Big ole summary

If you find any errors, feel free to send me a message at v.hofmann.1@student.rug.nl

Facts about various functions

$\frac{1}{2}||Ax-b||^2$

- convex
- L-smooth with L = the largest eigenvalue of A^TA
- if A is nonsingular, then f is strongly convex with μ the smallest eigenvalue of A^TA
- gradient is $A^T(Ax-b)$
- hessian is A^TA, so constant. It is also positive semidefinite.
- is differentiable in the sense of Fréchet
- If $\ker(A)=\{0\}$, then $\operatorname{argmin} f$ is a singleton set
- is strictly convex when A is full column rank
- is strongly convex when $A\in\mathbb{R}^{N\times M}$ has full rank and $N<M$.
- satisfies the PŁ Inequality, even if $\ker(A)$ is nontrivial.
- $f(x)=||Ax||$ is coercive
- The symmetric matrix A^TA is positive definite since $\ker(A^TA)=\ker(A)=\{0\}$

Common gradients

- gradient of $||x||^2$ is $2x$
- gradient of $\frac{1}{2}||Ax-b||^2$ is $A^T(Ax-b)$
- gradient of $Ax-b$ is A
- gradient of b^Tx is b
- gradient of $\frac{1}{2}x^TAx$ is $\frac{A^T+A}{2}x$

Chapter 1: Introduction

Definition: Optimization problem
Finding the minimal or maximal value that a function can have on a given set, and the points at which attains at that value, whenever they exist. There are two types: minimization and maximization problems

$\min_{x\in C} f(x)$ is a minimization problem where f is the objective function and C is the feasible set $\subseteq\mathbb{R}^N$. Points in C are called feasible points.

For a maximization problem, we note that $\max_{x\in C} f(x)=-\min_{x\in C}(-f(x))$, so we can turn maximization problems into minimization problems.

Optimization problems are constrained when the feasible set is $C \subsetneq \mathbb{R}^N$, and unconstrained when the feasible set is $\mathbb{R}^N$.

Definition: (global) minimizer
A point $\hat{x} \in C$ such that $f(\hat{x}) \leq f(x)$ for all $x \in C$.
The set of minimizers can be described as $\operatorname{argmin}_C f = \bigcap_{\gamma > \inf_C f} [f \leq \gamma]$

Definition: local minimizer
A point $\hat{x} \in C$ such that $f(\hat{x}) \leq f(x)$ for all $x \in C$ close to $\hat{x}$ give the lowest value of $\mathcal{L}_{\mathcal{D}}$.
(where $\mathcal{L}_{\mathcal{D}}$ is the area around $\hat{x}$)

Notation: $[f \leq \gamma]$
This is the γ-sublevel set of f, aka the set $\{x \in X : f(x) \leq \gamma\}$

Proposition: minimizers for open intervals
Let I be an open intervals and let $f : I \to R$ be differentiable.

- If there is $\overline{x} \in I$ such that $f(\overline{x}) \leq f(x)$ for all $x \in I$, then $f'(\overline{x}) = 0$.
- Assume also that f is twice differentiable, with $f''(x) > 0$ for all $x \in I$. Then, if there is $\overline{x} \in I$ such that $f'(\overline{x}) = 0$, then $f(\overline{x}) < f(x)$ for every $x \in I \setminus \{\overline{x}\}$.

Example: a broken fastest path
Principle of Least Time:
The path taken by a ray between two given points is the one that can be traveled in the shortest time.

Law of Refraction:
For a pair of given media, the ratio of the sines of the angles of incidence and refraction equals the ratio of the refractive indices of the two media.

Examples from data analysis
These were simply examples, we don't have to learn them for the course.

Chapter 2: Existence of minimizers

Metric and Topological structure of $\mathbb{R}^N$

Definition: Euclidean norm
The Euclidean norm of $x \in \mathbb{R}^N$ is $||x|| = \sqrt{x_1^2 + ... + x_N^2}$.

Norm properties
Any norm must satisfy these:

1. $||x|| > 0$ for all $x \neq 0$ and $||0|| = 0$
2. $||\alpha x|| = |\alpha| ||x||$ for all $x \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$
3. $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in \mathbb{R}^N$. This is the triangle inequality.

1-norm and ∞-norm

- $||x||_1 = |x_1| + ... + |x_N|$
- $||x||_\infty = \max\{|x_1|, ..., |x_N|\}$

Comparison of various norms

For every $x \in \mathbb{R}^N$, we have:
$||x||_\infty \leq ||x|| \leq ||x||_1 \leq \sqrt{N}||x|| \leq N||x||_\infty$

Definition: distance

The distance between two vectors $x, y \in \mathbb{R}^N$ is dist$(x, y) = ||x - y||$

Definition: open/closed balls:

Open ball centered at $x \in \mathbb{R}^N$ with radius $r > 0$:
$B(x; r) = \{y \in \mathbb{R}^N : \text{dist}(x, y) < r\}$.

Closed ball centered at $x \in \mathbb{R}^N$ with radius $r > 0$:
$\overline{B}(x; r) = \{y \in \mathbb{R}^N : \text{dist}(x, y) \leq r\}$

Some ball knowledge:

- Balls are convex.
- Open balls are open sets; closed balls are closed sets.
- If B_1, B_2 are balls and $\lambda \in (0, 1)$, then $\lambda B_1 + (1 - \lambda)B_2$ is also a ball.

Definition: open set

A subset $S \subset \mathbb{R}^N$ is open if for each $x \in S$, there is $r > 0$ such that $B(x, r) \subset S$.

Intersections and unions of open sets

- The intersection of a finite number of open sets is open.
- Any union (finite or infinite) of open sets is open.

Definition: interior

The interior of a set $S \subset \mathbb{R}^N$, denoted by int(S), is the largest open set containing S. Equivalently, it is the union of all open subsets of S.

The elements of $\text{int}(S)$ are called interior points of S.

$x \in \text{int}(S) \iff$ there is $r > 0$ such that $B(x, r) \subset S$.

Interiors of open sets

S is open $\iff \text{int}(S) = S$

Defintion: convergence of sequences, limits

A sequence (x_k) in $\mathbb{R}^N$ converges to a limit $\overline{x} \in \mathbb{R}^N$ if $\forall \varepsilon > 0, \exists k_0 \in \mathbb{N}$ s.t. $x_k \in B(\overline{x}; \varepsilon) \forall k \geq k_0$.

Definition: closed set
A subset $S \subset \mathbb{R}^N$ is closed if its complement is open.

S closed $\iff$ every convergent sequence of points in S has its limit in S.

Definition: closure
The closure of a set S, denoted by $\overline{S}$ or $\mathrm{cl}(S)$ is the smallest closed set that contains S. It is also the intersection of all closed sets containing S, and the set of the limits of all convergent sequences of elements in S.

Closures of closed sets
S is closed $\iff \overline{S} = S$

Definition: bounded
A set $B \subset \mathbb{R}^N$ is bounded if it is contained in a ball.

Bolzano-Weierstrass Theorem
Every bounded sequence in $\mathbb{R}^N$ has a convergent subsequence.

Definition: compact
A set is compact if it is both closed and bounded.

Sequences in compact sets
Every sequence in a compact set has a subsequence that converges to a point in the set.

Continuous functions and the existence of minimizers

Definition: continuous at a point
A function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is continuous at $\overline{x} \in D$ if $\lim_{x \to \overline{x}} f(x) = f(\overline{x})$:
$\forall \varepsilon > 0, \exists \delta > 0$ s.t. $||f(\overline{x}) - f(x)| < \varepsilon \forall x \in B(\overline{x}; \delta)$

Definition: continuous
f is continuous on D if is is continuous at every point of D

Another characterization:
$f : D \subset \mathbb{R}^N \to \mathbb{R}$ is continuous at $\overline{x} \in D$ iff for every sequence (x_k) converging to $\overline{x}$, we have $\lim_{k \to \infty} f(x_k) = f(\overline{x})$

Weierstrass Theorem
If $C \subset \mathbb{R}^N$ is compact and $f : C \to \mathbb{R}$ is continuous, there exist $x_{\min}, x_{\max} \in C$ such that $f(x_{\min)} \leq f(x) \leq f(x_{\max})$ for all $x \in C$.

Definition: coercivity
A function $f : \mathbb{R}^N \to \mathbb{R}$ is coercive if $\lim_{||x|| \to \infty} f(x) = \infty$:
$\forall R > 0, \exists \rho > 0$ s.t. $f(x) > R \, \forall x \notin B(0; \rho)$

Another characterization (exercise):
$f : \mathbb{R}^N \to \mathbb{R}$ is coercive $\iff$ for every sequence (x_k) in $\mathbb{R}^N$ such that $\lim_{k\to\infty} ||x_k|| = \infty$, we have $\lim_{k\to\infty} f(x_k) = \infty$.

Minimizers of continuous, coercive functions
If $f : \mathbb{R}^N \to \mathbb{R}$ is continuous and coercive then argmin$(f) \neq \emptyset$

Definition: (sub)-level sets
The (sub)level sets of $f : D \to \mathbb{R}$:

- γ-sublevel set: $[f \leq \gamma] = \{x \in D | f(x) \leq \gamma\}, \gamma \in \mathbb{R}$
- γ-level set: $[f = \gamma] = \{x \in D : f(x) = \gamma\}, \gamma \in \mathbb{R}$

Some facts:

1. the level and sublevel sets of a continuous function are closed
2. the sublevel sets of a coercive function must be bounded

Proposition: minimizers of continuous and coercive functions
If $f : \mathbb{R}^N \to \mathbb{R}$ is continuous and coercive, then $\operatorname{argmin}(f) \neq \emptyset$.

Definition: semi-continuity
$f : D \subset \mathbb{R}^N \to \mathbb{R}$ is lower semi-continuous at $x \in D$ if for every $\varepsilon > 0$, there is $\delta > 0$ such that $f(y) \geq f(x) - \varepsilon$ for every $y \in B(x, \delta) \cap D$.
If f is lower semicontinuous at every point of D, it is lower semi-continuous in D, or just lower semi-continuous.

Every continuous function is lower semi-continuous.

Another characterization:
f is lower semi-continuous at $x \in D \iff f(x) \leq \lim_{k\to\infty} \inf f(x_k)$ for every sequence (x_k) in D that converges to x.

Definition: epigraph
The epigraph of a function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is $\operatorname{epi}(f) = \{(x, t) \in \mathbb{R}^N \times \mathbb{R} : t \geq f(x)\}$

Minimizers of lower semi-continuous functions
Consider $f : D \subset \mathbb{R}^N \to \mathbb{R}$. If:

1. $C \subset D$ is closed
2. f is lower semi-continuous
3. Either C is bounded or f is coercive

Then: $\operatorname{argmin}_C(f) \neq \emptyset$.

Chapter 3: Linear and geometric structure of $\mathbb{R}^N$

Geometry of $\mathbb{R}^N$

Definition: dot product
The dot product (also inner product) of $x, y \in \mathbb{R}^N$ is $x \cdot y = x^T y = x_1 y_1 + ... + x_N y_N$

Dot products have the following properties:

- $x \cdot x = ||x||^2$ for all $x \in \mathbb{R}^N$
- $x \cdot y = y \cdot x$ for all $x, y \in \mathbb{R}^N$
- $(\alpha x + z) \cdot y = \alpha(x \cdot y) + z \cdot y$ for all $x, y, z \in \mathbb{R}^N$ and all $\alpha \in \mathbb{R}^N$

Cauchy-Schwartz inequality
For every $x, y \in \mathbb{R}^N$, we have $|x \cdot y| \leq ||x||||y||$.

