Introduction to Optimization

Lecture 3: Introduction to Continuous Optimization

September 22, 2017

TC2 - Optimisation

Université Paris-Saclay

INVENTORS FOR THE DIGITAL WORLD

Dimo Brockhoff

Inria Saclay – Ile-de-France

Course Overview

1	Mon, 18.9.2017	first lecture
	Tue, 19.9.2017	groups defined via wiki
		everybody went (actively!) through the Getting Started part of github.com/numbbo/coco
2	Wed, 20.9.2017	lecture: "Benchmarking", final adjustments of groups everybody can run and postprocess the example experiment (~1h for final questions/help during the lecture)
3	Fri, 22.9.2017	today's lecture "Introduction to Continuous Optimization"
4	Fri, 29.9.2017	lecture "Gradient-Based Algorithms"
5	Fri, 6.10.2017	lecture "Stochastic Algorithms and DFO"
6	Fri, 13.10.2017	lecture "Discrete Optimization I: graphs, greedy algos, dyn. progr." deadline for submitting data sets
	Wed, 18.10.2017	deadline for paper submission
7	Fri, 20.10.2017	final lecture "Discrete Optimization II: dyn. progr., B&B, heuristics"
	Thu, 26.10.2017 / Fri, 27.10.2017	oral presentations (individual time slots)
	after 30.10.2017	vacation aka learning for the exams
	Fri, 10.11.2017	written exam

All deadlines:
23:59pm Paris time

Details on Continuous Optimization Lectures

Introduction to Continuous Optimization

- examples (from ML / black-box problems)
- typical difficulties in optimization

Mathematical Tools to Characterize Optima

- reminders about differentiability, gradient, Hessian matrix
- unconstraint optimization
  - first and second order conditions
  - convexity
- constraint optimization

Gradient-based Algorithms

- quasi-Newton method (BFGS)
- [DFO trust-region method]

Learning in Optimization / Stochastic Optimization

- CMA-ES (adaptive algorithms / Information Geometry)
- PhD thesis possible on this topic

method strongly related to ML / new promising research area

interesting open questions

Continuous Optimization

- Optimize f: $\begin{cases} \Omega \subset \mathbb{R}^n \to \mathbb{R} \\ x = (x_1, \dots, x_n) \to f(x_1, \dots, x_n) \end{cases}$

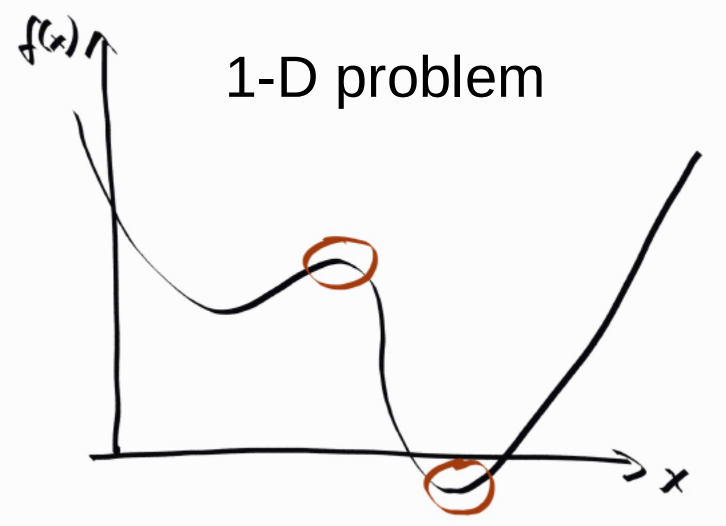
  ($x_1 \in \mathbb{R}$)

  unconstrained optimization

- Search space is continuous, i.e. composed of real vectors $x \in \mathbb{R}^n$

- $n = \begin{cases} \text{dimension of the problem} \\ \text{dimension of the search space } \mathbb{R}^n \text{ (as vector space)} \end{cases}$

