Optimization :: CHEAT SHEET

Note: The lecture uses many examples from maths, statistics and machine learning. Therefore notation may overlap, but the context should make clear how to uncerstand the notation. If notation is unclear nevertheless, please contact the instructors.

General Mathematical Notation (I)

$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$: natural, integer, and real numbers

$|a|$: absolute value of a scalar $a \in \mathbb{R}$

$\mathcal{I} = [a, b] \subseteq \mathbb{R}$: interval ranging from $a \in \mathbb{R}$ to $b \in \mathbb{R}$, $a < b$

$\mathbf{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$: real vector in $\mathbb{R}^n$

vectors in **lowercase bold**; refer to i-th element by **sub**script i.

$\mathbf{e}_k$: k-th unit vector

vector which is zero at all positions $\neq k$, and one at position k

$\mathbf{x}^{[t]}$: t-th iteration of a sequence $\left(\mathbf{x}^{[1]}, \mathbf{x}^{[2]}, \mathbf{x}^{[3]}, ...\right)$

$t \in \{1, ..., T\}$ (finite sequence) or $t \in \mathbb{N}$ (infinite sequence)

$\mathbf{x}^{(i)}$: i-th element of an indexed family $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}\right\}$

collection of vectors, indexed by indices i of an index set

$\boldsymbol{A} = (A_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$: matrix in $\mathbb{R}^{n \times m}$

matrices in **uppercase bold**, **sub**script (i, j) denotes row i, column j

$\mathbf{I}_n$: $n \times n$ identity matrix

$\mathbf{x}^\top \in \mathbb{R}^{1 \times n}, \boldsymbol{A}^\top \in \mathbb{R}^{m \times n}$: Transpose of a vector / matrix

$\boldsymbol{A}^{-1} \in \mathbb{R}^{n \times n}$: Inverse of a quadratic matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ (if it exists)

$\|\mathbf{x}\|, \|\boldsymbol{A}\|$: vector norm, (induced) matrix norm

$\|.\|_p$ denotes a p-norm, $\|\boldsymbol{A}\|_F$ denotes the Frobenius matrix norm

General Mathematical Notation (II)

$f : \mathcal{S} \to \mathbb{R}^m$: function with domain $\mathcal{S}$ and codomain $\mathbb{R}^m$

$f(T)$: image of $T \subseteq \mathcal{S}$ under f

range of values y for which there is an $\mathbf{x} \in \mathcal{S}$ such that $f(\mathbf{x}) = y$.

$\mathcal{C}^k$: class of k-times continuously differentiable functions

if $f \in \mathcal{C}^k$, the k-th derivative of f exists and it is continuous; $\mathcal{C}^\infty$ means that the function is infinitely often continuously differentiable

$f', f'', f^{(3)}, ...$: derivatives for a function f (if it exists)

this notation is only valid for univariate functions $f : \mathcal{S} \to \mathbb{R}$, $\mathcal{S} \subseteq \mathbb{R}$

$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, ..., \frac{\partial f(\mathbf{x})}{\partial x_d}\right)^\top \in \mathbb{R}^d$: gradient of f

$D_v f(\mathbf{x})$: directional derivative for f in the direction of $v \in \mathbb{R}^d$

$\nabla^2 f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j}\right)_{i,j=1,...,d}$: Hessian of f (if it exists)

often also denoted by $\mathbf{H}(\mathbf{x})$

Probability Theory and Statistics

X : (discrete or continuous) random variable

$\boldsymbol{\theta}, \hat{\theta} \in \Theta$: parameter $\boldsymbol{\theta}$ of a distribution and its estimate $\hat{\theta}$

F_X, f_X : distribution/density function of a continuous RV X

$\mathbb{P}(X = k)$: density of a discrete random variable X

ϕ, Φ : density, distribution function of a standard normal random variable

$X \sim U(a, b)$: X is sampled uniformly from the intervall $[a, b]$

Machine Learning

$\mathcal{X}, \mathcal{Y}$: input space, and label / output space

$\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), ..., \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$: dataset

$\mathbf{x}^{(i)} \in \mathcal{X}$ is called input and $y^{(i)} \in \mathcal{Y}$ is called output

$f : \mathcal{X} \to \mathbb{R}^g$: model

usually, f is parameterized by a parameter $\boldsymbol{\theta} \in \Theta$; we write $f = f(\cdot \mid \boldsymbol{\theta})$ (e.g. $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}$)

$\boldsymbol{\theta}, \hat{\theta} \in \Theta$: model parameter and its estimate

$L : \mathcal{Y} \times \mathbb{R}^g \to \mathbb{R}_0^+$: loss function

measures the "goodness" of a prediction $\hat{y}$; for example, the squared loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$

$\mathcal{R} : \Theta \to \mathbb{R}$: empirical risk function

the (empirical) risk maps a parameter $\boldsymbol{\theta}$, which parameterizes the model $f(\cdot \mid \boldsymbol{\theta})$, to the sum of losses $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right))$