Young's Inequality (used to prove Cauchy-Schwartz)
$|x \cdot y| \leq \frac{1}{2t}||x||^2 + \frac{t}{2}||y||^2$ for $t > 0$
Also known as the Peter-Paul Inequality

Parallelogram Identity
For every $x, y \in \mathbb{R}^N$, we have $||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2)$.

Definition: Perpendicular vectors
Also called orthogonal vectors.
Vectors x, y s.t. $x \cdot y = 0$. This is denoted $x \perp y$.

Definition: Parallel vectors
Vectors $x, y \in \mathbb{R}^n \setminus \{0\}$ such that $|x \cdot y| = ||x||||y||$. This is denoted $x||y$.

Another characterization:
$x||y \iff |x \cdot y| = ||x||||y||$

Definition: orthogonal projection
Given $x, y \in \mathbb{R}^N \setminus \{0\}$, the (orthogonal) projection of x onto y is $\text{Proj}_y(x) = \min_{z||y} ||z - x||$
$= \frac{x \cdot y}{||y||^2} y$

Orthogonal projection properties
$\text{Proj}_y(x)||y$ and $x - \text{Proj}_y(x) \perp y$

Definition: angle between two vectors
The angle θ between $x \neq y$ and $y \neq 0$ is $\theta = \cos^{-1}(\frac{x \cdot y}{||x||||y||})$

Law of Cosines and Pythagoras' Theorem
For every $x, y \in \mathbb{R}^N$, we have

1. $||x + y||^2 = ||x||^2 + ||y||^2 + 2||x||||y|| \cos(\theta)$
2. $x \perp y \iff ||x + y||^2 = ||x||^2 + ||y||^2$

Linear transformations and matrices

Definition: linear combination
A linear combination of vectors in $B \subset V$, where V is a vector space is a vector $v = \alpha_1 v_1 + ... + \alpha_k v_k$, with $\alpha_i \in \mathbb{R}$ and $v_i \in B$

Definition: linearly independent
Vectors $v_1, ..., v_k$ are linearly independent if $\sum_{i=1}^{k} \alpha_i v_i = 0$ implies $\alpha_i = 0$

Definition: dimension
The dimension of $B \subset V$, with V a real vector space, is the maximum possible size of linearly independent sets of vectors in B.

Definition: basis
A basis for $B \subset V$, with V a real vector space, is a maximal linearly independent set of vectors in B

Definition: coordinates
The coordinates of a point with respect to a basis are the scalars corresponding to the linear combination of basis vectors that results in the point.

Linear transformations and their representing matrices

Definition: linear transformation
Given, vector spaces U, V, a linear transformation is a map $\Phi : U \to V$ such that $\Phi(u + v) = \Phi(u) + \Phi(v)$ and $\Phi(cv) = c\Phi(v)$ for all $u, v \in U$ and $c \in \mathbb{R}$.

Representation of a linear transformation as a matrix
For any linear transformation $\Phi : \mathbb{R}^N \to \mathbb{R}^M$, there exists a unique matrix $A \in \mathbb{R}^{M \times N}$ such that $\Phi(x) = Ax$ for all $x \in \mathbb{R}^N$. Conversely, every real matrix induces a linear transformation.

Bijectivity necessity
A matrix can only define a bijective function if it is square.

Decomposition of square matrices
Any square matrix Q can be expressed as the sum of a symmetric part and a skew-symmetric part, by writing:
$Q = \frac{1}{2}(Q + Q^T) + \frac{1}{2}(Q - Q^T)$### Distance properties

- dist$(x, y) > 0$ for all $x \neq y$ and dist$(x, x) = 0$
- dist$(x, y) =$ dist(y, x)
- dist$(x, y) \leq$ dist$(x, z)+$ dist(z, y)

Definition: range
Of a linear transformation:
The range of a linear transformation $\Phi : U \to V$ is the space ran$(\Phi) := \{\Phi(u) | u \in U\} \subseteq V$

Of a matrix:
The span of the columns of the matrix.

Definition: span
The span of $B \subset V$, where V is a vector space is $\{\alpha_1 v_1 + ... + \alpha_k v_k | \alpha_i \in \mathbb{R}, v_i \in B\}$

Definition: rank
The rank of a linear transformation $\Phi : U \to V$ is rank$(\Phi) := \dim(\text{ran}(\Phi))$

Definition: kernel
Of a linear transformation:
The kernel of a linear transformation $\Phi : U \to V$ is the space $\ker(\Phi) := \{u \in U | \Phi(u) = 0\} \subseteq U$

Of a matrix:
The orthogonal space to the rows of A $= \text{ran}(A^T)^\perp$

Definition: nullity
The nullity of a linear transformation $\Phi : U \to V$ is $\text{null}(\Phi) := \dim(\ker(\Phi))$.

Some facts about the matrices

- $\ker(A^T A) = \ker(A)$ and $\text{ran}(AA^T) = \text{ran}(A)$
- $\text{ran}(A)$ is the span of the columns of A
- $\ker(A)$ is the orthogonal space to the rows of A
- $\text{ran}(A^T)$ is the span of the rows of A
- $\ker(A) = \text{ran}(A^T)^\perp$

Definition: adjoint
The (real) adjoint of a linear transformation $\phi : \mathbb{R}^N \to \mathbb{R}^M$ is the unique linear transformation $\phi^* : \mathbb{R}^M \to \mathbb{R}^N$ that satisfies
$\phi(u) \cdot v = u \cdot \phi^*(v)$ for every $u \in U$ and $v \in V$.

If ϕ is represented by A, then ϕ^* is represented by A^T.

Rank-nullity theorem
Let $A \in \mathbb{R}^{M \times N}$. Then, each $x \in \mathbb{R}^N$ admits a unique decomposition of the form $x = u + v$, with $u \in \ker(A)$ and $v \in \text{ran}(A^T)$. Moreover, $\text{rank}(A) + \text{null}(A) = N$.

Square matrices

Definition: symmetric matrix
A square matrix Q such that $Q^T = Q$

Defintion: skew-symmetric matrix
A square matrix $Q \in \mathbb{R}^{N \times N}$ such that $x^T Q x = 0$ for every $x \in \mathbb{R}^N$

Definition: eigenvalue
A number $\lambda \in \mathbb{C}$ is an eigenvalue of a square matrix $Q \in \mathbb{C}^{N \times N}$ if there exists a nonzero vector $v \in \mathbb{C}^N$ such that $Qv = \lambda v$.
Note that λ can be complex even if $Q \in \mathbb{R}^{N \times N}$.

Definition: eigenvector
A vector $v \in \mathbb{C}^N$ is an eigenvector for a square matrix $Q \in \mathbb{C}^{N \times N}$ if $Qv = \lambda v$ for some eigenvalue λ of Q. v is then an eigenvector associated to the eigenvalue λ.
Note that v can be complex even if $Q \in \mathbb{R}^{N \times N}$.

Definition: eigenspace
The eigenspace of an eigenvalue λ of some matrix Q is the span of all the eigenvectors associated to λ.

Invariance of eigenspaces
Each eigenspace associated to an eigenvalue of the symmetric matrix Q is invariant under the action of Q. This means that if v is in E_λ, then so is Qv.

Spectral Theorem
Let $Q \in \mathbb{R}^{N \times N}$ be symmetric.
There exist eigenvectors $v_1, ..., v_N \in \mathbb{R}^N$, associated to eigenvalues $\lambda_1, ..., \lambda_N \in \mathbb{R}$, such that $\{v_1, ..., v : N\}$ is an orthonormal basis of $\mathbb{R}^N$, and $Q = \sum_{i=1}^N \lambda_i v_i v_i^T$.

Definition: positive semidefinite matrix
A symmetric matrix $Q \in \mathbb{R}^{N \times N}$ such that $x^T Q x \geq 0$ for all $x \in \mathbb{R}^N$
Notation: $Q \succeq 0$
$Q \succeq \mu I$ means $Q - \mu I$ is positive semidefinite.

Definition: positive definite matrix
A symmetric matrix $Q \in \mathbb{R}^{N \times N}$ such that $x^T Q x > 0$ for all nonzero $x \in \mathbb{R}^N$
Notation: $Q \succ 0$
$Q \succ \mu I$ means $Q - \mu I$ is positive definite.

Definiteness and eigenvalues
Let $Q \in \mathbb{R}^{N \times N}$ be symmetric.

- Q is positive semidefinite $\iff$ all of its eigenvalues are nonnegative.
- Q is positive definite $\iff$ all its eigenvalues are strictly positive.

Inequality with smallest and largest eigenvalue
Let μ be the smallest eigenvalue of A, and $Q = A^T A$. Then:
$||Ax||^2 = x^T A^T A x = x^T Q x \geq \mu ||x||^2$ for all $x \in \mathbb{R}^N$

Similarly, if L is the largest eigenvalue of A, we have:
$||Ax||^2 = x^T Q x \leq L ||x||^2$ for all $x \in \mathbb{R}^N$

Condition number of a matrix:
The ratio $\kappa = \frac{L}{\mu}$, where μ and L are the smallest and largest eigenvalues of the matrix respectively, provide $\mu > 0$.

Definition: invertible matrix
A square matrix Q for which we can find a Q^{-1} such that $QQ^{-1} = Q^{-1}Q = I$

Properties of A^TA
Given $A \in \mathbb{R}^{M \times N}$, the matrix $Q = A^TA$ is symmetric and positive semidefinite.
Furthermore:

- For every $x \in \mathbb{R}^N$, we ave $x \cdot Qx = ||A||^2$
- For every $x \in \mathbb{R}^N$, we have $\mu||x||^2 \leq ||Ax||^2 \leq L||x||^2$, where μ is the smallest eigenvalue of Q and L is the largest eigenvalue of Q
  (see also: condition number)

Various equivalences

Injectivity equivalences:
For a matrix $A \in \mathbb{R}^{M \times N}$, the following are equivalent:

- The function $x \to Ax$ is injective
- $\ker(A) = \{0\}$
- The columns of A are linearly independent.
  In this case, necessarily $N \leq M$.

Surjectivity equivalences:
For a matrix $A \in \mathbb{R}^{M \times N}$, the following are equivalent:

- The function $x \to Ax$ is surjective
- $\ker(A^T) = \{0\}$
- The rows of A are linearly independent.
  In this case, necessarily $N \geq M$.

Invertability equivalences:
Let Q be a square matrix. The following are equivalent:

- Q is invertible (nonsingular)
- the function $x \mapsto Qx$ is bijective
- $\ker(Q) = \{0\}$
- 0 is not an eigenvalue of Q
- The rows of Q are linearly independent
- $\det(Q) \neq 0$
- The columns of Q are linearly independent
  - $\mu > 0$, where μ is the smallest eigenvalue of Q

- If $Q = A^T A$, $\ker(A) = \{0\}$

(Also, $\ker(Q) \neq \{0\} \iff 0$ is an eigenvalue of Q)

Chapter 4: Differentiable Calculus

Differentiable functions and a first order necessity for optimality

Definition: differentiability at a point in the sense of Gâteaux
Let $D \subset \mathbb{R}^N$ be nonempty and open. A function $f : D \to \mathbb{R}$ is differentiable at $x \in D$ in the sense of Gâteaux if the directional derivative $f'(x;h) = \lim_{t \to 0} \frac{f(x+th)-f(x)}{t}$ exists for all $h \in \mathbb{R}^N$,
and there is a $g \in \mathbb{R}^N$ such that $g \cdot h = f'(x;h)$ for all $h \in \mathbb{R}^N$.

