Calculus

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2013

Outline

Outline

Functions and Limits

Caution!

The functions f (or $\boldsymbol{f}$) discussed here are not required to be linear anymore.

Definition (Limit of a Function)

A function $\boldsymbol{f}:\mathcal{V}\to\mathbb{R}^m$, $\mathcal{V}\subseteq\mathbb{R}^n$, has a ***limit*** $\boldsymbol{f}^*(\boldsymbol{a})$ at the point $\boldsymbol{a}\in\mathcal{V}$ if given any $\varepsilon\in\mathbb{R}$, $\varepsilon>0$, there exists $\delta\in\mathbb{R}$, $\delta>0$ such that for all $\boldsymbol{x}\in\mathcal{V}$, $0<\|\boldsymbol{x}-\boldsymbol{a}\|<\delta$, we have $\|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}^*(\boldsymbol{a})\|<\varepsilon$. This is denoted by $\lim_{\boldsymbol{x}\to\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{f}^*(\boldsymbol{a})$.

Definition (Continuity)

A function $\boldsymbol{f}:\mathcal{V}\to\mathbb{R}^m$, $\mathcal{V}\subseteq\mathbb{R}^n$, is ***continuous*** at $\boldsymbol{a}$ iff $\boldsymbol{f}^*(\boldsymbol{a})=\boldsymbol{f}(\boldsymbol{a})$; that is, given any $\varepsilon>0$, there exists $\delta>0$ such that for all $\boldsymbol{x}\in\mathcal{V}$, $0<\|\boldsymbol{x}-\boldsymbol{a}\|<\delta$, we have $\|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{a})\|<\varepsilon$.

Sequences and Convergence (1/2)

- A ***sequence*** of vectors $\{\boldsymbol{x}^{(k)}\}_k$ can be think of as the $\mathcal{R}(\boldsymbol{f})$ for some $\boldsymbol{f}:\mathbb{N}\to\mathbb{R}^n$
  - A sequence is ***increasing*** iff $\boldsymbol{x}^{(1)} < \boldsymbol{x}^{(2)} < \cdots$, and ***nondecreasing*** iff $\boldsymbol{x}^{(1)} \leqslant \boldsymbol{x}^{(2)} \leqslant \cdots$
  - Nondecreasing and nonincreasing sequences are called ***monotone sequences***

Definition (Limit of a Sequence)

A sequence $\{\boldsymbol{x}^{(k)}\}_k$ has a ***limit*** $\boldsymbol{x}^*$ if given any $\varepsilon\in\mathbb{R}$, $\varepsilon>0$, there exists $K\in\mathbb{N}$, such that for all $k>K$, we have $\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|<\varepsilon$. This is denoted by $\lim_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*$.

- A sequence having a limit is called a ***convergent*** sequence
- Given a sequence $\{\boldsymbol{x}^{(k)}\}_k$ convergent to $\boldsymbol{a}$, we can see that a function $\boldsymbol{f}:\mathcal{V}\to\mathbb{R}^m$ is continuous at $\boldsymbol{a}$ iff $\lim_{k\to\infty}\boldsymbol{f}(\boldsymbol{x}^{(k)})=\boldsymbol{f}(\boldsymbol{a})$ [Proof: Using definitions and the fact that $\lim_{k\to\infty}\boldsymbol{f}(\boldsymbol{x}^{(k)})=\boldsymbol{f}(\lim_{k\to\infty}\boldsymbol{x}^{(k)})$]

Sequences and Convergence (2/2)

- Given a sequence $\{\boldsymbol{x}^{(k)}\}_k$ and an increasing sequence of nature numbers $\{m_k\}_k$, we call $\{\boldsymbol{x}^{(m_k)}\}_k$ the ***subsequence*** of $\{\boldsymbol{x}^{(k)}\}_k$
  - A subsequence is obtained by neglecting some elements of a given sequence
- If a sequence converges to $\boldsymbol{x}^*$, then all its subsequences converge to $\boldsymbol{x}^*$ too [Proof]

Extreme Value Theorem

Theorem

Let $f : \Omega \to \mathbb{R}$ *be a continuous function over a compact set* $\Omega \subseteq \mathbb{R}^n$. *There exist* $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \Omega$ *such that* $f(\boldsymbol{x}_0) \leqslant f(\boldsymbol{x}) \leqslant f(\boldsymbol{x}_1), \forall \boldsymbol{x} \in \Omega$*; that is,* $f(\boldsymbol{x}_0) = \min_{\boldsymbol{x}\in\Omega}(f(\boldsymbol{x}))$ *and* $f(\boldsymbol{x}_1) = \max_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$.

- We say f is ***bounded*** on Ω iff there exists $l, h \in \mathbb{R}$ such that $l \leqslant f(\boldsymbol{x}) \leqslant h, \forall \boldsymbol{x} \in \Omega$
  - The above theorem says that f is bounded on Ω if Ω is compact

Min, Max, Inf, and Sup

- Given a subset $\mathcal{S}$ (e.g., $[0,1)$ or $\{2,4,6,\cdots\}$) of $\mathbb{R}$ (or any other ordered set where elements can be compared with each other), we have:

Definition (Supremum)

An point $p \in \mathbb{R}$ is called the ***supremum***, denoted by $\sup_{s\in\mathcal{S}} s$, iff a) $s \leqslant p, \forall s \in \mathcal{S}$; b) for any $\varepsilon > 0$, there exist $s \in \mathcal{S}$ such that $s > p - \varepsilon$.

- p is called the maximum iff $p \in \mathcal{S}$

Definition (Infimum)

An point $p \in \mathbb{R}$ is called the ***infimum***, denoted by $\inf_{s\in\mathcal{S}} s$, iff a) $s \geqslant p, \forall s \in \mathcal{S}$; b) for any $\varepsilon > 0$, there exist $s \in \mathcal{S}$ such that $s < p + \varepsilon$.

