APM462
Lecture Notes

Yuchen Wang

October 11, 2019

Contents

1 Matrix Calculus

Row v.s. Column Vector Our default rule is that every vector is a column vector unless explicitly stated otherwise.
This is also known as the numerator layout.
Special case: For $f : \mathbb{R}^n \to \mathbb{R}$, Df is a $1 \times n$ matrix or row vector.

1.1 Matrix Multiplication

Definition 1.1.1 Let A be $m \times n$, and B be $n \times p$, and let the product AB be

$$C = AB$$

then C is a $m \times p$ matrix, with element (i, j) given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

for all $i = 1, 2, \ldots, m, j = 1, 2, \ldots, p$.

Proposition 1.1.2 Let A be $m \times n$, and x be $n \times 1$, then the typical element of the product

$$z = Ax$$

is given by

$$z_i = \sum_{k=1}^{n} a_{ik} x_k$$

for all $i = 1, 2, \ldots, m$.
Similarly, let y be $m \times 1$, then the typical element of the product

$$z^T = y^T A$$

is given by

$$z_i = \sum_{k=1}^{n} a_{ki} y_k$$

for all $i = 1, 2, \ldots, n$.
Finally, the scalar resulting from the product

$$\alpha = y^T Ax$$

is given by

$$\alpha = \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} y_i x_k$$

1.2 Partitioned Matrices

Proposition 1.2.1 Let A be a square, nonsingular matrix of order m. Partition A as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

so that A_{11} and A_{22} are invertible.
Then

$$A^{-1} = \begin{bmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ -A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{bmatrix}$$

proof:
Direct multiplication of the proposed A^{-1} and A yields

$$A^{-1}A = I$$

■

1.3 Matrix Differentiation

Proposition 1.3.1

$$\frac{\partial A}{\partial x} = \frac{\partial A^T}{\partial x}$$

Proposition 1.3.2 Let

$$y = Ax$$

where y is $m \times 1$, x is $n \times 1$, A is $m \times n$, and A does not depend on x. Suppose that x is a function of the vector z, while A is independent of z. Then

$$\frac{\partial y}{\partial z} = A\frac{\partial x}{\partial z}$$

Proposition 1.3.3 Let the scalar α be defined by

$$\alpha = y^T A x$$

where y is $m \times 1$, x is $n \times 1$, A is $m \times n$, and A is independent of x and y, then

$$\frac{\partial \alpha}{\partial x} = y^T A$$

and

$$\frac{\partial \alpha}{\partial y} = x^T A^T$$

Proposition 1.3.4 For the special case where the scalar α is given by the quadratic form

$$\alpha = x^T A x$$

where x is $n \times 1$, A is $n \times n$, and A does not depend on x, then

$$\frac{\partial \alpha}{\partial x} = x^T (A + A^T)$$

proof:
By definition

$$\alpha = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$$

Differentiating with respect to the kth element of x we have

$$\frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_J + \sum_{i=1}^{n} a_{ik} x_i$$

for all $k = 1, 2, \ldots, n$, and consequently,

$$\frac{\partial \alpha}{\partial x} = x^T A^T + x^T A = x^T (A^T + A)$$

■

Proposition 1.3.4 For the special case where A is a symmetric matrix and

$$\alpha = x^T A x$$

where x is $n \times 1$, A is $n \times n$, and A does not depend on x, then

$$\frac{\partial \alpha}{\partial x} = 2 x^T A$$

Proposition 1.3.5 Let the scalar α be defined by

$$\alpha = y^T x$$

where y is $n \times 1$, x is $n \times 1$, and both y and x are functions of the vector z. Then

$$\frac{\partial \alpha}{\partial z} = x^T \frac{\partial y}{\partial z} + y^T \frac{\partial x}{\partial z}$$

Proposition 1.3.6 Let the scalar α be defined by

$$\alpha = x^T x$$

where x is $n \times 1$, and x is a functions of the vector z. Then

$$\frac{\partial \alpha}{\partial z} = 2x^T \frac{\partial y}{\partial z}$$

Proposition 1.3.7 Let the scalar α be defined by

$$\alpha = y^T A x$$

where y is $m \times 1$, A is $m \times n$, x is $n \times 1$, and both y and x are functions of the vector z, while A does not depend on z. Then

$$\frac{\partial \alpha}{\partial z} = x^T A^T \frac{\partial y}{\partial z} + y^T A \frac{\partial x}{\partial z}$$

Proposition 1.3.8 Let A be an invertible, $m \times m$ matrix whose elements are functions of the scalar parameter α. Then

