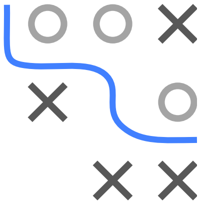
Optimization in Machine Learning



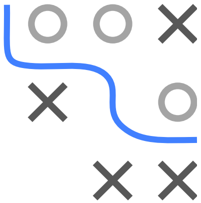
Mathematical Concepts
Matrix Calculus

δ

Learning goals

- Rules of matrix calculus
- Connection of gradient, Jacobian and Hessian

SCOPE



- $\mathcal{X}/\mathcal{Y}$ denote space of **independent/dependent** variables
- Identify dependent variable y with a **function**
  $f : \mathcal{X} \to \mathcal{Y}, x \mapsto f(x)$
- Assume y sufficiently smooth
- In matrix calculus, x and y can be **scalars**, **vectors**, or **matrices**
- We denote vectors/matrices in **bold** lowercase/uppercase letters

Type	scalar x	vector $\mathbf{x}$	matrix $\mathbf{X}$
scalar y	dy/dx	$dy/d\mathbf{x}$	$dy/d\mathbf{X}$
vector $\mathbf{y}$	$d\mathbf{y}/dx$	$d\mathbf{y}/d\mathbf{x}$	–
matrix $\mathbf{Y}$	$d\mathbf{Y}/dx$	–	–

- This notation is also referred to as *Leibniz notation*

LEIBNIZ NOTATION CONVENTION

- Instead of writing $f(x)$ everywhere, we replace the function f with the variable y.
- This helps clarify relationships when multiple functions or variables are involved, especially in contexts like partial derivatives or matrix calculus.
- Also applicable to partial derivatives: For $y = f : \mathbb{R}^n \longrightarrow \mathbb{R},\ \mathbf{x} \mapsto f(\mathbf{x}) = f(x_1, x_2, \dots, x_n)$, the partial derivative of y w.r.t. x_i is dy/dx_i.
- **Examples:**
  - $y = x^3 + 5x \Longrightarrow \dfrac{dy}{dx} = 3x^2 + 5$
  - $y = x_1^2 + 3x_2 \Longrightarrow \dfrac{dy}{dx_1} = 2x_1, \quad \dfrac{dy}{dx_2} = 3$

DERIVATIVES OF SCALAR-VALUED FUNCTIONS

Type	scalar x	vector $\mathbf{x}$	matrix $\mathbf{X}$
scalar y	dy/dx	$dy/d\mathbf{x}$	$dy/d\mathbf{X}$
vector $\mathbf{y}$	$d\mathbf{y}/dx$	$d\mathbf{y}/d\mathbf{x}$	–
matrix $\mathbf{Y}$	$d\mathbf{Y}/dx$	–	–

- $dy/d\mathbf{x}$ is the gradient from the previous slide deck
- When the input is a matrix the concept remains the same, i.e. for $y = f : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R},\ \mathbf{X} \mapsto f(\mathbf{X})$

$$\frac{dy}{d\mathbf{X}} = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix} \in \mathbb{R}^{m\times n}$$

DERIVATIVES: UNIVARIATE AND JACOBIAN

Type	scalar x	vector $\mathbf{x}$	matrix $\mathbf{X}$
scalar y	dy/dx	$dy/d\mathbf{x}$	$dy/d\mathbf{X}$
vector $\mathbf{y}$	$d\mathbf{y}/dx$	$d\mathbf{y}/d\mathbf{x}$	–
matrix $\mathbf{Y}$	$d\mathbf{Y}/dx$	–	–

- dy/dx is the univariate derivative y'
- $d\mathbf{y}/d\mathbf{x}$ is the Jacobian from the previous slide deck

