_i \in \mathbb{R}^N$ is a vector for which the i^{th} component equals 1 and the other components are 0 ($u_i \in \mathbb{R}^M$ is defined analogously). For all $x \in S$, $f(x)$ is a ***linear combination*** of $\mathcal{U}$ and some set of functions $\{f_1, \ldots, f_M\}$ s.t. $f_i : S \to \mathbb{R}$ with

$$f(x) = \sum_{i=1}^{M} f_i(x) u_i$$

Defining f this way ensures that addition and scalar multiplication are well-defined operation on $\mathbb{R}^M$-valued functions f and g. Specifically, $\forall \alpha, \beta \in \mathbb{R}$,

$$\alpha f(x) + \beta g(x) = \alpha \left(\sum_{i=1}^{M} f_i(x) u_i \right) + \beta \left(\sum_{i=1}^{M} g_i(x) u_i \right) = \sum_{i=1}^{M} \left[\alpha f_i(x) + \beta g_i(x) \right] u_i$$

that is, scalar multiplication and vector addition are performed component-wise.

Definition 3. Let $S \subseteq \mathbb{R}^N$, let x be an interior point of S, and let $1 \leq j \leq N$. The ***partial derivative*** of a function $f : S \to \mathbb{R}^M$ in the x_j variable at x is

$$\frac{\partial}{\partial x_j} f(x) = \lim_{t \to 0} \frac{f(x + t \cdot e_j) - f(x)}{t}$$

We say that f is ***continuously differentiable*** if all the partial derivatives of f exist and are continuous.

Note that partial derivatives needn't imply anything about the behavior of the function overall. Take, for instance, $f : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x, y) = \begin{cases} \dfrac{x^2 y^4}{x^4 + y^8} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0) \end{cases}$$

The partial derivatives at 0 are all 0 (crucially, we take the limits one dimension at a time, so it's not that $(x, y) \to (0, 0)$, but rather $x \to 0$ and $y \to 0$ separately. Now take any $y_m \to 0$ and $x_m = y_m^2$ s.t. $y_m \neq 0$ for all n.

$$\lim_{m \to \infty} f(x_m, y_m) = \frac{1}{2} \neq 0 = f(0, 0)$$

which means the function is not even continuous at 0.

Theorem 9 (Schwarz's[2] Theorem)**.** *Let $S \subseteq \mathbb{R}^N$ be open and let $f : S \to \mathbb{R}^M$ be twice continuously*

[2] I have more commonly seen this referred to as Clairaut's Theorem; however, according to Wikipedia, Schwarz published

differentiable (i.e., f is continuously differentiable and all the partial derivatives $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_N}$ are themselves continuously differentiable). Then

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

that is, mixed partials are symmetric.

Definition 4. The ***gradient*** of $f : S \subseteq \mathbb{R}^N \to \mathbb{R}$ at $x \in S$ is

$$(\nabla f)(x) = \begin{bmatrix} \frac{\partial}{\partial x_1}f(x) & \cdots & \frac{\partial}{\partial x_N}f(x) \end{bmatrix}$$

The gradient can also be denoted as $(Df)(x)$ or $D_{x'}f(x)$.

Definition 5. The ***Hessian*** of $f : S \subseteq \mathbb{R}^N \to \mathbb{R}$ at $x \in S$ is

$$(D^2 f)(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2}f(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_N}f(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_N \partial x_1}f(x) & \cdots & \frac{\partial^2}{\partial x_N^2}f(x) \end{bmatrix}$$

Where D^2 denotes the application of the D operator twice.

Definition 6. Let $f : S \subseteq \mathbb{R}^N \to \mathbb{R}^M$ and $x \in S$; the ***Jacobian*** of f is the $M \times N$ matrix with i, j entry equal to $\frac{\partial}{\partial x_j}f_i(x)$, denoted $Df(x)$. That is,

$$Df(x) = \begin{pmatrix} (\nabla f_1)(x) \\ (\nabla f_2)(x) \\ \vdots \\ (\nabla f_M)(x) \end{pmatrix}$$

2 Implicit Function Theorem (IFT)

2.1 Motivation

Consider a function $f(x, y) = 0$. How do we characterize a function relating x to y? We can write that for $f : \mathbb{R}^2 \to \mathbb{R}$ continuously differentiable on an open set O with $f(x, y) = 0$, there exists some function h s.t.

$$f(x, h(x)) = 0 \quad \text{and} \quad \frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y}$$

One classic example is how to characterize the slope of a tangent line at a point (x, y) of some circle of radius r centered at $(0, 0)$. That is, $x^2 + y^2 = r^2$. We can write

$$f(x, y) = x^2 + y^2 - r^2 = 0$$

the first officially accepted proof of the result.