1-D problem



2-D level sets

Reminder: Different Notions of Optimum

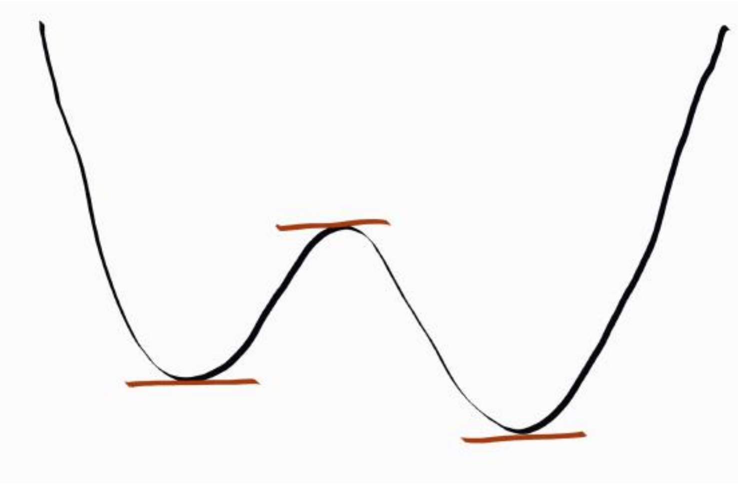
Unconstrained case

- local vs. global
  - local minimum $\boldsymbol{x}^*$: ∃ a neighborhood V of $\boldsymbol{x}^*$ such that
    $\forall \boldsymbol{x} \in \mathrm{V}: f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$
  - global minimum: $\forall \boldsymbol{x} \in \Omega: f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$
- strict local minimum if the inequality is strict

Mathematical Characterization of Optima

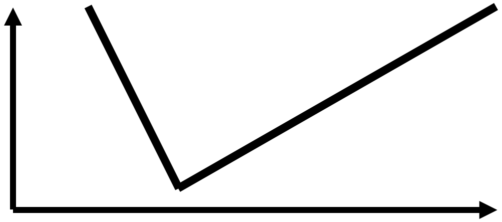
Objective: Derive general characterization of optima

Example: if $f: \mathbb{R} \to \mathbb{R}$ differentiable,
$f'(x) = 0$ at optimal points



- generalization to $f: \mathbb{R}^n \to \mathbb{R}$?
- generalization to constrained problems?

Remark: notion of optimum independent of notion of derivability

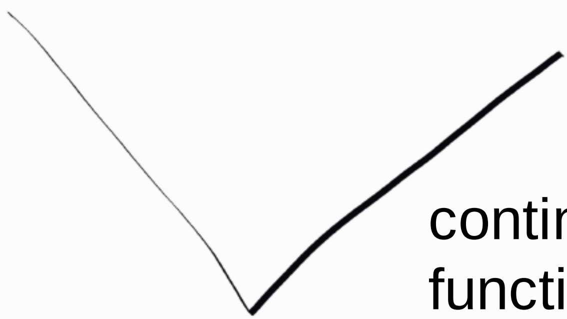


optima of such function can be easily approached by certain type of methods

Reminder: Continuity of a Function

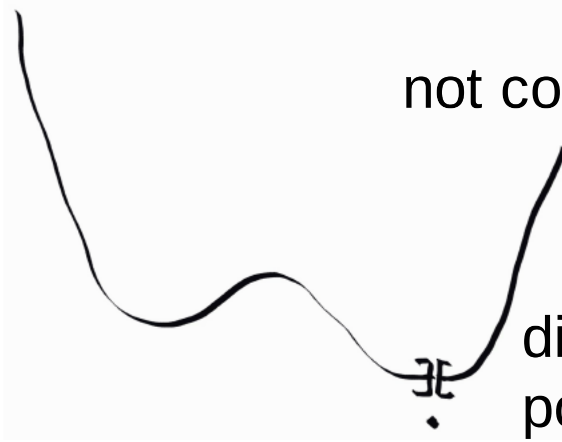
$f: (V, || \, ||_V) \longrightarrow (W, || \, ||_W)$ is continuous in $x \in V$ if