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

proof:
Start with the definition of the inverse

$$A^{-1} A = I$$

and differentiate, yielding

$$A^{-1} \frac{\partial A}{\partial \alpha} + \frac{\partial A^{-1}}{\partial \alpha} A = 0$$

rearranging the terms yields

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

■

Vector-by-vector Differentiation Identities 1.3.9

Condition	Expression	Numerator layout, i.e. by y and x^T	Denominator layout, i.e. by y^T and x
a is not a function of **x**	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
A is not a function of **x**	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^\top$
A is not a function of **x**	$\frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^\top$	$\mathbf{A}$
a is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	
v = *v*(**x**), **u** = **u**(**x**)	$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} =$	$v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial v}{\partial \mathbf{x}}$	$v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}\mathbf{u}^\top$
A is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^\top$
u = **u**(**x**), **v** = **v**(**x**)	$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	
u = **u**(**x**)	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$
u = **u**(**x**)	$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

Young's Theorem 1.3.10 i.e. Symmetry of second derivatives

$$[\nabla_{xy} f(x,y)]^T = \nabla_{yx} f(x,y)$$

proof:
This is straightforward by writing out the elements of the matrix. ■

2 Second-year Calculus Review

functions $\mathbb{R} \to \mathbb{R}$

2.1 Mean Value Theorem in 1 Dimension

$g \in C^1$ on $\mathbb{R}$

$$\frac{g(x+h)-g(x)}{h} = g'(x+\theta h)$$

where $\theta \in (0,1)$
Or equivalently,

$$g(x+h) = g(x) + hg'(x+\theta h)$$

2.2 1st Order Taylor Approximation

$g \in C^1$ on $\mathbb{R}$

$$g(x+h) = g(x) + hg'(x) + o(h)$$

where $o(h)$ is "little o" of h, the error term.
Say a function $f(h) = o(h)$, this means $\lim_{h\to 0} \frac{f(h)}{h} = 0$
For example, for $f(h) = h^2$, we can say $f(h) = o(h)$,
since $\lim_{h\to 0} \frac{f(h)}{h} = \lim_{h\to 0} \frac{h^2}{h} = \lim_{h\to 0} h = 0$
proof: (Use MVT):
WTS : $g(x+h) - g(x) - hg'(x) = o(h)$

$$\begin{aligned}
\lim_{h\to 0} \frac{[g(x+h)-g(x)] - hg'(x)}{h} &= \lim_{h\to 0} \frac{[hg'(x+\theta h)] - hg'(x)}{h} \\
&= \lim_{h\to 0} g'(x+\theta h) - g'(x) \\
&= \lim_{h\to 0} g'(x) - g'(x) \\
&= 0
\end{aligned}$$

■

2.3 2nd Order Mean Value Theorem

$g \in C^2$ on $\mathbb{R}$

$$g(x+h) = g(x) + hg'(x) + \frac{h^2}{2} g'(x+\theta h)$$

for some $\theta \in (0,1)$

proof:

WTS: $g(x+h) - g(x) - hg'(x) - \frac{h^2}{2}g''(x) = o(h^2)$

$$\begin{aligned}
\lim_{h\to 0}\frac{g(x+h) - g(x) - hg'(x) - \frac{h^2}{2}g''(x)}{h^2} &= \lim_{h\to 0}\frac{[\frac{h^2}{2}g'(x+\theta h)] - \frac{h^2}{2}g''(x)}{h^2} \\
&= \lim_{h\to 0}\frac{1}{2}(g''(x+\theta h) - g''(x)) \\
&= \lim_{h\to 0}\frac{1}{2}(g''(x) - g''(x)) \\
&= 0
\end{aligned}$$

■

multivariate functions: $\mathbb{R}^n \to \mathbb{R}$

2.4 Recall: Definition of gradient

Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ (denoted $\nabla f(x)$) if exists is a vector characterized by the property:

$$\lim_{\mathbf{v}\to\mathbf{0}}\frac{f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x}) - \nabla f(\mathbf{x})\cdot\mathbf{v}}{||\mathbf{v}||} = 0$$

In Cartesian coordinates, $\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial \mathbf{x}_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial \mathbf{x}_n}(\mathbf{x}))$

2.5 Mean Value Theorem in n dimension

$f \in C^1$ on $\mathbb{R}^n$, then for any $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$,

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x}+\theta\mathbf{v})\cdot\mathbf{v}$$

for some $\theta \in (0,1)$

<u>*proof:*</u> Reduce to 1-dimension case
$g(t) := f(\mathbf{x}+t\mathbf{v}), t \in \mathbb{R}$

$$\begin{aligned}
g'(t) &= \frac{d}{dt}f(\mathbf{x}+t\mathbf{v}) \\
&= \sum_{i=1}^{n}\frac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x}+t\mathbf{v})\cdot\frac{d(\mathbf{x}+t\mathbf{v})_i}{dt} && \text{(by Chain Rule)} \\
&= \sum \frac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x}+t\mathbf{v})\cdot\frac{d(\mathbf{x}_i+t\mathbf{v}_i)}{dt} \\
&= \sum \frac{\partial f}{\partial \mathbf{x}_i}(\mathbf{x}+t\mathbf{v})\cdot\mathbf{v}_i \\
&= \nabla f(\mathbf{x}+t\mathbf{v})\cdot\mathbf{v} && (*)
\end{aligned}$$