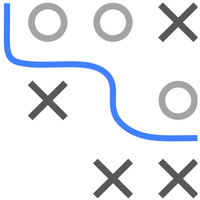
DERIV. OF FUNCTIONS WITH SCALAR INPUTS

Type	scalar x	vector $\mathbf{x}$	matrix $\mathbf{X}$
scalar y	dy/dx	$dy/d\mathbf{x}$	$dy/d\mathbf{X}$
vector $\mathbf{y}$	$d\mathbf{y}/dx$	$d\mathbf{y}/d\mathbf{x}$	–
matrix $\mathbf{Y}$	$d\mathbf{Y}/dx$	–	–

- Here, for univariate $f_{ij} : \mathbb{R} \to \mathbb{R}$, $\mathbf{y}$ ($n = 1$) or $\mathbf{Y}$ ($n > 1$) are equal to a function $f : \mathbb{R} \longrightarrow \mathbb{R}^{m \times n}$, $x \mapsto \big(f_{ij}(x)\big)_{i=1,\dots,m;\, j=1,\dots,n}$ and the derivatives are, respectively given by

$$\frac{d\mathbf{y}}{dx} = \begin{pmatrix} \frac{\partial f_1}{\partial x} \\ \vdots \\ \frac{\partial f_m}{\partial x} \end{pmatrix} \in \mathbb{R}^m; \quad \frac{d\mathbf{Y}}{dx} = \begin{pmatrix} \frac{\partial f_{11}}{\partial x} & \dots & \frac{\partial f_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{m1}}{\partial x} & \dots & \frac{\partial f_{mn}}{\partial x} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

MULTIVARIATE DIFFERENTIATION RULES



- Basic rules from single-variable calculus still apply.
- But, for $\boldsymbol{x} \in \mathbb{R}^n$: gradients are vectors/matrices (order matters).

Key Rules:

- **Sum:**

$$\frac{d}{d\boldsymbol{x}}\big(f + g\big) = \frac{df}{d\boldsymbol{x}} + \frac{dg}{d\boldsymbol{x}}$$

- **Product:**

$$\frac{d}{d\boldsymbol{x}}\big(fg\big) = \frac{df}{d\boldsymbol{x}}\, g + f\, \frac{dg}{d\boldsymbol{x}}$$

- **Chain:**

$$\frac{d}{d\boldsymbol{x}}\Big((f \circ g)(\boldsymbol{x})\Big) = \frac{d}{d\boldsymbol{x}}(f(g(\boldsymbol{x}))) = \frac{df}{dg}\,\frac{dg}{d\boldsymbol{x}}$$

DETAILS ON THE CHAIN RULE I

- Suppose
  - we have functions $\boldsymbol{g} : S \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{f} : T \subseteq \mathbb{R}^m \to \mathbb{R}^\ell$
  - $\boldsymbol{a} \in S$ is a point such that $\boldsymbol{g}(\boldsymbol{a}) \in T \Rightarrow \boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))$ is well-defined for all $\boldsymbol{x}$ close to $\boldsymbol{a}$
- Then, if $\boldsymbol{g}$ is differentiable at $\boldsymbol{a}$ and $\boldsymbol{f}$ is differentiable at $\boldsymbol{g}(\boldsymbol{a})$

  $\Rightarrow \boldsymbol{f} \circ \boldsymbol{g}$ is differentiable at $\boldsymbol{a}$, and the derivative $\frac{d\boldsymbol{f}}{d\boldsymbol{g}}\frac{d\boldsymbol{g}}{d\boldsymbol{x}}$, is equal to

$$\nabla_{\boldsymbol{a}}\boldsymbol{f} \circ \boldsymbol{g} \mathrel{\hat{=}} \boldsymbol{J}_{\boldsymbol{f}\circ\boldsymbol{g}}(\boldsymbol{a}) = \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{g}(\boldsymbol{a}))\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{a}) \mathrel{\hat{=}} \nabla_{\boldsymbol{g}(\boldsymbol{a})}\boldsymbol{f}\ \nabla_{\boldsymbol{a}}\boldsymbol{g} \in \mathbb{R}^{l\times n}$$