NOTE: differentiability in the sense of Gâteaux does NOT imply continuity

In all what follows, differentiable means differentiable in the sense of Gâteaux, unless otherwise stated.

Definition: directional derivative
Let $f : D \to \mathbb{R}$ and $D \subset \mathbb{R}^N$ be nonempty and open. Then the directional derivative in the direction of $h \in \mathbb{R}^N$ is
$f'(x;h) = \lim_{t \to 0} \frac{f(x+th)-f(x)}{t}$

Definition: gradient
The gradient of $f : D \to \mathbb{R}$, with $D \subset \mathbb{R}^N$ is $\nabla f(x) = g$, where g is such that $g \cdot h = f'(x;h)$ for all $h \in \mathbb{R}^N$. The components of the gradient are the partial derivatives:

$$\nabla f(x) = (\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_N}(x))$$

The gradient is orthogonal to the level sets of the function.

Definition: differentiability
A function is differentiable on a nonempty and open $D \subset \mathbb{R}^N$ if it is so at every point of D.

Direction of maximum ascent
Let $D \subset \mathbb{R}^N$ be open and let $f : D \to \mathbb{R}$ be differentiable at $x \in D$. If $\nabla f(x) \neq 0$, then $\nabla f(x)$ points in the direction of maximum ascent. More precisely,
$\frac{\nabla f(x)}{||\nabla f(x)||} \in \operatorname{argmax}\{f'(x;v) : ||v|| = 1\}$

Direction of maximum descent
Let $D \subset \mathbb{R}^N$ be open, and let $f : D \to \mathbb{R}$ be differentiable at $x \in D$. If $\nabla f(x) \neq 0$, the descent directions for f at x contain the half-space $\{d \in \mathbb{R}^N | d \cdot \nabla f(x) < 0\}$.
Moreover, $-\nabla f(x)$ points in the direction of maximum descent.

Fermat's Rule I
Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$ and let $\hat{x} \in \operatorname{argmin}(f) \cap \operatorname{int}(D)$. If f is differentiable at $\hat{x}$, then $\nabla f(\hat{x}) = 0$.

This theorem remains true if D is not necessarily open and f is not necessarily differentiable in all of D, as long as $\hat{x} \in \operatorname{int}(D)$ and f is differentiable at $\hat{x}$.

Stronger first order regularity

Differentiability at a point in the sense of Fréchet:
Let $D \subset \mathbb{R}^N$ be nonempty and open. $f : D \to \mathbb{R}$ is differentiable at $x \in D$ in the sense of Fréchet if there is a vector $G(x) \in \mathbb{R}^N$ such that
$\lim_{h \to \infty} \frac{|f(x+h) - f(x) - G(x) \cdot h|}{||h||} = 0$.

Implications of differentiability in the sense of Fréchet
Let $D \subset \mathbb{R}^N$ be nonempty and open. If $f : D \to \mathbb{R}$ is differentiable at $x \in D$ in the sense of Fréchet, then f is continuous at x and differentiable there in the sense of Gâteaux, with $G(x) = \nabla f(x)$.

Definition: L-smooth
A function $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is L-smooth if it is differentiable and its gradient is Lipschitz-continuous with constant L.

Definition: Lipschitz-continuous:
The gradient ∇f of a functin f is Lipschitz-continuous with constant L if $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for al $x, y \in D$.

Descent lemma
Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$ be L-smooth.
If the line segment joining x and y is contained in D, then

$$|f(y) - f(x) - \nabla f(x) \cdot (y - x)| \leq \frac{L}{2}||x - y||^2$$

If we set $\ell(y) = f(x) + \nabla f(x) \cdot (y - x)$ and $q(x) = \frac{L}{2}||y - x||^2$, then the above result means that $\ell - q \leq f \leq \ell + q$.

An algorithmic consequence is:

$$f(x - t\nabla f(x)) - f(x) \leq -t||\nabla f(x)||^2 + \frac{Lt^2}{2}||\nabla f(x)||^2 = t[\frac{Lt}{2} - 1]||\nabla f(x)||^2$$

If $t < 2/L$, the right hand side is negative. Its smallest possible value is attained when $t = 1/L$, giving: $f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}||\nabla f(x)||^2$

L-smooth and differentiability and continuity
If $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is L-smooth, then it is differentiable in the sense of Fréchet. This immediately implies f is continuous.

Second order optimality conditions

Definition: twice differentiable
Let $D \subset \mathbb{R}^N$ be open and let $f : D \to \mathbb{R}$ be differentiable. We say f is twice differentiable at $x \in D$ if there is a matrix H of size $N \times N$ such that $Hv = \lim_{t \to 0} \frac{\nabla f(xtv) - \nabla f(x)}{t}$ for all $v \in \mathbb{R}^N$.

Definition: Hessian
The Hessian of a twice differentiable function f is $\nabla^2 f(x) = H$ (see above for what H is). If f is twice continuously differentiable, $\nabla^2 f(x)$ must be symmetric for each $x \in D$.

Definition: twice continuously differentiable
A function $f : D \to \mathbb{R}$, with $D \subset \mathbb{R}^N$ open, is twice continuously differentiable (of class $\mathcal{C}^2$) if the function $x \to \nabla^2 f(x)$ is continuous. In this case, $\nabla^2 f(x)$ must be symmetric for each $x \in D$.

Proposition: Second order Taylor approzimation
Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$ be of class $\mathcal{C}^2$, and let $x \in D$. For each $v \in \mathbb{R}^N$,
$\lim_{t \to 0} \frac{1}{t^2} |f(x + tv) - f(x) - t \nabla f(x) \cdot v - \frac{t^2}{2} \nabla^2 f(x) v \cdot v| = 0$

This is a local property, which allows us to understand the geometry of the function in a (possibly small) neighbourhood of a point.

Second-order optimality conditions
Let $D \subset \mathbb{R}^N$ be open, let $f : D \to \mathbb{R}$ be of class $\mathcal{C}^2$, and let $\hat{x} \in D$.

- If $\hat{x} \in \operatorname{argmin}(f)$, then $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(\hat{x})$ is positive semidefinite.
- If $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(\hat{x})$ is positive definite, there is $\varepsilon > 0$ such that $f(\hat{x}) < f(y)$ for every $y \in B(\hat{x}, \varepsilon)$.

Positive semidefinite Hessian
Let $D \subset \mathbb{R}^N$ be open and convex, let $f : D \to \mathbb{R}$ be of class $\mathcal{C}^2$, and let $\hat{x} \in D$. If $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(y)$ is positive semidefinite for every $y \in D$, then $\hat{x} \in \operatorname{argmin}(f)$.

Chapter 5: Convexity

Convex sets

Definition: segment
The segment joining $x, y \in \mathbb{R}^N$ is the set $\text{seg}[x, y] = \{\lambda x + (1 - \lambda) y | \lambda \in (0, 1)\}$.

Definition: convex set
A set $C \subset \mathbb{R}^N$ is convex if $\text{seg}[x, y] \subset C$ whenever $x, y \in C$.

Definition: convex combinatiom
Take points $x_1, ..., x_n$ and let $\lambda_1, ..., \lambda_n \in [0, 1]$ such that $\sum_{i=1}^n \lambda_i = 1$. Then $\lambda_1 x_1 + ... + \lambda_n x_n$

is a convex combination of $x_1, ..., x_n$.

Convex combinations in convex sets
Let C be a convex set. Any convex combination of its elements is also in C.

Intersections and unions of convex sets
The intersection of any (possibly infinite) collection of convex sets is convex.
The union of convex sets is not necessarily convex.

Preservation of convexity
The convexity of a set is preserved under intersection, linear combination, cartesian products and images and inverse images under linear transformations.
Moreover, the interior and closure of a convex set are convex.

Operations on convex sets
Let $\alpha \in \mathbb{R}$, $C_1, C_2 \subset \mathbb{R}^N$ and $C_3 \subset \mathbb{R}^M$, and a linear function $A : \mathbb{R}^N \to \mathbb{R}^M$. If C_1, C_2, C_3 are convex, then so are the following:

- $\alpha C_1 + C_2$
- $C_1 \times C_3$
- $A(C_1)$
- $A^{-1}(C_3)$

Interior and closure of convex sets

- The interior and closure of convex sets are convex.
  Let C be convex with a nonempty interior. Then:
- $x \in \text{int}(C)$ and $y \in \overline{C}$, then $\lambda x + (1 - \lambda)y \in \text{int}(C)$ for all $\lambda \in (0, 1)$.
- $\overline{\text{int}(C)} = \overline{C}$
- $\text{int}(C) = \text{int}(\overline{C})$

Definition: convex hull
The convex hull of a set $S \subset \mathbb{R}^N$ is
$\text{conv}(S) = \{\sum_{j=1}^{J} \lambda_j x_j | x_1, ..., x_j \in S, \lambda_j \in [0, 1] \text{ and } \sum_{j=1}^{J} \lambda_j = 1\}$
(So the convex hull of S contains all convex combinations of the elements of S.)

Another characterization: it is the intersection of all convex sets containing S, which means $\text{conv}(S)$ contains S.

Definition: $(N-1)$-dimensional simplex
Let $e_1, ..., e_N$ denote the canonical vectors of $\mathbb{R}^N$. The set $\text{conv}(\{e_1, ..., e_N\})$ is known as the $(N-1)$-dimensional simplex.
The 1-dimensional simplex is a line; the 2-dimensional simplex is the triangle with vertices $(0, 0, 1), (0, 1, 0), (1, 0, 0)$; the 3-dimensional simplex is a tetrahedron.

Convex hull of a convex set
S is convex $\iff S = \text{conv}(S)$

Hyperplanes in convex sets
If C is a nonempty and convex subset of $\mathbb{R}^N$ not containing the origin, there is $v \in \mathbb{R}^N \setminus \{0\}$ such that $v \cdot x \leq 0$ for each $x \in C$.

Parallel subspaces to convex sets
Let $C \subset \mathbb{R}^N$ be nonempty, open and convex. If $N \geq 2$ and $0 \notin C$, there is a nontrivial subspace of R^N that does not intersect C.

Some examples of convex sets

Definition: box
A set $\prod_{i=1}^{N}[a_i, b_i]$
Boxes are also sometimes called N-dimensional rectangles or parallelipipeds.
Boxes are convex.

Definition: unit cube
The unit cube is the box $[-1, 1<]^N$.

Definition: affine subspace
A "shifted" subspace: $C = \{x_0\} + V = \{x_0 + v : v \in V\}$, where V is a subspace and x_0 is some point.
Affine subspaces are convex.
The dimension of C is equal to the dimension of V.

Definition: open halfspace
$[v < \gamma] = \{x \in \mathbb{R}^N : v \cdot x < \gamma\}$
Open halfspaces are convex.

Definition: closed halfspace
$[v \leq \gamma] = \{x \in \mathbb{R}^N : v \cdot x \leq \gamma\}$
Closed halfspaces are convex.

Definition: hyperplane
The space $[v = \gamma] = \{x \in \mathbb{R}^N | v \cdot x = \gamma\}$
Hyperplanes are convex.

Definition: polyhedron
A set $\{x | Ax \leq b\} = \cap_{j=1}^{J}[v_{j \leq \gamma_j}]$
Polyhedra are convex.