- p is called the minimum iff $p \in \mathcal{S}$

Convergence of Functions (1/2)

- Given a set of data $\{\boldsymbol{x}^{(t)}\}_{t=1}^{N}$, suppose we use an ML algorithm to train a model, say $\boldsymbol{f}^{(N)}$
  - Usually, we want to know how the ML algorithm works when $N \to \infty$
  - We can think of $\{\boldsymbol{f}^{(N)}\}_N$ as a sequence, and then investigate the properties of its limit $\boldsymbol{f}^*$

Definition (Pointwise Convergence)

A sequence of functions $\{\boldsymbol{f}^{(N)}\}_N$, where $\boldsymbol{f}^{(N)} : \mathcal{V} \to \mathbb{R}^m$ and $\mathcal{V} \subseteq \mathbb{R}^n$, ***converges pointwise*** to a function $\boldsymbol{f}^* : \mathcal{V} \to \mathbb{R}^m$ iff for any $\boldsymbol{x} \in \mathcal{V}$, we have $\lim_{N\to\infty} \boldsymbol{f}^{(N)}(\boldsymbol{x}) = \boldsymbol{f}^*(\boldsymbol{x})$.

- Unfortunately, pointwise convergence is not strong enough to guarantee "reasonable" relations between $\boldsymbol{f}^{(N)}$ and $\boldsymbol{f}^*$
  - E.g., for all $x \in [0,1]$, a sequence of continuous function $f^{(N)}(x) = x^N$ converges pointwise to $f^*(x) = \begin{cases} 0, & 0 \leqslant x < 1 \\ 1, & x = 1 \end{cases}$, which is obviously not continuous

Convergence of Functions (2/2)

Definition (Uniform Convergence)

A sequence of functions $\{\boldsymbol{f}^{(N)}\}_N$, where $\boldsymbol{f}^{(N)}: \mathcal{V} \to \mathbb{R}^m$ and $\mathcal{V} \subseteq \mathbb{R}^n$, ***converges uniformly*** to a function $\boldsymbol{f}^*: \mathcal{V} \to \mathbb{R}^m$ iff given any $\varepsilon > 0$, there exists $K \in \mathbb{N}$ such that for all $N \geqslant K$, we have $\left\|\boldsymbol{f}^{(N)} - \boldsymbol{f}^*\right\| < \varepsilon$ for all $\boldsymbol{x} \in \mathcal{V}$.

- Intuitively, $\boldsymbol{f}^{(N)}$ can be fitted into any given "ε-tube" around $\boldsymbol{f}^*$as long as N is large enough

Theorem

If a sequence of continuous functions $\{\boldsymbol{f}^{(N)}\}_N$ converges uniformly to $\boldsymbol{f}^*$, then $\boldsymbol{f}^*$ will be continuous.

- Can be proved by either the "$\varepsilon/3$ trick" or the "ε-tube" intuition

Outline

Derivative (1/2)

Definition (Derivative)

A function $f : [s,t] \to \mathbb{R}$, $s,t \in \mathbb{R}$, is ***differentiable*** at $a \in (s,t)$ iff $\lim_{\delta \to 0} \frac{f(a+\delta)-f(a)}{\delta}$ (or equivalently, $\lim_{x \to a} \frac{f(x)-f(a)}{x-a}$) exists. The limit is called the ***derivative*** of f at a, and is denoted by $f'(a)$, $f^{(1)}(a)$, or $\frac{df}{dx}(a)$.

- "d" means the infinitesimal difference, and $f'(a)$ is the slope of a tangent line to f at $f(a)$
- If a function f is differentiable at a, then it is continuous at a (converse is not true, as evidenced by $f(x) = |x|$ and $a = 0$)
- f is said to be differentiable iff it is differentiable at any point of its domain
- f is said to be ***continuously differentiable*** iff f is differentiable and f' is continuous

Derivative (2/2)

- If f is differentiable, we can think of f' as a function too (although may not be continuous/differentiable)
  - E.g., given $f(x) = e^x$, we have $f'(x) = \lim_{\delta\to 0}\frac{e^{(x+\delta)}-e^x}{\delta} = \lim_{\delta\to 0}\frac{e^x(e^\delta - 1)}{\delta}$. Let $t = e^\delta - 1$, then $f'(x) = \lim_{t\to 0}\frac{e^x t}{\ln(1+t)} = e^x \lim_{t\to 0}\frac{1}{\ln(1+t)^{1/t}} = e^x \frac{1}{\ln\left(\lim_{t\to 0}(1+t)^{1/t}\right)} = e^x \frac{1}{\ln e} = e^x$
- $f \in \mathcal{C}^n$ denotes that f is n-times continuously differentiable

Rolle's and Mean Value Theorem

Theorem (Rolle's Theorem)

Given a function $f : [s,t] \to \mathbb{R}$, where $s,t \in \mathbb{R}$, $f \in \mathcal{C}^1$, and $f(s) = f(t)$. There exists some $u \in (s,t)$ such that $f'(u) = 0$.

- Starting from $f(s)$, f must change its direction at some point when it is getting to $f(t)$
- We can "rotate" the above theorem to get a new one:

Theorem (Mean Value Theorem)

Given a function $f : [s,t] \to \mathbb{R}$, where $s,t \in \mathbb{R}$ and $f \in \mathcal{C}^1$. There exists some $u \in (s,t)$ such that $f'(u) = \frac{f(t)-f(s)}{t-s}$.

Partial Derivative for Multivariate Functions

Definition (Partial Derivative)

The ***partial derivative*** of a function $f : \mathcal{V} \to \mathbb{R}$, $\mathcal{V} \subseteq \mathbb{R}^n$, in the direction along the ith component at point $\boldsymbol{a}$ is defined as $\lim_{\delta \to 0} \frac{f(a_1, \cdots, a_i + \delta, \cdots, a_n) - f(a_1, \cdots, a_n)}{\delta}$, denoted by $\frac{\partial f}{\partial x_i}(\boldsymbol{a})$.