So we know that for some $h(x)$, $f(x, h(x)) = 0$. Furthermore,

$$\frac{\partial f}{\partial x} = 2x \quad \frac{\partial f}{\partial y} = 2y \implies \frac{dy}{dx} = -\frac{x}{y}$$

We will look at the general version of the theorem. We often work with a parameter space and a variable space, and we want to express the variables in terms of the parameters (or with exogenous and endogenous variables, and we want to express the endogenous variables in terms of the exogenous variables). Take

$$(\theta_1, \ldots, \theta_N)^\top = \theta \in \mathbb{R}^N$$

to be the parameters (or exogenous variables) and

$$(x_1, \ldots, x_M)^\top = x \in \mathbb{R}^M$$

to be the variables (or endogenous). It's not common to have an explicit expression for the latter in terms of the former given by the problem, but often we will encounter an implicit relation of the form

$$f(\theta, x) = 0$$

For example, some system of equations

$$\begin{aligned} f_1(\theta, x) &= 0 \\ &\vdots \\ f_M(\theta, x) &= 0 \end{aligned}$$

The IFT gives a result we can apply to these types of problems.

2.2 The IFT

Theorem 10 (Implicit Function Theorem)**.** *Suppose $U \subseteq \mathbb{R}^{N+M}$ is open and $f : U \to \mathbb{R}^M$ is continuously differentiable on U. Fix a point $(\theta_0, x_0) \in U$ with $f(\theta_0, x_0) = 0 \in \mathbb{R}^M$. If $Df(\theta_0, x_0)$ is non-singular (i.e. full-rank, or has a non-zero determinant) then there exists an open set $V \subseteq \mathbb{R}^N$ and a function $g : V \to \mathbb{R}^M$ such that*

1. *$\theta_0 \in V$*
2. *$g(\theta_0) = x_0$*
3. *$f(\theta, g(\theta)) = 0$ for all $\theta \in V$*

Moreover, g is continuously differentiable on U and differentiating with respect to θ' yields

$$\begin{aligned} D_{\theta'} f(\theta, g(\theta)) + D_{x'} f(\theta, g(\theta)) D_{\theta'} g(\theta) &= 0 \\ D_{\theta'} g(\theta) &= -\left[D_{x'} f(\theta, g(\theta))\right]^{-1} D_{\theta'} f(\theta, g(\theta)) \end{aligned}$$

2.2.1 Example

Take a simplified version of the IS-LM model

$$\begin{aligned} Y &= C + I + G \\ C &= C(Y - T) \\ I &= I(r) \\ M^S &= M^D(Y, r) \end{aligned}$$

with

$$0 < C'(x) < 1 \qquad I'(r) < 0 \qquad \frac{\partial M^D}{\partial Y} > 0 \qquad \frac{\partial M^D}{\partial r} < 0$$

National income must equal consumption plus investment (savings) plus government spending; consumption is some function of income minus taxes, the level of investment is determined by the interest rate, and money supply must equal money demand. We have that

$$\begin{aligned} Y - C(Y - T) - I(r) - G &= 0 \\ M^S - M^D(Y, r) &= 0 \end{aligned}$$

which is the exact type of problem the IFT can help us solve. We have endogenous variables $x = (Y, r)$, national income and the interest rate, and exogenous variables $\theta = (M^S, G, T)$, money supply, government spending, and taxes. Hence

$$f(\theta, x) = \begin{bmatrix} f_1(\theta, x) \\ f_2(\theta, x) \end{bmatrix} = \begin{bmatrix} Y - C(Y - T) - I(r) - G \\ M^S - M^D(Y, r) \end{bmatrix} = 0 \tag{1}$$