Optimization under convex constraints

Fermat's Rule II
Let $D \subset \mathbb{R}^N$ be open, and let $C \subset D$ be nonempty, closed and convex. Consider a

differentiable function $f : D \to \mathbb{R}$ and a point $\hat{x} \in \operatorname{argmin}_C f$. Then, $\nabla f(\hat{x}) \cdot (y - \hat{x}) \geq 0$ for every $y \in C$.

Result of Fermat's Rule II for affine spaces
Let $D \subset \mathbb{R}^N$ be open, let $f : D \to \mathbb{R}$, and let $C = \{x_0\} + V$ be an affine subspace of $\mathbb{R}^N$ contained in D. If $\hat{x} \in \operatorname{argmin}_C(f)$, then $\nabla f(\hat{x}) \perp V$.

Projection of points onto sets
Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex. For each $x \in \mathbb{R}^N$, there is a unique point $P_C(x) \in C$, called the projection of x onto C, such that

$$||x - P_C(x)|| = \min\{||x - z|| : z \in C\} = \operatorname{dist}(x, C)$$

Moreover, $P_C(x)$ is the only point in C that satisfies the inequality

$$(x - P_C(x)) \cdot (z - P_C(x)) \leq 0$$

for all $z \in C$.

The function $P_{\mathbb{C}} : \mathbb{R}^N \to \mathbb{R}^N$ is nonexpansive, so $||P_C(x) - P_C(y)|| \leq ||x - y|$ for all $x, y \in \mathbb{R}^N$.

If $C = \{x \in \mathbb{R}^N : Ax = b\}$ then $P_C(x) - x \in [\ker(A)]^{\perp} + \operatorname{ran}((A^T)$.

Projection onto affine subspaces
If $C = \{x_0\} + V$ is an affine subspace of $\mathbb{R}^N$, then $P_C(x) - x \perp V$.

Definition: nonexpansive function
A function f such that $||f(x) - f(y)|| \leq ||x - y||$ for every x, y in the domain.

Decomposition of complex sets
Let C be closed and convex. Then there exists a family $(H_i)_{i \in \mathcal{I}}$ of closed halfspaces such that $C = \cap_{i \in \mathcal{I}} H_i$.

Convex functions

Definition: convex function
A function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is convex if D is convex and $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y)$ for all $x, y \in D$ and all $\lambda \in (0, 1)$.
Example: the function $\frac{1}{2}||Ax - b||^2$ is convex.

Another characterization:
$f : D \subset \mathbb{R}^N \to \mathbb{R}$ is convex $\iff$ $\operatorname{epi}(f)$ is a convex subset of $\mathbb{R}^N \times \mathbb{R}$.

Convex functions and convex combinations
Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex function. For all $k \geq 1$, for all $x_1, ..., x_k \in R^N$ and $\lambda_1, ..., \lambda_k \in [0, 1]$ satisfying $\lambda_1 + ... + \lambda_k = 1$, it holds that $f(\lambda_1 x_1 + ... + \lambda_k x_k)) \leq \lambda_1 f(x_1) + ...$

$+ \lambda_k f(x_k)$

Properties of convex functions

- Every sublevel set of a convex function is convex.
- If f is convex, then $\operatorname{argmin}(f)$ is convex.
- If f is convex and bounded from above, then f must be constant. More generally, if $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is bounded from above on an affine subspace $V \subseteq D$ of $\mathbb{R}^N$, then f is constant on V.
- Let $f_i : D_i \subset \mathbb{R}^N \to \mathbb{R}$ be convex for $i \in \{1, 2\}$ and let $a \geq 0$. Then the function $f : D_1 \cap D_2 \to \mathbb{R}$, defined by $f = af_1 + f_2$, is convex.

Definition: superadditive
$f : D \to \mathbb{R}$ is superadditive on D if for every $x, y \in D$, we have $f(x) + f(y) \leq f(x + y)$.

Definition: midpoint-convex
A function is midpoint-convex if $f(\frac{x+y}{2}) \leq \frac{f(x)+f(y)}{2}$ for all $x, y \in \mathbb{R}^N$.
If f is also continuous, midpoint-convexity implies convexity.

Interior of the epigraph
If $D \subset \mathbb{R}^N$ has a nonempty interior, and f has at least one point of continuity, then $\operatorname{epi}(f)$ has a nonempty interior.

Continuity on the interior of convex functions
If $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is convex, then for each $x \in \operatorname{int}(D)$ there exists $L_x, r_x > 0$ such that $|f(z) - f(y)| \leq L_x||z - y||$ for all $z, y \in B(x, r_x)$. In particular, f is continuous on $\operatorname{int}(D)$.

Convexity equivalences
Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$ be differentiable. The following are equivalent:

- f is convex
- for all $x, y \in D$, $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$;
- for all $x, y \in D$, $(\nabla f(y) - \nabla f(x)) \cdot (y - x) \geq 0$
  If f is twice differentiable, the three statements above are equivalent to:
- for all $x \in D$, $\nabla^2 f(x)$ is positive semidefinite

Baillon-Haddad Theorem
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and differentiable, and let $L > 0$. The following statements are equivalent:

- f is L-smooth
- $f(x) + \nabla f(x) \cdot (z - x) + \frac{1}{2L}||\nabla f(z) - \nabla f(x)||^2 \leq f(z)$, for all $x, z \in \mathbb{R}^N$
- $f(z) \leq f(x) + \nabla f(x) \cdot (z - x) + \frac{L}{2}||z - x||^2$ for all $x, z \in \mathbb{R}^N$
- $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{1}{L}||\nabla f(x) - \nabla f(y)||^2$, for all $x, y \in \mathbb{R}^N$.

Convex and L-smooth functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth and suppose $\operatorname{argmin}(f) \neq \emptyset$. Then for all $x \in \mathbb{R}^N$, we have $f(x) - \min f \geq \frac{1}{2L}||\nabla f(x)||^2$

Strict and strong convexity

Definition: strict convexity
A function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is strictly convex if D is convex and $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$ for all $x, y \in D, x \neq y$, and all $\lambda \in (0,1)$

Strictly convex functions are coercive.

See also μ-strongly convex

Minimizers of strictly convex functions
A strictly convex function can have at most one minimizer.

Definition: μ-strongly convex
Given $\mu > 0$, a function $f : D \subset \mathbb{R}^N \to \mathbb{R}$ is μ-strongly convex, if D is convex and $f(\lambda x + (1\ \lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)||x-y||^2$ for every $x, y \in D$ and all $\lambda \in (0,1)$.
Strongly convex functions are strictly convex, which implies they have at most one minimum.

Strongly convex functions cannot be Lipschitz continuous.

Every strongly convex function is coercive.

Strong and normal convexity

- f is μ-strongly convex iff $g(x) = f(x) - \frac{\mu}{2}||x||^2$ is convex.
- If f is convex and g is strictly (resp. strongly) convex, then $f + g$ is strictly (resp. strongly) convex

μ-strongly convex equivalences
Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$ be differentiable. The following are equivalent:

- f is μ-strongly convex
- for all $x, y \in D$, $f(y) \geq f(x) + \nabla f(x) \cdot (y-x) + \frac{\mu}{2}||x-y||^2$
- for all $x, y \in D$, $(\nabla f(y) - \nabla f(x)) \cdot (y-x) \geq \mu||x-y||^2$
  If f is twice differentiable, the three statements above are equivalent to
- for all $x \in D$, $\nabla^2 f(x)$ is positive definite, with $v \cdot \nabla^2 f(x) v \geq \mu||v||^2$ for every $v \in \mathbb{R}^N$

L-smooth and μ-strongly convex inequalities
If $f : \mathbb{R}^N \to \mathbb{R}$ is L-smooth and μ-strongly convex, then

- $(\nabla f(x) - \nabla f(y)) \cdot (x-y) \geq \lambda\mu||x-y||^2 + \frac{1-\lambda}{L}||\nabla f(x) - \nabla f(y)||^2$, for every $x, y \in \mathbb{R}^N$ and $\lambda \in [0,1]$.

- $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{\mu L}{L+\mu} ||\nabla f(x) - \nabla f(y)||^2$, for every $x, y \in \mathbb{R}^N$.
- $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{\mu L}{L+\mu} ||x + y||^2 + \frac{1}{L+\mu} ||\nabla f(x) - \nabla f(y)||^2$ for all $x, y \in \mathbb{R}^N$

Chapter 6: Numerical approach to smooth unconstrained minimization

Iterative algorithms

Definition: iterative algorithm
An algorithm that produces a sequence (x_k) of approximate solutions to an optimization problem. It involves:

- An initial guess x_0
- Possible a sequence (u_k) in $\mathbb{R}^M$ of parameters, either defined by the user or computed by a subroutine
- An operator $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ used to compute x_{k+1}, given (x_k, u_k) as $x_{k+1} = F(x_k, u_k)$.
- A stopping rule that is activated and terminates the algorithm, when the approximation is considered sufficiently good.

Definition: complexity
How much of a given resource (no. of iterations, computational time) an algorithm needs in order to produce some outcome

Definition: merit function
A merit function for the problem of minimizing f is a usually continuous function $\phi : \mathbb{R}^N \to \mathbb{R}_+$ such that

- $\phi(\hat{x}) = 0 \iff \hat{x} \in \operatorname{argmin}(f)$, and
- z is a better approximate solution than w if, and only if, $\phi(z) < \phi(w)$.

Some common merit functions

1. $\phi(z) = f(z) - \inf(f)$
2. $\phi(z) = \operatorname{dist}(z, \operatorname{argmin}(f))$ or $\phi(z) = \operatorname{dist}(z, \operatorname{argmin}(f))^2$
3. $\phi(z) = ||z - \hat{x}||$ or $\phi(z) = ||z - \hat{x}||^2$, if $\operatorname{argmin}(f) = \{\hat{x}\}$
4. $\phi(z) = ||\nabla f(z)||$ or $\phi(z) = ||\nabla f(z)||^2$, if f is convex & differentiable

Definition: convergence
Whether or not an algorithm is able to produce a sufficiently good approximate solution
An algorithm converges if $\lim_{k\to\infty} \phi(x_k) = 0$ for every sequence (x_k) it generates.

Definition: convergence rate
A convergence rate expresses how fast an algorithm converges. More precisely, it is a

description of how fast a nonnegative sequence (ϕ_k) converges to zero.

Some useful notation:

- Let $(\rho_k) \to 0$ be positive. We say $\phi_k = O(\rho_k)$ if $\phi_k \to 0$ at least as fast as ρ_k, namely if $C := \sup_{k \geq 0}[\frac{\phi_k}{\rho_k}] < \infty$
- Let $(\rho_k) \to 0$ be positive. We say $\phi_k = o(\rho_k)$ if $\phi_k \to 0$ at strictly faster than ρ_k, namely if:
  $\lim_{k \to \infty}[\frac{\phi_k}{\rho_k}] = 0$
  $\phi_k = o(\rho_k)$ implies $\phi_k = O(\rho_k)$

The convergence rate of a sequence (x_k) in $\mathbb{R}^N$ to a point $\hat{x} \in \mathbb{R}^N$ is defined as the convergence rate of $||x_k - \hat{x}||$ to zero.

Definition: Linear convergence
An algorithm converges linearly if there exists $C > 0$ and $Q \in (0, 1)$ such that $\phi(x_k)$ $\leq CQ^k$ for every $k \geq 0$.

In this case, the algorithm produces an x_ε such that $\phi(x_\varepsilon) < \varepsilon$ in at most $\lceil \log_Q(\frac{\varepsilon}{C}) \rceil$ iterations (see complexity).

