02611 Optimization for Data Science (F25)

Introduction and basic concepts

Martin S. Andersen

Technical University of Denmark

Practical information

Format

- 5 ECTS (1 ECTS $\sim$ 28 hours on average)
- Lectures and exercises, one assignment (20%)
- Final exam (80%) — written exam

Instructor

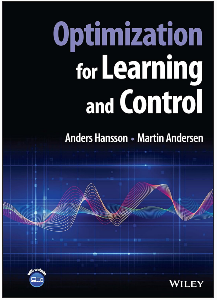
- Martin S. Andersen (mskan), DTU Compute

Teaching assistants

- Martin Sæbye Carøe (msaca), DTU Compute
- Philip Kishimoto Hohwy (s204790)

Post your questions on the DTU Learn discussion board; use email for personal matters only.

Textbook



Optimization for Learning and Control
by Anders Hansson & Martin Andersen

Wiley, 2023
ISBN: 978-1-119-80918-0

E-book available through DTU Library
Print copies available through Polyteknisk Boghandel
Errata available here

Learning objectives

A student who has met the objectives of the course will be able to:

- explain fundamental concepts in convex analysis, including convex sets and functions, conjugate functions, and subdifferentiability
- characterize optimization problems based on their mathematical properties (e.g., smooth/nonsmooth, convex/nonconvex, continuous/discrete, unconstrained/constrained) and recognize the implications of these properties on problem-solving approaches
- formulate optimization problems and derive optimality conditions and Lagrange dual problems
- explain how changes in constraints impact the optimal solution
- apply surrogation and convex relaxation techniques

Learning objectives (cont.)

- explain explicit and implicit regularization techniques
- analyze and apply stochastic optimization methods
- implement scalable algorithms for solving optimization problems within data science
- apply hyperparameter tuning strategies
- compare various optimization algorithms and evaluate trade-offs in terms of convergence speed, robustness, and scalability

Tentative schedule

1. Introduction to optimization – basic concepts (4.1)
2. Convex sets and functions (4.2-4.3)
3. Subdifferentiability and convex optimization problems (4.4-4.5)
4. Duality and optimality conditions (4.6-4.7)
5. Optimization problems (5)
6. Optimization methods I (6.1-6.3)
7. Optimization methods II (6.4-6.6)
8. Optimization methods III (6.7-6.11)
9. Applications of optimization in data science I (9.1, 9.3, 9.8-9.11, 9.14)
10. Project work
11. Applications of optimization in data science II (10.1-10.8)
12. Bayesian optimization (lecture notes)
13. Review

Notation

- natural numbers: $\mathbb{N} = \{1, 2, \ldots\}$
- natural numbers up to n: $\mathbb{N}_n = \{1, \ldots, n\}$
- integers: $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- real numbers: $\mathbb{R}$
- real n-dimensional vectors: $\mathbb{R}^n$
- real $m \times n$ matrices: $\mathbb{R}^{m \times n}$
- subset of (entrywise) nonnegative elements: $\mathbb{Z}_+$, $\mathbb{R}_+$, $\mathbb{R}^n_+$, $\mathbb{R}^{m \times n}_+$
- subset of (entrywise) positive elements: $\mathbb{R}_{++}$, $\mathbb{R}^n_{++}$, $\mathbb{R}^{m \times n}_{++}$

(vectors are interpreted as column vectors, unless otherwise stated)

Notation (cont.)