- We can only look at one component a time
- Given $f(x_1, x_2) = (x_1 + x_2)^2$, we have $\frac{\partial f}{\partial x_1}(x_1, x_2) = \lim_{\delta \to 0} \frac{(x_1 + \delta + x_2)^2 - (x_1 + x_2)^2}{\delta} = \lim_{\delta \to 0} \frac{\delta(2x_1 + 2x_2)}{\delta} = 2x_1 + 2x_2$
  - Simply treat x_2 as constant here

Integral

- Given a function $f : \mathcal{V} \to \mathbb{R}$, $\mathcal{V} \subseteq \mathbb{R}^n$, the set $\{[\boldsymbol{x}^\top, f(\boldsymbol{x})]^\top : \boldsymbol{x} \in \mathcal{V}\}$ is called the ***graph*** of f
- Now consider a function $f : [s,t] \to \mathbb{R}$, $s,t \in \mathbb{R}$, how do you approximate the area between the curve $y = f(x)$ and the x-axis in the graph of f?
  1. Partition $[s,t]$ evenly into n segments of width $\frac{t-s}{n}$, and let h_i (or l_i), $1 \leqslant i \leqslant n$, be the highest (or lowest) value of f in each segment
  2. We can approximate the area by $H(n) = \sum_{i=1}^n h_i(\frac{t-s}{n})$ (or $L(n) = \sum_{i=1}^n l_i(\frac{t-s}{n})$)
- The larger the value of n, the more precise the approximation

Definition (Integral)

A function $f : [s,t] \to \mathbb{R}$, $s,t \in \mathbb{R}$, is ***integrable*** iff both $\lim_{n\to\infty} H(n)$ and $\lim_{n\to\infty} L(n)$ exist and are equal to each other. The limit is called the ***integral*** of f, denoted by $\int_s^t f(x)dx$.

Fundamental Theorem of Calculus

Theorem (Fundamental Theorem of Calculus)

Given a function $f : [s, t] \to \mathbb{R}$, $s, t \in \mathbb{R}$. *We have:*
a) The function $F : [s, t] \to \mathbb{R}$, $F(x) = \int_s^x f(z)dz$, *is differentiable and* $\frac{dF}{dx}(x) = \frac{d}{dx}\int_s^x f(z)dz = f(x)$;
b) If there exists a differentiable function $G : [s', t'] \to \mathbb{R}$, $[s, t] \subseteq [s', t']$, *such that* $\frac{dG}{dx}(x) = f(x)$ *for every* $x \in [s, t]$, *then* $\int_s^t f(x)dx = G(t) - G(s)$.

- F is called the ***indefinite integral*** (or ***antiderivative***) of f and is a function of "accumulated area" from s
  - F can be "reversed" by differentiation and f is the "rate of accumulation"
- $\int_s^t f(x)dx$ is called the ***definite integral*** formally and represents an area
  - Definite integral can be computed by using indefinite integrals (which are usually easier to get)

Outline

Isn't It Straightforward?

- Recall that the derivative of a real-valued function f at a is defined as $f'(a) = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}$
  - This is not applicable to $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$, as we cannot divide vectors
- We need a more general definition where the vectors can be fitted in with
  - Note that $f'(a) = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}$ iff $0 = \lim_{x\to a} \frac{f(x)-f(a)}{x-a} - f'(a) = \lim_{x\to a} \frac{f(x)-f(a)-f'(a)(x-a)}{x-a}$
  - Since the limit equals 0, the sign of numerator and denominator at the right hand side does not matter; that is, the above equation is equivalent to $0 = \lim_{x\to a} \frac{|f(x)-f(a)-f'(a)(x-a)|}{|x-a|} = \lim_{x\to a} \frac{|f(x)-(f'(a)(x-a)+f(a))|}{|x-a|}$
  - Now we can replace $|\cdot|$ with a vector norm $\|\cdot\|$
- But what does it mean?

Isn't It Straightforward?

- Recall that the derivative of a real-valued function f at a is defined as $f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$
    - This is not applicable to $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$, as we cannot divide vectors
- We need a more general definition where the vectors can be fitted in with
    - Note that $f'(a) = \lim_{x \to a} \frac{f(x)-f(a)}{x-a}$ iff $0 = \lim_{x \to a} \frac{f(x)-f(a)}{x-a} - f'(a) = \lim_{x \to a} \frac{f(x)-f(a)-f'(a)(x-a)}{x-a}$
    - Since the limit equals 0, the sign of numerator and denominator at the right hand side does not matter; that is, the above equation is equivalent to $0 = \lim_{x \to a} \frac{|f(x)-f(a)-f'(a)(x-a)|}{|x-a|} = \lim_{x \to a} \frac{|f(x)-(f'(a)(x-a)+f(a))|}{|x-a|}$
    - Now we can replace $|\cdot|$ with a vector norm $\|\cdot\|$
- But what does it mean? In the graph of f, $f'(a)(x-a)+f(a)$ is a tangent line to f at $f(a)$

Derivative of Vector-Valued Functions

- The notion of a tangent line can be generalized into an ***affine function*** $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + \boldsymbol{c}$, where $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation and $\boldsymbol{c} \in \mathbb{R}^m$
  - An affine function is a "point" in an affine space

Theorem (Derivative)

