

Introduction to Optimization

Lecture 01: Motivation and examples. Optimization problems and their solutions.



university of
 groningen

Nothing takes place in the world whose meaning is not that of some maximum or minimum.

Leonhard Euler (1707 - 1783)

Principle of Least Time

The path taken by a ray between two given points is the path that can be traveled in the least time.

Law of Refraction

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

Material-efficient cans



Material-efficient cans



Some terminology

An **optimization problem** is usually written as

$$\min_{x \in C} f(x),$$

where

- f is the **objective function**.
- C is the **feasible set**, or the **set of constraints**. Points in C are **feasible**.

The problem is **constrained** if $C \subsetneq \mathbb{R}^N$, and **unconstrained** otherwise.

A point $\hat{x} \in C$ is a **global minimizer** if $f(\hat{x}) \leq f(x)$ for all $x \in C$. It is a **local minimizer** if $f(\hat{x}) \leq f(x)$ for all $x \in C$ close to \hat{x} .

The minimizer is **unique** or **strict** if the corresponding inequality is strict.

Course description

Introductory course on optimization, where the fundamental concepts, techniques and tools are presented, discussed and put into practice by means of modelling and analysis.

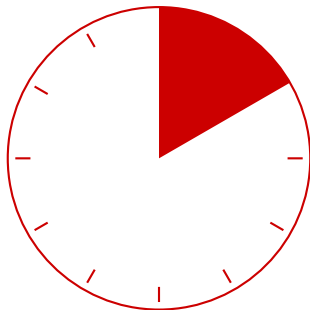
Assessment:

- 40% Written exam
- 15% Computational Exercises
- 45% Homework Assignments

Learning outcomes

- ➊ Identify and state unconstrained and constrained optimization problems, and assess whether they have solutions
- ➋ Characterize and calculate solutions to optimization problems by means of optimality conditions
- ➌ Prove the convergence of optimization algorithms and establish their convergence rates
- ➍ Implement optimization algorithms to approximate solutions numerically

Break



Motivation from data analysis

We begin with a **data set**

$$\mathcal{D} = \{(a_j, y_j) \in \mathbb{R}^K \times \mathbb{R}^M : j = 1, 2, \dots, J\},$$

where the a_j 's are **features** and the y_j 's are **observations** or **labels**.

The aim is to **discover** or **learn** a function $\phi : \mathbb{R}^K \rightarrow \mathbb{R}^M$ such that

$$\phi(a_j) \simeq y_j$$

for, let us say, many j 's.

The optimization point of view

Typically, ϕ belongs to a **family** of functions and, within that family, it is defined by some unknown **parameter** $x \in \mathbb{R}^N$.

We define a **loss function**

$$\mathcal{L}_{\mathcal{D}}(x) = \frac{1}{J} \sum_{j=1}^J \ell(a_j, y_j, x),$$

in such a way that the parameters **most consistent** with the data set are the ones that give the lowest values of $\mathcal{L}_{\mathcal{D}}$.

Regularization

The term **regularization** refers loosely to changes in the problem that either simplify the resolution or induce properties in the solutions.

Examples

- To find the minimum of a nonsmooth function, one may replace the function by a smooth one that is **similar**, in some sense.
- Approximate solutions with fewer nonzero entries may be preferred, for **storage** or **explainability** reasons (feature selection).

Regularization also helps reduce **overfitting**.

Least squares

If $\phi : \mathbb{R}^N \rightarrow \mathbb{R}$ is linear, namely $y = \phi(a) = x^* a$, we find x by

$$\min_{x \in \mathbb{R}^N} \frac{1}{2J} \sum_{j=1}^J (x^* a_j - y_j)^2 = \min_{x \in \mathbb{R}^N} \frac{1}{2J} \|Ax - y\|^2.$$

The minimum is attained at $x = (A^* A)^{-1} A^* y$, if $y \in \text{ran}(A)$.

Exercise


Is the minimum attained if $y \notin \text{ran}(A)$? If so, where?

Regularization: Ridge, LASSO, TV...




Support vector machines



Classification problem: Individual j is labeled $y_j \in \{-1, 1\}$. A vector a_j , with features of j , is known. We wish to predict the label of a new individual.

A **support vector machine** finds a hyperplane that separates the points labeled 1 from those labeled -1 . 

If this is possible, we may want to maximize the distance to the closest point on each side to obtain a more **robust** classification.

Nonlinear transformations of the space are possible. 

Other examples

- Matrix factorization
- Matrix completion (a.k.a. the *Netflix Prize* Problem) 
- Logistic regression
- Image processing: deblurring, in-painting
 - Topic of the computational project
- Face, voice and pattern recognition
- (Deep) learning 

What these problems have in common

- They can be formulated as minimizing functions of **one or more real variables** (sometimes in a huge number).
- The functions involved are continuous. They are either smooth, or they have **simple** or **structured** forms of nonsmoothness.
- The function to be minimized is often a sum of simple functions that depend either on few data points or involve few variables.
- Usually, the nonsmooth parts can be separated from the rest (additive structure).

Introduction to Optimization

Lecture 02: Calculus in \mathbb{R}^N .



Real vectors and their norms

\mathbb{R}^N is the (real) vector space of N -tuples of real numbers (columns)

$$x \in \mathbb{R}^N \quad \Longleftrightarrow \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}, \quad x_1, \dots, x_N \in \mathbb{R}.$$

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

Properties

- $\|x\| > 0$ for all $x \neq 0$ and $\|0\| = 0$.
- $\|\alpha x\| = |\alpha| \|x\|$ for all $x \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}$.
- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^N$.