Then for some g, we can write

$$\begin{aligned} g(\theta) &= \begin{bmatrix} Y(M^S, G, T) \\ r(M^S, G, T) \end{bmatrix} \\ D_{\theta'} f(\theta, g(\theta)) + D_{x'} f(\theta, g(\theta)) D_{\theta'} g(\theta) &= 0 \end{aligned}$$

(**Note**: ***Stop here in class; the rest is a homework problem.***) We have that

$$D_{\theta'} f(\theta, g(\theta)) = \begin{bmatrix} 0 & -1 & C'(Y(\cdot) - T) \\ 1 & 0 & 0 \end{bmatrix}$$

$$D_{x'} f(\theta, g(\theta)) = \begin{bmatrix} \dfrac{\partial f_1}{\partial Y} = 1 - C'(Y(\cdot) - T) & \dfrac{\partial f_1}{\partial r} = -I'(r(\cdot)) \\ \dfrac{\partial f_2}{\partial Y} = -\dfrac{\partial M^D}{\partial Y} & \dfrac{\partial f_2}{\partial r} = -\dfrac{\partial M^D}{\partial r} \end{bmatrix}$$

$$[D_{x'} f(\theta, g(\theta))]^{-1} = \frac{1}{\det(D_{x'} f(\theta, g(\theta)))} \begin{bmatrix} -\dfrac{\partial M^D}{\partial r} & I'(r(\cdot)) \\ \dfrac{\partial M^D}{\partial Y} & 1 - C'(Y(\cdot) - T) \end{bmatrix}$$

$$D \equiv \det(D_{x'} f(\theta, g(\theta))) = \underbrace{\underbrace{-\overbrace{\frac{\partial M^D}{\partial r}}^{<0}}_{>0} \underbrace{(1 - C'(Y(\cdot) - T))}_{>0}}_{>0} \underbrace{\underbrace{-\overbrace{I'(r(\cdot))}^{<0}}_{>0} \underbrace{\frac{\partial M^D}{\partial Y}}_{>0}}_{>0} \implies D > 0$$

A non-zero determinant implies that the inverse exists. Hence we find that

$$\begin{aligned} D_{\theta'} g(\theta) &= \begin{bmatrix} \dfrac{\partial Y}{\partial M^S} & \dfrac{\partial Y}{\partial G} & \dfrac{\partial Y}{\partial T} \\ \dfrac{\partial r}{\partial M^S} & \dfrac{\partial r}{\partial G} & \dfrac{\partial r}{\partial T} \end{bmatrix} = -\frac{1}{D} \begin{bmatrix} I'(r(\cdot)) & \dfrac{\partial M^D}{\partial r} & -\dfrac{\partial M^D}{\partial r} C'(Y(\cdot) - T) \\ 1 - C'(Y(\cdot) - T) & -\dfrac{\partial M^D}{\partial Y} & \dfrac{\partial M^D}{\partial Y} C'(Y(\cdot) - T) \end{bmatrix} \\ &= -\frac{1}{D} \begin{bmatrix} < 0 & < 0 & > 0 \\ > 0 & < 0 & > 0 \end{bmatrix} = \frac{1}{D} \begin{bmatrix} > 0 & > 0 & < 0 \\ < 0 & > 0 & < 0 \end{bmatrix} \end{aligned}$$

Which means that for some $x = (Y, r), \theta = (M^S, G, T)$ that satisfies (1) there is some local neighborhood around θ where we can characterize the behavior of (Y, r) with respect to each of the variables in θ. In particular, income reacts positively to increases money supply or government spending but negatively to taxes, while the interest rate goes down with increases in the money supply or taxes but goes up with increases in government spending.

3 Unconstrained Optimization

Definition 7. Let $A \subseteq \mathbb{R}^N$ for some $N \in \mathbb{N}$ and let $f : A \to \mathbb{R}$ be a function.

1. $x \in A$ is a ***local maximum*** of f if $\exists \varepsilon > 0$ s.t.

$$y \in B_\varepsilon(x) \cap A \implies f(x) \geq f(y)$$

   The local maximum is ***strict*** if the inequality is strict. (Note the intersection: For example, $f(x) = x$ has no local maximum on $\mathbb{R}$, but every point is a local maximum if we define the function over $x \in \mathbb{N}$.)