$g \in C^1$ on $\mathbb{R}$
Using MVT in $\mathbb{R}$:

$$\begin{aligned} f(\mathbf{x}+\mathbf{v}) &= g(1) \\ &= g(0+1) \\ &= g(0) + 1g'(0+\theta 1) && (\theta \in (0,1)) \\ &= g(0) + g'(\theta) \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x}+\theta\mathbf{v}) \cdot \mathbf{v} && (\text{by } (*)) \end{aligned}$$

■

2.6 1st Order Taylor Approximation in $\mathbb{R}^n$

$f \in C^1$ on $\mathbb{R}^n$

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{v} + o(||\mathbf{v}||)$$

proof:

$$\begin{aligned} \lim_{||\mathbf{v}|| \to 0} \frac{[f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x})] - \nabla f(\mathbf{x}) \cdot \mathbf{v}}{||\mathbf{v}||} &= \lim_{||\mathbf{v}|| \to 0} \frac{[\nabla f(\mathbf{x}+\theta\mathbf{v}) \cdot \mathbf{v}] - \nabla f(\mathbf{x}) \cdot \mathbf{v}}{||\mathbf{v}||} \\ &= \lim_{||\mathbf{v}|| \to 0} [\nabla f(\mathbf{x}+\theta\mathbf{v}) - \nabla f(\mathbf{x})] \cdot \frac{\mathbf{v}}{||\mathbf{v}||} \\ &= 0 \quad (\tfrac{\mathbf{v}}{||\mathbf{v}||} \text{ is a unit vector, remains } 1) \end{aligned}$$

■

2.7 2nd Order Mean Value Theorem in $\mathbb{R}^n$

$f \in C^2$ on $\mathbb{R}^n$

$$f(\mathbf{x}-\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{v} + \frac{1}{2}\mathbf{v}^T \nabla^2 f(\mathbf{x}+\theta\mathbf{v}) \cdot \mathbf{v}$$

Remarks In this course, ∇^2 means Hessian, not Laplacian.

$$\nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\right)_{1 \le i,j \le n} (\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial_1^2} & \frac{\partial f}{\partial_1 \partial_2} & \cdots \\ \frac{\partial f}{\partial_2 \partial_1} & \cdots & \\ \vdots & & \end{pmatrix}$$

The Hessian matrix is symmetric. This is sometimes called Clairaut's Theorem.
note: $\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} = \sum_{1 \le i,j \le n} \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j} f(\mathbf{x}) \mathbf{v}_i \mathbf{v}_j$

2.8 2nd Order Taylor Approximation in $\mathbb{R}^n$

$f \in C^2$ on $\mathbb{R}^n$

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{v} + \frac{1}{2}\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} + o(||\mathbf{v}||^2)$$

proof:

$$\begin{aligned}
\lim_{||\mathbf{v}|| \to 0} \frac{[f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x})] - \nabla f(\mathbf{x}) \cdot \mathbf{v} - \frac{1}{2}\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v}}{||\mathbf{v}||^2} &= \lim_{||\mathbf{v}|| \to 0} \frac{[\frac{1}{2}\mathbf{v}^T \nabla^2 f(\mathbf{x}+\theta\mathbf{v}) \cdot \mathbf{v}] - \frac{1}{2}\mathbf{v}^T \nabla^2 f(\mathbf{x}) \cdot \mathbf{v}}{||\mathbf{v}||^2} \\
&\quad \text{(By 2nd MVT)} \\
&= \lim_{||\mathbf{v}|| \to 0} \frac{1}{2}(\frac{\mathbf{v}}{||\mathbf{v}||})^T [\nabla^2 f(\mathbf{x}+\theta\mathbf{v}) - \nabla^2 f(\mathbf{x})](\frac{\mathbf{v}}{||\mathbf{v}||}) \\
&= 0
\end{aligned}$$

■

2.9 Geometric Meaning of Gradient

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

Rate of change of f at $\mathbf{x}$ in direction $\mathbf{v}$ $(||\mathbf{v}|| = 1) = \frac{d}{dt}|_{t=0} f(\mathbf{x}+t\mathbf{v})$

$$\begin{aligned}
\frac{d}{dt}|_{t=0} f(\mathbf{x}+t\mathbf{v}) &= \nabla f(\mathbf{x}+t\mathbf{v}) \cdot \mathbf{v}|_{t=0} \\
&= \nabla f(\mathbf{x}) \cdot \mathbf{v} \\
&= |\nabla f(\mathbf{x})||\mathbf{v}| \cos\theta \\
&= |\nabla f(\mathbf{x})| \cos\theta \qquad (||\mathbf{v}|| = 1)
\end{aligned}$$

maximized at $\theta = 0$

So $\nabla f(\mathbf{x})$ points in the direction of steepest ascent.