A function $\boldsymbol{f} : \mathcal{V} \to \mathbb{R}^m$, $\mathcal{V} \subseteq \mathbb{R}^n$, is ***differentiable*** *at $\boldsymbol{a} \in \mathcal{V}$ iff there exists a linear transformation $\mathcal{L}(\boldsymbol{a}) : \mathbb{R}^n \to \mathbb{R}^m$ such that $\lim_{\boldsymbol{\delta} \to \boldsymbol{0}} \frac{\|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{\delta}) - (\mathcal{L}(\boldsymbol{a})(\boldsymbol{\delta}) + \boldsymbol{f}(\boldsymbol{a}))\|}{\|\boldsymbol{\delta}\|} = 0$ (or equivalently, $\lim_{\boldsymbol{x} \to \boldsymbol{a}} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - (\mathcal{L}(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a}) + \boldsymbol{f}(\boldsymbol{a}))\|}{\|\boldsymbol{x}-\boldsymbol{a}\|} = 0$). $\mathcal{L}(\boldsymbol{a})$ is called the* ***derivative*** *of $\boldsymbol{f}$ at $\boldsymbol{a}$, denoted by $\boldsymbol{f}'(\boldsymbol{a})$.*

- Since $\boldsymbol{f}'(\boldsymbol{a})$ is linear, it can be represented by a matrix $\boldsymbol{J}_{\boldsymbol{a}}$
  - How does $\boldsymbol{J}_{\boldsymbol{a}}$ look like?

Jacobian Matrices (1/2)

- Any function $\boldsymbol{f}:\mathcal{V}\to\mathbb{R}^m$, $\mathcal{V}\subseteq\mathbb{R}^n$ and $\boldsymbol{f}(\boldsymbol{v})=\boldsymbol{w}$ can be rewritten as:

$$\left(\begin{array}{c} f_1(v_1,\cdots,v_n)=w_1 \\ \vdots \\ f_m(v_1,\cdots,v_n)=w_m \end{array}\right)$$

- If $\boldsymbol{f}$ is linear, each real-valued f_i, $1\leqslant i\leqslant m$, can be represented by a row vector $\boldsymbol{j}_i\in\mathbb{R}^n$ such that $\boldsymbol{j}_i^\top\boldsymbol{v}=w_i$ (Remember the system of linear equations?)
  - Let $\boldsymbol{J}_{\boldsymbol{a}}=\begin{bmatrix}\boldsymbol{j}_1^\top \\ \vdots \\ \boldsymbol{j}_m^\top\end{bmatrix}$ be the matrix representation of $\boldsymbol{f}'(\boldsymbol{a})$

Jacobian Matrices (2/2)

- Let $\boldsymbol{\delta} = \delta \boldsymbol{e}_j$, where $1 \leqslant j \leqslant n$ and $\delta \in \mathbb{R}$
- Looking at the definition $\lim_{\boldsymbol{\delta} \to \boldsymbol{0}} \frac{\|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{\delta})-(\mathcal{L}(\boldsymbol{a})(\boldsymbol{\delta})+\boldsymbol{f}(\boldsymbol{a}))\|}{\|\boldsymbol{\delta}\|} = 0$ row-by-row, we have $\lim_{\delta \to 0} \frac{f_i(\boldsymbol{a}+\delta \boldsymbol{e}_j)-(\delta \boldsymbol{j}_i^\top \boldsymbol{e}_j + f_i(\boldsymbol{a}))}{\delta} = 0$ for each i and j
  - This implies that $\lim_{\delta \to 0} \frac{f_i(\boldsymbol{a}+\delta \boldsymbol{e}_j)-f_i(\boldsymbol{a})}{\delta} = \boldsymbol{j}_i^\top \boldsymbol{e}_j$
  - The right hand side denotes the element of $\boldsymbol{J}_{\boldsymbol{a}}$ at the ith row and the jth column
  - The left hand side is the partial derivative $\frac{\partial f_i}{\partial x_j}(\boldsymbol{a})$ by definition
- $\boldsymbol{J}_{\boldsymbol{a}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{a}) \end{bmatrix}$ is called the ***Jacobian matrix*** (or ***derivative matrix***) of $\boldsymbol{f}$ at $\boldsymbol{a}$

Gradient (1/2)

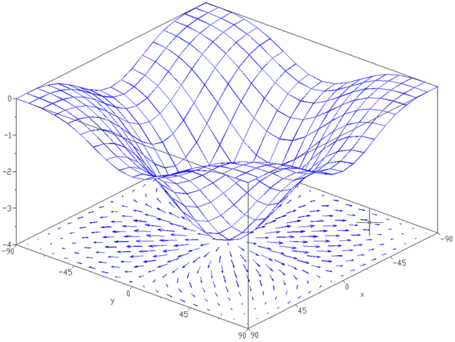
Definition

If a function $f : \mathcal{V} \to \mathbb{R}$, $\mathcal{V} \subseteq \mathbb{R}^n$, is differentiable, then the ***gradient*** of f is defined by $\nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \cdots, \frac{\partial f}{\partial x_n}(\boldsymbol{x})\right]^\top = f'(\boldsymbol{x})^\top$.

- The Jacobian matrix of $\boldsymbol{f}$ at $\boldsymbol{a}$ can be rewritten as $\boldsymbol{J_a} = \begin{bmatrix} \nabla f_1(\boldsymbol{a})^\top \\ \vdots \\ \nabla f_m(\boldsymbol{a})^\top \end{bmatrix}$
- From the previous page, we can see that in the graph of f_i, $\{\boldsymbol{x} : \nabla f_i(\boldsymbol{a})^\top(\boldsymbol{x} - \boldsymbol{a}) + f_i(\boldsymbol{a})\}$ is a tangent hyperplane to f_i at $f_i(\boldsymbol{a})$
  - The norm of $\nabla f_i(\boldsymbol{a})$ is the "slope" of the tangent hyperplane
  - $\nabla f_i(\boldsymbol{a})$ also acts as the direction that for a given small displacement from $\boldsymbol{a}$, f_i increase more in the direction of $\nabla f_i(\boldsymbol{a})$ than in any other direction [Proof]

Gradient (2/2)

- The gradient $\nabla f(\boldsymbol{x})$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$ and can be pictured as a ***vector field***
  - Each vector in the field is the direction of maximum rate of increase of f
  - E.g., consider $f(x_1, x_2) = -(\cos^2 x_1 + \cos^2 x_2)^2$:



Hessian Matrices (1/2)

Definition (Hessian Matrix)

Given a differentiable function $f : \mathcal{V} \to \mathbb{R}$, $\mathcal{V} \subseteq \mathbb{R}^n$. If $\nabla f(\boldsymbol{x})$ is differentiable, we have the derivative of $\nabla f(\boldsymbol{x})$ as:

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

which is called the ***Hessian matrix*** of f at $\boldsymbol{x}$.