2. The ***local minimum*** definition is analogous.

3. $x \in A$ is a ***global maximum*** of f if $\forall y \in A$, $f(x) \geq f(y)$. It is a strict global maximum if whenever $x \neq y$ we have $f(x) > f(y)$.

4. The ***global minimum*** definition is analogous.

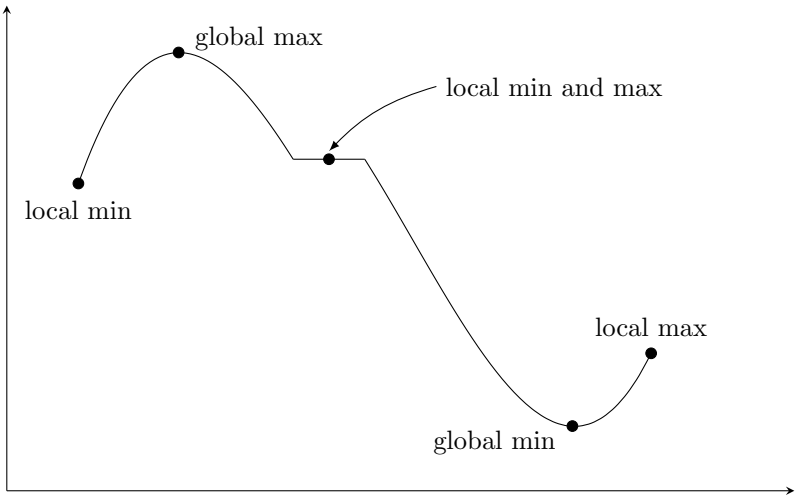


Figure 4: Examples of local and global maxima and minima

Definition 8. The arg max of a function f is the set

$$\arg\max_{x \in A} f(x) = \{x \in A : f(x) \geq f(y) \quad \forall y \in A\}$$

The arg min is analogously defined.

3.1 Quick Linear Algebra Review

Definition 9. A square $N \times N$ matrix S with elements in $\mathbb{R}$ is ***positive semidefinite*** (PSD) if $\forall p \in \mathbb{R}^N$

$$p^T S p \geq 0$$

and ***positive definite*** (PD) if the inequality is strict whenever $p \neq 0$.

Definition 10. A square $N \times N$ matrix S with elements in $\mathbb{R}$ is ***negative semidefinite*** (NSD) if $\forall p \in \mathbb{R}^N$

$$p^T S p \leq 0$$

and ***negative definite*** (ND) if the inequality is strict $p \neq 0$.

Definition 11. Let S be a $N \times N$ matrix with elements in $\mathbb{R}$. If $\exists p_1, p_2 \in \mathbb{R}^N$ s.t.

$$p_1^T S p_1 > 0 \quad \text{and} \quad p_2^T S p_2 < 0$$

then we say S is ***indefinite***.

Example 3. Consider the matrix

$$S = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

and take $p_1 = (1, 0, 0), p_2 = (-2, 1, 0)$. Then

$$p_1^T S p_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 1 > 0$$

$$p_2^T S p_2 = \begin{bmatrix} -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} = -3 < 0$$

So S is indefinite. In general we only need one counterexample for indefiniteness, but we need to check that *every* vector p gives a positive or negative quadratic form for definiteness. It turns out for symmetric matrices there are rules definite matrices have to follow that will help us determine their definiteness.

Definition 12. Let S be a $N \times N$ matrix with elements in $\mathbb{R}$. A kth order ***principal submatrix*** is the submatrix of S obtained by removing $N-k$ rows and the corresponding columns of S. A kth order ***principal minor*** is the determinant of a kth order principal submatrix.

It's easiest to talk about the principal minors using examples: Take

$$S = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

The 1st order principal submatrices are

$$[1] \qquad [5] \qquad [9]$$

and the principal minors are the determinants therein. The 2nd order principal submatrices are

$$\begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix} \qquad \begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix} \qquad \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix}$$

and the 2*nd* order principal minors are their determinants:

$$\det \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix} = -3 \qquad \det \begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix} = -12 \qquad \det \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix} = -3$$

Finally, the 3rd order principal submatrix is just the matrix S itself, and the 3rd order principal minor is the determinant of S (in this case, 0).