Distances and balls

The **distance** between $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ is $\text{dist}(x, y) = \|x - y\|$.

Properties

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

The **open ball** centered at $x \in \mathbb{R}^N$ with radius $r > 0$ is

$$B(x; r) = \{y \in \mathbb{R}^N : \text{dist}(x, y) < r\}.$$

The **closed ball** centered at $x \in \mathbb{R}^N$ with radius $r > 0$ is

$$\bar{B}(x; r) = \{y \in \mathbb{R}^N : \text{dist}(x, y) \leq r\}.$$

Topology I

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

A subset $C \subset \mathbb{R}^N$ is **closed** if its complement is open.

Example

Open balls are open sets. Closed balls are closed sets. 

Proposition

If a sequence in a closed set is convergent, its limit has to be in the set. 

Topology II

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

Proposition

Every bounded sequence in \mathbb{R}^N has a convergent subsequence.

Finally, a subset $K \subset \mathbb{R}^N$ is **compact** if it is both closed and bounded.

Proposition

Every sequence in a compact subset of \mathbb{R}^N has a convergent subsequence. The limits of all convergent subsequences must lie in the set.

Dot product

The **dot product** (or also **inner product**) of $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ is

$$x \cdot y = x_1 y_1 + \cdots + x_N y_N.$$

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 $\alpha \in \mathbb{R}$.

Other notations: $x \cdot y = \langle x, y \rangle = \langle x | y \rangle = x^T y$ (product of matrices).

Perpendicularity and parallelism

Proposition

$$|x \cdot y| \leq \|x\| \|y\| \text{ for all } x, y \in \mathbb{R}^N.$$

The vectors $x, y \in \mathbb{R}^N \setminus \{0\}$ are **perpendicular** or **orthogonal** if $x \cdot y = 0$. In that case, we write $x \perp y$.

The vectors $x, y \in \mathbb{R}^N \setminus \{0\}$ are **parallel** if there is $\alpha \in \mathbb{R}$ such that $x = \alpha y$. We write $x \parallel y$.

Exercise

Show that $x \parallel y$ if, and only if, $|x \cdot y| = \|x\| \|y\|$.

Angles and triangles

The **angle** θ between $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^N$ is

$$\cos^{-1} \left(\frac{x \cdot y}{\|x\| \|y\|} \right).$$

Law of cosines


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



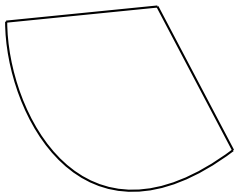
Pythagoras's Theorem

$$x \perp y \text{ if, and only if, } \|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

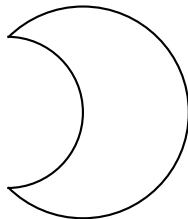
Convex sets

A subset $C \subset \mathbb{R}^N$ is **convex** if $\lambda x + (1 - \lambda)y \in C$ whenever $x, y \in C$ and $\lambda \in [0, 1]$. In other words, if the segment joining any two points of C also belongs to C . 

This set
is convex



This one
is not



Projection

Theorem

Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex. For each $x \in \mathbb{R}^N$, there is a unique point $\hat{x} \in C$ such that

$$\text{dist}(x, \hat{x}) = \min\{\text{dist}(x, y) : y \in C\}.$$

Moreover, \hat{x} is the only point in C that satisfies the inequality

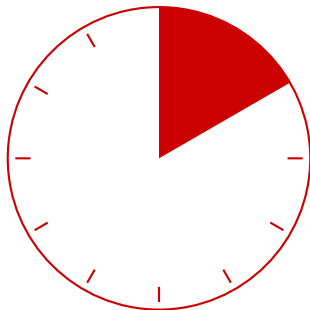
$$(x - \hat{x}) \cdot (y - \hat{x}) \leq 0$$

for all $y \in C$.



The point \hat{x} is the **projection** of x onto C , and is denoted by $P_C(x)$.

Break



Differentiability and gradient

Let $A \subset \mathbb{R}^N$ be nonempty and open. A function $f : A \rightarrow \mathbb{R}$ is **differentiable** at $x \in A$ (in the sense of Gâteaux) if the **directional derivative**

$$f'(x; h) = \lim_{t \rightarrow 0} \frac{f(x + th) - f(x)}{t}$$

exists for all $h \in \mathbb{R}^N$, and there is $g \in \mathbb{R}^N$ such that

$$g \cdot h = f'(x; h)$$

for all $h \in \mathbb{R}^N$. In this case, the **gradient** of f at x is $\nabla f(x) = g$.

As usual, f is **differentiable** on A if it is so at every point of A .

More about the gradient

Remark

Let f be differentiable at x , and let $g = \nabla f(x)$ be its gradient at that point. If e_i denotes the i -th canonical vector in \mathbb{R}^N , then

$$g_i = \nabla f(x) \cdot e_i = f'(x; e_i) = \lim_{t \rightarrow 0} \frac{f(x + te_i) - f(x)}{t} = \frac{\partial f}{\partial x_i}(x).$$

Example

Let us compute the gradient of the function $f : \mathbb{R}^N \rightarrow \mathbb{R}$, defined by

$$f(x) = \frac{1}{2} \|Ax - b\|^2,$$

where A is a real matrix of size $M \times N$ and $b \in \mathbb{R}^M$.



