Optimization in Machine Learning

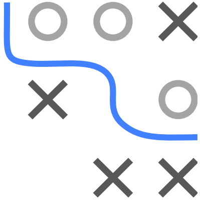
Mathematical Concepts
Matrix Calculus

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 with a **function** $y : \mathcal{X} \to \mathcal{Y}, x \mapsto y(x)$
- Assume y sufficiently smooth
- In matrix calculus, x and y can be **scalars**, **vectors**, or **matrices**:

Type	scalar x	vector $\mathbf{x}$	matrix $\mathbf{X}$
scalar y	$\partial y/\partial x$	$\partial y/\partial \mathbf{x}$	$\partial y/\partial \mathbf{X}$
vector $\mathbf{y}$	$\partial \mathbf{y}/\partial x$	$\partial \mathbf{y}/\partial \mathbf{x}$	–
matrix $\mathbf{Y}$	$\partial \mathbf{Y}/\partial x$	–	–

- We denote vectors/matrices in **bold** lowercase/uppercase letters

NUMERATOR LAYOUT

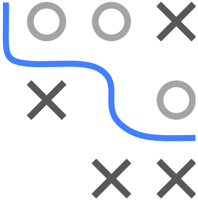
- **Matrix calculus:** collect derivative of each component of dependent variable w.r.t. each component of independent variable
- We use so-called **numerator layout** convention:

$$\frac{\partial y}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial x_1}, \cdots, \frac{\partial y}{\partial x_d}\right) = \nabla y^T \in \mathbb{R}^{1\times d}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \left(\frac{\partial y_1}{\partial x}, \cdots, \frac{\partial y_m}{\partial x}\right)^T \in \mathbb{R}^{m}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \mathbf{x}} \end{pmatrix} = \left(\frac{\partial \mathbf{y}}{\partial x_1} \cdots \frac{\partial \mathbf{y}}{\partial x_d}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_d} \end{pmatrix} = J_{\mathbf{y}} \in \mathbb{R}^{m\times d}$$

SCALAR-BY-VECTOR



Let $\mathbf{x} \in \mathbb{R}^d$, $y, z : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{A}$ be a matrix.

- If y is a **constant** function: $\frac{\partial y}{\partial \mathbf{x}} = \mathbf{0}^T \in \mathbb{R}^{1 \times d}$
- **Linearity**: $\frac{\partial (a \cdot y + z)}{\partial \mathbf{x}} = a \frac{\partial y}{\partial \mathbf{x}} + \frac{\partial z}{\partial \mathbf{x}}$ (a constant)
- **Product** rule: $\frac{\partial (y \cdot z)}{\partial \mathbf{x}} = y \frac{\partial z}{\partial \mathbf{x}} + \frac{\partial y}{\partial \mathbf{x}} z$
- **Chain** rule: $\frac{\partial g(y)}{\partial \mathbf{x}} = \frac{\partial g(y)}{\partial y} \frac{\partial y}{\partial \mathbf{x}}$ (g scalar-valued function)
- **Second** derivative: $\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^T} = \nabla^2 y^T$ ($= \nabla^2 y$ if $y \in \mathcal{C}^2$) (Hessian)
- $\frac{\partial (\mathbf{x}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T (\mathbf{A} + \mathbf{A}^T)$
- $\frac{\partial (\mathbf{y}^T \mathbf{A} \mathbf{z})}{\partial \mathbf{x}} = \mathbf{y}^T \mathbf{A} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \mathbf{z}^T \mathbf{A}^T \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ ($\mathbf{y}$, $\mathbf{z}$ vector-valued functions of $\mathbf{x}$)

VECTOR-BY-SCALAR

Let $x \in \mathbb{R}$ and $\mathbf{y}, \mathbf{z} : \mathbb{R} \to \mathbb{R}^m$.

- If $\mathbf{y}$ is a **constant** function: $\frac{\partial \mathbf{y}}{\partial x} = \mathbf{0} \in \mathbb{R}^m$
- **Linearity**: $\frac{\partial (a \cdot \mathbf{y} + \mathbf{z})}{\partial x} = a \frac{\partial \mathbf{y}}{\partial x} + \frac{\partial \mathbf{z}}{\partial x}$ (a constant)
- **Chain** rule: $\frac{\partial \mathbf{g}(\mathbf{y})}{\partial x} = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial x}$ ($\mathbf{g}$ vector-valued function)
- $\frac{\partial (\mathbf{A}\mathbf{y})}{\partial x} = \mathbf{A} \frac{\partial \mathbf{y}}{\partial x}$ ($\mathbf{A}$ matrix)

VECTOR-BY-VECTOR

Let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y}, \mathbf{z} : \mathbb{R}^d \to \mathbb{R}^m$.

- If $\mathbf{y}$ is a **constant** function: $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{0} \in \mathbb{R}^{m \times d}$
- $\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I} \in \mathbb{R}^{d \times d}$
- **Linearity**: $\frac{\partial (a \cdot \mathbf{y} + \mathbf{z})}{\partial \mathbf{x}} = a \frac{\partial \mathbf{y}}{\partial \mathbf{x}} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ (a constant)
- **Chain** rule: $\frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{x}} = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ ($\mathbf{g}$ vector-valued function)
- $\frac{\partial (\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}$, $\frac{\partial (\mathbf{x}^T \mathbf{B})}{\partial \mathbf{x}} = \mathbf{B}^T$ ($\mathbf{A}, \mathbf{B}$ matrices)

EXAMPLE

Consider $f : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(\mathbf{x}) = \exp\left(-(\mathbf{x}-\mathbf{c})^T \mathbf{A}(\mathbf{x}-\mathbf{c})\right),$$

where $\mathbf{c} = (1, 1)^T$ and $\mathbf{A} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$.

Compute $\nabla f(\mathbf{x})$ at $\mathbf{x}^* = \mathbf{0}$:

1. Write $f(\mathbf{x}) = \exp(g(\mathbf{u}(\mathbf{x})))$ with $g(\mathbf{u}) = -\mathbf{u}^T\mathbf{A}\mathbf{u}$ and $\mathbf{u}(\mathbf{x}) = \mathbf{x} - \mathbf{c}$
2. **Chain** rule: $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \exp(g(\mathbf{u}(\mathbf{x})))\frac{\partial g(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}}$
3. $\mathbf{u}^* := \mathbf{u}(\mathbf{x}^*) = (-1, -1)^T$, $g(\mathbf{u}^*) = -3$
4. $\frac{\partial g(\mathbf{u})}{\partial \mathbf{u}} = -2\mathbf{u}^T\mathbf{A}$, $\frac{\partial g(\mathbf{u}^*)}{\partial \mathbf{u}} = (3, 3)$
5. **Linearity**: $\frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}-\mathbf{c})}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}}{\partial \mathbf{x}} - \frac{\partial \mathbf{c}}{\partial \mathbf{x}} = \mathbf{I}_2$
6. $\nabla f(\mathbf{x}^*) = \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}}^T = (\exp(-3) \cdot (3, 3) \cdot \mathbf{I}_2)^T = \exp(-3)\begin{pmatrix} 3 \\ 3 \end{pmatrix}$