Using O-notation: if $\phi_k = O(q^k)$ for some $q \in (0, 1)$, we say ϕ_k converges linearly to zero with rate q, or that it converges q-linearly to zero.

Definition: sublinear convergence
If $\phi_k \neq O(q^k)$ for every $q \in (0, 1)$ it converges sublinearly. Often, we still have $\phi_k = O(k^{-p})$ for some $p > 0$.

Definition: superlinear convergence
If $\phi_k = o(q^k)$ for all $q \in (0, 1)$, it converges superlinearly.

Definition: convergence in the sense of the ratio test
(ϕ_k) converges to zero with order $p > 0$ and rate $R \in (0, 1)$ in the sense of the ratio test if $\lim_{k \to \infty} \sup[\frac{\phi_{k+1}}{\phi_k^p}] = R$

If (ϕ_k) converges to zero with order 1 and rate $\rho \in (0, 1)$ in the sense of the ratio test, it converges ρ-linearly to zero.

Definition: polynomial convergence
$\phi_(x_k) \leq \frac{C}{k^p}$ for $p > 0$

Definition: complexity bound
Consider a problem with merit function ϕ and a sequence (x_k) produced by an algorithm.
For every $\varepsilon > 0$, which represents a tolerance or precision level, we have the nonempty set of ε-approximate solutions $\sum_\varepsilon := \{k : \phi(x_k) < \varepsilon\}$.

A complexity bound is a description, estimation or approximation of the set $\sum_{\varepsilon}$.
We can express a complexity bound by stating that $\phi(x_k) < \varepsilon$ for every $k \geq k_{\varepsilon}$, where k_{ε} is a number that can be computed as a function of ε.

Descent directions and the Gradient Method

Definition: descent direction and descent step sizes
Let $D \subset \mathbb{R}^N$ be nonempty and open, and let $f : D \to \mathbb{R}$.
A vector $d \neq 0$ is a descent direction for f at x if there is $\Gamma_d > 0$ such that $f(x + \gamma d) < f(x)$ for all $\gamma \in (0, \Gamma_d)$.
The numbers $\gamma \in (0, \Gamma_d)$ are descent step sizes.

Steepest descent direction
If f is differentiable at x and $\nabla f(x) \neq 0$, then $-\nabla f(x)$ is the steepest descent direction for f at x.

Finding descent directions
If f is differentiable at x_0 and $\nabla f(x_0) \neq 0$, then every vector $d \neq 0$, satisfying $\nabla f(x_0) \cdot d < 0$, is a descent direction for f at x.

Definition: descent method
A descent method iterates $x_{k+1} = x_k + \gamma_k d_k$ for $k \leq 0$, where d_k is a descent direction and γ_k is a descent step size.
An obvious choice is $d_k = -\nabla f(x_k)$. If f is L-smooth, every $\gamma \in (0, \frac{2}{L})$ is a descent step size.

The Gradient Method (gradient desent)
Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth, let $x_0 \in \mathbb{R}^N$ and let $\gamma \in (0, \frac{2}{L})$. The iterative algorithm defined by $x_{k+1} = x_k - \gamma \nabla f(x_k)$ is known as the Gradient Method.

We have that $f(x_{k+1}) \leq f(x_k) + \gamma(\frac{\gamma L}{2} - 1)||\nabla f(x_k)||^2 \leq f(x_k)$.

Gradient method for L-smooth bounded functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below, let $\gamma \in (0, \frac{2}{L})$, and let $x_0 \in \mathbb{R}^N$. For $k \geq 0$, iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Then,

- $\lim_{k \to \infty} f(x_k)$ exists and $\lim_{k \to \infty} ||\nabla f(x_k)|| = 0$;
- For every $k \geq 0$, we have $\min\{||\nabla f(x_j)|| : 0 \leq j \leq k\} \leq \sqrt{\frac{2[f(x_0) - \inf(f)]}{\gamma(2 - \gamma L)k + 1}}$
- Every subsequential limit point of x_k is critical, which means that if $x_{m_k} \to \hat{x}$, then $\nabla f(\hat{x}) = 0$; and
- If f has no critical points, then $\lim_{k \to \infty} ||x_k|| = +\infty$.

Gradient method for L-smooth, bounded and convex functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth, convex and bounded from below, let $\gamma \in (0, \frac{2}{L})$ and let $x_0 \in \mathbb{R}^N$. For $k \geq 0$, iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Then,

1. $\lim_{k\to\infty} f(x_k) = \inf(f)$
2. If f has minimizers, then x_k converges to one of them with $f(x_k) - \min(f) \leq \frac{\text{dist}(x_0, \text{argmin}(f))^2}{\gamma(2-\gamma L)(k+1)}$ and $\lim_{k\to\infty} k[f(x_k) - \min(f)] = 0$.

Remark: the best theoretical convergence rate is achieved when $\gamma = \frac{1}{L}$.

Lemmas used to prove the above theorem

1. Let $(a_k), (b_k)$ and (ε_k) be nonnegative sequences such that (ε_k) is summable and $a_{k+1} - a_k + b_k \leq \varepsilon_{k+1}$ for every $k \geq 0$. Then: $\lim_{k\to\infty} a_k$ exists and (b_k) is summable.
   Note: *Summable* means that the sum $\sum_{i=1}^{k} \varepsilon_k$ converges.
2. Let (c_k) be nonnegative, nonincreasing and summable. Then $\lim_{k\to\infty} kc_k = 0$.

Gradient method for μ-strongly convex functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and μ-strongly convex, let $\gamma \in (0, \frac{2}{L+\mu}]$ and let $x_0 \in \mathbb{R}^N$. For $k \geq 0$, iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Then,

$$||x_k - \hat{x}|| \leq (1 - \frac{2\gamma\mu L}{L+\mu})^{k/2} ||x_0 - \hat{x}||$$

where $\hat{x}$ is the unique minimizer of f. In particular, taking $\gamma = \frac{2}{L+\mu}$, we obtain $||x_k - \hat{x}|| \leq (\frac{L-\mu}{L+\mu})^k ||x_0 - \hat{x}||$.

Polyak-Łojasiewicz Inequality:
A function $f : \mathbb{R}^N \to \mathbb{R}$ satisfies this inequality with constant $\mu > 0$ (or is $\mu - P$Ł) if

$$2\mu(f(x) - \min(f)) \leq ||\nabla f(x)||^2$$

for all $x \in \mathbb{R}^N$.

Properties:

- If $f : \mathbb{R}^N \to \mathbb{R}$ is μ-strongly convex, it is $\mu-$PŁ.
- If $g : \mathbb{R}^M \to \mathbb{R}$ is μ-strongly convex and $A \in \mathbb{R}^{M\times N}$, then $g \circ A$ is PŁ.

Gradient Method for PŁ functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and satisfy the Polyak-Łojasiewicz inequality with constant $\mu > 0$. Pick $\gamma \in (0, \frac{2}{L})$, and iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Then, $f(x_k)$ converges linearly to $\min(f)$, with
$f(x_k) - \min(f) \leq (1 - \mu\gamma(2 - \gamma L))^k (f(x_0) - \min(f))$ for all $k \geq 1$.

In particular, for $\gamma = \frac{1}{L}$, we obtain $f(x_k) - \min(f) \leq (1 - \frac{\mu}{L})^k (f(x_0) - \min(f))$.

Step size selection

Fixed or constant step sizes
These are the step sizes discussed in the above section. If the objective function f is L-

smooth, we can fix any $\gamma \in (0, \frac{2}{L})$, where the best theoretical choice is $\gamma = \frac{1}{L}$ and define $\gamma_k = \gamma$.
Drawbacks: this strategy requires knowledge or at least a good estimation of L.

Vanishing step sizes (also called open loop strategies)
Instead of a fixed or constant step size, we have vanishing step sizes $\gamma_k \to 0$. No matter how large L is, the step sizes will eventually fall in the correct interval. However, if γ_k tends to 0 too fast, the algorithm may converge slowly, or even fail to converge.

Theorem:
Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below, let (γ_k) be a positive sequence such that $\lim_{k\to\infty} \gamma_k = 0$ and $\sum_{k\geq 0} \gamma_k = \infty$. Starting from $x_0 \in \mathbb{R}^N$, iterate $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ for $k \geq 0$. Then,
$\lim_{k\to\infty} f(x_k)$ exists, $\lim_{k\to\infty} ||\nabla f(x_k)|| = 0$ and every subsequential limit points of (x_k) is critical (i.e. if $x_{m_k} \to \hat{x}$ then $\nabla f(\hat{x}) = 0$.)

Line search

Definition: line search strategies
Line search strategies select a descent step size once a descent direction has been chosen. At each iteration a decision is made and the current step size is chosen based on the available information. In this case, they are feedback strategies.

Greedy strategy, also exact line search strategy
One way to choose the step size is to choose the one that gives the smallest function value along the ray defined by the descent direction, namely:
$\gamma_k := \text{argmin}_{\gamma>0} f(x_k + \gamma d_k)$, or $\gamma_k = \text{argmin}_{\gamma\in(0,\bar{\gamma})]} f(x_k + \gamma d_k)$.

Backtracking
Instead of solving the minimization problem in the greedy strategy analytically at each step, it is common to apply algorithms to find approximate solutions.
First, fix an initial guess $\bar{\gamma} > 0$ that defines the part of the ray that will be tested, along with a shrinking factor $\beta \in (0, 1)$. At the k-ith iteration, we have computed x_k and selected d_k. Then defines the step size as:
$\gamma_k = \beta^{m_k}\bar{\gamma}$,
where m_k is the smallest nonnnegative integer such that $\gamma = \beta^m\bar{\gamma}$ satisfies a given set of conditions.

The process of finding m_k is called backtracking because it involves choosing a larger step size and then shrinking it.

The most popular condition is Armijo's Rule, which requires $f(x_k + \gamma d_k) \leq f(x_k) + \sigma\gamma\nabla f(x_k) \cdot d_k$, where $\sigma \in (0, 1)$, typically $\sim 10^{-4}$

We can add additional conditions, such as:

- Goldstein's Rule
  $f(x_k + \gamma d_k) \geq f(x_k) + (1-\sigma)\gamma \nabla f(x_k) \cdot d_k$.
- The Weak Wolfe Rule
  $\nabla f(x_k + \gamma d_k) \cdot d_k \geq \tau \nabla f(x_k) \cdot d_k$, where $\tau \in (\sigma, 1)$, typically in $(0.1, 0.9)$
- The Strong Wolfe Rule
  $|\nabla f(x_k + \gamma d_k) \cdot d_k| \leq \tau |\nabla f(x_k) \cdot d_k|$

The combination of Armijo's and Goldstein's Rule can be expressed as
$\sigma \leq \frac{f(x_k+\gamma d_k)-f(x_k)}{\gamma \nabla f(x_k) \cdot d_k} \leq 1-\sigma$
The denominator is negative, so we prefer smaller values of σ.

Acceleration methods

The gradient method works, but converges slowly. This section discusses methods that aim to improve on the basic gradient descent. Apart from the $\nabla f(x_k)$ term, they also work with a momentum term $(x_k - x_{k+1})$.

Heavy-Ball algorithm
We use the following two-step algorithm to iterate over
$\begin{cases} y_k = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_k - \alpha \nabla f(x_k) \end{cases}$

Note that the lecture slides use $\beta = \theta_k$ and $\alpha = \gamma_k$

The step to define y_k is an extrapolation or momentum step.
Note that the following theorem only proves convergence for quadratic functions

Convergence of the Heavy-Ball algorithm for quadratic functions
Let $0 < \mu < L$ and let $f : \mathbb{R}^N \to \mathbb{R}$ be a μ-strongly convex, L-smooth quadratic function.