- diagonal matrix with diagonal entries $x = (x_1, \ldots, x_n)$

$$\mathrm{diag}(x) = \begin{bmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{bmatrix}$$

- $\mathbb{1} = (1, \ldots, 1)$ denotes the vector of all ones (size inferred from context)
- $I = \mathrm{diag}(\mathbb{1})$ denotes the identity matrix (size inferred from context)
- $Q \in \mathbb{R}^{n\times n}$ is orthogonal iff $Q^T Q = I$
- the range and nullspace of $A \in \mathbb{R}^{m\times n}$ are $\mathcal{R}(A)$ and $\mathcal{N}(A)$, respectively
- (standard) inner product on $\mathbb{R}^n$: $\langle a, b\rangle = a^T b = \sum_{i=1}^n a_i b_i$
- Frobenius inner product on $\mathbb{R}^{m\times n}$: $\langle A, B\rangle = \mathrm{tr}(A^T B) = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ij}$

Extended reals

Extension of the real numbers with two elements

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$$

- same total order as on $\mathbb{R}$ but includes $-\infty < x < +\infty$ for all $x \in \mathbb{R}$
- arithmetic operations can (partially) be extended to $\overline{\mathbb{R}}$, *e.g.*,

$$x + \infty = +\infty,\ x \in \mathbb{R} \cup \{+\infty\}, \qquad x - \infty = -\infty,\ x \in \mathbb{R} \cup \{-\infty\}$$

- indeterminate forms are left undefined (*e.g.*, $0 \cdot \pm\infty$, $\infty - \infty$)
- unbounded sequences have limits in $\overline{\mathbb{R}}$, *e.g.*, we may write

$$\lim_{n\to\infty} x_n = +\infty, \qquad \lim_{n\to\infty} x_n = -\infty$$

Functions

A function f from a set $\mathcal{X}$ to a set $\mathcal{Y}$ is denoted as

$$f \colon \mathcal{X} \to \mathcal{Y}$$

- $\mathcal{X}$ is the domain of f
- $\mathcal{Y}$ is the codomain of f
- when $\mathcal{Y} = \overline{\mathbb{R}}$, the effective domain of f is defined as

$$\operatorname{dom} f = \{x \in \mathcal{X} \mid f(x) < \infty\}$$

- f is proper if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \operatorname{dom} f$ (otherwise, f is improper)
- we will sometimes use the notation $f \in \mathcal{Y}^{\mathcal{X}}$

Sublevel sets and epigraph

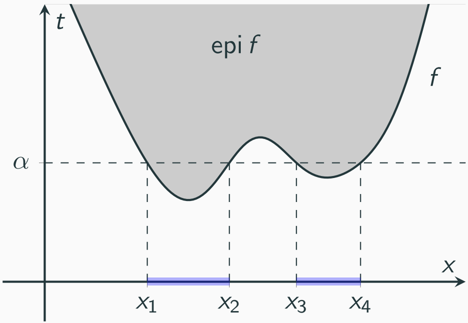
α-sublevel set of f

$$S_\alpha(f) = \{x \in \mathcal{X} \mid f(x) \leq \alpha\}$$

Epigraph of f

$$\operatorname{epi} f = \{(x, t) \in \mathcal{X} \times \mathbb{R} \mid f(x) \leq t\}$$

f is closed if epi f is closed



$$S_\alpha(f) = [x_1, x_2] \cup [x_3, x_4]$$

Differentiability

Gradient of continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

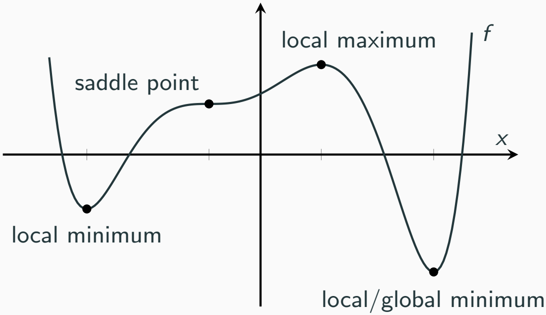
Directional derivative of f at x in the direction d

$$\frac{d}{dt} f(x + td)\Big|_{t=0} = \nabla f(x)^T d$$

Extrema

$x \in \mathbb{R}^n$ is a stationary point of a continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ iff

$$\nabla f(x) = 0$$



Infimum and supremum

Given $A \subseteq \overline{\mathbb{R}}$, we define

- the infimum of A, $\inf A$, as the largest a such that $a \leq x$ for all $x \in A$
- the supremum of A, $\sup A$, as the smallest a such that $x \leq a$ for all $x \in A$