First order optimality condition

Theorem (Fermat's Rule)

Let $f : A \subset \mathbb{R}^N \rightarrow \mathbb{R}$ and let $\emptyset \neq C \subset A$ be convex. If $\hat{x} \in C$ is such that $f(\hat{x}) \leq f(y)$ for all $y \in C$, and if f is differentiable at \hat{x} , then

$$\nabla f(\hat{x}) \cdot (y - \hat{x}) \geq 0$$

for all $y \in C$.

If, moreover, $\hat{x} \in \text{int}(C)$, then $\nabla f(\hat{x}) = 0$.

Introduction to Optimization

Lecture 03: Optimality conditions. Examples. Convex functions.



university of
groningen

First order optimality condition

Theorem (Fermat's Rule)

Let $f : A \subset \mathbb{R}^N \rightarrow \mathbb{R}$ and let $\emptyset \neq C \subset A$ be convex. If $\hat{x} \in C$ is such that $f(\hat{x}) \leq f(y)$ for all $y \in C$, and if f is differentiable at \hat{x} , then

$$\nabla f(\hat{x}) \cdot (y - \hat{x}) \geq 0$$

for all $y \in C$. If, moreover, $\hat{x} \in \text{int}(C)$, then $\nabla f(\hat{x}) = 0$.

Question

What if C is **affine**?

Example

Compute the maximum value of the expression

$$\sum_{i=1}^N \alpha_i \ln(x_i)$$

subject to the constraint that


$$\sum_{i=1}^N x_i = b,$$

where $\alpha_1, \dots, \alpha_N, b > 0$.

Second order conditions

Theorem (Second order optimality conditions)

Let $f : A \subset \mathbb{R}^N \rightarrow \mathbb{R}$ be twice differentiable at $\hat{x} \in \text{int}(A)$.

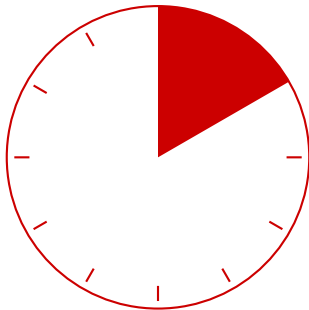
- i) If \hat{x} is a local minimizer of f , then $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(\hat{x})$ is positive semidefinite ($\nabla^2 f(\hat{x})d \cdot d \geq 0$ for all $d \in \mathbb{R}^N$). 
- ii) If $\nabla f(\hat{x}) = 0$ and $\nabla^2 f(\hat{x})$ is positive definite ($\nabla^2 f(\hat{x})d \cdot d > 0$ for all $d \neq 0$), then \hat{x} is a strict local minimizer of f .

Lemma (Taylor's Approximation)

Let $f : A \subset \mathbb{R}^N \rightarrow \mathbb{R}$ be of class C^2 , and let $x \in A$. For each $d \in \mathbb{R}^N$,

$$\lim_{t \rightarrow 0} \frac{1}{t^2} \left| f(x + td) - f(x) - t \nabla f(x) \cdot d - \frac{t^2}{2} \nabla^2 f(x) d \cdot d \right| = 0.$$

Break



Convex functions

A function $f : D \subset \mathbb{R}^N \rightarrow \mathbb{R}$ 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 $\lambda \in (0, 1)$ (or $[0, 1]$, if you prefer).



Proposition

The function $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex.

Proposition

Local minimizers of convex functions are global minimizers.



Characterizations of differentiable convex functions

Proposition

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

- ① f is convex;
- ② for all $x, y \in D$, $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$;
- ③ 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

- ④ for all $x \in D$, $\nabla^2 f(x)$ is positive semidefinite.

Introduction to Optimization

Lecture 04: Strict and strong convexity. Iterative algorithms. Descent methods.



Characterizations of differentiable convex functions

Proposition

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

- ① f is convex;
- ② for all $x, y \in D$, $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$;
- ③ 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

- ④ for all $x \in D$, $\nabla^2 f(x)$ is positive semidefinite.

Sketch of the proof

Assuming just differentiability:

- $(1) \Rightarrow (2)$: use definition of convexity and directional derivative;
- $(2) \Rightarrow (3)$: replace $x \leftrightarrow y$ and sum;
- $(3) \Rightarrow (1)$: show that $\phi(\lambda) = f(\lambda x + (1 - \lambda)y) - \lambda f(x) + (1 - \lambda)f(y)$ is increasing.

In the twice differentiable case:

- $(3) \Rightarrow (4)$: similar to $(1) \Rightarrow (2)$ above;
- $(4) \Rightarrow (1)$: similar to $(3) \Rightarrow (1)$ above.

Strict and strong convexity

A function $f : D \subset \mathbb{R}^N \rightarrow \mathbb{R}$ 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 $\lambda \in (0, 1)$, and it is **strongly convex** if D is convex and there is $\mu > 0$ such that

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2}\lambda(1 - \lambda)\|x - y\|^2$$

for all $x, y \in D$ and all $\lambda \in (0, 1)$.