  (See Chapter 2.3 of the UofT course *MAT237 - Multivariable Calculus* for proof)

- We can also write $\boldsymbol{f}$ as a function of $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$, and $\boldsymbol{g}$ as a function of $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then for each $k = 1, \ldots, \ell$ and $j = 1, \ldots, n$

$$[\boldsymbol{J}_{\boldsymbol{f}\circ\boldsymbol{g}}(\boldsymbol{a})]_{kj} = \frac{\partial}{\partial x_j}\left(f_k \circ \boldsymbol{g}\right)(\boldsymbol{a}) = \sum_{i=1}^{m} \frac{\partial f_k}{\partial y_i}(\boldsymbol{g}(\boldsymbol{a}))\frac{\partial g_i}{\partial x_j}(\boldsymbol{a})$$

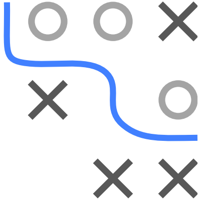
HELPFUL CALCULATION RULES

Let $\boldsymbol{a}$, $\boldsymbol{b}$ denote vectors, $\boldsymbol{X}$, $\boldsymbol{A}$ matrices, and $f(\boldsymbol{X})^{-1}$ the inverse of $f(\boldsymbol{X})$ if it exists.

- $\frac{d\boldsymbol{x}^\top \boldsymbol{a}}{d\boldsymbol{x}} = \boldsymbol{a}^\top$, $\frac{d\boldsymbol{a}^\top \boldsymbol{x}}{d\boldsymbol{x}} = \boldsymbol{a}^\top$
- $\frac{d\boldsymbol{X}\boldsymbol{a}}{d\boldsymbol{a}} = \boldsymbol{X}$, $\frac{d\boldsymbol{a}^T \boldsymbol{X}}{d\boldsymbol{a}} = \boldsymbol{X}^T$
- $\frac{d\boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}}{d\boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^\top$
- $\frac{d\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{d\boldsymbol{x}} = \boldsymbol{x}^\top \left(\boldsymbol{A} + \boldsymbol{A}^\top\right)$
- For a symmetric matrix $\boldsymbol{W}$,
  $\frac{d}{d\boldsymbol{s}}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W}\boldsymbol{A}$
- $\frac{d}{d\boldsymbol{X}}\boldsymbol{f}(\boldsymbol{X})^\top = \left(\frac{d\boldsymbol{f}(\boldsymbol{X})}{d\boldsymbol{X}}\right)^\top$
- $\frac{d}{d\boldsymbol{X}}\boldsymbol{f}(\boldsymbol{X})^{-1} = -\boldsymbol{f}(\boldsymbol{X})^{-1}\frac{d\boldsymbol{f}(\boldsymbol{X})}{d\boldsymbol{X}}\boldsymbol{f}(\boldsymbol{X})^{-1}$
- $\frac{d\boldsymbol{a}^\top \boldsymbol{X}^{-1}\boldsymbol{b}}{d\boldsymbol{X}} = -\left(\boldsymbol{X}^{-1}\right)^\top \boldsymbol{a}\boldsymbol{b}^\top \left(\boldsymbol{X}^{-1}\right)^\top$

Note: to compute gradients of matrices with respect to vectors (or other matrices) we need *tensors*, see chapter 5.4 of Deisenroth for more.

EXAMPLE: LOGISTIC REGRESSION I



- Let's say, for data in $\mathbb{R}^{n\times m}$ we're trying to minimize the risk in logistic regression by finding the gradient for negative log loss:

$$-\ell(\theta) = \sum_{i=1}^{n} -y^{(i)} \log\left(\pi\left(\mathbf{x}^{(i)} \mid \theta\right)\right) - \left(1-y^{(i)}\right)\log\left(1-\pi\left(\mathbf{x}^{(i)} \mid \theta\right)\right)$$

where $\pi(\mathbf{x} \mid \theta) = s(f(\theta, \mathbf{x}))$

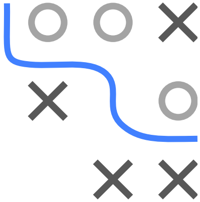
with $f(\theta, \mathbf{x}) = \theta^\top \mathbf{x}$ and $s(x) = \frac{1}{1+\exp(-x)}$.