Definition 13. Let S be a $N \times N$ matrix with elements in $\mathbb{R}$. The kth ***leading principal minor*** is the principal minors obtained by removing the "last" $N - k$ columns and rows of S.

In our example above, these are the determinants of the matrix S itself (3rd leading principal minor), and

$$\text{1st } \to \det[1] = 1 \qquad \text{2nd } \to \det\begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix} = -3 \qquad \text{3rd } \to \det\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = 0$$

Definition 14. A matrix S is ***symmetric*** if $S = S^T$; that is, for S with entries $s(i,j)$ (where $i, j \in \{1, \ldots, N\}$) we have

$$s(i,j) = s(j,i)$$

Theorem 11. *Let S be a $N \times N$ symmetric matrix.*

1. *If all the leading principal minors are strictly positive, then S is positive definite.*
2. *If for every $k \leq N$ the kth order leading principal minor has sign $(-1)^k$ (that is, positive for k even and negative for k odd), then S is negative definite.*
3. *If all principal minors of S are weakly positive (≥ 0) then S is positive semidefinite.*
4. *If for every $k \leq N$ the kth order principal minors are ≤ 0 when k is odd and ≥ 0 when k is even, then S is negative semidefinite.*
5. *If the principal minors do not fit any of the above patterns then S is indefinite.*

Example 4. Consider the matrices

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

A, the identity, is positive definite. Consider any non-zero p:

$$p^T A p = p^T p = \sum_i p_i^2 > 0$$

if $p_i \neq 0$ for some i. Note all the leading principal minors are 1. For B, we similarly have

$$p^T B p = -p^T p = -\sum_i p_i^2 < 0$$

if $p_i \neq 0$ for some i. Note the leading principal minors are $-1, 1, -1$. For C, take $p_1 = (0,1,0), p_2 = (0,0,-1)$:

$$p_1^T C p_1 = 1 > 0 \qquad p_2^T C p_2 = -1 < 0$$

so C is indefinite. Note the leading principal minors are all 0, but the non-leading principal minors do not obey the pattern that gives semi-definiteness. In this case, the 1st-order principal minors are 0 (leading), 1, and -1, which is already an issue since the sign flips within a given set of principal minors.

3.2 First Order Conditions (FOC)

Theorem 12. *Let $f : A \to \mathbb{R}$ be a continuously differentiable function on an open set $A \subseteq \mathbb{R}^N$. If $x^* \in A$ is a local minimum or maximum, then*

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

that is, the first-order partials evaluated at x^* equal 0.

Remark 1. In general the converse need not be true. For instance $f(x) = x^3$. We have $f'(x) = 3x^2 = 0$ if $x = 0$. However the function does not have a local minimum or maximum at 0. Hence $Df(x) = 0$ is a necessary but not sufficient condition.

Some examples

1. Take $f(x) = 2x^3 - 3x^2$. We have
$$Df(x) = 6x^2 - 6x$$
$Df(x) = 0$ at $x = 0, 1$, so if f has local maxima or minima they must occur at those points, but we don't yet know how to check whether they are local maxima or minima.

2. $f(x, y) = x^3 - y^3 + 9xy$, so $f : \mathbb{R}^2 \to \mathbb{R}$. We have
$$Df(x, y) = \begin{bmatrix} 3x^2 + 9y & -3y^2 + 9x \end{bmatrix}$$
$Df(x, y) = 0$ at $(0, 0)$ and $(3, -3)$. To see this, consider the following
$$\begin{aligned} 0 = 3x^2 + 9y &\implies x^2 = -3y \\ 0 = -3y^2 + 9x &\implies 3y^2 = 9x \end{aligned}$$
Squaring the top equation and plugging in the bottom equation yields:
$$x^4 = 9y^2 = 3(3y^2) = 3(9x)$$
So
$$\begin{aligned} x^4 = 27x &\implies x = 3 \\ 27 + 9y = 0 &\implies y = -3 \end{aligned}$$

3.3 Second Order Conditions (SOC)

Theorem 13. *Let $f : A \to \mathbb{R}$ be a twice continuously differentiable function on an open set $A \subseteq \mathbb{R}^N$ with*

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

for some $x^* \in A$. If $D^2 f(x^*)$, the Hessian at x^*, is negative definite, then x^* is a local maximum. Similarly, if $D^2 f(x^*)$ is positive definite, then it is a local minimum.