2.10 Implicit Function Theorem

$f : \mathbb{R}^{n+1} \to \mathbb{R} \in C^1$

Fix $(\mathbf{a}, b) \in \mathbb{R}^n \times \mathbb{R}$ s.t. $f(\mathbf{a}, b) = 0$.

If $\nabla f(\mathbf{a}, b) \neq 0$, then $\{(\mathbf{x}, y) \in (\mathbb{R}^n \times \mathbb{R}) | f(\mathbf{x}, \mathbf{y}) = 0\}$ is locally (near $(\mathbf{a}, b)$) the graph of a function.

2.11 Level Sets of f

c-level set of $f := \{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) = c\}$

Fact gradient $\nabla f(\mathbf{x}_0) \perp$ level curve (through $\mathbf{x}_0$)

3 Convex Set & Functions

3.1 Definitions

Definition of Convex Set $\Omega \subseteq \mathbb{R}^n$ is a convex set if $\mathbf{x}_1, \mathbf{x}_2 \in \Omega \Rightarrow s\mathbf{x}_1 + (1-s)\mathbf{x}_2 \in \Omega$ where $s \in [0,1]$

Definition of Convex Function A function f : convex $\Omega \subseteq \mathbb{R}^n$ is convex if

$$f(s\mathbf{x}_1 + (1-s)\mathbf{x}_2) \leq sf(\mathbf{x}_1) + (1-s)f(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ and all $s \in [0,1]$

Remarks Second line above (or equal to) the graph

Definition of Concave Function A function f is concave if $-f$ is convex.

3.2 Basic Properties of convex functions

Let $\Omega \subseteq \mathbb{R}^n$ be a convex set.

1. f_1, f_2 are convex functions on $\Omega \Rightarrow f_1 + f_2$ is a convex function on Ω.
2. f is a convex function, $a \geq 0 \Rightarrow af$ is a convex function.
3. f is a convex on $\Omega \Rightarrow$ The sublevel sets of f, $SL_c := \{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) \leq c\}$ is convex.

proof of (3):
Let $x_1, x_2 \in SL_C$, so that $f(x_1) \leq c$ and $f(x_2) \leq c$.
WTS: $sx_1 + (1-s)x_2 \in SL_c$ for any $s \in [0,1]$

$$\begin{aligned} f(sx_1 + (1-s)x_2) &\leq sf(x_1) + (1-s)f(x_2) && (f \text{ is convex}) \\ &\leq sc + (1-s)c \\ &= c \end{aligned}$$

$$\Rightarrow sx_1 + (1-s)x_2 \in SL_c$$

■

Example of a convex function Let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$
Let $x_1, x_2 \in \mathbb{R}, s \in [0, 1]$
Then

$$\begin{aligned} f(sx_1 + (1-s)x_2) &= |sx_1 + (1-s)x_2| \\ &\le |sx_1| + |(1-s)x_2| \qquad \text{(by Triangle Inequality)} \\ &= s|x_1| + (1-s)|x_2| \\ &= sf(x_1) + (1-s)f(x_2) \end{aligned}$$

Then f is a convex function.

Theorem - Characterization of C^1 convex functions Let f : convex subset of $\mathbb{R}^n$ $\Omega \to \mathbb{R}$ be a C^1 function.
Then,
f is convex $\iff f(y) \ge f(x) + \nabla f(x) \cdot (y - x)$ for all $x, y \in \Omega$

Remarks Tangent line below the graph.

proof:
($\Rightarrow$)
f is convex, then by definition,

$$\begin{aligned} f(s\mathbf{x}_1 + (1-s)\mathbf{x}_2) &\le sf(\mathbf{x}_1) + (1-s)f(\mathbf{x}_2) \\ f(s\mathbf{x}_1 + (1-s)\mathbf{x}_2) - f(\mathbf{x}_2) &\le s(f(\mathbf{x}_1) - f(\mathbf{x}_2)) \\ \frac{f(s\mathbf{x}_1 + (1-s)\mathbf{x}_2) - f(\mathbf{x}_2)}{s} &\le f(\mathbf{x}_1) - f(\mathbf{x}_2) \\ \lim_{s\to 0} \frac{f(\mathbf{x}_2 + s(\mathbf{x}_1 - \mathbf{x}_2)) - f(\mathbf{x}_2)}{s} &\le f(\mathbf{x}_1) - f(\mathbf{x}_2) \\ \nabla f(\mathbf{x}_2) \cdot (\mathbf{x}_1 - \mathbf{x}_2) &\le f(\mathbf{x}_1) - f(\mathbf{x}_2) \\ &\text{(since } \tfrac{d}{ds}|_{s=0} f(\mathbf{x}_2 + s(\mathbf{x}_1 - \mathbf{x}_2)) = \nabla f(\mathbf{x}_2) \cdot (\mathbf{x}_1 - \mathbf{x}_2)) \\ f(\mathbf{x}_2) + \nabla f(\mathbf{x}_2) \cdot (\mathbf{x}_1 - \mathbf{x}_2) &\le f(\mathbf{x}_1) \\ f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) &\le f(\mathbf{y}) \end{aligned}$$