$\Rightarrow$ We want to find

$$\nabla_\theta - \ell(\theta) = -\nabla_\theta \ell(\theta) = -\sum_{i=1}^{n} \underbrace{y^{(i)} \log(s(f(\theta, \mathbf{x}^{(i)})) + (1-y^{(i)}) \log(1 - s(f(\theta, \mathbf{x}^{(i)}))}_{=: y_i \log(s_i) + (1-y_i)\log(1-s_i)}$$

- $s_i = s(f_i)$; $h_i := -[y_i \log(s_i) + (1-y_i)\log(1-s_i)]$; $f_i := \theta^\top \mathbf{x}^{(i)}$

EXAMPLE: LOGISTIC REGRESSION II



- We can now directly apply the chain rule:

$$-\nabla_\theta \ell(\theta) = \sum_{i=1}^{n} \nabla_\theta h_i \circ s_i \circ f_i = \sum_{i=1}^{n} \frac{dh_i}{ds_i}\frac{ds_i}{df_i}\frac{df_i}{d\theta}$$

- Given that, $\forall i \in \{1, \ldots, n\}$

$$\frac{dh_i}{ds_i} = \frac{1-y_i}{1-s_i} - \frac{y_i}{s_i}; \quad \frac{ds_i}{df_i} = s(f_i)\big(1 - s_i(f_i)\big); \quad \frac{df_i}{d\theta} = \left(\mathbf{x}^{(i)}\right)^\top$$

- we get that $-\nabla_\theta \ell(\theta) = \sum_{i=1}^{n} \frac{dh_i}{ds_i}\frac{ds_i}{df_i}\frac{df_i}{d\theta}$ equals

$$\sum_{i=1}^{n} \left[\frac{1-y^{(i)}}{1-s(f(\theta,\mathbf{x}^{(i)}))} - \frac{y^{(i)}}{s(f(\theta,\mathbf{x}^{(i)}))}\right] \cdot \left[s(f(\theta,\mathbf{x}^{(i)}))(1-s(f(\theta,\mathbf{x}^{(i)})))\right] \left(\mathbf{x}^{(i)}\right)^\top$$

$$= \sum_{i=1}^{n} \left(s\left(f\left(\theta,\mathbf{x}^{(i)}\right)\right) - y^{(i)}\right)\left(\mathbf{x}^{(i)}\right)^\top \in \mathbb{R}^{1\times m}$$

EX: LOG. REGR. – ALTERNATIVE NOTATION

This example highlights how using Leibniz notation can make things easier.
Alternatively, we could write this example out as follows:

- Given the sum rule, it suffices to find the derivative of

$$-[y \log(s(f(\theta, \mathbf{x})) + (1 - y) \log(1 - s(f(\theta, \mathbf{x}))]$$

- Define $h(y, z) := -[y \log(z) + (1 - y) \log(1 - z)]$.
- Now, what we are looking for is

$$\nabla_\theta h(y, \cdot) \circ s \circ f(\cdot, \mathbf{x}).$$

  where f is a function from $\mathbb{R}^p$ to $\mathbb{R}$ and both h and s are functions from $\mathbb{R}$ to $\mathbb{R}$ $\Longrightarrow$ by the chain rule

$$\nabla_\theta h(y, \cdot) \circ s \circ f(\cdot, \mathbf{x}) = \frac{dh}{ds}\frac{ds}{df}\frac{df}{d\theta} \in \mathbb{R}^{1\times p}.$$