If we set $\alpha = \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}$ and $\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} \right)^2$
there exists a constant $C > 0$, independent of μ and L, such that
$f(x_k) - \min f \leq Ck\beta^k ||x_0 - x_*||^2$
where (x_k) is defined by the Heavy-Ball algorithm above and $x_* = \operatorname{argmin} f$

Nesterov accelerated gradient descent
Similar to the heavy-ball algorithm, but modified slightly.
$\begin{cases} y_k = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_k - \alpha \nabla f(y_k) \end{cases}$
Note that here for x_{k+1} we use the gradient of y_k instead of x_k.
The lecture slides use $\beta = \theta_k$ and $\alpha = \gamma_k$.

Convergence of Nesterov's method (lecture slides)
Let $f : \mathbb{R}^N \to \mathbb{R}$ be an L-smooth convex function with minimizers, and let (x_k, y_k) be generated by Nesterov's method with convenient γ_k, θ_k. Let $S = \operatorname{argmin}(f)$.

- Then, $f(x_k) - \min(f) \leq \frac{L\text{dist}(x_0,S)^2}{(k+1)^2}$ for all $k \geq 1$. In addition, $\lim_{k\to\infty} k^2(f(x_k) - \min(f)) = 0$.
- If, moreover, f is μ-strongly convex, then
  $f(x_k) - \min(f) \leq L\text{dist}(x_0, S)^2(1 - \sqrt{\frac{\mu}{L}})^k$ for all $k \geq 1$.

Results for the Nesterov method
Let $0 < \mu < L$ and let $F : \mathbb{R}^N \to \mathbb{R}$ be a μ-strongly convex L-smooth function. By setting $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$
then
$f(x_k) - \min f \leq 2(1 - \frac{\sqrt{\mu}}{\sqrt{L}})^k(f(x_0) - \min f)$
where x_k is calculated acoording to the Nesterov method.

Quadratic function: Conjugate Gradient Method

Definition: quadratic function
Functions of the form $f(x) = \frac{1}{2}x^TQx - b^Tx + c$,
where $Q \in \mathbb{R}^{N\times N}$ satisfies $Q \succ \mu I$ for some $\mu > 0$, and $b \in \mathbb{R}^N$.
$\nabla f(x) = Qx - b$, since Q is symmetric.
These functions are strongly convex.

Note that $\frac{1}{2}||Ax - b||^2$ is a quadratic function.
If $\ker(A) = \{0\}$ and $b \in \text{ran}(A)$, then the function has a unique minimizer. That minimizer is also the unique minimizer of $\phi(x) = \frac{1}{2}x \cdot Qx - c \cdot x$, where $Q = A^TA$ (symmetric and positive definite) and $c = A^Tb$.

Definition: conjugate
Vectors $\{d_0, ..., d_{N-1}\}$ are conjugate with respect to Q if $d_i \cdot Qd_j = 0$ whenever $i \neq j$,
where we use the conjugate inner product.
Conjugate vectors are linearly independent.

Definition: Conjugate inner product
$< x, y >_Q := x \cdot Qy$ is an inner product in $\mathbb{R}^N$. Being conjugate means being orthogonal with respect to this inner product.

Conjugate gradient method (expanding subspace minimization)
Intuitive idea: we do line search along one dimension of the space each time, picking the minimum every time until we find the absolute minimum.

$$\begin{cases} x_{k+1} &= x_k + \alpha_k p_k \\ r_{k+1} &= r_k + \alpha_k Q p_k \\ p_{k+1} &= -r_{k+1} + \gamma_k p_k \end{cases}$$

Where we define:

- $\alpha_{-1} = 0$
- $\gamma_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$

- $r_k = \nabla f(x_k) = Qx_k - b$
- $\alpha_k = -\frac{p_k^T r_k}{p_k^T Q p_k}$

Stopping condition: $r_k = 0$.

We choose α_k such that $f(x_{k+1}) = f(x_k + \alpha_k p_k)$ is minimized (so using greedy line search)
We choose γ_k such that the search directions p_k and p_{k+1} are conjugate with respect to Q. Thus, the search directions form a linearly independent set and x_k is the minimizer over f over the affine subspace $x_0 + \text{span}(p_0, ..., p_k)$.

If Q has exactly k distinct eigenvalues, the algorith terminates in at most k steps (so searching along one eigenvector is enough to "search" all of the ones for that eigenvalue.)

Nonlinear extensions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.
Given $x_0 \in \mathbb{R}^N$, set $p_0 = -\nabla f(x_0)$.
At step n, we know x_n, $r_n = \nabla f(x_n)$ and p_n.
If $r_n = 0$, we stop; otherwise, we compute (backtracking) α_n satisfying
$f(x_n + \alpha_n p_n) \leq f(x_n) + c_1 \alpha_n r_n \cdot p_n$
$|\nabla f(x_n + \alpha_n d_n) \cdot p_n| \leq -c_2 r_n \cdot p_n$
with $0 < c_1 < c_2 < \frac{1}{2}$ (strong Wolfe conditions)

Then, we update
$x_{n+1} = x_n + \alpha_n p_n$
$p_{n+1} = -r_{n+1} + \beta_n p_n$
where several choices for β_n are possible, such as:

- Fletcher-Reeves: $\frac{||r_{n+1}||^2}{||r_n||^2}$
- Polak-Ribière: $\frac{r_{n+1} \cdot (r_{n+1} - r_n)}{||r_n||^2}$
- Hestenes-Stiefel: $\frac{r_{n+1} \cdot (r_{n+1} - r_n)}{p_n \cdot (r_{n+1} - r_n)}$
- Dai-Yuan: $\frac{||r_{n+1}||^2}{p_n \cdot (r_{n+1} - r_n)}$

Second-order methods

Newton's method
For convex functions, we are looking for a point that minimizes a given function f, such that $\nabla f(x) = 0$. By Taylor's approzimation, we know that, for any point y close to x, $\nabla f(x) \approx \nabla f(y) + \nabla^2 f(y)(x - y)$.

We cannot solve for $\nabla f(x) = 0$ analytically, but we can find the roots of the approximation analytically: $x = y - [\nabla^2 f(y)]^{-1} \nabla f(y)$. We can iterate this with an initial

point $x_0 \in \mathbb{R}^N$, giving us:

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

Convergence of Newton's method
Consider $f : \mathbb{R}^N \to \mathbb{R}$ and $\hat{x} \in \mathbb{R}^N$ such that $\nabla f(\hat{x}) = 0$.

Suppose $||\nabla^2 f(x) - \nabla^2 f(y)|| \leq L||x - y||$ and $||\nabla^2 f(x)^{-1}|| \leq M$ for all x, y in a neighbourhood of $\hat{x}$.

Then, there is $\delta > 0$ such that if $||x_0 - \hat{x}|| < \delta$, then $||x_n - \hat{x}|| \leq r^{2^n}$ for some $r \in (0, 1)$ and all $n \geq 0$.

The actual convergence rate is a bit too complicated for this course, but it is a lot faster than the standard gradient descent. On the other hand, it costs a lot to compute the Hessian and its inverse each iteration, especially for large N.

Quasi-Newton methods

The secant condition
The key inequality defining Newton's method is the secant condition:

$$\nabla^2 f(x_n)^{-1}[\nabla f(x_{n+1}) - \nabla f(x_n)] = x_{n+1} - x_n$$

Quasi-Newton methods
These methods are similar to Newton's method, but they compute a matrix $D_k \sim \nabla^2 f(x_k)^{-1}$ each iteration to save on computational costs.

For example, some methods choose $p \in \mathbb{N}$ and define $D_n = \nabla^2 f(x_{kp})^{-1}$ for $n = kp, kp + 1, ..., (k+1)p - 1$. This is called **periodic evaluation**.

D_{k+1} is constructed to be symmetric, like the Hessian, and to satisfy the secant condition: $D_{k+1} g_k = s_k$, where $g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ and $s_k = x_{k+1} - x_k$

Some popular methods to compute D_{k+1} are:

- DFP: Davidson (1959), Fletcher and Powell (1987)
  $D_{n+1} = D_n - \frac{(D_n g_n)(D_n g_n)^T}{g_n \cdot D_n g_n} + \rho_n (s_n s_n^T)$, $\rho_n = \frac{1}{g_n \cdot s_n}$
- BFGS: Broyden, Fletcher, Goldfarb and Shanno (1970)
  $D_{n+1} = (I - \rho_n s_n g_n^T) D_n (I - \rho_n s_n g_n^T)^T + \rho_n (s_n s_n^T)$, $\rho_n = \frac{1}{g_n \cdot s_n}$

Convergence of quasi-Newton methods
The convergence rate does suffer from the approximations, but because of the computational costs being a lot less, these methods are still faster than Newton's.

Let $f : \mathbb{R}^N \to \mathbb{R}$ be μ-strongly convex and $L-$smooth, and let D_0 be positive definite. Then, x_n converges to the minimizer of f. It does so in at most N steps if f is quadratic.

Chapter 7: Minimization of nonsmooth convex functions

Subgradients and subdifferential of a convex function

Definition: subgradient
For f convex and differentiable at x, we had $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$ for all $y \in \mathbb{R}^N$.
Now, we say a vector $v \in \mathbb{R}^N$ is a subgradient of f at x if
$f(y) \geq f(x) + v \cdot (y - x)$ for all $y \in \mathbb{R}^N$.

Definition: subdifferential
The subdifferential of f at x, denoted by $\partial f(x)$, is the set of all the subgradients of f at x.

Monotonicity of ∂f
If $f : \mathbb{R}^N \to \mathbb{R}$ is convex, then ∂f is a monotone operator. Specfically, for all $x, y \in \mathbb{R}^N$ and $u \in \partial f(x)$, $v \in \partial f(y)$, it holds that $(x - y) \cdot (u - v) \geq 0$.

Fermat's Rule III
Let $f : \mathbb{R}^N \to \mathbb{R}$. Then $\hat{x} \in \operatorname{argmin}(f)$ if, and only if, $0 \in \partial f(\hat{x})$

Examples of some subdifferentials
For $f(x) = |x|, \partial f(x) = \begin{cases} \{1\} & x > 0 \\ [-1, 1] & x = 0 \\ \{-1\} & x < 0 \end{cases}$
For $f(x) = ||x||$, $\partial f(x) = \begin{cases} \{\frac{x}{||x||}\} & x \neq 0 \\ \overline{B}(0, 1) & x = 0 \end{cases}$.
For $f(x) = |x_1| + ... + |x_N|$ (the $1-$norm). See "piecewise subdifferentials" and note that each "piece" is the absolute value function, which we did above.
For $f(x) = \max\{f_1(x), ..., f_M(x)\}, \text{ with } f_1, ..., f_M : \mathbb{R}^N \to \mathbb{R}$ convex, let $a(x) = \{i : f(x) = f_i(x)\}$. Then, ∂f(x) = conv$(\cup_{i \in a(x)} \partial f_i(x))$

Properties of subdifferentials
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and let $x \in \mathbb{R}^N$. Then the set $\partial f(x)$ is closed, convex, nonempty and bounded.
(The domain here is important. Closedness and convexity hold for all cases, however.)

Subdifferential for differentiable functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex. If f is differentiable at $x \in \mathbb{R}^N$, then $\partial f(x) = \{\nabla f(x)\}$
(The domain here is important.)