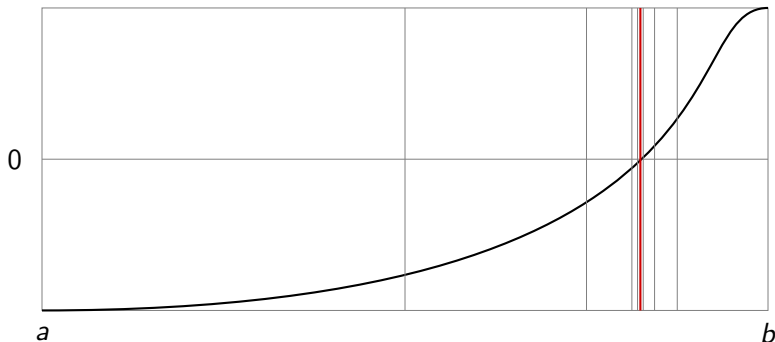


Exercises

- 1 Determine the values of $p \in \mathbb{R}$ for which the expression $f(x) = x^p$ represents a strictly/strongly convex function (on $(0, \infty)$ if $p < 0$).
- 2 When is the function $f(x) = \frac{1}{2}\|Ax - y\|^2$ strictly/strongly convex?
- 3 Prove that every strictly convex function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ has at most one minimizer, and every strongly convex function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ has exactly one minimizer.
- 4 Can you obtain characterizations of strict and strong convexity of f in terms of properties of ∇f and $\nabla^2 f$?

Introduction to iterative algorithms

Example: Bisection method to solve $g(x) = 0$



After k iterations, the distance to a solution is $|x_k - \hat{x}| \leq \frac{b - a}{2^k}$.

Introduction to iterative algorithms

An **iterative algorithm** is a procedure that computes a sequence (x_n) of points in \mathbb{R}^N that approximate a solution to a problem. It requires:

- An initial guess x_0 .
- A sequence (p_k) of parameters (typically $p_k \in \mathbb{R}^M$ for all $k \geq 0$).
- An operator $T : \mathbb{R}^N \times \mathbb{R}^M \rightarrow \mathbb{R}^N$ used to compute x_{k+1} , given x_k :

$$x_{k+1} = T(p_k, x_k).$$

- A stopping rule that is activated when the approximation is **sufficiently good**.

Important questions: **convergence** and **complexity**.

Stopping rules in minimization problems

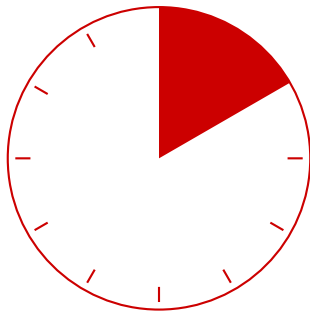
Ideally, the algorithm should stop when this is when

- x_k is close to a minimizer, ✗
- $f(x_k)$ is close to the optimal value $\inf(f)$, ✗
- $\|\nabla f(x_k)\|$ is small. ✓

Common implementable rules:

- | | | |
|---|---|---|
| • $\ \nabla f(x_k)\ < \varepsilon$ | • $f(x_{k+1}) - f(x_k) < \varepsilon$ | • $\ x_{k+1} - x_k\ < \varepsilon$ |
| • $\frac{\ \nabla f(x_k)\ }{\ \nabla f(x_0)\ } < \varepsilon$ | • $\frac{f(x_{k+1}) - f(x_k)}{f(x_1) - f(x_0)} < \varepsilon$ | • $\frac{\ x_{k+1} - x_k\ }{\ x_1 - x_0\ } < \varepsilon$ |

Break



Descent methods

Many algorithms are based on the idea of (sufficient) descent: given x_k , find x_{k+1} such that

$$(1) \quad f(x_{k+1}) \leq f(x_k) - \delta_k^2.$$

One way is to find $d_k \in \mathbb{R}^N$ and $\alpha_k > 0$, such that (1) holds with

$$x_{k+1} = x_k - \alpha_k d_k.$$

We say $-d_k$ is a **descent direction**, and α_k is the **step size**, **step length** or **learning rate** (in ML).

Motivation: 3 case studies

Let us analyze the behavior in the following cases:

① $f(x) = x^2$

② $f(x) = 1/x$, $\text{dom}(f) = (0, \infty)$.

③ $f(x) = 1/x$, $\text{dom}(f) = (-\infty, 0)$.

Introduction to Optimization

Lecture 05: Gradient descent. Convex functions.



university of
 groningen

Recall: descent methods

Most algorithms are based on the idea of (sufficient) descent: given x_k , find x_{k+1} such that

$$(1) \quad f(x_{k+1}) \leq f(x_k) - \delta_k^2.$$

One way is to find $d_k \in \mathbb{R}^N$ and $\alpha_k > 0$, such that (1) holds with

$$x_{k+1} = x_k - \alpha_k d_k.$$

We say $-d_k$ is a **descent direction**, and α_k is the **step size**, **step length** or **learning rate** (in ML).

L -smoothness

A differentiable function $f : A \subset \mathbb{R}^N \rightarrow \mathbb{R}$ is **L -smooth**, with $L > 0$, if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for all $x, y \in A$.

Proposition (Descent Lemma)

If f is L -smooth and A is convex, then

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

for all $x, y \in A$.



Identifying descent directions

The L -smooth case

By the Descent Lemma, we have

$$f(x_n - \alpha_n d_n) \leq f(x_n) - \alpha_n \nabla f(x_n) \cdot d_n + \frac{\alpha_n^2 L}{2} \|d_n\|^2.$$

It is therefore sufficient for α_n and d_n to satisfy

$$0 < \alpha_n L \|d_n\|^2 < 2 \nabla f(x_n) \cdot d_n.$$

Gradient-consistency: $\tau \|d_n\|^2 \leq \|\nabla f(x_n)\|^2 \leq \sigma \nabla f(x_n) \cdot d_n$.

