Optimization in Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Outline

Outline

1. Course overview
2. Optimization in machine learning
3. Linear algebra
4. Analysis
5. Matrix Calculus

Course setup

Grading Policy:

- Homework, 40%
- Final project, 60%

Website:

- 超星泛雅: `https://ecnu.fanya.chaoxing.com`

Notice:

- Using LLM for homework and project is not allowed
- Plagiarism is prohibited

What can I learn in this course?

Traditional convex optimization problems such as linear programming.
NOT HERE: we mainly study optimization problems in machine learning.

In practice, libraries are available, algorithms are treated as “black box”.
NOT HERE: we look inside the optimization algorithms and try to understand why and how fast they work.

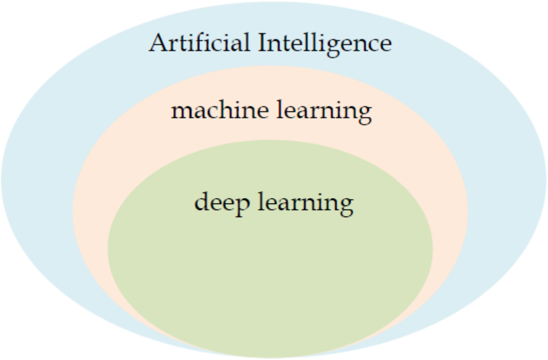
Prerequisite course: calculus, linear algebra, probability, Python/Matlab.

It would be better if you have learnt: machine learning, convex optimization.

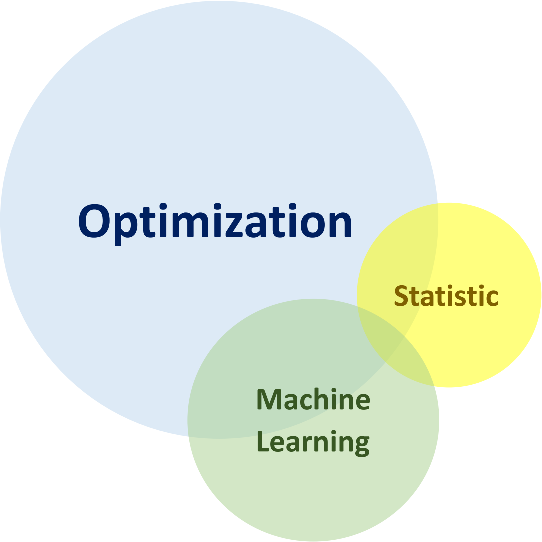
Outline

What is machine learning?

Machine learning studies how to empower computers to automatically improve their own abilities by utilizing data.



What is optimization?



Why is optimization important?

Pedro Domingos (AAAI Fellow, Prof. of UW):

Machine Learning = Representation + Evaluation + Optimization

History of optimization

- 1847: Cauchy proposes gradient descent
- 1950s: Linear Programming, soon followed by non-linear, Stochastic Gradient Descent (SGD)
- 1980s: General optimization, convergence theory
- 2005-2015: Large scale optimization (mostly convex) for machine learning
- 2015-today: Optimization methods for deep learning

Optimization problems

General optimization problem:

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

- $\mathcal{X} \subseteq \mathbb{R}^d$: feasible set
- f: objective function
- usually f is continuous in machine learning problems

Classifications of optimization problems in machine learning

The description of the feasible set:

- unconstrained vs. constrained

The properties of the objective function:

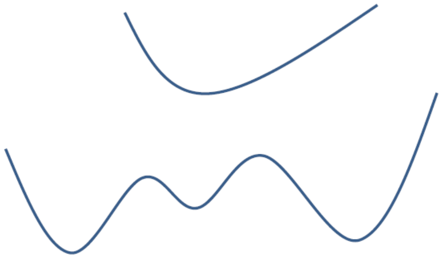
- linear vs. nonlinear
- smooth vs. nonsmooth
- convex vs. nonconvex

The settings in real application:

- deterministic vs. stochastic
- non-distributed vs. distributed

Convex vs. Nonconvex

"In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity." by R. T. Rockfeller



No-free-lunch theorem for optimization

D. H. Wolpert and W. G. Macready (1997):

- There is no universally better algorithms exist.
- If algorithm A performs better than algorithm B for some optimization functions, then B will outperform A for other functions.
- If averaged over all possible function space, both algorithms A and B will perform on average equally well.