$\forall \epsilon > 0, \exists \eta > 0$ such that $\forall y \in V$: $||x - y||_V \leq \eta$; $||f(x) - f(y)||_W \leq \epsilon$



continuous function

not continuous



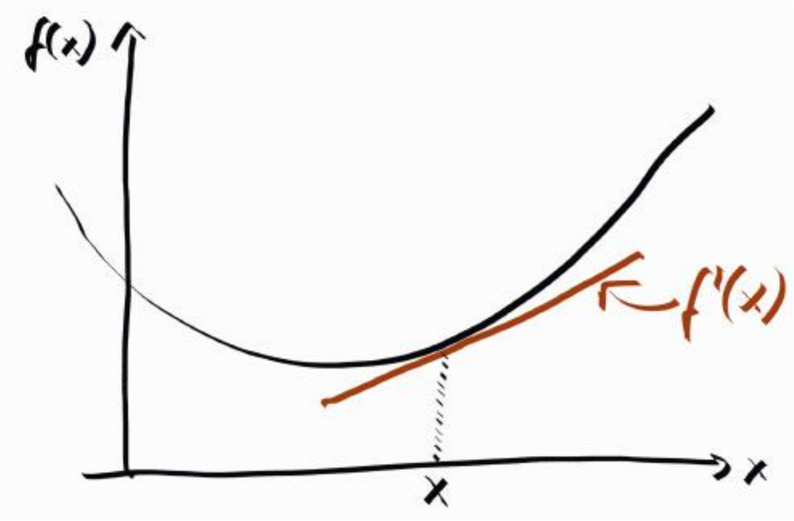
discontinuity point

$f: \mathbb{R} \to \mathbb{R}$ is differentiable in $x \in \mathbb{R}$ if

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \text{ exists, } h \in \mathbb{R}$$

Notation:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



The derivative corresponds to the slope of the tangent in x.

Reminder: Differentiability in 1D (n=1)

Taylor Formula (Order 1)

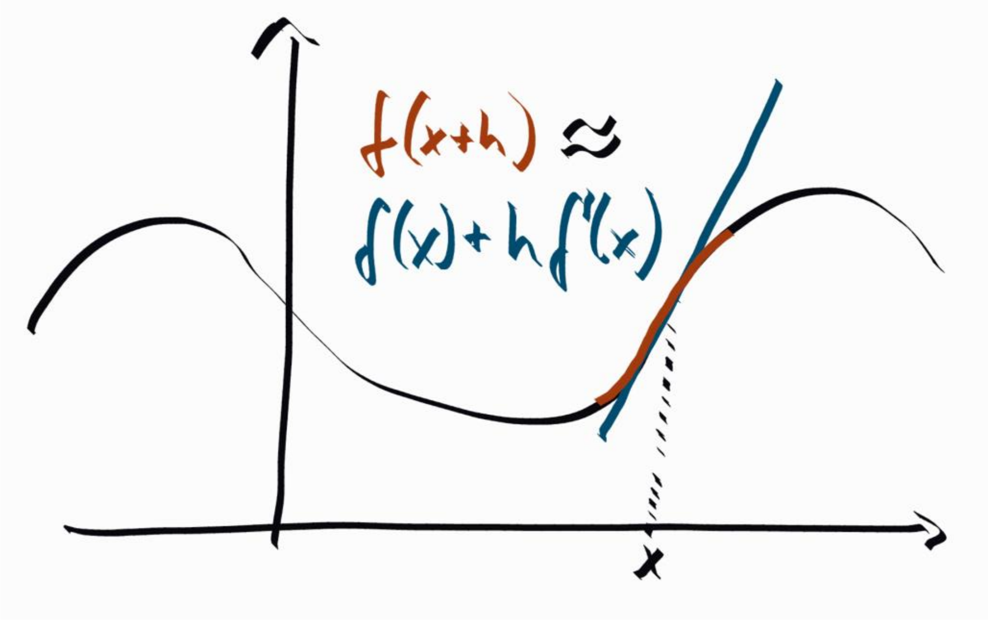
If f is differentiable in x then

$$f(x+h) = f(x) + f'(x)h + o(||h||)$$

i.e. for h small enough, $h \mapsto f(x+h)$ is approximated by $h \mapsto f(x) + f'(x)h$

$h \mapsto f(x) + f'(x)h$ is called a first order approximation of $f(x+h)$

Geometrically:



The notion of derivative of a function defined on $\mathbb{R}^n$ is generalized via this idea of a linear approximation of $f(x+h)$ for h small enough.

How to generalize this to arbitrary dimension?