Sufficient condition for descent: $0 < \inf \alpha_n \leq \sup \alpha_n < \frac{2\tau}{\sigma L}$.

Gradient-consistent methods

Constant stepsize $\alpha_n \equiv \alpha$, for simplicity

Proposition

Let f be L -smooth and bounded from below. Iterate $\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \mathbf{d}_n$, where \mathbf{d}_n is gradient-consistent and $0 < \alpha < \frac{2\tau}{\sigma L}$. Then,

- ① $\exists \lim_{n \rightarrow \infty} f(\mathbf{x}_n) \in \mathbb{R}$, and $\lim_{n \rightarrow \infty} \|\nabla f(\mathbf{x}_n)\| = 0$.
- ② Cluster points are critical: if $\mathbf{x}_{k_n} \rightarrow \hat{\mathbf{x}}$, then $\nabla f(\hat{\mathbf{x}}) = 0$.
- ③ If f has no critical points, then $\lim_{n \rightarrow \infty} \|\mathbf{x}_n\| = +\infty$.
- ④ There is $C > 0$ such that $\min \{\|\nabla f(\mathbf{x}_i)\| : 1 \leq i \leq n\} \leq \frac{C}{\sqrt{n}}$.

Convex functions and the gradient method

Theorem

Let f be convex and L -smooth. Iterate $x_{n+1} = x_n - \alpha \nabla f(x_n)$ with $0 < \alpha < \frac{2}{L}$.

① $\lim_{n \rightarrow \infty} f(x_n) = \inf(f).$

② If f has minimizers, then x_n converges to one of them, and

$$f(x_n) - \min(f) \leq \frac{D^2}{\alpha(2 - \alpha L)n},$$

where D is the distance from x_0 to its closest minimizer.

Moreover, $\lim_{n \rightarrow \infty} n [f(x_n) - \min(f)] = 0.$

An important tool

Proposition (Baillon-Haddad Lemma)

If f is convex and L -smooth, then

$$\frac{1}{L} |\nabla f(y) - \nabla f(x)|^2 \leq (\nabla f(y) - \nabla f(x)) \cdot (y - x)$$

for all $x, y \in A$.

This will be proved in the tutorial.

Sketch of the proof

First, use

$$(2) \quad \|x_{n+1} - u\|^2 = \alpha^2 \|\nabla f(x_n)\|^2 + \|x_n - u\|^2 - 2\alpha \nabla f(x_n) \cdot (x_n - u)$$

and Baillon-Haddad Lemma to show that

$$\frac{\alpha}{L}(2 - \alpha L) \sum_{n=0}^k \|\nabla f(x_n)\|^2 \leq \|x_0 - u\|^2.$$

Then, use (2) and convexity to prove that

$$2\alpha(f(x_n) - \min(f)) \leq \|x_n - u\|^2 - \|x_{n+1}\|^2 + \alpha^2 \|\nabla f(x_n)\|^2.$$

Sum over n , and combine the two inequalities, to conclude that

$$2\alpha k(f(x_k) - \min(f)) \leq \|x_0 - u\|^2 \left(1 + \frac{\alpha L}{2 - \alpha L}\right).$$

Sketch of the proof, continued

For the convergence, use that cluster points are critical, and that $\|x_n - u\|$ is nonincreasing, to deduce that (x_n) cannot have more than one cluster point. This is sufficient for convergence because (x_n) is bounded.

For the last statement, use the following lemma with $e_n = f(x_n) - \min(f)$:

Lemma

Let (e_n) be a positive, nonincreasing sequence such that $\sum_{n=0}^{\infty} e_n < +\infty$. Then $\lim_{n \rightarrow \infty} ne_n = 0$.

Other choices for α_n

- **Constant:** $\alpha_n \equiv \alpha \in (0, 2/L)$.
- **Exact minimization:** $\min_{\alpha > 0} f(x_n - \alpha d_n)$.
- **Limited minimization:** $\min_{\alpha \in (0, A]} f(x_n - \alpha d_n)$.
- **Vanishing:** $\alpha_n \rightarrow 0, \sum \alpha_n = \infty$.
- **Backtracking** (Armijo, Goldstein).



Some remarks

The simplest example, revisited

We had applied the gradient method to the function $f(x) = x^2$. Were the hypotheses on α and the rate of convergence consistent with the previous theorem?

Strong convexity

If the objective function is strongly convex, can we expect the gradient method to converge faster?

Introduction to Optimization

**Lecture 06: Gradient descent, continued.
Geometry and convergence rates. Computational exercise.**



The gradient method

f is L -smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

- ① If $\inf(f) > -\infty$, $\lim_{n \rightarrow \infty} \|\nabla f(x_n)\| = 0$, and $\min_{1 \leq i \leq n} \{\|\nabla f(x_i)\|\} \leq \frac{C}{\sqrt{n}}$.
- ② If f is convex, $\lim_{n \rightarrow \infty} f(x_n) = \inf(f)$ and every cluster point of (x_n) is a minimizer of f .
- ③ If, moreover, f has minimizers, then x_n converges to one of them,
 $f(x_n) - \min(f) \leq \frac{\text{dist}(x_0, S)^2}{\alpha(2 - \alpha L)n}$, and $\lim_{n \rightarrow \infty} n [f(x_n) - \min(f)] = 0$.