Optimization in machine learning

The most common optimization problem in machine learning is the empirical risk minimization:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}; \mathbf{a}_i, b_i) + \lambda R(\mathbf{x}), \quad \lambda \geq 0.$$

where $\mathbf{a}_i$ is the data point, b_i is the corresponding label and $\mathbf{x}$ is the parameter of the model.

$R(x)$ is called the regularization term.

Loss functions

Some traditional loss functions:

- squared loss (least square regression):

$$\ell(\mathbf{x}; \mathbf{a}_i, b_i) = (\mathbf{a}_i^\top \mathbf{x} - b_i)^2$$

- hinge loss (support vector machine):

$$\ell(\mathbf{x}; \mathbf{a}_i, b_i) = \max\{1 - b_i \mathbf{a}_i^\top \mathbf{x}, 0\}$$

- logistic loss (logistic regression):

$$\ell(\mathbf{x}; \mathbf{a}_i, b_i) = \ln(1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x}))$$

- neural network (deep learning): highly nonconvex

Regularization terms

Some popular regularization terms:

- Ridge regularization:
$$R(\mathbf{x}) \triangleq \|\mathbf{x}\|_2^2$$
- Lasso regularization:
$$R(\mathbf{x}) \triangleq \|\mathbf{x}\|_1$$

Outline

Notations

We use x_i to denote the entry of the n-dimensional vector $\mathbf{x}$ such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

We use a_{ij} to denote the entry of matrix $\mathbf{A}$ with dimension $m \times n$ such that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Vector norms

A **norm** of a vector $\mathbf{x} \in \mathbb{R}^n$ written by $\|\mathbf{x}\|$, is informally **a measure of the length of the vector**. For example, we have the commonly-used Euclidean norm (or ℓ_2 norm),

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Formally, a norm is any function $\mathbb{R}^n \to \mathbb{R}$ that satisfies four properties:

1. For all $\mathbf{x} \in \mathbb{R}^n$, we have $\|\mathbf{x}\| \geq 0$ (non-negativity).
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (definiteness).
3. For all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have $\|t\mathbf{x}\| = |t| \, \|\mathbf{x}\|$ (homogeneity).
4. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

Vector norms

Some examples for $\mathbf{x} \in \mathbb{R}^n$:

- The ℓ_1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
- The ℓ_2-norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- The ℓ_p-norm: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p \geq 1$
- The ℓ_∞-norm: $\|\mathbf{x}\|_\infty = \max_i |x_i|$

Vector inner product

The inner product on $\mathbb{R}^n$ is given by:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

We have following properties:

- $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|_2^2$
- $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ (Cauchy–Schwarz inequality)

Matrix norms

General matrix norm is any function $\mathbb{R}^{m \times n} \to \mathbb{R}$ that satisfies:

1. For all $\mathbf{A} \in \mathbb{R}^{m \times n}$, we have $\|\mathbf{A}\| \geq 0$ (non-negativity).
2. $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$ (definiteness).
3. For all $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $t \in \mathbb{R}$, we have $\|t\mathbf{A}\| = |t| \, \|\mathbf{A}\|$ (homogeneity).
4. For all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, we have $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ (triangle inequality).