Remark 2. The converse need not hold. For instance, take $f(x) = x^4, Df(x) = 4x^3, D^2f(x) = 12x^2$. At $x^* = 0$, we have $Df(0) = 0$, but $D^2f(0) = 0$ is neither positive nor negative definite. Hence a strictly definite Hessian is a sufficient condition for a local maximum or minimum, but it is not necessary.

Theorem 14. *Let $f : A \to \mathbb{R}$ be a twice continuously differentiable function on an open set $A \subseteq \mathbb{R}^N$. If $x^* \in A$ is a local maximum, then*

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

and $D^2f(x^*)$ is negative semidefinite. If $x^* \in A$ is a local minimum, then $Df(x^*) = 0$ and $D^2f(x^*)$ is positive semidefinite.

Remark 3. Again, the converse need not be true. In our previous example, $x^* = 0$ is actually a local minimum, and we can check that $Df(x^*) = 4(0^3) = 0$ and $D^2f(x^*) = 12(0^2) = 0 \geq 0$ (positive semidefinite). However, $D^2f(x^*) \leq 0$ means that it is negative semidefinite as well, but that does not imply a local maximum. Hence the condition is necessary but not sufficient.

Let us take $f(x) = 2x^3 - 3x^2$ again. We saw that $Df(x) = 6x^2 - 6x = 0$ at $x = 0, 1$. Now we have

$$\begin{gathered} D^2f(x) = 12x - 6 \\ D^2f(0) = -6 < 0 \implies \text{local max} \\ D^2f(1) = 6 > 0 \implies \text{local min} \end{gathered}$$

What about $f(x, y) = x^3 - y^3 + 9xy$? Recall

$$Df(x, y) = \begin{bmatrix} 3x^2 + 9y & -3y^2 + 9x \end{bmatrix} = 0 \impliedby (x, y) = (0, 0) \text{ or } (3, -3)$$

We find that

$$D^2f(x, y) = \begin{bmatrix} 6x & 9 \\ 9 & -6y \end{bmatrix}$$

The second order principal minor is the determinant of the Hessian itself. The 1st order principal minor is $6x$. Note the determinant of the Hessian is

$$\left|D^2f(x, y)\right| = -36xy - 81$$

1. For $(0, 0)$, we have $6(0) = 0$ and $-36(0)(0) - 81 = -81 < 0$. The 2nd leading principal minor is negative, so the Hessian at that point cannot be positive definite. Further, $(-1)^2$ is positive, so it cannot be negative definite either. Hence the Hessian at $(0, 0)$ is indefinite.
2. For $(3, -3)$, we have $6(3) = 18 > 0$ and $-36(3)(-3) - 81 = 324 - 81 = 243 > 0$. The 1st and 2nd leading principal minors are both positive, which means the Hessian at that point is positive definite and $(3, -3)$ is a local minimum.

Note: This is exactly the second-derivative test you may have learned in early calculus. For a function $f : \mathbb{R}^2 \to \mathbb{R}$, the Hessian is given by

$$D^2 f(x,y) = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}$$

with f_{xy} denoting the partials with respect to x and then y.

1. The second-oder leading principal minor has to be positive for the Hessian to be definite, positive or negative. This is the determinant of the Hessian itself, or $f_{xx}f_{yy} - 2f_{xy} > 0$.
2. The first-oder leading principal minor has to be positive for a min, or $f_{xx} > 0$, which gives PD.
3. The first-oder leading principal minor has to be negative for a max, or $f_{xx} < 0$, which gives ND.