Given $f : \mathcal{X} \to \overline{\mathbb{R}}$ and $C \subseteq \mathcal{X}$, we define

$$\inf_{x \in C} f(x) = \inf\{f(x) \,|\, x \in C\}, \qquad \sup_{x \in C} f(x) = \sup\{f(x) \,|\, x \in C\}$$

Minimum and maximum

Given $A \subseteq \overline{\mathbb{R}}$, we define

- the minimum element of A, $\min A = \inf A$, if $\inf A \in A$ (undefined otherwise)
- the maximum element of A, $\max A = \sup A$, if $\sup A \in A$ (undefined otherwise)

Given $f : \mathcal{X} \to \overline{\mathbb{R}}$ and $C \subseteq \mathcal{X}$, we define

$$\operatorname*{argmin}_{x \in C} f(x) = \{x \in C \mid f(x) = \inf_{y \in C} f(y)\}$$

$$\operatorname*{argmax}_{x \in C} f(x) = \{x \in C \mid f(x) = \sup_{y \in C} f(y)\}$$

and, if infimum/supremum is attained, the minimum/maximum value

$$\min_{x \in C} f(x) = \min \{f(x) \mid x \in C\}, \qquad \max_{x \in C} f(x) = \max \{f(x) \mid x \in C\}.$$

Examples

Let $f(x)\colon \mathbb{R}_{++} \to \overline{\mathbb{R}}$ be defined as $f(x) = \ln(x)$.

- If $C = \mathbb{R}_{++}$, then $f(C) = \mathbb{R}$ and

  $$\inf_{x \in C} f(x) = -\infty, \qquad \sup_{x \in C} f(x) = +\infty,$$

  minimum and maximum do not exists.
- If $C = [1, e)$, then $f(C) = [0, 1)$ and

  $$\inf_{x \in C} f(x) = 0, \qquad \sup_{x \in C} f(x) = 1,$$

  minimum $\min_{x \in C} f(x) = 0$ is attained at $x = 1$, maximum does not exist.

Extreme value theorem

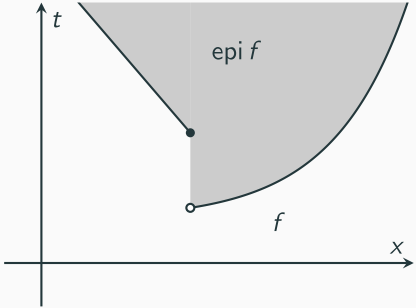
Extreme value theorem (Weierstrass)

Let $f : \mathcal{X} \to \mathbb{R}$ be continuous on a compact set $\mathcal{X} \subset \mathbb{R}^n$. Then f attains its minimum and maximum on $\mathcal{X}$.

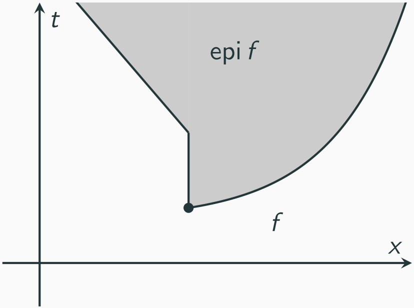
Implies that f attains its minimum if the following conditions are met:

- $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper and closed ($\operatorname{epi} f$ is closed)
- there exists $\alpha \in \mathbb{R}$ such that α-sublevel set $S_\alpha(f)$ is nonempty and bounded

Example



f is not closed



f is closed

Symmetric matrices

Symmetric matrices of order n

$$\mathbb{S}^n = \{A \in \mathbb{R}^{n\times n} \,|\, A = A^T\}$$

- symmetric positive semidefinite matrices of order n

$$\mathbb{S}^n_+ = \{A \in \mathbb{S}^n \,|\, x^T A x \geq 0 \ \forall x\}$$

- symmetric positive definite matrices of order n

$$\mathbb{S}^n_{++} = \{A \in \mathbb{S}^n \,|\, x^T A x > 0 \ \forall x \neq 0\}$$

Symmetric matrices (cont.)