- $\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})$ means taking the partial derivative of f in the direction x_j first, and then x_i

Hessian Matrices (2/2)

Theorem (Clairaut's/Schwarz's Theorem)

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable at $\mathbf{x}$, then its Hessian matrix at $\mathbf{x}$ is symmetric.

- If the second partial derivatives of f is not continuous, then there is no such a guarantee
- Here is an example:

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{(x_1^2 + x_2^2)}, & [x_1, x_2]^\top \neq [0,0]^\top \\ 0, & [x_1, x_2]^\top = [0,0]^\top. \end{cases}$$

The Hessian matrix of f at the point $[0,0]^\top$ is not symmetric [Homework]

Outline

Differentiation Rules

Theorem (Chain Rule)

Let $\boldsymbol{f} : (s,t) \to \mathcal{D}$ and $g : \mathcal{D} \to \mathbb{R}$ be differentiable functions, where $\mathcal{D} \subseteq \mathbb{R}^n$ is an open set and $s,t \in \mathbb{R}$. The composite function $g \circ \boldsymbol{f} : (s,t) \to \mathbb{R}$ is differentiable and $(g \circ \boldsymbol{f})'(x) = g'(\boldsymbol{f}(x))\boldsymbol{f}'(x) = \nabla g(\boldsymbol{f}(x))^\top \begin{bmatrix} f_1'(x) \\ \vdots \\ f_n'(x) \end{bmatrix}$.

- g' and $\boldsymbol{f}'$ are derivatives but with respect to different variables

Theorem (Product Rule)

Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable functions. Then the function $h : \mathbb{R}^n \to \mathbb{R}$, $h(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x})$, is differentiable and $h'(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}'(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{f}'(\boldsymbol{x})^\top$.

Outline

Level Sets

Definition (Level Set)

The ***level set*** of a function $f : \mathbb{R}^n \to \mathbb{R}$ at level $c \in \mathbb{R}$ is the set of points $S = \{\mathbf{x} : f(\mathbf{x}) = c\}$.

- When $n = 2$, S is a plane curve
- When $n = 3$, S is a surface

Level Sets and Gradients

Theorem

The vector $\nabla f(\boldsymbol{a})$ is orthogonal to the tangent vector to an arbitrary curve passing through $\boldsymbol{a}$ on a level set at level $f(\boldsymbol{a})$.

Proof.

Let S be the level set at level $f(\boldsymbol{a})$ and $\boldsymbol{\gamma}:\mathbb{R}\to\mathbb{R}^n$ be a curve lying on S passing through $\boldsymbol{a}$; that is, there exists c such that $\boldsymbol{\gamma}(c)=\boldsymbol{a}$. Suppose $\boldsymbol{\gamma}'(c)=\boldsymbol{v}\neq\boldsymbol{0}$ so $\boldsymbol{v}$ is a tangent vector to $\boldsymbol{\gamma}$ at $\boldsymbol{a}$. We have $(f\circ\boldsymbol{\gamma})'(c)=f'(\boldsymbol{\gamma}(c))\boldsymbol{\gamma}'(c)=f'(\boldsymbol{a})\boldsymbol{v}$. Since $\boldsymbol{\gamma}$ lies on S, we have $f\circ\boldsymbol{\gamma}(t)=f(\boldsymbol{a})$ for all $t\in\mathbb{R}$. The function $f\circ\boldsymbol{\gamma}$ is a constant, implying that $(f\circ\boldsymbol{\gamma})'(c)=0$. So $f'(\boldsymbol{a})\boldsymbol{v}=\nabla f(\boldsymbol{a})^\top\boldsymbol{v}=0$. $\square$

Outline

Taylor's Theorem (1/2)

Theorem

Given a function $f:[a,b]\to\mathbb{R}$ *and* $f\in\mathcal{C}^m$. *We have*
$f(b)=f(a)+\frac{f^{(1)}(a)}{1!}(b-a)+\frac{f^{(2)}(a)}{2!}(b-a)^2+\cdots+\frac{f^{(m-1)}(a)}{(m-1)!}(b-a)^{m-1}+R_m,$
where $R_m=\frac{f^{(m)}(c)}{m!}(b-a)^m$ *for some* $c\in(a,b)$.

Proof.

Define $R(x)=$
$f(b)-f(x)-\frac{f^{(1)}(x)}{1!}(b-x)-\frac{f^{(2)}(x)}{2!}(b-x)^2-\cdots-\frac{f^{(m-1)}(x)}{(m-1)!}(b-x)^{m-1}$.
We show that $R(a)=\frac{f^{(m)}(c)}{m!}(b-a)^m$ for some $c\in(a,b)$. Note that
$R^{(1)}(x)=$
$-f^{(1)}(x)+\left[f^{(1)}(x)-\frac{f^{(2)}(x)}{1!}(b-x)\right]+\left[\frac{f^{(2)}(x)}{1!}(b-x)-\frac{f^{(3)}(x)}{2!}(b-x)^2\right]+$
$\cdots+\left[\frac{f^{(m-1)}(x)}{(m-2)!}(b-x)^{m-2}-\frac{f^{(m)}(x)}{(m-1)!}(b-x)^{m-1}\right]=-\frac{f^{(m)}(x)}{(m-1)!}(b-x)^{m-1}$.
Define $g(x)=R(x)-(\frac{b-x}{b-a})^mR(a)$. It's easy to check that $g(a)=g(b)=0$.
By Rolle's theorem there exists some $c\in(a,b)$ such that $0=g^{(1)}(c)=$
$R^{(1)}(c)+\frac{m(b-c)^{m-1}}{(b-a)^m}R(a)=-\frac{f^{(m)}(c)}{(m-1)!}(b-c)^{m-1}+\frac{m(b-c)^{m-1}}{(b-a)^m}R(a)$,
implying $R(a)=\frac{f^{(m)}(c)}{m!}(b-a)^m$. □