Frobenius norm of $m \times n$ matrix $\mathbf{A}$:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j}^2}$$

Induced matrix norms

Given vector norm $\|\cdot\|$, the corresponding **induced matrix norm** of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| .$$

For example, we define

$$\|\mathbf{A}\|_1 = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1$$

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$$

$$\|\mathbf{A}\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty .$$

Some properties of norms

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m \times m}$, we have

$$\|\mathbf{QA}\|_2 = \|\mathbf{A}\|_2, \qquad \|\mathbf{QA}\|_F = \|\mathbf{A}\|_F.$$

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we have

$$\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}^\top \mathbf{A})} = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^\top)}$$

Symmetric eigenvalue decomposition

The eigenvalue decomposition (EVD) of a **symmetric** matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top,$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is orthogonal and $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$ is a diagonal matrix, i.e., $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ where λ_i are eigenvalues of $\mathbf{A}$.

Usually we order the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We use $\lambda_i(\mathbf{A})$ to denote the i-th largest eigenvalue of $\mathbf{A}$.

Singular value decomposition

The singular value decomposition (SVD) of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ is orthogonal, $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is rectangular diagonal matrix with non-negative real numbers on the diagonal and $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal.

Usually we order the eigenvalues as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}}$. We use $\sigma_i(\mathbf{A})$ to denote the i-th largest singular value of $\mathbf{A}$.

Singular value decomposition

The term sometimes refers to the **compact SVD**, a similar decomposition

$$\mathbf{A} = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^\top$$

in which $\boldsymbol{\Sigma}_r$ is square diagonal of size $r \times r$, where $r \leq \min\{m, n\}$ is the rank of $\mathbf{A}$, and has only the non-zero singular values. In this variant, $\mathbf{U}_r$ is an $m \times r$ column orthogonal matrix and $\mathbf{V}_r$ is an $n \times r$ column orthogonal matrix such that $\mathbf{U}_r^\top \mathbf{U}_r = \mathbf{V}_r^\top \mathbf{V}_r = \mathbf{I}$.

Some properties of SVD

- Suppose the rank of $\mathbf{A}$ is r, then

$$\|\mathbf{A}\|_2 = \sigma_1, \quad \|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}.$$

- The nonzero singular values of $\mathbf{A}$ are the square roots of the nonzero eigenvalues of $\mathbf{A}^\top \mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$.
- If $\mathbf{A} = \mathbf{A}^\top$, then the singular values of $\mathbf{A}$ are the absolute values of the eigenvalues of $\mathbf{A}$.

Pseudo-inverse of general matrices

Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(\mathbf{A}) = r$. We define the pseudo-inverse of $\mathbf{A}$ as

$$\mathbf{A}^\dagger = \mathbf{V}\Sigma^{-1}\mathbf{U}^\top \in \mathbb{R}^{n \times m}.$$

Quadratic forms

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the scalar $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is called a quadratic form and we have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

We often implicitly assume that the matrices appearing in a quadratic form are symmetric.

Definiteness

1. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive definite (PD)** if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$. This is usually denoted by $\mathbf{A} \succ \mathbf{0}$.
2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive semi-definite (PSD)** if for all vectors $\mathbf{x} \in \mathbb{R}^n$ holds that $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$. This is usually denoted by $\mathbf{A} \succeq \mathbf{0}$.

Outline

Q-convergence Rates

Assume the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$. We define the sequence of errors to be

$$z_k = \|\mathbf{x}_k - \mathbf{x}^*\| .$$

We say the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$ with rate r and rate constant C if

$$\lim_{k \to +\infty} \frac{z_{k+1}}{z_k^r} = C \quad \text{for some } C \in \mathbb{R}.$$

- linear: $r = 1$, $0 < C < 1$; Q-linear
- sublinear: $r = 1$, $C = 1$;
- superlinear: $r = 1$, $C = 0$;
- quadratic: $r = 2$.

Q-convergence rates

Examples:

- $x_k = 1/k^2$
- $x_k = 10^{-k}$
- $x_k = 10^{-2^k}$

Convergence rates

Consider the example

$$x_k = \begin{cases} 1 + 2^{-k}, & \text{if } k \text{ is even,} \\ 1, & \text{if } k \text{ is odd.} \end{cases}$$

It should converge to $x^* = 1$ linearly, however,

$$\lim_{k \to +\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|}$$

does not exist!