where $0 \le s \le 1$
($\Leftarrow$)
Fix $\mathbf{x}_0, \mathbf{x}_1 \in \Omega$ and $s \in (0, 1)$
Let $x = s\mathbf{x}_0 + (1-s)\mathbf{x}_1$

$$\begin{cases} f(\mathbf{x}_0) & \geq f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (\mathbf{x}_0 - \mathbf{x}) \\ & = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (1-s)(\mathbf{x}_0 - \mathbf{x}_1) \\ f(\mathbf{x}_1) & \geq f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (\mathbf{x}_1 - \mathbf{x}) \\ & = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot s(\mathbf{x}_1 - \mathbf{x}_0) \end{cases}$$

$$\begin{cases} sf(x_0) & \geq sf(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (1-s) \cdot s(\mathbf{x}_0 - \mathbf{x}_1) \\ (1-s)f(\mathbf{x}_1) & \geq (1-s)f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (1-s) \cdot s(\mathbf{x}_1 - \mathbf{x}_0) \end{cases}$$

Then

$$sf(\mathbf{x}_0) + (1-s)f(\mathbf{x}_1) \geq f(x) + 0$$

Then f is convex. ■

3.3 Criterions for convexity

C^1 criterion for convexity

$$f : \Omega \to \mathbb{R} \text{ is convex } \iff f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$$

for all $x, y \in \Omega$

Theorem: C^2 criterion for convexity Let $f \in C^2$ on $\Omega \subseteq \mathbb{R}^n$ (here we assume $\Omega \subseteq \mathbb{R}^n$ is a convex set containing an interior point)
Then

$$f \text{ is convex on } \Omega \iff \nabla^2 f(x) \geq 0$$

for all $x \in \Omega$

Remark 1 Let A be an $n \times n$ matrix.
"$A \geq 0$" means A is positive semi-definite:

$$v^T A v \geq 0$$

for all $v \in \mathbb{R}^n$

Remark 2 In $\mathbb{R}$,

$$f \text{ is convex } \iff f'(x) \geq 0$$

for all $x \in \Omega$
("concave up" in first year calculus)

<u>*proof for Theorem:*</u>
Recall 2nd order MVT:

$$f(y) = f(x) + \nabla f(x) \cdot (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x + s(y-x)) \cdot (y-x)$$

for some $s \in [0,1]$
($\Leftarrow$)
Since $\nabla^2 f(x) \geq 0$, then

$$\frac{1}{2}(y-x)^T \nabla^2 f(x + s(y-x)) \cdot (y-x) \geq 0$$

Then

$$f(y) \geq f(x) + \nabla f(x) \cdot (y-x)$$

for all $x, y \in \Omega$.
Then by C^1 criterion, f is convex.
($\Rightarrow$)
Assume f is convex on Ω.
Suppose for contradiction that $\nabla^2 f(x)$ is not positive semi-definite at some $x \in \Omega$.
Then $\exists v \neq 0$ s.t. $v^T \nabla^2 f(x) v < 0$ v could be arbitrarily small and > 0
Let $y = x + v$, then

$$(y-x)^T \nabla^2 f(x + s(y-x)) \cdot (y-x) < 0$$

for all $s \in [0,1]$
Then by MVT,

$$f(y) < f(x) + \nabla f(x) \cdot (y-x)$$

for some $x, y \in \Omega$, and this contradicts the C^1 criterion. ■

3.4 Minimization and Maximization of Convex Functions

Theorem f : convex $\Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function.
Suppose $\Gamma := \{x \in \Omega | f(x) = \min_{\Omega} f(x)\} \neq \emptyset$
(i.e. minimizer exists)
Then Γ is a convex set, and any local minimum of f is a global minimum of f.
<u>*proof:*</u>

Let $m = \min_{\Omega} f(x)$.

$$\Gamma = \{x \in \Omega | f(x) = m\} = \{x \in \Omega | f(x) \leq m\}$$

(sublevel set)
Then by Basic Properties of Convex Sets, Γ is convex.
Let x be a local minimum of f.
Suppose for contradiction that $\exists y$ s.t. $f(y) < f(x)$
(i.e. x is not a global minimum)

$$\begin{aligned} f(sy + (1-s)x) &\leq sf(y) + (1-s)f(x) \\ &< sf(x) + (1-s)f(x) && (f(y) < f(x)) \\ &= f(x) \end{aligned}$$

for all $s \in (0,1)$
As s approaches 0, s approaches x.
Then we have $\lim\limits_{s \to 0} f(sy + (1-s)x) = f(x) < f(x)$.
which is a contradiction. ■