Remark 4. One way to think about why it is that definiteness of the Hessian gives local extrema is to use a "multivariate" version of Taylor's theorem. Let $f : A \to \mathbb{R}$ be a three times continuously differentiable function on an open set with $Df(x^*) = 0$ for some x^*. Note any x can be written as $x^* + \alpha z$ for some unit vector z and some α. Note all such x are at most α away from x^* (i.e. $\|x - x^*\| = |\alpha|$). Let $g(\alpha) = f(x^* + \alpha z)$ and consider the second-order Taylor expansion of g around $\alpha = 0$:

$$g(\alpha) = g(0) + g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 + R(\alpha)\alpha^2$$
$$f(x^* + \alpha z) = f(x^*) + \alpha Df(x^*)z + \frac{\alpha^2}{2} z^T D^2 f(x^*) z + R(\alpha)\alpha^2$$

with $R(\alpha) \to 0$ as $\alpha \to 0$. We know $Df(x^*) = 0$ by premise. If $D^2 f(x^*)$ is positive definite, it turns out there exists some λ s.t. $z^T D^2 f(x^*) z \geq \lambda z^T z = \lambda > 0$ for any unit vector z.[3] Hence

$$f(x^* + \alpha z) = f(x^*) + \frac{\alpha^2}{2} z^T D^2 f(x^*) z + R(\alpha)\alpha^2 > f(x^*) + \left(\frac{\lambda}{2} + R(\alpha)\right)\alpha^2$$

for *any* z s.t. $x^* + \alpha z \in B_\delta(x^*)$ and $\|z\| = 1$. Since $R(\alpha) \to 0$, there should be an α small enough to make the last term above positive; thus for some $\alpha > 0$ we have

$$f(x^* + \alpha z) > f(x^*)$$

for any such z; the last step is to remark once again $x = x^* + \alpha z \in B_\alpha(x^*)$. Hence if $D^2 f(x^*)$ is PD we have a local min for the α-neighborhood. The steps for a local max are analogous (note λ will be negative).

3.4 Concavity and Convexity

Definition 15. A function $f : A \to \mathbb{R}$ is ***concave*** if for any $\alpha \in [0,1]$ and $x, y \in A$

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

[3] We will discuss this in the last lecture, but a symmetric matrix is PD iff all its eigenvalues are strictly positive. Further, a symmetric matrix S can be decomposed into $C\Lambda C^\top$ with Λ a diagonal matrix of eigenvalues and C an orthonormal matrix of eigenvectors. Hence $z^\top S z = z^\top C \Lambda C^\top z = p^\top \Lambda p = \sum_i \lambda_i p_i^2$ for $p = C^\top z$. Let $\lambda = \min_i \lambda_i$ and we have the result, since $z^\top S z \geq \lambda p^\top p = \lambda (z^\top C C^\top z) = \lambda z^\top z$ since C is orthonormal. Lastly, a symmetric matrix is ND iff all its eigenvalues are strictly negative, and the analogous steps give the result with the inequality flipped.

It is ***strictly concave*** if the above holds strictly for $\alpha \in (0,1)$ and $x \neq y$.

Definition 16. A function $f : A \to \mathbb{R}$ is ***convex*** if for any $\alpha \in [0,1]$ and $x, y \in A$

$$\alpha f(x) + (1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y)$$

It is ***strictly convex*** if the above holds strictly for $\alpha \in (0,1)$ and $x \neq y$.

Theorem 15. *Let $f : A \to \mathbb{R}$ be a twice continuously differentiable function on an open set $A \subseteq \mathbb{R}^N$.*

1. *$f(x)$ is concave $\iff$ $D^2 f(x)$ is negative semidefinite. (Negative definite $\implies$ strictly concave.)*
2. *$f(x)$ is convex $\iff$ $D^2 f(x)$ is positive semidefinite. (Positive definite $\implies$ strictly convex.)*

Remark 5. As was pointed out during the lecture, $f(x) = x^4$ is convex but the Hessian is not positive definite everywhere. In particular, $D^2 f(0) = 0$. Hence while a positive definite Hessian implies concavity, the converse only gives weak concavity.