R-convergence rates

Suppose that $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$. The sequence is said to converge R-linearly to $\mathbf{x}^*$ if there exists a sequence $\{\epsilon_k\}$ such that

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \epsilon_k$$

for all k and $\{\epsilon_k\}$ converges Q-linearly to zero.

The sequence

$$x_k = \begin{cases} 1 + 2^{-k}, & \text{if } k \text{ is even,} \\ 1, & \text{if } k \text{ is odd.} \end{cases}$$

R-linearly converges to one.

Outline

Topology in Euclidean space

- A subset $\mathcal{S}$ of $\mathbb{R}^n$ is called **open**, if for every $\mathbf{x} \in \mathcal{S}$ there exists $\delta > 0$ such that the ball $\mathcal{B}_\delta(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 \leq \delta\}$ is included in $\mathcal{S}$.
  Example: $\{x|a < x < b\}$, $\{\mathbf{x}|\mathbf{x} > 0\}$, $\{\mathbf{x}|\, \|\mathbf{x} - \mathbf{a}\| < 1\}$.
- A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called **closed**, if its complement $\mathcal{C}^{\mathrm{c}} = \mathbb{R}^n \backslash \mathcal{C}$ is open.
  Example: $\{x|a \leq x \leq b\}$, $\{\mathbf{x}|\mathbf{x} \geq 0\}$, $\{\mathbf{x}|\, \|\mathbf{x} - \mathbf{a}\| \leq 1\}$.
- A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called **bounded**, if there exists $r > 0$ such that $\|\mathbf{x}\|_2 < r$ for all $\mathbf{x} \in \mathcal{C}$.
  Example: $\{x|a \leq x < b\}$, $\{\mathbf{x}|1 > \mathbf{x} \geq 0\}$, $\{\mathbf{x}|\, \|\mathbf{x} - \mathbf{a}\| < 1\}$.
- A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called **compact**, if it is both bounded and closed.
  Example: $\{x|a \leq x \leq b\}$, $\{\mathbf{x}|1 \geq \mathbf{x} \geq 0\}$, $\{\mathbf{x}|\, \|\mathbf{x} - \mathbf{a}\| \leq 1\}$.

Topology in Euclidean space

1. The **interior** of $\mathcal{C} \in \mathbb{R}^n$ is defined as

$$\mathcal{C}^\circ = \{\mathbf{y} : \text{there exist } \varepsilon > 0 \text{ such that } \mathcal{B}_\varepsilon(\mathbf{y}) \subset \mathcal{C}\}$$

2. The **closure** of $\mathcal{C} \in \mathbb{R}^n$ is defined as

$$\overline{\mathcal{C}} = \mathbb{R}^n \backslash (\mathbb{R}^n \backslash \mathcal{C})^\circ.$$

3. The **boundary** of $\mathcal{C} \in \mathbb{R}^n$ is defined as $\overline{\mathcal{C}} \backslash \mathcal{C}^\circ$.

Derivative (导数)

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{x} \in (\mathrm{dom}\ f)^\circ$. The derivative at $\mathbf{x}$ is

$$\mathrm{D}f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

This matrix is also called Jacobian matrix.

Gradient (梯度)

When f is real-valued, i.e., $f : \mathbb{R}^n \to \mathbb{R}$, the gradient of f is:

$$\nabla f(\mathbf{x}) = \mathrm{D}f(\mathbf{x})^\top = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

Gradient of matrix functions

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$. Then the gradient of f with respect to $\mathbf{X}$ is

$$\nabla f(\mathbf{X}) = \frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Example:

$$f(\mathbf{X}) = \|\mathbf{X}\|_F^2$$

Examples

Example 1. $\mathbf{X} \in \mathbb{R}^{m\times n}$, $f(\mathbf{X}) = \|\mathbf{X}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2$

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} 2x_{11} & \cdots & 2x_{1n} \\ \vdots & \ddots & \vdots \\ 2x_{m1} & \cdots & 2x_{mn} \end{bmatrix} = 2\mathbf{X}$$

Example 2. $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} = \sum_{i=1}^n a_i x_i$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}$$

Examples (Cont.)