Gradient Definition Via Partial Derivatives

- In $(\mathbb{R}^n, || \ ||_2)$ where $||\boldsymbol{x}||_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ is the Euclidean norm deriving from the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y}$

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

- Reminder: partial derivative in x_0

$$\mathrm{f_i}\colon y \rightarrow f\left(x_0^1, \dots, x_0^{i-1}, y, x_0^{i+1}, \dots, x_0^n\right)$$

$$\frac{\partial f}{\partial x_i}(x_0) = f_i{}'(x_0)$$

Exercise: Gradients

Exercise:

Compute the gradients of

a) $f(x) = x_1$ with $x \in \mathbb{R}^n$

b) $f(x) = a^T x$ with $\mathrm{a}, x \in \mathbb{R}^n$

c) $f(x) = x^T x$ $(= ||\mathrm{x}||^2)$ with $x \in \mathbb{R}^n$

Exercise: Gradients

Exercise:

Compute the gradients of

a) $f(x) = x_1$ with $x \in \mathbb{R}^n$

b) $f(x) = a^T x$ with $a, x \in \mathbb{R}^n$

c) $f(x) = x^T x$ $(= ||x||^2)$ with $x \in \mathbb{R}^n$

Some more examples:

- in $\mathbb{R}^n$, if $f(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$, then $\nabla f(\boldsymbol{x}) = (A + A^T)\boldsymbol{x}$
- in $\mathbb{R}$, $\nabla f(\boldsymbol{x}) = f'(\boldsymbol{x})$

Gradient: Geometrical Interpretation

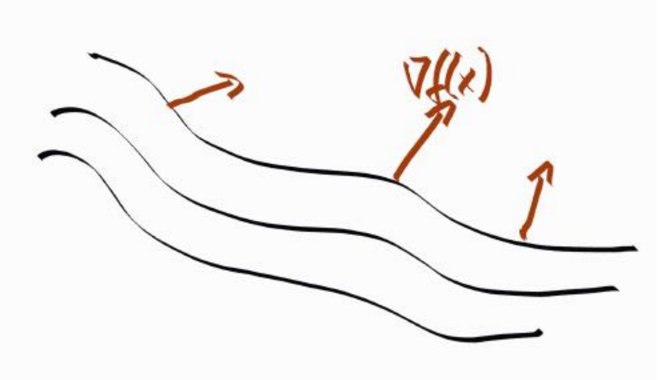
Exercise:

Let $L_c = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = c\}$ be again a level set of a function $f(\boldsymbol{x})$. Let $\boldsymbol{x}_0 \in L_c \neq \emptyset$.

Compute the level sets for $f_1(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ and $f_2(\boldsymbol{x}) = ||\boldsymbol{x}||^2$ and the gradient in a chosen point x_0 and observe that $\nabla f(\boldsymbol{x_0})$ is ***orthogonal*** to the level set in x_0.

Again: if this seems too difficult, do it for two variables (and a concrete $\boldsymbol{a} \in \mathbb{R}^2$ and draw the level sets and the gradients.

More generally, the gradient of a differentiable function is orthogonal to its level sets.



Differentiability in $\mathbb{R}^n$

Taylor Formula – Order One

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \left(\nabla f(\boldsymbol{x})\right)^T \boldsymbol{h} + o(||\boldsymbol{h}||)$$

Reminder: Second Order Derivability in 1D

- Let $f: \mathbb{R} \to \mathbb{R}$ be a derivable function and let $f': x \to f'(x)$ be its derivative function.
- If f' is derivable in x, then we denote its derivative as $f''(x)$
- $f''(x)$ is called the *second order derivative* of f.