Subdifferentials of a sum
Let $f, g : \mathbb{R}^N \to \mathbb{R}$ be convex and set $h = f + g$. Then for every $x \in \mathbb{R}^N$, we have $\partial h(x) = \partial f(x) + \partial g(x)$.
Here, we define addition of sets as $A + B = \{a + b : a \in A, b \in B\}$.
Note that again the domain is important here.

Piecewise subdifferentials
Similarly, if $f(x) = \sum_{i=1}^{N} f_i(x_i)$, then $\partial f(x) = \prod_{i=1}^{N} \partial f_i(x_i)$, where that product is the Cartesian product.

Subdifferentials for μ-strongly convex functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be μ-strongly convex and let $x, y \in \mathbb{R}^N$.

1. For every $x^* \in \partial f(x)$, we have
   $f(y) \geq f(x) + x^* \cdot (y - x) + \frac{\mu}{2}||x - y||^2$
2. For every $x^* \in \partial f(x)$ and $y^* \in \partial f(y)$, we have
   $(x^* - y^*) \cdot (x - y) \geq \mu||x - y||^2$

Subgradient Method

Definition subgradient method
The subgradient method starts from $x_0 \in \mathbb{R}^N$ and then for $k \geq 0$ iterates $x_{k+1} = x_k - \gamma d_k$, with $d_k \in \partial f(x_k)$

Convergence of the subgradient method
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and Lipschitz-continuous with constant M.
Suppose $S := \operatorname{argmin}(f) \neq \emptyset$ and let (x_k) be defined by the subgradient method.
Let $\overline{x}_k = \frac{1}{k+1} \sum_{j=0}^{k} x_j$. Then,
$f(\overline{x}_k) - \min(f) \leq \frac{\gamma M^2}{2} + \frac{\operatorname{dist}(x_0, S)^2}{2\gamma(k+1)}$

Extended real-valued functions

Definition: extended real numbers
The set of extended real numbers is $\mathbb{R} \cup \{+\infty\}$, with the followeing conventions:

- the numbers are ordered as usual, with $+\infty > \lambda$ for all $\lambda \in \mathbb{R}$
- summation: $+\infty + (+\infty) = +\infty + \lambda = +\infty$ for all $\lambda \in \mathbb{R}$
- multiplication: $\lambda(+\infty) = +\infty$ for all $\lambda > 0$ and $0(+\infty) = 0$

Definition: extended real-valued function
A function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$

Definition: effective domain
The domain of an extended real-valued function is
$\operatorname{dom}(f) = \{x \in \mathbb{R}^N : f(x) < +\infty\} \neq \emptyset$
(Note that we always assume that the domain is nonempty)

Definition: epigraph
The epigraph is defined the same as for real-valued functions, namely:
$\operatorname{epi}(f) = \{(x, t) \in \mathbb{R}^N \times \mathbb{R} : t \geq f(x)\}$

Other notions

- convexity is defined as usual (adapted to the new definition of the domain)
- subgradients and subdifferentials are defined in the same way as before

Fermat's Rule IV
Let $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$. Then $\hat{x} \in \operatorname{argmin}(f)$ if and only if $0 \in \partial f(\hat{x})$

Properites of extended real-valued functions
For $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, we have the following:

1. f is convex $\iff$ $\operatorname{epi}(f)$ is convex
2. f is convex $\implies$ $\operatorname{dom}(f)$ and $\operatorname{argmin}(f)$ are convex. Also, for every $\gamma \in \mathbb{R}$, the sublevel set $[f \leq \gamma]$ is convex
3. f strictly convex $\implies$ $\operatorname{argmin}(f)$ is either empty or a singleton

Definition: sum of functions
Given $f, g : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, we can define the sum $f + g : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ as $(f + g)(x) = f(x) + g(x)$ whenever $\operatorname{dom}(f) \cap \operatorname{dom}(g) \neq \emptyset$ (otherwise, $f + g = +\infty$)

Sum of convex functions
The sum of two convex functions is convex. If one of them is strictly convex, so is the sum.

Definition: closed function
A function $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ is closed if $\operatorname{epi}(f)$ is closed.

Closed equivalences
Let $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$. The following statements are equivalent:

1. f is closed;
2. for every $\gamma > \inf(f)$, the set $[f \leq \gamma]$ is closed;
3. the function $f|_{\operatorname{dom}(f)} : \operatorname{dom}(f) \subset \mathbb{R}^N \to \mathbb{R}$ is lower-semicontinuous
4. for every $x \in \mathbb{R}^N$, and every sequence (x_k) that converges to x, we have $f(x) \leq \operatorname{liminf}_{k\to\infty} f(x_k)$

Existence of an affine minorant
If $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ is closed and convex, there exist $c \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$ such that $f(x) \geq c \cdot x + \alpha$ for every $x \in \mathbb{R}^N$

Definition: indicator function
The indicator function of a nonempty set $C \subset \mathbb{R}^N$ is the function $\iota_C : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ is defined as
$\iota_C = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise} \end{cases}$
We see that $\operatorname{dom}(\iota_C) = C$ and $\operatorname{epi}(\iota_C) = C \times [0, +\infty)$. ι_C is closed if and only if C is closed. It is convex if and only if C is convex.

We also have, for every $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$:
$\min\{f(x) : x \in C\} = \min\{f(x) + \iota_C(x) : x \in \mathbb{R}^N\}$

Definition: normal cone
The set $\partial\iota_C(x)$ is the normal cone to C at x, and is denoted by $N_C(x)$

Subdifferentials of sums of functions
For extended real-valued functions $f, g : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$, we have $\partial f(x) + \partial g(x) \subset \partial(f + g)(x)$
The converse is not true in general; the converse is true if at least one of the functions is continuous at a point in the domain of the other.

Subdifferentials of strongly convex functions
Let $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ be closed and strongly convex. For every $y^* \in \mathbb{R}^N$ there is $x* \in \mathbb{R}^N$ such that $y^* \in \partial f(x^*)$.

The proximity operator

Minimizer of closed, μ-strongly convex functions
If $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ is closed and μ-strongly convex, it has exactly one minimizer.
Moreover, if $v \in \partial f(x)$, then
$f(y) \geq f(x) + v \cdot (y - x) + \frac{\mu}{2}||y - x||^2$ for all $y \in \mathbb{R}^N$

Definition: proximity operator or resolvent
Let $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex, and let $z \in \mathbb{R}^N$. The **proximity operator** associated to f is the function $\text{prox}_f : \mathbb{R}^N \to \mathbb{R}^N$ defined by

$$\text{prox}_f(z) = \text{argmin}\{f_z(x)\} = \text{argmin}\{f(x) + \frac{1}{2}||x - z||^2\}$$

The relationship between z and $\text{prox}_f(z)$ is given by $z - \text{prox}_f(z) \in \partial f(\text{prox}_f(z))$

See below for information about f_z

f_z of closed and convex function
Let $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex, and let $z \in \mathbb{R}^N$. The function $f_z : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$, defined by $f_z(x) = f(x) + \frac{1}{2}||x - z||^2$ is closed and strongly convex.
$\partial f_z(x) = \partial f(x) + x - z$ for every $x, z \in \mathbb{R}^N$
Moreover, its unique minimizer $\hat{x}_z$ satisfies $0 \in \partial f_z(\hat{x}_z)$. Thus, we see that $z \in \hat{x}_z + \partial f(\hat{x}_z)$.

Definition: resolvent of ∂f
Since $\text{prox}_f(z)$ is uniquely determined by z, we often write $\text{prox}_f(z) = (I + \partial f)^{-1}(z)$.
The function $\text{prox}_f = (I + \partial f)^{-1}$ is known as the resolvent of the operator ∂f.

Proximal operator on closed convex domain
If $C \subset \mathbb{R}^N$ is nonempty, closed and convex, then $\text{prox}_{\iota_C}(x) = \text{Proj}_C(x)$

Continuity of proximal operators
For every closed convex function $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$, we have
$||\text{prox}_f(x) - \text{prox}_f(y)|| \leq ||x - y||$ for all $x, y \in \mathbb{R}^N$
In particular, prox_f is a continuous function.

Proximal Point Method

Consider $f : \mathbb{R}^d \to \mathbb{R}^N \cup \{+\infty\}$ which is proper, convex, lower semi-continuous and which has minimizers.
The Proximal Point Method (PPM) consists of the iteration $x_{k+1} = \text{prox}_f(x_k)$

Proximal-Gradient Method

Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and smooth, and let $g : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex, and let the optimisation problem be $\min\{f(x) + g(x) : x \in \mathbb{R}^N\}$
The Proximal-Gradient Method iterates
$x_{k+1} = \text{argmin}\left\{g(z) + f(x_k) + \nabla f(x_k) \cdot (z - x_k) + \frac{1}{2\gamma}||z - x_k||^2\right\} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$

Convergence of the Proximal-Gradient Method
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth, let $g : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex, and set $\gamma \leq \frac{1}{L}$. Define (x_k) by the proximal-gradient method. If $\text{argmin}(f + g) \neq \emptyset$, then x_k converges, as $k \to \infty$, to a minimizer of $f + g$. Moreover, for every $k \geq 1$, we have:

$$(f + g)(x_k) - \min(f + g) \leq \frac{\text{dist}(x_0, \text{argmin}(f + g))^2}{2\gamma k}$$

Polyak-Łojasiewicz Inequality for extended real valued functions
A function $\phi\mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ satisfies the Polyak-Łojasiewicz (PŁ) Inequality with constant $\mu > 0$ if
$2\mu(f(x) - \min(f)) \leq \min\{||v||^2 : v \in \partial f(x)\}$
for all $x \in \text{dom}(\partial f)$.

As in the finite case, strongly convex functions satisfy this inequality.

Convergence of Polyak-Łojasiewicz functions
Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth, let $g : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex and set $\gamma < \frac{2}{L}$. Define (x_k) by the proximal-gradient method. If $\phi := f + g$ satisfies the Polyak-Łojasiewicz Inequality with constant $\mu > 0$, and $\text{argmin}(\phi) \neq \emptyset$, then

$$\phi(x_k) - \min(\phi) \leq \frac{\phi(x_0) - \min(\phi)}{(1 + \eta)^k}$$

where $\eta = \mu\gamma(2 - \gamma L)$

Chapter 8: Duality

Conjugate functions

Definition: Fenchel conjugate
The Fenchel conjugate of a closed convex function $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ is the function
$f^*(x^*) = \sup_{x \in \mathbb{R}^N} \{x^* \cdot x - f(x)\}$

Definition: biconjugate
Let $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$. The biconjugate of f is $f^{**} := (f^*)^*$.
$f^{**} \leq f$, with equality if f is closed, convex and not identically $+\infty$ (see below)

Fenchel conjugate properties

1. If $f \leq g$, then $f^* \geq g^*$

If f is closed and convex...

1. $f^*(y) > -\infty$ for every $y \in \mathbb{R}^N$ and f^* is not identically $+\infty$ (meaning it is not the constant function that always gives the value $+\infty$)

If f is not identically $+\infty$...

1. f^* is closed and convex

If f is closed, convex and not identically $+\infty$...

1. f^* is closed, convex and not identically $+\infty$
2. Fenchel-Young Inequality: $f(x) + f^*(x^*) \geq x^* \cdot x$. There is equality if and only if $x^* \in \partial f(x)$
3. $f^{**} = f$
4. Legendre-Fenchel Reciprocity Formula: $x^* \in \partial f(x)$ if and only if $x \in \partial f^*(x^*)$
5. Let $\mu\ell = 1$. Then, f is μ-strongly convex if and only f^* is ℓ-smooth

Conjugates of affine functions
Given $c \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$, define $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ as $f(x) = c \cdot x + \alpha$.
Then $f^*(y) = \iota_c(y) - \alpha$.