Theorem If $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function, and Ω is convex and compact, then

$$\max_{\Omega} f = \max_{\partial\Omega} f$$

Remarks Maximum value of f is attained (also) on the boundary of Ω
proof:
Since Ω is closed, $\partial\Omega \subseteq \Omega$, so $\max\limits_{\Omega} f \geq \max\limits_{\partial\Omega} f$.
Suppose $f(x_0) = \max\limits_{\Omega} f$ for some $x_0 \notin \partial\Omega$. Let L be an arbitrary line through x_0.
By convexity and compactness of Ω, L meets $\partial\Omega$ at two points x_1, x_2.
Let $x_0 + sx_1 + (1-s)x_2$ for $s \in (0,1)$

$$\begin{aligned} f(x_0) &= f(sx_1 + (1-s)x_2) \\ &\leq sf(x_1) + (1-s)f(x_2) && (f \text{ convex}) \\ &\leq \max\{f(x_1), f(x_2)\} \\ &\leq \max_{\partial\Omega} f \\ &\leq \max_{\Omega} f = f(x_0) \end{aligned}$$

This implies that

$$\max_{\Omega} f = \max_{\partial\Omega} f$$

as wanted. ■

Example

$$|ab| \leq \frac{1}{p}|a|^p + \frac{1}{q}|b|^q$$

where $p, q > 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$.
Special cases:

1. $$p = q = 2, |ab| \leq \frac{|a|^2 + |b|^2}{2}$$

2. $$p = 3, q = \frac{3}{2}, |ab| \leq \frac{1}{3}|a|^3 + \frac{2}{3}|b|^{\frac{3}{2}}$$

proof:
Since function $f(x) = -\log(x)$ is convex, then

$$\begin{aligned}
(-\log)|ab| &= (-\log)|a| + (-\log)|b| \\
&= \frac{1}{p}(-\log)|a|^p + \frac{1}{q}(-\log)|b|^q \\
&\geq (-\log)(\frac{1}{p}|a|^p + \frac{1}{q}|b|^q) \\
(-\log)|ab| &\geq (-\log)(\frac{1}{p}|a|^p + \frac{1}{q}|b|^q) \\
\log|ab| &\leq \log(\frac{1}{p}|a|^p + \frac{1}{q}|b|^q) \\
|ab| &\leq \frac{1}{p}|a|^p + \frac{1}{q}|b|^q \qquad \text{(exponential function is increasing)}
\end{aligned}$$

■

4 Basics of Unconstrained Optimization

4.1 Extreme Value Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is continues, and compact set $K \subseteq \mathbb{R}^n$ Then the problem

$$\min_{x \in K} f(x)$$

has a solution.

Recall

1. $$K \subseteq \mathbb{R}^n \text{ compact} \iff K \text{ closed and bounded}$$

2. If $h_1, \dots, h_k$ and $g_1, \dots, g_m$ are continuous functions on $\mathbb{R}^n$, then the set of all points $x \in \mathbb{R}^n$ s.t.
$$\begin{cases} h_i(x) = 0 & \text{for all } i \\ g_j(x) \leq 0 & \text{for all } j \end{cases}$$
is a closed set.

3. If such a set is also bounded, then it is compact.

Example
$$\{(x, y) \in \mathbb{R}^2 | x^2 + y^2 - 1 = 0\}$$
by (2), this is a closed set
by (3), this is a compact set.

Remarks $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ convex does not imply f is continuous.

4.2 Unconstrained Optimization

$$\min_{x \in \Omega \subseteq \mathbb{R}^n} f(x)$$

typically

1. $\Omega \subseteq \mathbb{R}^n$
2. $\Omega = \mathbb{R}^n$
3. $\Omega =$ open
4. $\Omega = \overline{\text{open}}$

Remark

1. $\max f(x) = -(\min -f(x))$
2. $\min f(x) = -(\max -f(x))$

Definition: local minimum We say that f has a local minimum at a point $x_0 \in \Omega$ if

$$f(x_0) \leq f(x)$$

for all $x \in B_\Omega^\varepsilon(x_0)$, where $B_\Omega^\varepsilon(x_0) = \{x \in \Omega : |x - x_0| < \varepsilon\}$ which is an open ball around x_0 inside Ω of radius $\varepsilon > 0$.
We say that f has a strict local minimum at a point $x_0 \in \Omega$ if

$$f(x_0) < f(x)$$

for all $x \in B_\Omega^\varepsilon(x_0) \setminus \{x_0\}$

4.3 1st order necessary condition for local minimum

Theorem Let f be a C^1 function on $\Omega \subseteq \mathbb{R}^n$. If $x_0 \in \Omega$ is a local minimum of f, then

$$\nabla f(x_0) \cdot v \geq 0$$

for all feasible directions v at x_0

Definition: feasible direction $v \in \mathbb{R}^n$ is a feasible direction at $x_0 \in \Omega$ if

$$x_0 + sv \in \Omega$$

for all $0 \leq s \leq \bar{s}$ where $\bar{s} \in \mathbb{R}$

Remarks Feasible directions go into the set.