Strong convexity and the gradient method

The simplest example, one more time

For $f(x) = x^2$, we obtained $x_n = (1 - 2\alpha)^n x_0$, and so

$$f(x_n) - \min(f) = (1 - 2\alpha)^{2n} (f(x_0) - \min(f)).$$

Part of an exercise from Lecture 04

If f is μ -strongly convex, it has exactly one minimizer, and

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

An important consequence

If f is μ -strongly convex, then $2\mu(f(x) - \min(f)) \leq \|\nabla f(x)\|^2$.

Linear convergence of the gradient method

f is L -smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

If f has minimizers and $2\mu(f(x) - \min(f)) \leq \|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^N$, then

$$f(x_n) - \min(f) \leq [1 - \alpha\mu(2 - \alpha L)]^n (f(x_0) - \min(f)).$$

Questions

- How does this compare with the example?
- What value of α gives the best rate?
- How does this relate to the homework assignment?

Uniform growth and Łojasiewicz inequalities

Proposition

If f is convex, has minimizers, and has quadratic growth:

$$c \operatorname{dist}(x, S)^2 \leq f(x) - \min(f)$$

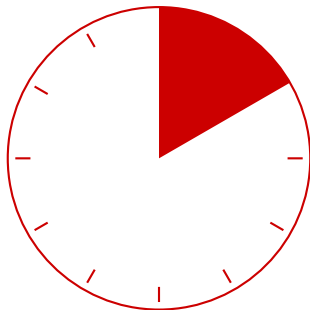
for all $x \in \mathbb{R}^N$, then $c^2(f(x) - \min(f)) \leq \|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^N$.

Exercise

Suppose f has minimizers, and there is $p > 0$ such that, for all $x \in \mathbb{R}^N$, we have $c \operatorname{dist}(x, S)^p \leq f(x) - \min(f)$.

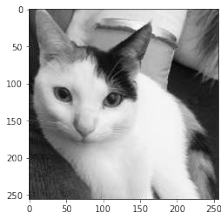
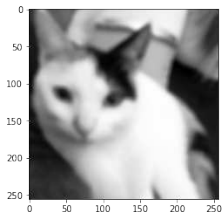
- 1 Which values of p are compatible with f being convex?
- 2 How fast does the gradient method converge in that case?

Break



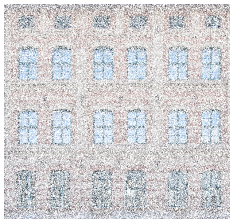
Computational exercise

A simple example of image processing



Deblurring

In-painting



Warm up

We will solve a problem of the form

$$\min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|Ax - b\|^2 + \rho \|Lx\|_1 \right\},$$

but we do not have the tools yet.

Let us begin with something simpler: set $N, M \in \mathbb{N}$, $A \in \mathbb{R}^{M \times N}$, $b \in \mathbb{R}^M$, and consider the least squares minimization problem:

$$\min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|Ax - b\|^2 \right\}.$$

Exercise

- 1 Define a (pseudo)random matrix A of size $M \times N$.
- 2 Write the optimality conditions and find a formula for a solution \bar{x} (it depends on whether $b \in \text{ran}(A)$ or not).
- 3 Set $b = A\mathbf{1}$ and compute \bar{x} with the formula.
- 4 Now compute \bar{x} by applying gradient descent with constant (in the appropriate range) and step sizes, as well as using exact minimization and backtracking.
- 5 Compare the execution times for different combinations of M and N .

Reminder

Choices for α_n

- **Constant:** $\alpha_n \equiv \alpha \in (0, 2/L)$.
- **Vanishing:** $\alpha_n \rightarrow 0$, $\sum \alpha_n = \infty$.
- **Exact minimization:** $\min_{\alpha > 0} f(x_n - \alpha \nabla f(x_n))$.
- **Limited minimization:** $\min_{\alpha \in (0, A]} f(x_n - \alpha \nabla f(x_n))$.
- **Backtracking:** Pick $\alpha_0 > 0$ and $\sigma, \beta \in (0, 1)$, and set

$$m_n := \min \{j \in \mathbb{N} : f(x_n - \alpha_0 \beta^j \nabla f(x_n)) \leq \alpha_0 \beta^j \sigma \|\nabla f(x_n)\|^2\},$$

and then define $x_{n+1} = x_n - \alpha_0 \beta^{m_n} \nabla f(x_n)$.

Introduction to Optimization

Improving convergence rates.

Lecture 07: Quadratic functions and finite termination.

Lecture 08: Inertial algorithms. Stochastic gradient.



university of
groningen

Systems of linear equations

Let $\mathcal{A} \in \mathbb{R}^{N \times N}$ be invertible (to simplify), and let $\beta \in \mathbb{R}^N$. The problem

$$\mathcal{A}x = \beta$$

has a unique solution, which is the unique minimizer of

$$f(x) = \frac{1}{2} \|\mathcal{A}x - \beta\|^2,$$

and also the unique minimizer of

$$\phi(x) = \frac{1}{2} x \cdot Ax - b \cdot x,$$

where $A = \mathcal{A}^T \mathcal{A}$ (symmetric and positive definite) and $b = \mathcal{A}^T \beta$.