A symmetric matrix $A \in \mathbb{S}^n$ has a the spectral decomposition

$$A = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

where $Q^T Q = I$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \ldots \geq \lambda_n$.

- $A \in \mathbb{S}^n_+$ iff $\lambda_i(A) \geq 0$ for $i \in \mathbb{N}_n$

$$x^T A x = \sum_{i=1}^{n} \lambda_i (q_i^T x)^2$$

- $A \in \mathbb{S}^n_{++}$ defines weighted inner product $\langle x, y \rangle_A = x^T A y$ and quadratic norm on $\mathbb{R}^n$

$$\|x\|_A = \sqrt{\langle x, x \rangle_A} = \sqrt{x^T A x}$$

Quadratic forms

Let $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ be defined as

$$f(u, v) = \begin{bmatrix} u \\ v \end{bmatrix}^T \underbrace{\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}}_{X} \begin{bmatrix} u \\ v \end{bmatrix} = u^T A u + v^T C v + 2u^T B v$$

where $A \in \mathbb{S}^{n_1}$, $C \in \mathbb{S}^{n_2}$, and $B \in \mathbb{R}^{n_1 \times n_2}$.

$X \in \mathbb{S}_+^{n_1+n_2}$ iff $f(u, v) \geq 0$ for all (u, v), which implies the following:

- $A \in \mathbb{S}_+^{n_1}$ and $C \in \mathbb{S}_+^{n_2}$ since $f(u, 0) \geq 0 \ \forall u$ and $f(0, v) \geq 0 \ \forall v$
- $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ and $\mathcal{R}(B^T) \subseteq \mathcal{R}(C)$ since

$$\mathcal{R}(B) \not\subseteq \mathcal{R}(A) \implies \exists v \colon Bv \neq 0 \wedge Bv \notin \mathcal{R}(A) \implies f(-tBv, v) = v^T C v - 2t\|Bv\|_2^2$$

Optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i \in \mathbb{N}_m \\ & h_i(x) = 0, \quad i \in \mathbb{N}_p \end{array}$$

- $f_0 : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the objective function and $x \in \mathbb{R}^n$ is the optimization variable
- $f_i : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the ith inequality constraint function
- $h_i : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the ith equality constraint function
- the problem is unconstrained if $m = p = 0$ and constrained otherwise
- the domain of the optimization problem is $\mathcal{D} = (\cap_{i=0}^m \operatorname{dom} f_i) \cap (\cap_{i=1}^p \operatorname{dom} h_i)$
- the feasible set is $\mathcal{F} = \{x \in \mathcal{D} \,|\, f_i(x) \le 0, i \in \mathbb{N}_m, h_i(x) = 0, i \in \mathbb{N}_p\}$
- the optimal value is $p^\star = \inf \{f_0(x) \,|\, x \in \mathcal{F}\}$
- the optimal value is attained if $\exists x^\star \in \mathcal{F}$ such that $f_0(x^\star) = p^\star$
- the problem is unbounded if $p^\star = -\infty$ and infeasible if $\mathcal{F} = \emptyset$ (we let $p^\star = +\infty$)

Local optimum

A point $x \in \mathbb{R}^n$ is said to be locally optimal if there exists $r > 0$ such that

$$f_0(x) = \inf_z \{f_0(z) \,|\, x \in \mathcal{F} \cap B_2(x, r)\}$$

where $B_2(c, r)$ is the Euclidean ball centered at $c \in \mathbb{R}^n$ with radius r, *i.e.*,

$$B_2(c, r) = \{z \in \mathbb{R}^n \,|\, \|z - c\|_2 \leq r\}.$$

A globally optimal point $x^\star$ is also locally optimal but the converse is not necessarily true.