Corollary Special case: If $\Omega = \mathbb{R}^n$ is an open set, then any direction is a feasible direction. Then x_0 is a local minimum of f on Ω implies that $\nabla f(x_0) \cdot v \geq 0$ for all $v \in \mathbb{R}^n$.

$$\begin{cases} \nabla f(x_0) \cdot v \geq 0 \\ \nabla f(x_0) \cdot (-v) \geq 0 \iff \nabla f(x_0) \cdot v \leq 0 \end{cases} \implies \nabla f(x_0) \cdot v = 0 \text{ for all } v \in \mathbb{R}^n$$

$$\implies \nabla f(x_0) = 0$$

proof: [ ]

Definition: local minimum We say that f has a local minimum at a point $x_0 \in \Omega$ if

$$f(x_0) \leq f(x)$$

for all $x \in B_\Omega^\varepsilon(x_0)$, where $B_\Omega^\varepsilon(x_0) = \{x \in \Omega : |x - x_0| < \varepsilon\}$ which is an open ball around x_0 inside Ω of radius $\varepsilon > 0$.
We say that f has a strict local minimum at a point $x_0 \in \Omega$ if

$$f(x_0) < f(x)$$

for all $x \in B_\Omega^\varepsilon(x_0) \setminus \{x_0\}$

4.3 1st order necessary condition for local minimum

Theorem Let f be a C^1 function on $\Omega \subseteq \mathbb{R}^n$. If $x_0 \in \Omega$ is a local minimum of f, then

$$\nabla f(x_0) \cdot v \geq 0$$

for all feasible directions v at x_0

Definition: feasible direction $v \in \mathbb{R}^n$ is a feasible direction at $x_0 \in \Omega$ if

$$x_0 + sv \in \Omega$$

for all $0 \leq s \leq \bar{s}$ where $\bar{s} \in \mathbb{R}$

Remarks Feasible directions go into the set.

Corollary Special case: If $\Omega = \mathbb{R}^n$ is an open set, then any direction is a feasible direction. Then x_0 is a local minimum of f on Ω implies that $\nabla f(x_0) \cdot v \geq 0$ for all $v \in \mathbb{R}^n$.

$$\begin{cases} \nabla f(x_0) \cdot v \geq 0 \\ \nabla f(x_0) \cdot (-v) \geq 0 \iff \nabla f(x_0) \cdot v \leq 0 \end{cases} \implies \nabla f(x_0) \cdot v = 0 \text{ for all } v \in \mathbb{R}^n$$

$$\implies \nabla f(x_0) = 0$$

proof: [ ]

4.4 2nd order necessary condition for local minimum

$f \in C^2, \Omega \subseteq \mathbb{R}^n$
If $x_0 \in \Omega$ is a local minimum of f on Ω, then

1. $\nabla \cdot v \geq 0$ for al feasible directions v at x_0
2. If $\nabla f(x_0) \cdot v = 0$, then $v^T \nabla^2 f(x_0) v \geq 0$ (function curves up)

proof: [ ]

Remark If x_0 is an interior point of Ω, then

$$\nabla f(x_0) = 0, \quad \nabla^2 f(x_0) \geq 0$$

$$f'(x_0) = 0, \quad f''(x_0) \geq 0$$

4.5 Definition: positive definiteness

A $n \times n$ matrix A is

6

Equality Constraints

Definition 1: surface

$$M = \text{“surface”} = \{x \in \mathbb{R}^n | h_1(x) = 0, \ldots, h_k(x) = 0\}$$

where $h_i \in C^1$

Definition 2: differentiable curve on surface A differentiable curve on surface $M \subseteq \mathbb{R}^n$ is a C^1 function

$$x : (-\epsilon, \epsilon) \to M : s \mapsto \lambda(s)$$

???

Definition 4: tangent space Tangent space to the surface M at point x_0 is

$$T_{x_0}M = \{\text{all tangent vectors to } M \text{ at } x_0\} = \{v \in \mathbb{R}^n : v = \frac{d}{ds}|_{s=0}\, x(s)\}$$

where $x(s)$ is a differentiable curve on M s.t. $x(0) = x_0$

Remarks The zero vector is contained in all tangent spaces.