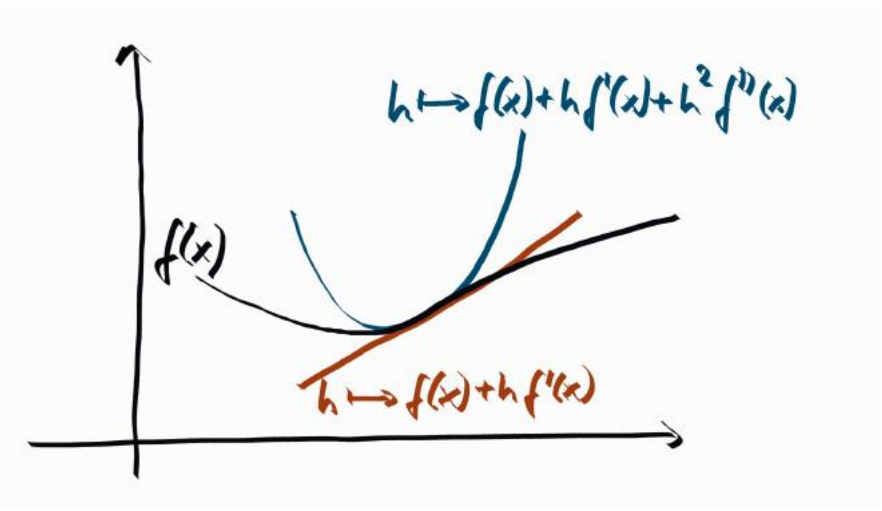
Taylor Formula: Second Order Derivative

- If $f: \mathbb{R} \rightarrow \mathbb{R}$ is two times differentiable then

$$f(x+h) = f(x) + f'(x)h + f''(x)h^2 + o(||h||^2)$$

  i.e. for h small enough, $h \rightarrow f(x) + hf'(x) + h^2 f''(x)$ approximates $h + f(x+h)$

- $h \rightarrow f(x) + hf'(x) + h^2 f''(x)$ is a quadratic approximation (or order 2) of f in a neighborhood of x



- The second derivative of $f: \mathbb{R} \rightarrow \mathbb{R}$ generalizes naturally to larger dimension.

Hessian Matrix

In $(\mathbb{R}^n, \langle x, y \rangle = x^T y)$, $\nabla^2 f(x)$ is represented by a symmetric matrix called the Hessian matrix. It can be computed as

$$\nabla^2(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Exercise:

Let $f(\boldsymbol{x}) = \boldsymbol{x}^T A\, \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$.

Compute the Hessian matrix of f.

If it is too complex, consider f: $\begin{cases} \mathbb{R}^2 \to \mathbb{R} \\ \boldsymbol{x} \to \boldsymbol{x}^T A\, \boldsymbol{x} \end{cases}$ with $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$

Second Order Differentiability in $\mathbb{R}^n$

Taylor Formula – Order Two

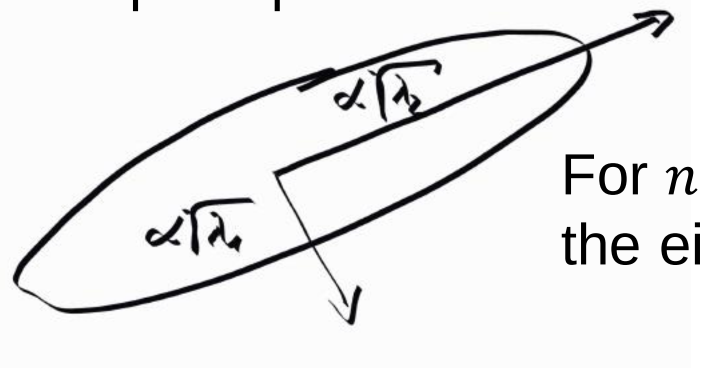
$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \left(\nabla f(\boldsymbol{x})\right)^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T\left(\nabla^2 f(\boldsymbol{x})\right) \boldsymbol{h} + o(||\boldsymbol{h}||^2)$$

Back to Ill-Conditioned Problems

We have seen that for a convex quadratic function

$f(x) = \frac{1}{2}(x - x_0)^T A (x - x_0) + b$ of $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, A SPD, $b \in \mathbb{R}^n$:

1) The level sets are ellipsoids. The eigenvalues of A determine the lengths of the principle axes of the ellipsoid.



For $n = 2$, let λ_1, λ_2 be the eigenvalues of A.

2) The Hessian matrix of f equals to A.

Ill-conditioned convex quadratic problems are problems with large ratio between largest and smallest eigenvalue of A which means large ratio between longest and shortest axis of ellipsoid.

This corresponds to having an ill-conditioned Hessian matrix.

