Optimization in Machine Learning

Mathematical Concepts:
Matrix Calculus

δ

Learning goals

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

SCOPE

- Let $\mathcal{X}$ be the space of independent variables and $\mathcal{Y}$ be the output space of dependent variables.
- A dependent variable can be identified with a function $y : \mathcal{X} \to \mathcal{Y}, x \mapsto y(x)$
- In matrix calculus, $\mathcal{X}$ and $\mathcal{Y}$ can each be a space of scalars, vectors or matrices:

Types	Scalar y	Vector $\mathbf{y}$	Matrix $\mathbf{Y}$
Scalar x	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector $\mathbf{x}$	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	/
Matrix $\mathbf{X}$	$\frac{\partial y}{\partial \mathbf{X}}$	/	/

- Here, we denote vectors and matrices in bold lowercase and bold uppercase, respectively

NUMERATOR LAYOUT

- Matrix calculus describes how we collect the derivative of each component of the dependent variable with respect to each component of the independent variable
- Here, we use the so-called numerator layout convention:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_d} \end{pmatrix} = \nabla y^\top \in \mathbb{R}^{1\times d} \text{ (transpose of the gradient)}$$

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

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{pmatrix} \in \mathbb{R}^m$$

- In the following we assume that all partial derivatives exist

DEPENDENT: SCALAR, INDEPENDENT: VECTOR

Let $x \in \mathbb{R}^d, y, u, v : \mathbb{R}^d \to \mathbb{R}, a \in \mathbb{R}, g : \mathbb{R} \to \mathbb{R}$.

Some useful rules:

- If y is a constant function: $\frac{\partial y}{\partial \mathbf{x}} = \mathbf{0}^\top \in \mathbb{R}^{1\times d}$
- Linearity: $\frac{\partial (a \cdot u + v)}{\partial \mathbf{x}} = a\frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$
- Product rule: $\frac{\partial (uv)}{\partial \mathbf{x}} = v\frac{\partial u}{\partial \mathbf{x}} + u\frac{\partial v}{\partial \mathbf{x}}$
- Chain rule: $\frac{\partial g(y)}{\partial \mathbf{x}} = \frac{\partial g(y)}{\partial y}\frac{\partial y}{\partial \mathbf{x}}$
- Second derivative: $\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \nabla^2 y^\top$ (transpose of the Hessian)
- Second derivative: $\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \nabla^2 y$ (Hessian, if PDs are continuous)
- $\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$ with constant $\mathbf{A} \in \mathbb{R}^{d\times d}$
- $\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^\top \mathbf{A}$ with constant, symmetric $\mathbf{A} \in \mathbb{R}^{d\times d}$
- $\frac{\partial (\mathbf{z}^\top \mathbf{C} \mathbf{v})}{\partial \mathbf{x}} = \mathbf{z}^\top \mathbf{C}\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \mathbf{C}^\top \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ with constant $\mathbf{C} \in \mathbb{R}^{p\times m}$

DEPENDENT: VECTOR, INDEPENDENT: VECTOR

Let $\mathbf{x} \in \mathbb{R}^d, \mathbf{y}, \mathbf{u}, \mathbf{v} : \mathbb{R}^d \to \mathbb{R}^m, \mathbf{z} : \mathbb{R}^d \to \mathbb{R}^p, a \in \mathbb{R}, \mathbf{g} : \mathbb{R}^m \to \mathbb{R}^m$.

Some useful rules:

- 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: $\frac{\partial (a \cdot \mathbf{u} + \mathbf{v})}{\partial \mathbf{x}} = a \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}, a \in \mathbb{R}$
- Chain rule: $\frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{x}} = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$
- $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}, \frac{\partial \mathbf{x}^\top \mathbf{B}}{\partial \mathbf{x}} = \mathbf{B}^\top$ with constant $\mathbf{A} \in \mathbb{R}^{m \times d}, \mathbf{B} \in \mathbb{R}^{d \times m}$

DEPENDENT: VECTOR, INDEPENDENT: SCALAR

Let $x \in \mathbb{R}, \mathbf{y}, \mathbf{u}, \mathbf{v} : \mathbb{R}^m \to \mathbb{R}^m, a \in \mathbb{R}, \mathbf{g} : \mathbb{R} \to \mathbb{R}^m$.

Some useful rules:

- If $\mathbf{y}$ is a constant function: $\frac{\partial \mathbf{y}}{\partial x} = \mathbf{0} \in \mathbb{R}^m$
- Linearity: $\frac{\partial (a \cdot \mathbf{u} + \mathbf{v})}{\partial x} = a \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$
- Chain rule: $\frac{\partial \mathbf{g}(\mathbf{y})}{\partial x} = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial x}$
- $\frac{\partial \mathbf{A}\mathbf{y}}{\partial x} = \mathbf{A} \frac{\partial \mathbf{y}}{\partial x}$ with constant $\mathbf{A} \in \mathbb{R}^{m \times d}$

EXAMPLE

Let $f : \mathbb{R}^2 \to \mathbb{R}, \mathbf{x} \mapsto \exp\left(-(\mathbf{x}-\mathbf{c})^\top \mathbf{A}(\mathbf{x}-\mathbf{c})\right)$ with
$\mathbf{c} = (1,1)^\top, \mathbf{A} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
We want to compute the gradient of f at $\mathbf{x} = \mathbf{0}$:

1. We can write $f = \exp(g(\mathbf{u}))$ with
$u : \mathbb{R}^2 \to \mathbb{R}^2, \mathbf{x} \mapsto \mathbf{x} - \mathbf{c}, g : \mathbb{R}^2 \to \mathbb{R}, \mathbf{u} \mapsto -\mathbf{u}^\top \mathbf{A}\mathbf{u}$
2. Via the chain rule it follows that $\frac{\partial f}{\partial \mathbf{x}} = \exp(g(\mathbf{u}))\frac{\partial g}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$
3. $\frac{\partial g}{\partial \mathbf{u}} = -2\mathbf{u}^\top \mathbf{A} = -(-1 \; -1)\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = (3,3)$
4. From linearity it follows that $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}-\mathbf{c})}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}}{\partial \mathbf{x}} - \frac{\partial \mathbf{c}}{\partial \mathbf{x}} = \mathbf{I} - \mathbf{0}$
5. $\nabla f(0,0) = \frac{\partial f}{\partial \mathbf{x}}(0,0)^\top = f(0,0) \cdot (3,3)^\top = \exp(-3) \cdot (3,3)^\top$