Taylor's Theorem (2/2)

- An well-known application is Taylor series:
  - $e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ for all x (expending e^x at $a = 0$)
  - $\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$ for $|x| < 1$, which implies $\ln(1+x) \approx x$ for $|x| \ll 1$
- A function $f : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is an open interval, is said to be ***analytic*** iff for any $x, a \in D$, f can be written as $f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n$ for some $c_n \in \mathbb{R}$
  - An analytic f is easy to analyze, e.g., $c_n = \frac{f^{(n)}(a)}{n!}$
  - f is analytic iff give any x, the Taylor series at a converges to $f(x)$

Order of Convergence

- Consider $\boldsymbol{f}(\boldsymbol{x}) : \mathcal{V} \to \mathbb{R}^m$ and $\boldsymbol{g}(\boldsymbol{x}) : \mathcal{V} \to \mathbb{R}^m$, where $\mathcal{V} \subseteq \mathbb{R}^n$ includes $\mathbf{0}$
- We denote $\boldsymbol{f}(\boldsymbol{x}) = o(\boldsymbol{g}(\boldsymbol{x}))$ iff $\boldsymbol{f}(\boldsymbol{x})$ goes to $\mathbf{0}$ faster than $\boldsymbol{g}(\boldsymbol{x})$ does
  - Specifically, $\lim_{\boldsymbol{x}\to\mathbf{0}} \frac{\|\boldsymbol{f}(\boldsymbol{x})\|}{\|\boldsymbol{g}(\boldsymbol{x})\|} = 0$
- We denote $\boldsymbol{f}(\boldsymbol{x}) = O(\boldsymbol{g}(\boldsymbol{x}))$ iff $\boldsymbol{f}(\boldsymbol{x})$ goes to $\mathbf{0}$ faster than or equal to $\boldsymbol{g}(\boldsymbol{x})$ does
  - Specifically, for a sufficiently small $\delta \in \mathbb{R}$, there exists $c \in \mathbb{R}$ such that if $\|\boldsymbol{x}\| < \delta$, then $\frac{\|\boldsymbol{f}(\boldsymbol{x})\|}{\|\boldsymbol{g}(\boldsymbol{x})\|} \leqslant c$
- E.g., $x^2 = o(x)$, $x = O(x)$, $[x^3, 2x^2]^\top = o([x, 0]^\top)$
- Don't mix this up with the order of growth

Taylor Theorem for Multivariate Functions (1/2)

- Recall that a function $f : [a,b] \to \mathbb{R}$, $f \in \mathcal{C}^m$, can be written as $f(b) = f(a) + \frac{f^{(1)}(a)}{1!}(b-a) + \frac{f^{(2)}(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(m-1)}(a)}{(m-1)!}(b-a)^{m-1} + \frac{f^{(m)}(a+\delta(b-a))}{m!}(b-a)^m$, where $\delta \in (0,1)$ is a constant
- By the continuity of $f^{(m)}$, we have $\lim_{(b-a)\to 0} f^{(m)}(a+\delta(b-a)) = f^{(m)}(a) \Rightarrow \lim_{(b-a)\to 0} \frac{f^{(m)}(a+\delta(b-a)) - f^{(m)}(a)}{1} = 0$; that is, $f^{(m)}(a+\delta(b-a)) - f^{(m)}(a) = o(1) \Rightarrow f^{(m)}(a+\delta(b-a)) = f^{(m)}(a) + o(1)$
- We can rewrite f as $f(b) = f(a) + \frac{f^{(1)}(a)}{1!}(b-a) + \frac{f^{(2)}(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(m)}(a)}{m!}(b-a)^m + o((b-a)^m)$, since $o(1)\frac{(b-a)^m}{m!} = o((b-a)^m)$
- If $f \in \mathcal{C}^{m+1}$, we can further rewrite f as $f(b) = f(a) + \frac{f^{(1)}(a)}{1!}(b-a) + \frac{f^{(2)}(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(m)}(a)}{m!}(b-a)^m + O((b-a)^{m+1})$
  - $R_{m+1} = \frac{f^{(m+1)}(c)}{(m+1)!}(b-a)^{m+1} = O((b-a)^{m+1})$, as $f^{(m+1)}$ is bound on the compact set $[a,b]$ and therefore can be regarded as a constant

Taylor Theorem for Multivariate Functions (2/2)

Theorem

Given $f : \mathcal{V} \to \mathbb{R}$, where $\mathcal{V} \subseteq \mathbb{R}^n$ is an open set and $f \in \mathcal{C}^2$. For any $\boldsymbol{x}, \boldsymbol{a} \in \mathcal{V}$, there exists $\boldsymbol{c} = \boldsymbol{a} + c(\boldsymbol{x}-\boldsymbol{a})/\|\boldsymbol{x}-\boldsymbol{a}\|$ for some $c \in (0, \|\boldsymbol{x}-\boldsymbol{a}\|)$ such that $f(\boldsymbol{x}) = f(\boldsymbol{a}) + \frac{1}{1!}\nabla f(\boldsymbol{a})^\top(\boldsymbol{x}-\boldsymbol{a}) + R_2$, where $R_2 = \frac{1}{2!}(\boldsymbol{x}-\boldsymbol{a})^\top \boldsymbol{H}(\boldsymbol{c})(\boldsymbol{x}-\boldsymbol{a})$.

Proof.