Definition 1: T-space

$$T_{x_0} = \{x \in \mathbb{R}^n : x^T \nabla h_i(x_0) = 0\, \forall i\} = Span\{\nabla h_1(x_0), \ldots, \nabla h_k(x_0)\}^\perp$$

Definition 2: regular point $x_0 \in M$ is a regular point (of the constraints) if $\{\nabla h_1(x_0), \ldots . \nabla h_k(x_0)\}$ are linearly independent.

When does the T-space equivalent to the tangent space? When x_0 is a regular point (of the constraints).

Theorem 3 Suppose x_0 is a regular point s.t. $M = \{x \in real^n : h_i(x) = 0\, \forall i\}$. Then

$$T_{x_0}M = T_{x_0}$$

Lemma 4 $f, h_1, \ldots, h_k \in C^1$ on open $\Omega \subseteq \mathbb{R}^n$
$M = \{x \in real^n : h_i(x) = 0\, \forall i\}$
Suppose $x_0 \in M$ is a local minimum of f on M, then

$$\nabla f(x_0) \perp T_{x_0}M \iff \nabla f(x_0) \cdot v = 0$$

for all $v \in T_{x_0}M$

5.1 Lagrange Multipliers: 1st order necessary condition for local minimum

$f, h_1, \ldots, h_k \in C^1$ on open $\Omega \subseteq \mathbb{R}^n$.
Let x_0 be a regular point of the constraints $M = \{x \in real^n : h_i(x) = 0\, \forall i\}$.
Suppose x_0 is a local minimum of f on M, then $\exists \lambda_1, \ldots, \lambda_k \in \mathbb{R}$ s.t.

$$\nabla f(x_0) + \lambda_1 \nabla h_1(x_0) + \ldots + \lambda_k \nabla h_k(x_0) = 0$$

5.2 2nd order necessary condition for local minimum

$f, h_1, \dots, h_k \in C^2$ on open $\Omega \subseteq \mathbb{R}^n$.
Let x_0 be a regular point of the constraints $M = \{x \in real^n : h_i(x) = 0 \, \forall i\}$.
Suppose x_0 is a local minimum of f on M, then

1. $$\nabla f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla h_i(x_0) = 0$$

   for some $\lambda_i \in \mathbb{R}$

2. $$\nabla^2 f(x_0) + \sum \lambda_i \nabla^2 h_i(x_0) \geq 0$$

   on $T_{x_0}M$

5.3 2nd order sufficient condition for local minimum

$f, h_1, \dots, h_k \in C^2$ on open $\Omega \subseteq \mathbb{R}^n$.
Let x_0 be a regular point of the constraints $M = \{x \in real^n : h_i(x) = 0 \, \forall i\}$.
If $\exists \lambda_i \in \mathbb{R}$ s.t.

1. $$\nabla f(x_0) + \sum \lambda_i \nabla h_i(x_0) = 0$$

2. $$\nabla^2 f(x_0) + \sum \lambda_i \nabla^2 h_i(x_0) > 0$$

   on $T_{x_0}M$

Then x_0 is a strict local minimum.

6 Inequality Constraints

Problem open $\Omega \subseteq \mathbb{R}^n$
$f : \Omega \to \mathbb{R}$
$h_1, \dots, h_k : \Omega \to \mathbb{R}$
$g_1, \dots, g_l : \Omega \to \mathbb{R}$

$$\min f(x)$$

$$x \in \Omega \text{ subject to } \begin{cases} h_1(x) = 0, \dots, h_k(x) = 0 \\ g_1(x) \leq 0, \dots, g_l(x) \leq 0 \end{cases}$$

Definition 1: activeness Let x_0 satisfy the constraints. We say that the constraint $g_i(x) \leq 0$ is active at x_0 if $g_i(x_0) = 0$. It is inactive at x_0 if $g_i(x_0) < 0$

Definition 2: regular point Suppose for some $l' \leq l$:

$$g_1(x) \leq 0, \ldots, g_{l'}(x) \leq 0; \, g_{l'+1}(x) \leq 0, \ldots, g_l(x) \leq 0$$

where $g_1, \ldots g_{l'}$ active and the rest inactive. We say that x_0 is a regular point of the constraints if $\{\nabla h_1(x_0), \ldots, \nabla h_k(x_0), \nabla g_1(x_0), \ldots, \nabla g_{l'}(x_0)\}$ is linearly independent.

6.1 Kuhn-Tucker conditions: 1st order necessary condition for local minimum

open $\Omega \subseteq \mathbb{R}^n$
$f : \Omega \to \mathbb{R}$
$h_1, \ldots, h_k, g_1, \ldots, g_l : C^1 \in \Omega$
Suppose $x_0 \in \Omega$ is a regular point of the constraints which is a local minimum, then

1. $$\nabla f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla h_i(x_0) + \sum_{j=1}^{l} \mu_j \nabla g_j(x_0) = 0$$

   for some $\lambda_i \in \mathbb{R}$

2. $\mu_j g_j(x_0) = 0$ for all f with some $\mu_j \geq 0$