Example 3. $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m\times n}$, $f(\mathbf{X}) = \mathrm{tr}(\mathbf{A}^\top \mathbf{X}) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \cdot x_{ij}$

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \mathbf{A}$$

Example 4. For $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n$, we have $\dfrac{\partial\, \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial\, \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

If $\mathbf{A}$ is symmetric, we have $\dfrac{\partial\, \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial\, \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

We can find more results in the matrix cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Chain rules

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in \mathrm{dom}\ f$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(\mathbf{x}) \in (\mathrm{dom}\ g)^\circ$. Define the composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(\mathbf{z}) = g(f(\mathbf{z}))$. Then h is is differentiable at $\mathbf{x}$ and

$$\mathrm{D}h(\mathbf{x}) = \mathrm{D}(g(f(\mathbf{x})))\mathrm{D}(f(\mathbf{x})).$$

Examples:

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and $h(\mathbf{x}) = g(f(\mathbf{x}))$. Then

$$\nabla h(\mathbf{x}) = g'(f(\mathbf{x}))\nabla f(\mathbf{x}).$$

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$. Define $h : \mathbb{R}^p \to \mathbb{R}$ as $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then,

$$\nabla h(\mathbf{x}) = \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

Gradient of logistic regression

What is the gradient of the following function?

$$f(\mathbf{x}) = \ln\left(\sum_{i=1}^{m} \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)\right) \tag{1}$$

Let $h(\mathbf{y}) = \sum_{i=1}^m \exp(y_i)$, $\ell(x) = \log(x)$ and $g(\mathbf{y}) = \ell(h(\mathbf{y}))$. Then we have $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$, where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]^\top$ and $\mathbf{b} = [b_1, \ldots, b_m]^\top$. By the chain rule, we can obtain:

$$\nabla g(\mathbf{y}) = \ell'(h(\mathbf{y}))\nabla h(\mathbf{y}) = \frac{1}{\sum_{i=1}^m \exp(y_i)} \begin{bmatrix} \exp(y_1) \\ \exp(y_2) \\ \vdots \\ \exp(y_m) \end{bmatrix}.$$

Gradient of logistic regression (Cont.)

Thus we can write the gradient of $f(\mathbf{x})$ as:

$$\nabla f(\mathbf{x}) = \mathbf{A}^T \nabla g(\mathbf{A}\mathbf{x} + \mathbf{b})$$
$$= \mathbf{A}^T \frac{1}{\sum_{i=1}^{n} \exp(\mathbf{a}_i^T \mathbf{x} + b_i)} \begin{bmatrix} \exp(\mathbf{a}_1^T \mathbf{x} + b_1) \\ \exp(\mathbf{a}_2^T \mathbf{x} + b_2) \\ \vdots \\ \exp(\mathbf{a}_m^T \mathbf{x} + b_n) \end{bmatrix} = \frac{1}{\mathbf{1}^\top \mathbf{z}} \mathbf{A}^\top \mathbf{z}.$$

where $z_i = \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)$.

The Hessian matrix

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function that takes as input a matrix $\mathbf{x} \in \mathbb{R}^n$ and returns a real value. Then the Hessian matrix with respect to $\mathbf{x}$, written as $\nabla^2 f(\mathbf{x})$, which is defined as

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Taylor's expansion for multivariable function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

Chain rules for second derivative

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and $h(\mathbf{x}) = g(f(\mathbf{x}))$. Then

$$\nabla^2 h(\mathbf{x}) = g'(f(\mathbf{x}))\nabla^2 f(\mathbf{x}) + g''(f(\mathbf{x}))\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top.$$

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$. Define $h : \mathbb{R}^p \to \mathbb{R}$ as $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then,

$$\nabla^2 h(\mathbf{x}) = \mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{x} + \mathbf{b})\mathbf{A}.$$