Define $z : \mathbb{R} \to \mathbb{R}^n$ by $z(\delta) = \boldsymbol{a} + \delta(\boldsymbol{x}-\boldsymbol{a})/\|\boldsymbol{x}-\boldsymbol{a}\|$ and $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(\delta) = f \circ z(\delta) = f(\boldsymbol{a} + \delta(\boldsymbol{x}-\boldsymbol{a})/\|\boldsymbol{x}-\boldsymbol{a}\|)$, we can see that $f(\boldsymbol{x}) = \phi(\|\boldsymbol{x}-\boldsymbol{a}\|)$ and by Taylor's theorem,

$f(\boldsymbol{x}) = \phi(0) + \frac{\phi^{(1)}(0)}{1!}\|\boldsymbol{x}-\boldsymbol{a}\| + \frac{\phi^{(2)}(c)}{2!}\|\boldsymbol{x}-\boldsymbol{a}\|^2$. Note

$\phi^{(1)}(\delta) = f^{(1)}(z(\delta))z^{(1)}(\delta) = \nabla f(z(\delta))^\top \left(\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|}\right) = \left(\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|}\right)^\top \nabla f(z(\delta))$

and $\phi^{(2)}(\delta) = \left(\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|}\right)^\top \boldsymbol{H}(z(\delta))z^{(1)}(\delta) = \left(\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|}\right)^\top \boldsymbol{H}(z(\delta))\left(\frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|}\right)$.

Substituting $\phi^{(1)}(0)$ and $\phi^{(2)}(c)$ in $f(\boldsymbol{x})$ we have the proof. □

- We can also write $R_2 = \frac{1}{2!}(\boldsymbol{x}-\boldsymbol{a})^\top \boldsymbol{H}(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a}) + o(\|\boldsymbol{x}-\boldsymbol{a}\|^2)$ [Proof]
- Or $R_2 = \frac{1}{2!}(\boldsymbol{x}-\boldsymbol{a})^\top \boldsymbol{H}(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a}) + O(\|\boldsymbol{x}-\boldsymbol{a}\|^3)$ if $f \in \mathcal{C}^3$ [Proof]

Mean Value Theorem Revisited

Theorem (Mean Value Theorem)

Given a function $\boldsymbol{f}: \mathcal{V} \to \mathbb{R}^m$, *where* $\mathcal{V} \subseteq \mathbb{R}^n$ *is open and* $f \in \mathcal{C}^1$. *For any* $\boldsymbol{b}, \boldsymbol{a} \in \mathcal{V}$, *there exists* $\boldsymbol{M} = \begin{bmatrix} \nabla f_1(\boldsymbol{c}^{(1)})^\top \\ \vdots \\ \nabla f_m(\boldsymbol{c}^{(m)})^\top \end{bmatrix}$ *for some* $\boldsymbol{c}^{(1)}, \cdots, \boldsymbol{c}^{(m)} \in \mathcal{V}$ *such that* $f(\boldsymbol{b}) - f(\boldsymbol{a}) = \boldsymbol{M}(\boldsymbol{b} - \boldsymbol{a})$.

- Can be easily proved by the Taylor's Theorem [Proof]

Outline

Vector Derivatives

- Recall that the derivative of a function $\boldsymbol{f}:\mathbb{R}^n \to \mathbb{R}^m$ at a point $\boldsymbol{x} \in \mathbb{R}^n$ can be written as a Jacobian matrix $\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$
- Given $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$, define $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \in \mathbb{R}^{m\times n}$ such that $\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)_{i,j} = \frac{\partial y_i}{\partial x_j}$
  - We can express the above Jacobian matrix succinctly as $\frac{\partial \boldsymbol{f}(x)}{\partial x}$
  - $\frac{\partial}{\partial x}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}$ and $\frac{\partial x}{\partial x} = \boldsymbol{I}$
- $\frac{\partial \boldsymbol{y}}{\partial x} \in \mathbb{R}^{m\times 1}$ for $x \in \mathbb{R}$; and $\frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times n}$ for $y \in \mathbb{R}$
  - $\frac{\partial}{\partial x}(\boldsymbol{a}^\top \boldsymbol{x}) = \frac{\partial}{\partial x}(\boldsymbol{x}^\top \boldsymbol{a}) = \boldsymbol{a}^\top$ for any $\boldsymbol{a} \in \mathbb{R}^n$
  - $\frac{\partial}{\partial x}(\boldsymbol{x}^\top \boldsymbol{x}) = 2\boldsymbol{x}^\top$
- Differentiation rules are applicable
  - $\frac{\partial}{\partial x}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}^\top \frac{\partial}{\partial x}(\boldsymbol{A}\boldsymbol{x}) + (\boldsymbol{A}\boldsymbol{x})^\top \left(\frac{\partial x}{\partial x}\right)^\top = \boldsymbol{x}^\top(\boldsymbol{A}+\boldsymbol{A}^\top)$ for any $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ [Proof]

Matrix Derivatives

- Given $x \in \mathbb{R}$ and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, define
  - $\frac{\partial \boldsymbol{A}}{\partial x} \in \mathbb{R}^{m \times n}$ such that $\left(\frac{\partial \boldsymbol{A}}{\partial x}\right)_{i,j} = \frac{\partial a_{i,j}}{\partial x}$
  - $\frac{\partial x}{\partial \boldsymbol{A}} \in \mathbb{R}^{m \times n}$ such that $\left(\frac{\partial x}{\partial \boldsymbol{A}}\right)_{i,j} = \frac{\partial x}{\partial a_{i,j}}$
- x should be related to $\boldsymbol{A}$ (e.g., $a_{i,j}$, $tr(\boldsymbol{A})$, or $det(\boldsymbol{A})$, etc.)
  - $\frac{\partial \boldsymbol{A}}{\partial a_{i,j}}$ is a matrix whose element at the ith row and jth column equals 1, and others 0
- Although looked similar to vector derivatives, matrix derivatives have no obvious geometric implications and are used mainly to simplify the calculation of partial derivatives
- $\frac{\partial}{\partial x}(\boldsymbol{AB}) = \boldsymbol{A}\frac{\partial \boldsymbol{B}}{\partial x} + \frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{B}$ [Proof]
  - $\frac{\partial}{\partial x}(\boldsymbol{A}^{-1}) = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$ [Proof: $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ and apply the above]