Conjugate gradient methods

Conjugate gradient methods iterate


$$x_{n+1} = x_n + \alpha_n d_n$$

for convenient choices of d_n and α_n in such a way that

$$\{d_0, \dots, d_{N-1}\}$$

is a basis of \mathbb{R}^N and x_n minimizes ϕ on the affine subspace

$$x_0 + \text{span}\{d_0, \dots, d_{n-1}\}.$$

As a consequence, the solution is found in at most N steps for every initial point x_0 . 

Conjugate directions

Vectors $\{d_0, \dots, d_{N-1}\}$ are **conjugate with respect to A** if

$$d_i \cdot A d_j = 0$$

whenever $i \neq j$.

Remark

The expression

$$\langle x, y \rangle_A := x \cdot A y$$

is an inner product in \mathbb{R}^N . Being conjugate means being orthogonal with respect to this inner product. Conjugate vectors are linearly independent.

An abstract conjugate gradient method

Theorem

Let $\{d_0, \dots, d_{N-1}\}$ be conjugate, and let $x_{n+1} = x_n + \alpha_n d_n$, where

$$\alpha_n = \operatorname{Argmin}_{\alpha > 0} \phi(x_n + \alpha d_n) = -\frac{d_n \cdot \nabla \phi(x_n)}{\|d_n\|_A^2} = -\frac{d_n \cdot (Ax_n - b)}{\|d_n\|_A^2},$$

for $n = 0, \dots, N-1$. Then x_N minimizes ϕ , whence $Ax_N = b$.

Moreover, $\nabla \phi(x_n) \cdot d_j = 0$ for $j = 0, \dots, n-1$, and x_n minimizes ϕ over

$$x_0 + \operatorname{span}\{d_0, \dots, d_{n-1}\}.$$

This is known as *expanding subspace minimization*.

Implementation of a conjugate gradient method

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla\phi(x_0) = b - Ax_0$.

At step n , we know x_n , $g_n = \nabla\phi(x_n)$ and d_n .

If $g_n = 0$, we stop; otherwise, we compute $\alpha_n = \frac{\|g_n\|^2}{\|d_n\|_A^2}$, and then update

$$x_{n+1} = x_n + \alpha_n d_n$$

$$g_{n+1} = g_n + \alpha_n A d_n$$

$$d_{n+1} = -g_{n+1} + \beta_n d_n, \quad \text{with} \quad \beta_n = \frac{\|g_{n+1}\|^2}{\|g_n\|^2}.$$

Convergence

Theorem

The procedure described above produces a conjugate set $\{d_0, \dots, d_n\}$, with $n \leq N$, and terminates at x_n , where $Ax_n = b$.

Remark

If A has exactly k distinct eigenvalues, the algorithm terminates in at most k steps.

Nonlinear extensions

Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

At step n , we know x_n , $g_n = \nabla f(x_n)$ and d_n .

If $g_n = 0$, we stop; otherwise, we compute (backtracking) α_n satisfying

$$\begin{aligned} f(x_n + \alpha_n d_n) &\leq f(x_n) + c_1 \alpha_n g_n \cdot d_n \\ |\nabla f(x_n + \alpha_n d_n) \cdot d_n| &\leq -c_2 g_n \cdot d_n, \end{aligned}$$

with $0 < c_1 < c_2 < 1/2$ (**strong Wolfe conditions**).

Nonlinear extensions, continued

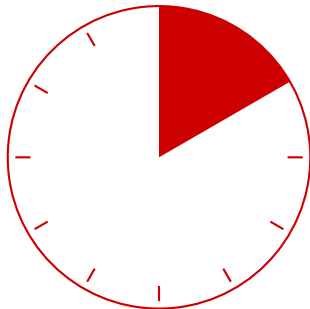
Then, we update

$$\begin{aligned}x_{n+1} &= x_n + \alpha_n d_n \\ d_{n+1} &= -g_{n+1} + \beta_n d_n,\end{aligned}$$

where several choices for β_n are possible, such as:

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

Break



Newton's method

Quadratic model of f given by Taylor's expansion at x_n :

$$f(x) \simeq f(x_n) + \nabla f(x_n) \cdot (x - x_n) + \frac{1}{2}(x - x_n) \cdot \nabla^2 f(x_n)(x - x_n).$$

Set x_{n+1} as the minimizer: $\nabla f(x_n) + \nabla^2 f(x_n)(x_{n+1} - x_n) = 0$, or

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

if $\nabla^2 f(x_n)$ is invertible.

Example: $f(x) = \frac{1}{2} \|Ax - b\|^2$.



Newton's method

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

Theorem

Consider $f : \mathbb{R}^N \rightarrow \mathbb{R}$ 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 neighborhood 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 $n \geq 0$.

Avoid the cost of $\nabla^2 f(x_n)^{-1}$

Define $x_{n+1} = x_n - \alpha_n D_n \nabla f(x_n)$, where

$$D_n \sim \nabla^2 f(x_n)^{-1}$$

in some sense, and is not as costly to compute.

One simple heuristic is **periodic evaluation**: Choose $p \in \mathbb{N}$ and define

$$D_n = \nabla^2 f(x_{kp})^{-1} \quad \text{for } n = kp, kp + 1, \dots, (k + 1)p - 1.$$

Another idea is to define D_{n+1} as a function of D_n (not $\nabla^2 f(x_n)$ or $\nabla^2 f(x_{n+1})$) while keeping the **essence** of Newton's method.

Quasi-Newton methods

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

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

Let us construct D_{n+1} so that it is symmetric (like the Hessian), **not too different** from D_n , and satisfies

$$D_{n+1} g_n = s_n,$$

where $g_n = \nabla f(x_{n+1}) - \nabla f(x_n)$ and $s_n = x_{n+1} - x_n$.