Remark 6. The univariate intuition for concave and convex functions generalizes. The first derivative of a concave function is decreasing (the function is either increasing at a decreasing rate or decreasing at an increasing rate), so the second derivative must be negative; the opposite for a convex function. To see why this intuition is sufficient for multivariate functions, consider

$$g(\alpha) = f(\alpha x + (1-\alpha)y)$$

and note

$$g''(\alpha) = (x-y)^T D^2 f(\alpha x + (1-\alpha)y)(x-y)$$

which is the square form for the Hessian (and exactly the form we want to check for definiteness). We will show f is concave iff g is concave. If f is concave,

$$\begin{aligned}
g(\gamma\alpha + (1-\gamma)\beta) &= f\left([\gamma\alpha + (1-\gamma)\beta]x + [1 - (\gamma\alpha + (1-\gamma)\beta)]y\right) \\
&= f\left(\alpha\gamma x + (1-\gamma)\beta x + \gamma y + (1-\gamma)y - \alpha\gamma y - (1-\gamma)\beta y\right) \qquad \text{Add and subtract } \gamma y \\
&= f\left(\gamma[\alpha x + (1-\alpha)y] + (1-\gamma)[\beta x + (1-\beta)y]\right) \\
&\geq \gamma f(\alpha x + (1-\alpha)y) + (1-\gamma) f(\beta x + (1-\beta)y) \\
&= \gamma g(\alpha) + (1-\gamma) g(\beta)
\end{aligned}$$

If g is concave,

$$f(\alpha x + (1-\alpha)y) = g(\alpha) \geq \alpha g(1) + (1-\alpha) g(0) = \alpha f(x) + (1-\alpha) f(y)$$

where we use the fact that $\alpha = 1 \cdot \alpha + (1-\alpha) \cdot 0$. Hence it is sufficient to show g has a negative second derivative, which follows from the univariate version of the theorem.

Theorem 16. *Let $f : A \to \mathbb{R}$ be a twice continuously differentiable function on an open set $A \subseteq \mathbb{R}^N$.*

1. *If f is concave and x^* is s.t. $Df(x^*) = 0$ then x^* is a global maximum. (If f is strictly concave the global maximum is unique.)*
2. *If f is convex and x^* is s.t. $Df(x^*) = 0$ then x^* is a global minimum. (If f is strictly convex the global minimum is unique.)*

Some examples:

1. Take the function $f(x, y) = x^2 + y^2$. We have

$$Df(x, y) = \begin{bmatrix} 2x & 2y \end{bmatrix} \qquad D^2 f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

   Note that $Df(0, 0) = (0, 0)$. Further, the 1st leading principal minor of the Hessian is $2 > 0$; for the 2nd leading principal minor we have $2(2) - 0(0) = 4 > 0$. Thus $D^2 f(x)$ is positive definite; this means f is convex and $(0, 0)$ is a global minimum.

2. What about $f(x) = x^4$? In this case

$$Df(x) = 4x^3 \qquad D^2 f(x, y) = 12x^2$$

   $Df(0) = 0$ and $D^2 f(0) = 0$. However, $12x^2 \geq 0$ for all x, so the function is convex, which means that 0 is a global minimum.

3. $f(x, y) = x^2 y^2$

$$Df(x, y) = \begin{bmatrix} 2xy^2 & 2x^2 y \end{bmatrix} \qquad D^2 f(x, y) = \begin{bmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{bmatrix}$$

   Note $Df(x, 0) = Df(0, y) = (0, 0)$, but the kth order principal minors are all 0 at $(x, 0)$ or $(0, y)$. More genearlly we have that while the 1st principal minors are $2y^2 \geq 0$ and $2x^2 \geq 0$, the determinant of the 2nd principal minor is

$$4x^2 y^2 - 16x^2 y^2 \leq 0$$

   Hence we cannot even say whether it is positive or negative semidefinite.

4 Fun Remarks

- Modern calculus was developed in 17th-century Europe by Isaac Newton and Gottfried Wilhelm Leibniz (independently of each other, first publishing around the same time) but elements of it first appeared in ancient Egypt and later Greece, then in China and the Middle East, and still later again in medieval Europe and India.
- Modern mathematics classes frequently teach derivatives before integration, but the theory of integration is actually what motivated the discovery of calculus with the theory of differentiation coming afterwards.
- Guillaume de l'Hôpital (also written l'Hospital) published Theorem 4 in his 1696 book *Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes* (literal translation: *Analysis of the Infinitely*

Small for the Understanding of Curved Lines), the first textbook on differential calculus. However, it is believed that the rule was discovered by the Swiss mathematician Johann Bernoulli, and l'Hôpital purchased the naming rights of the result.

Index