Gradient Direction Vs. Newton Direction

Gradient direction: $\nabla f(\boldsymbol{x})$

Newton direction: $\left(H(\boldsymbol{x})\right)^{-1} \cdot \nabla f(\boldsymbol{x})$

with $H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$ being the Hessian at $\boldsymbol{x}$

Exercise:

Let again $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\,\boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^2$, $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

Plot the gradient and Newton direction of f in a point $x \in \mathbb{R}^n$ of your choice (which should not be on a coordinate axis) into the same plot with the level sets, we created before.

Optimality Conditions for Unconstrained Problems

Optimality Conditions: First Order Necessary Cond.

For 1-dimensional optimization problems $\boldsymbol{f}$: $\mathbb{R} \to \mathbb{R}$

Assume f is differentiable

- $\boldsymbol{x}^*$ is a local optimum $\Longrightarrow f'(\boldsymbol{x}^*) = 0$

not a sufficient condition: consider $f(\boldsymbol{x}) = \boldsymbol{x}^3$

proof via Taylor formula: $f(\boldsymbol{x}^* + \boldsymbol{h}) = f(\boldsymbol{x}^*) + f'(\boldsymbol{x}^*)\boldsymbol{h} + o(||\boldsymbol{h}||)$

- points $\boldsymbol{y}$ such that $f'(\boldsymbol{y}) = 0$ are called critical or stationary points

Generalization to $\boldsymbol{n}$-dimensional functions

If $f: U \subset \mathbb{R}^n \longmapsto \mathbb{R}$ is differentiable

- necessary condition: If $\boldsymbol{x}^*$ is a local optimum of f, then $\nabla f(\boldsymbol{x}^*) = 0$

proof via Taylor formula

Second Order Necessary and Sufficient Opt. Cond.

If f is twice continuously differentiable

- Necessary condition: if $\boldsymbol{x}^*$ is a local minimum, then $\nabla f(\boldsymbol{x}^*) = 0$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive semi-definite

proof via Taylor formula at order 2

- Sufficient condition: if $\nabla f(\boldsymbol{x}^*) = 0$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, then $\boldsymbol{x}^*$ is a strict local minimum

Proof of Sufficient Condition:

- Let $\lambda > 0$ be the smallest eigenvalue of $\nabla^2 f(\boldsymbol{x}^*)$, using a second order Taylor expansion, we have for all $\boldsymbol{h}$:
- $f(\boldsymbol{x}^* + \boldsymbol{h}) - f(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^T \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{h} + o(||\boldsymbol{h}||^2)$

$$> \frac{\lambda}{2} ||\boldsymbol{h}||^2 + \mathrm{o}(||\boldsymbol{h}||^2) = \left(\frac{\lambda}{2} + \frac{o(||\boldsymbol{h}||^2)}{||\boldsymbol{h}||^2} \right) ||\boldsymbol{h}||^2$$

Convex Functions

Let U be a convex open set of $\mathbb{R}^n$ and $f: U \to \mathbb{R}$. The function f is said to be convex if for all $\boldsymbol{x}, \boldsymbol{y} \in U$ and for all $t \in [0,1]$

$$f\big((1-t)\boldsymbol{x} + t\boldsymbol{y}\big) \leq (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y})$$

Theorem

If f is differentiable, then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y}$

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \big(\nabla f(x)\big)^T(\boldsymbol{y} - \boldsymbol{x})$$

if $n = 1$, the curve is on top of the tangent

If f is twice continuously differentiable, then f is convex if and only if $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite for all $\boldsymbol{x}$.

Convex Functions: Why Convexity?

Examples of Convex Functions:

- $f(\boldsymbol{x}) = a^T\boldsymbol{x} + b$
- $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + a^T\boldsymbol{x} + b$, A symmetric positive definite
- the negative of the entropy function (i.e. $f(\boldsymbol{x}) = -\sum_{i=1}^{n} \boldsymbol{x}_i \ln(\boldsymbol{x}_i)$)

Exercise:

Let $f: U \to \mathbb{R}$ be a convex and differentiable function on a convex open U.
Show that if $\nabla f(\boldsymbol{x}^*) = 0$, then $\boldsymbol{x}^*$ is a global minimum of f

Why is convexity an important concept?