Encoding

u_i : i-th bit in a machine number

$\mathcal{M}$: Machine numbers

$x = S \cdot b^e \cdot \left(1 + \sum_{i=1}^{m} u_i b^{-i}\right)$: representation of a machine number

with sign $S \in \{-1, 1\}$, base $b > 1$, mantissa length m, exponent e

ϵ_m : machine epsilon

upper bound on the relative error

Optimization :: CHEAT SHEET

Numerical Analysis

$\mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^n$: in-disturbed / disturbed input

$\Delta\mathbf{x}, \delta\mathbf{x}$: absolute / relative error in the computation of $\mathbf{x}$

κ : smallest $\kappa \geq 0$, so that $\frac{|f(x+\Delta x)-y|}{|y|} \leq \kappa \frac{|\Delta x|}{|x|}$ for $\Delta x \to 0$.

κ is called condition number

$f, \tilde{f} : \mathbb{R}^n \to \mathbb{R}^m$: a problem f and an algorithm $\tilde{f}$ to solve f

Big-O

Note: in this chapter, all functions are real-valued

$f \in \mathcal{O}(g)$ for $x \to \infty$: $\limsup_{x\to\infty} \frac{f(x)}{g(x)} < \infty$

"Big-O" notation: say "f runs in the order of g"

$f \in \mathcal{O}(1)\}$: f has constant runtime complexity

$f \in \mathcal{O}(\log(n))$: f has logarithmic runtime complexity

$f \in \mathcal{O}(n^c)$: f has polynomial runtime complexity

linear for $c = 1$, quadratic for $c = 2$ and cubic for $c = 3$.

$f \in \mathcal{O}(c^n)$: f has exponential runtime complexity

Quadrature

Note: in this chapter, all functions are real-valued functions.

$I(f) = \int_a^b f(x)\, dx$: (Riemann) integral of f

$Q(f)$: numerical approximation of $I(f)$ based on a quadrature rule Q

Optimization (General)

$\min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}), f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}^m$: optimization problem

w.l.o.g. we only consider minimization

$\mathbf{x} \in \mathcal{S}$: decision variable $\mathbf{x}$ in the decision space $\mathcal{S}$

$f : \mathcal{S} \to \mathbb{R}^m$: objective function with domain $\mathcal{S} \subseteq \mathbb{R}^d$ and codomain $\mathbb{R}^m$

$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$: empirical risk minimization (ERM) problem

Note: We often consider the ERM problem in ML as an example of an optimization problem. In this case, we switch from the general optimization notation to the ML notation:

- We optimize the function $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ (instead of f); f instead denotes the ML model
- We optimize over $\boldsymbol{\theta} \in \Theta$ (instead over $\mathbf{x} \in \mathcal{S}$)

$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{S}} f(x)$, $y^* = \min_{\mathbf{x}\in\mathcal{S}} f(x)$

theoretical optimum $\mathbf{x}^*$ and optimal value $y^* = f(\mathbf{x}^*)$

$\hat{\mathbf{x}} \in \mathcal{S}, \hat{y} \in \mathbb{R}$: estimated optimum and optimal value

typically $(\hat{\mathbf{x}}, \hat{y}) = \mathcal{A}(f, \mathcal{S})$ is returned by an optimization algorithm $\mathcal{A}$

$\mathbf{x}^{[t]} \in \mathcal{S}$: t-th step of an optimizer in the decision space

Multivariate Optimization

$\alpha \in \mathbb{R}_+$: step-size / learning rate

$\boldsymbol{d} \in \mathbb{R}^d$: descent direction in $\mathbf{x}$

$\varphi \in [0, 1]$: momentum

Optimization (Constrained Optimization)

$\min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x})$ s.t. $h(\mathbf{x}) = 0, g(\mathbf{x}) \leq 0$: constrained optimization problem

Optimization (Evolutionary Algorithms)

P : population (of solution candidates)

$\mu \in \mathbb{N}$: size of a population

$\lambda \in \mathbb{N}$: offspring size

(μ, λ)-selection : survival selection strategy

the best μ individuals from λ candidates are chosen ($\lambda \geq \mu$ required)

$(\mu + \lambda)$-selection : survival selection strategy

the best μ individuals are chosen from the pool of the current population of size μ and the offspring of size λ.

$\mathbf{x}_{i:\lambda}$: i-th ranked candidate

λ solution candidates are ranked according to some criterion (e.g. by a fitness function); $\mathbf{x}_{i:\lambda}$ means that this candidate has rank i.

$\boldsymbol{m}^{[g]}, \boldsymbol{C}^{[g]}, \sigma^{[g]}$: configurations in generation g

superscript $[g]$ denotes the g-th generation

$\mathbf{x}^{[g](k)}$: k-th individual in the population in generation g

Optimization (Multi-Objective)

$\mathcal{P}$: Pareto set

Set of nondominated solutions (in the decision space $\mathcal{S}$)

$\mathcal{F}$: Pareto front

image of the Pareto set $\mathcal{P}$ under a multi-objective function f