Equivalent problems

Two optimization problems are said to be equivalent if the solution to one can readily be translated into a solution to the other and vice versa.

Example

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i \in \mathbb{N}_m \\ & h_i(x) = 0, \quad i \in \mathbb{N}_p \end{array}$$

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & f_0(x) - t \leq 0 \\ & f_i(x) \leq 0, \quad i \in \mathbb{N}_m \\ & h_i(x) = 0, \quad i \in \mathbb{N}_p \end{array}$$

What makes an optimization problem difficult?

"most optimization problems are unsolvable" — *Yu. Nesterov (2004)*

Example[1]

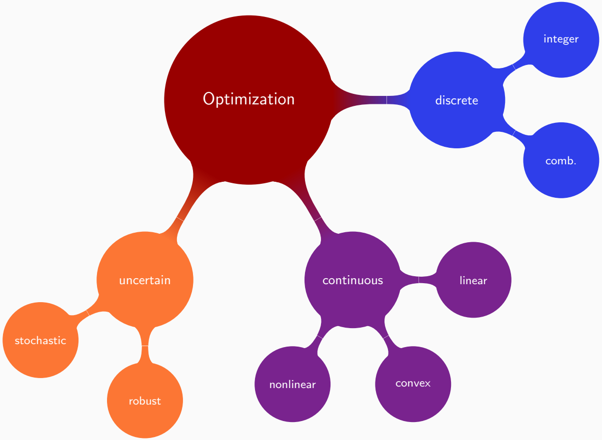
$$\begin{array}{ll} \text{minimize} & (x_1^{x_4} + x_2^{x_4} - x_3^{x_4})^2 + \sum_{i=1}^4 (\sin \pi x_i)^2 \\ \text{subject to} & x_1, x_2, x_3 \geq 1, x_4 \geq 3 \end{array}$$

Fermat's last theorem:

$$\nexists\, a, b, c, n \in \mathbb{N} \;:\; a^n + b^n = c^n,\; n \geq 3$$

[1] Guenin, Könemann, and Tunçel, *A gentle introduction to optimization*, 2014.

Taxonomy



- continuous vs. discrete variables
- deterministic vs. stochastic
- smooth vs. nonsmooth functions
- global vs. local optimization
- black/gray/white box model

Jacobian matrix

The Jacobian of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ is the matrix of partial derivatives

$$\frac{\partial f(x)}{\partial x^T} = \frac{\partial f}{\partial x^T} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

We will sometimes use the notation

$$\frac{\partial f(x)^T}{\partial x} = \left(\frac{\partial f(x)}{\partial x^T} \right)^T$$

Gradient and Hessian

The gradient of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at x is

$$\frac{\partial f}{\partial x} = \nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

The Hessian of a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at x is

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

$\nabla^2 f(x) \in \mathbb{S}^n$ if f is twice continuously differentiable at x (Clairaut/Schwarz/Young's theorem)

Stationary points of twice continuously differentiable functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and x a stationary point (*i.e.*, $\nabla f(x) = 0$).

From Taylor's theorem, we have

$$f(x+p) = f(x) + \underbrace{\nabla f(x)^T p}_{0} + \frac{1}{2} p^T \nabla^2 f(x) p + \epsilon(p) \|p\|_2^2$$

where $\epsilon(p) \to 0$ as $\|p\|_2 \to 0$.