Definition: support functional
Given a nonempty, closed convex set $C \subset \mathbb{R}^N$, the conjugate of the indicator function ι_C, namely
$\sigma_C(y) := \iota_C^*(y) = \sup\{y \cdot x : x \in C\}$
is called the support funcitonal of C

Fenchel-Rockafellar Duality

Definition: primal problem and dual problem
Let $A \in \mathbb{R}^{M \times N}$, and let $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$ and $g : \mathbb{R}^M \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex.

The primal problem is $\min_{x\in\mathbb{R}^N}\{f(x)+g(Ax)\}$, with optimal value $v\in\mathbb{R}$, and set of **primal solutions** $S\subset\mathbb{R}^N$.

The dual problem is $\min_{y\in\mathbb{R}^M}\{f^*(-A^Ty)+g^*(y)\}$ with optimal value $v^*\in\mathbb{R}$, and set of dual solutions $S^*\subset\mathbb{R}^M$.

Definition: duality gap
The quantity $v+v^*$

Definition: weak duality
The situation $v+v^*\geq 0$

Defintion: strong duality
The situation $v+v^*=0$

Strong dualitity equivalences
Let $\hat{x}\in\mathbb{R}^N$ and $\hat{y}\in\mathbb{R}^M$. The following statements are equivalent:

1. $\hat{x}\in S,\hat{y}\in S^*$;
2. $-A^T\hat{y}\in\partial f(\hat{x})$ and $\hat{y}\in\partial g(A\hat{x})$;
3. $f(\hat{x})+f^*(-A^T\hat{y})=-A^T\hat{y}\cdot\hat{x}$ and $g(A\hat{x})+g^*(\hat{y})=\hat{y}\cdot A\hat{x}$; and
4. $f(\hat{x})+g(A\hat{x})+f^*(-A^T\hat{y})+g^*(\hat{y})=0$.
   If the statements above hold, we have strong duality, so $v+v^*=0$.
   Moreover, if either f is continuous and $g\circ A$ is not identically $+\infty$, or that g is continuous. If $\hat{x}\in S$, there is $\hat{y}\in\mathbb{R}^M$ such that all four statements hold.

Primal-Dual Method

If f is not strongly convex, f^* is not smooth and the computations required for the proximal algorithm may be challening. The primal-dual algorithm solves the primal and dual problems simultaneously.

Primal-dual algorithm
We consider the problem $\min\{f(x)+g(Ax)+h(x)\}$, where
$A\in\mathbb{R}^{M\times N}$;
$f:R^N\to\mathbb{R}^N\cup\{+\infty\}$ and $g:\mathbb{R}^M\to\mathbb{R}^N\cup\{+\infty\}$ are closed and convex; and
$h:\mathbb{R}^N\to\mathbb{R}$ is ℓ-smooth and convex

Then we iterate:

$$\begin{cases} x_{k+1} &= \operatorname{prox}_{\tau f}(x_k-\tau\nabla h(x_k)-\tau A^Ty_k) \\ y_{k+1} &= \operatorname{prox}_{\sigma g^*}(y_k+\sigma A(2x_{k+1}-x_k)) \end{cases}$$

If $\tau\sigma||A||^2+\frac{\tau\ell}{2}\leq 1$, the algorithm converges.
Every subsequential limit point of (x_k,y_k) is a primal-dual solution. Also, (x_k,y_k) is a primal-dual solution iff $x_{k+1}=x_k$ and $y_{k+1}=y_k$.

To avoid having to work with g^*, we can use...

Moreau's Identity:
Let $g : \mathbb{R}^M \to \mathbb{R}^N \cup \{+\infty\}$ be closed and convex, and let $\sigma > 0$. For every $y \in \mathbb{R}^M$, we have $\text{prox}_{\sigma g^*}(y) = y - \sigma \text{prox}_{\sigma^{-1} g}(\sigma^{-1} y)$

A complementary approach to linear equality constraints

Consider a differentiable $f : \mathbb{R}^N \to \mathbb{R}^N \cup \{+\infty\}$, a matrix size $M \times N$ and a vector $b \in \mathbb{R}^M$, with the problem
$\min\{f(x) : Ax = b\}$.

Definition: Lagrangian
The Lagrangian associated to the above problem is the function $L : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$, defined by
$L(x, z) = f(x) + z^T(Ax - b)$.
Here, x is the primal variable and z is the dual variable, or the Lagrange multiplier.

If $\hat{x}$ is a solution to minimization problem and $\hat{z}$ the corresponding multiplier, we have $0 = \nabla L(\hat{x}, \hat{z}) = \begin{pmatrix} \nabla f(\hat{x}) + A^T \hat{z} \\ A\hat{x} - b \end{pmatrix}$

The equalites $A\hat{x} = b$ and $-A^T\hat{z} = \nabla f(\hat{x})$ are called primal and dual feasibility, respectively.

Definition: augmented Lagrangian
$\mathcal{L}_\alpha(x, \lambda) = f(x) + \lambda^T(Ax - b) + \frac{\alpha}{2}||Ax - b||^2$

Method of Multipliers
$\begin{cases} x_{k+1} & = & \text{argmin}\{\mathcal{L}_\alpha(x, \lambda_k) : x \in \mathbb{R}^N\} \\ \lambda_{k+1} & = & \lambda_k + \alpha(Ax_{k+1} - b) \end{cases}$

Dual feasibility is enforced: $\nabla f(x_{k+1}) + A^T\lambda_{k+1} = 0$

Limitations:

- x_{k+1} may not be well defined: if f is not strongly convex or A has a nontrivial kernel, the augmented Langrangian may not have a minimizer, and if it exists it may not be unique
- if f is not differentiable, computing x_{k+1} can be more complicated

Variants of the method of multipliers
Proximal version
If f is closed and convex, we can implement
$\begin{cases} x_{k+1} & = & \text{argmin}\{\mathcal{L}_\alpha(x, \lambda_k) + \frac{\alpha}{2}||x - x_k||^2 : x \in \mathbb{R}^N\} \\ \lambda_{k+1} & = & \lambda_k + \alpha(Ax_{k+1} - b) \end{cases}$

Tricky parts: the quadratic term, since there is no explicit formula for the prox of a sum of functions. This is solved by the:

Predictor corrector method

$$\begin{cases} p_{k+1} &= \lambda_k + \alpha(Ax_k - b) \\ x_{k+1} &= \text{argmin}\{L(x, p_{k+1}) + \frac{1}{2\alpha}||x - x_k||^2 : x \in \mathbb{R}^N\} \\ \lambda_{k+1} &= \lambda_k + \alpha(Ax_{k+1} - b) \end{cases}$$

(Note we use the nomal Lagrangian L here.)

Alternation direction method of multipliers (ADMM)

For the structured problem $\min\{f(x) + g(y) : (x, y) \in \mathbb{R}^{N_1 \times N_2}, Ax + By = c\}$,
We could use the multiplier method as it is, but the quadratic function $Ax + By = c$ mixes the variables. Instead, we do as follows:

$$\begin{cases} x_{k+1} &= \text{argmin}\mathcal{L}_\alpha(x, y_k, \lambda_k) : x \in \mathbb{R}^{N_1} \\ y_{k+1} &= \text{argmin}\{\mathcal{L}_\alpha(x_{k+1}, y, \lambda_k) : y \in \mathbb{R}^{N_2}\} \\ \lambda_{k+1} &= \lambda_k + \alpha(Ax_{k+1} + By_{k+1} - c) \end{cases}$$

Proximal and predictor-corrector variants can also be implemented here.

Optimality conditions: mathematical programming and Lagrangian duality

Definition: mathematical programming problem

Let $f, h_1, h_2, ..., h_M \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$, and consider the constrained problem

$(\mathcal{P}) : \min_{x \in C} f(x)$

where the feasible set C is given by $C = \{x \in \mathbb{R}^N : h_m(x) = 0, m = 1, .., M\}$

Step 1: No inequality constraints

Here we consider the case where there are no inequality constraints on the problem.

Definition: regular point (for a problem with no inequality constraints)

A feasible point $\hat{x} \in C$ is regular if the set $V = \{\nabla h_1(\hat{x}), ..., \nabla h_M(\hat{x})\}$ is linearly independent

Lagrange Multiplier Theorem (Necessary conditions for optimality)

Let $\hat{x}$ be a regular local minimizer of $(\mathcal{P})$. There is a unique $\hat{\lambda} \in \mathbb{R}^M$, called Lagrange multiplier vector, such that

$$\nabla f(\hat{x}) + \sum_{m+1}^{M} \hat{\lambda} \nabla h_m(\hat{x}) = 0$$

If, in addition, $f, h_1, h_2, ..., h_M \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R})$, then

$$y \cdot (\nabla^2 f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda} \nabla^2 h_m(\hat{x}))y \geq 0$$

for all $y \in V^{\perp}$.

Second order sufficient conditions
Let $f, h_1, ..., h_M : \mathbb{R}^N \to \mathbb{R}$ be twice continuously differentiable, and let $\hat{x} \in C$ and $\hat{\lambda} \in \mathbb{R}^M$ satisfy the conditions outlined in the Lagrange multiplier theorem with a *strict inequality* for all $y \in V^{\perp} \setminus \{0\}$. Then $\hat{x}$ is a strict local minimizer of $(\mathcal{P})$.

The general case

Let $f, g_j, h_m \in \mathcal{C}^1(\mathbb{R}^N; \mathbb{R})$, and consider the constrained problem
$(\mathcal{P})$ $\min_{x \in C} f(x)$
With $C = \{x \in \mathbb{R}^N : h_m(x) = 0, g_j(x) \leq 0, \forall m, j\}$

Definition: active constraints
The set of active inequality constraints at a local solution $\hat{x}$ is $A(\hat{x}) = \{j : g_j(\hat{x}) = 0\}$

Definition: regular point (problem with inequality constraints)
A feasible point $\hat{x} \in C$ is regular if the set $V = \{\nabla g_j(\hat{x}), \nabla h_m(\hat{x}) : j \in A(\hat{x}), m = 1, ..., M\}$ is linearly independent.

Karush-Kuhn-Tucker conditions (necessary conditions for optimality)
Let $\hat{x}$ be a regular local minimizer of $(\mathcal{P})$. There exist unique Lagrange multiplier vectors $\hat{\lambda} \in \mathbb{R}^M$ and $\hat{\mu} \in \mathbb{R}^J_+$, such that

$$\nabla f(\hat{x}() + \sum_{m=1}^{J} \hat{\mu}_j \nabla g_j(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla h_m(\hat{x}) = 0$$

and $\hat{\mu}_j = 0$ for all $j \notin A(\hat{x})$, which means that $\hat{\mu}_j g_j(\hat{x}) = 0$ for all j.

If, in addition, $f, g_j, h_m \in \mathcal{C}^2(\mathbb{R}^N; \mathbb{R})$, then

$$y \cdot (\nabla^2 f(\hat{x}) + \sum_{m=1}^{M} \hat{\lambda}_m \nabla^2 h_m(\hat{x}))y \geq 0$$

for all $y \in \mathcal{V}^{\perp}$, where $\mathcal{V} = \{\nabla g_j(\hat{x}), \nabla h_m(\hat{x}), j \in A(\hat{x}), m = 1, ..., M\}$