Outline

Derivatives of Traces

- $\frac{\partial}{\partial \boldsymbol{A}} tr(\boldsymbol{AB}) = \frac{\partial}{\partial \boldsymbol{A}} tr(\boldsymbol{BA}) = \boldsymbol{B}^\top$, as $\frac{\partial}{\partial a_{i,j}} tr(\boldsymbol{AB}) = \frac{\partial}{\partial a_{i,j}} \sum_{r=1}^{n} \sum_{s=1}^{n} a_{r,s} b_{s,r} = b_{j,i}$
  - $\frac{\partial}{\partial \boldsymbol{A}} tr(\boldsymbol{A}) = \frac{\partial}{\partial \boldsymbol{A}} tr(\boldsymbol{AI}) = \boldsymbol{I}$
- $\frac{\partial}{\partial \boldsymbol{A}} tr(\boldsymbol{A}^\top \boldsymbol{B}) = \boldsymbol{B}$ [Proof]
- $\frac{\partial}{\partial \boldsymbol{A}} tr(\boldsymbol{ABA}^\top) = \boldsymbol{A}(\boldsymbol{B} + \boldsymbol{B}^\top)$ [Proof]

Derivatives of Determinants (1/2)

Theorem

Given an invertible matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *and* $x \in \mathbb{R}$, *we have* $\frac{\partial}{\partial x}\ln(det(\boldsymbol{A})) = tr(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x})$.

Proof.

We only proof the case where $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ is symmetric here. We have $\frac{\partial}{\partial x}\ln(det(\boldsymbol{A})) = \frac{\partial}{\partial x}\ln(det(\boldsymbol{U})det(\boldsymbol{D})det(\boldsymbol{U})^{-1}) = \frac{\partial}{\partial x}\ln(det(\boldsymbol{D})) = \frac{\partial}{\partial x}\ln(\prod_{i=1}^{n}\lambda_i) = \sum_{i=1}^{n}\frac{\partial}{\partial x}\ln\lambda_i = \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x} = tr(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial x})$. Note $\boldsymbol{U}$ is orthogonal, and diagonalizable, so there exists an antisymmetric matrix $\boldsymbol{W} = \frac{1}{x}\ln\boldsymbol{U}$ such that $\boldsymbol{U} = e^{\boldsymbol{W}x}$. By the chain rule we have $\frac{\partial \boldsymbol{U}}{\partial x} = e^{\boldsymbol{W}x}(\frac{\partial}{\partial x}\boldsymbol{W}x) = \boldsymbol{U}\boldsymbol{W}$ and $\frac{\partial \boldsymbol{U}^\top}{\partial x} = \frac{\partial}{\partial x}e^{\boldsymbol{W}^\top x} = \frac{\partial}{\partial x}e^{-\boldsymbol{W}x} = -\boldsymbol{U}^\top\boldsymbol{W}$.
Therefore, $tr(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial x}) = tr((\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U})^{-1}\frac{\partial}{\partial x}(\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U})) =$
$tr((\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U})^{-1}(\frac{\partial \boldsymbol{U}^\top}{\partial x}\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}^\top\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{U} + \boldsymbol{U}^\top\boldsymbol{A}\frac{\partial \boldsymbol{U}}{\partial x})))=$
$tr((\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U})^{-1}(-\boldsymbol{U}^\top\boldsymbol{W}\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}^\top\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{U} + \boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U}\boldsymbol{W}) =$
$tr(-\boldsymbol{U}^\top\boldsymbol{A}^{-1}\boldsymbol{W}\boldsymbol{A}\boldsymbol{U}) + tr(\boldsymbol{U}^\top\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{U}) + tr(\boldsymbol{W})$, which can be simplified to $tr(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x})$ by the cyclic property of trace. $\square$

Derivatives of Determinants (2/2)

- $\frac{\partial}{\partial \boldsymbol{A}} \ln(det(\boldsymbol{A})) = (\boldsymbol{A}^{-1})^\top$
  - Let $a_{i,j}$ and $b_{i,j}$ be the elements of $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$ respectively, then $\frac{\partial}{\partial a_{i,j}} \ln(det(\boldsymbol{A})) = tr(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial a_{i,j}}) = \sum_{r=1}^{n} \sum_{s=1}^{n} b_{r,s} \frac{\partial a_{s,r}}{\partial a_{i,j}} = b_{j,i}$

Outline

Functionals

- Consider a function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = ax + b$ (or $f(x|a, b) = ax + b$)
  - x is an ***argument*** and a and b are parameters
- Let $\mathcal{S}$ be the set of functions $f : \mathcal{V} \to \mathcal{W}$, we can define a ***functional*** $F : \mathcal{S} \to \mathcal{W}$, $F[f]$, with f as the argument
- E.g., value of a function $f : \mathbb{R} \to \mathbb{R}$ at x: $F : \mathcal{S} \to \mathbb{R}$, $F[f] = f(x)$
  - x is a ***parameter***
  - We can write $F[f]$ as $F[f|x]$
- E.g., definite integral of a function $f : \mathbb{R} \to \mathbb{R}$: $I : \mathcal{S} \to \mathbb{R}$, $I[f] = \int_a^b f(x)dx$
  - a and b are parameters
- E.g., expectation of $f : \mathbb{R} \to \mathbb{R}$ defined over the values of a random variable X: $E : \mathcal{S} \to \mathbb{R}$, $E[f(X)] = \int f(x)p_X(x)dx$