- x is local minimum/maximum if $\nabla^2 f(x)$ is positive/negative semidefinite
- x is a strict local minimum/maximum if $\nabla^2 f(x)$ is positive/negative definite
- x is saddle point if $\nabla^2 f(x)$ is indefinite

Matrix-valued functions and functions of matrices

Differentiable matrix-valued function $F\colon \mathbb{R} \to \mathbb{R}^{m\times n}$

$$\frac{dF}{dx} = \begin{bmatrix} \frac{dF_{11}}{dx} & \cdots & \frac{dF_{1n}}{dx} \\ \vdots & \ddots & \vdots \\ \frac{dF_{m1}}{dx} & \cdots & \frac{dF_{mn}}{dx} \end{bmatrix}$$

Differentiable function of a matrix $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix}$$

Composite functions and products

Chain rule

Suppose $f = g \circ h$ with $g \colon \mathbb{R}^p \to \mathbb{R}^m$ and $h \colon \mathbb{R}^n \to \mathbb{R}^p$ differentiable

$$\frac{\partial f(x)}{\partial x^T} = \frac{\partial g(h(x))}{\partial x^T} = \left.\frac{\partial g(y)}{\partial y^T}\right|_{y=h(x)} \frac{\partial h(x)}{\partial x^T}$$

Product rule

Suppose $f = g \cdot h$ with $g \colon \mathbb{R}^n \to \mathbb{R}^m$ and $h \colon \mathbb{R}^n \to \mathbb{R}$ differentiable

$$\frac{\partial f(x)}{\partial x^T} = \frac{\partial g(x)}{\partial x^T} h(x) + g(x) \frac{\partial h(x)}{\partial x^T}$$

Affine transformation

Suppose $f = g \circ h$ with $g \colon \mathbb{R}^p \to \mathbb{R}^m$ differentiable and $h \colon \mathbb{R}^n \to \mathbb{R}^p$ given by $h(x) = Ax + b$.

$$\frac{\partial f}{\partial x^T} = \left.\frac{\partial g(y)}{\partial y^T}\right|_{y=Ax+b} A$$

Chain rule for gradients ($m = 1$)

$$\nabla f(x) = A^T \nabla g(Ax + b)$$

Chain rule for Hessians ($m = 1$, g twice differentiable)

$$\nabla^2 f(x) = A^T \nabla^2 g(Ax + b) A$$

Example: log-sum-exp function

Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(x) = \ln\left(\sum_{i=1}^{n} e^{x_i}\right)$$

Express f as composition $f = g \circ h$ where $g(y) = \ln(\mathbb{1}^T y)$ and $h(x) = (e^{x_1}, \ldots, e^{x_n})$

$$\frac{\partial g}{\partial y^T} = \frac{1}{\mathbb{1}^T y}\mathbb{1}^T, \qquad \frac{\partial h}{\partial x^T} = \operatorname{diag}(h(x))$$

Chain rule yields

$$\frac{\partial f}{\partial x^T} = \frac{1}{\mathbb{1}^T h(x)}\mathbb{1}^T \operatorname{diag}(h(x)) = \frac{1}{\mathbb{1}^T h(x)} h(x)^T$$

Cauchy and the gradient method

Unconstrained optimization problem with continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$

$$\text{minimize} \quad f(x)$$

Cauchy's method (1847) with initial guess $x_0 \in \mathbb{R}^n$:

for $k = 0, 1, 2, \ldots$

- compute steepest descent direction at x_k

$$p_k = \underset{v \in \mathbb{R}^n}{\operatorname{argmin}} \{\nabla f(x_k)^T v \mid \|v\|_2 \leq 1\}$$

- move in the direction of p_k

$$x_{k+1} = x_k + t_k p_k$$

Applications of optimization in data science

- unsupervised learning (*e.g.*, estimation, clustering, dimensionality reduction)
- supervised learning (*e.g.*, regression, classification)
- signal and image processing (*e.g.*, denoising, deconvolution, image registration)
- control (*e.g.*, system identification, reinforcement learning)
- model selection and hyperparameter tuning
- adversarial learning

Typical challenges

- large data sets
- large number of variables (parameters)
- continuous and discrete variables
- expensive/noisy function evaluations
- hyperparameter selection
- regularization

Algorithms

- gradient-based methods
- stochastic optimization
- proximal methods
- coordinate descent
- quasi-Newton methods
- trust region methods
- augmented Lagrangian methods
- interior point methods
- Bayesian optimization