Popular instances

- DFP: Davidon (1959), Fletcher and Powell (1987)

$$D_{n+1} = D_n - \frac{(D_n g_n)(D_n g_n)^T}{g_n \cdot D_n g_n} + \rho_n (s_n s_n^T), \quad \rho_n = \frac{1}{g_n \cdot s_n}.$$

- BFGS: Broyden, Fletcher, Goldfarb and Shanno (1970)

$$D_{n+1} = (I - \rho_n s_n g_n^T) D_n (I - \rho_n s_n g_n^T)^T + \rho_n (s_n s_n^T).$$

Theorem

Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ 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.

Introduction to Optimization

Improving convergence rates.

Lecture 07: Quadratic functions and finite termination.

Lecture 08: Inertial algorithms. Stochastic gradient.



university of
groningen

Momentum, inertia, acceleration

$x_{n+1} = x_n - \alpha_n \nabla f(x_n)$ is equivalent to

$$-\frac{x_{n+1} - x_n}{\alpha_n} = \nabla f(x_n),$$

which is an approximation of the **steepest descent** evolution equation

$$-\dot{x}(t) = \nabla f(x(t)).$$

Other dynamics are related to minimization of potentials. For example,

$$m\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0.$$

Discretization

We discretize

$$m\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0$$

to obtain

$$m \frac{x_{n+1} - 2x_n + x_{n-1}}{h_n^2} + \gamma_n \frac{x_n - x_{n-1}}{h_n} + \nabla f(y_n) = 0.$$

Equivalently,

$$x_{n+1} = x_n + \beta_n (x_n - x_{n-1}) - \alpha_n \nabla f(y_n),$$

with $\alpha_n = \frac{h_n^2}{m}$ and $\beta_n = 1 - \frac{\gamma_n h_n}{m}$.

Two popular choices

Polyak's **heavy ball** (1964)

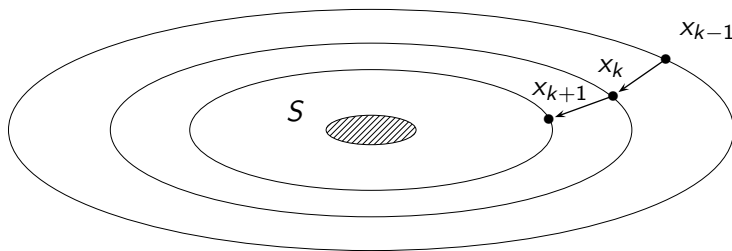
$$x_{n+1} = x_n + \beta_n (x_n - x_{n-1}) - \alpha_n \nabla f(x_n).$$

Nesterov's extrapolation (1983)

$$\begin{cases} y_n &= x_n + \beta_n (x_n - x_{n-1}) \\ x_{n+1} &= y_n - \alpha_n \nabla f(y_n). \end{cases}$$

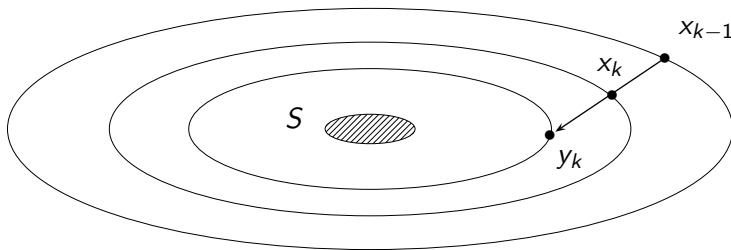
Nesterov's extrapolation

The main idea is the following: Instead of doing this



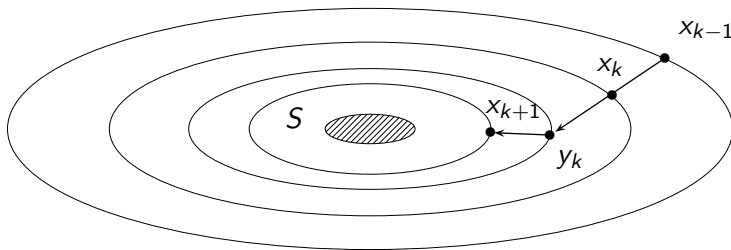
Nesterov's extrapolation

Better try this



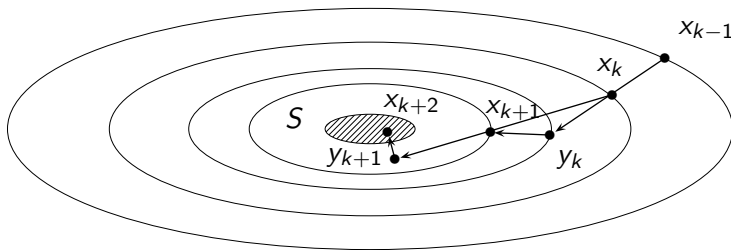
Nesterov's extrapolation

Better try this



Nesterov's extrapolation

Better try this



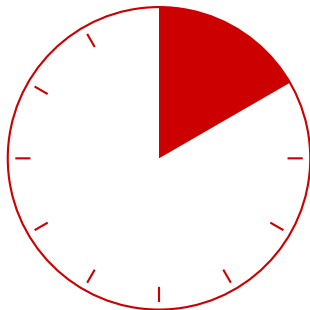
Convergence of Nesterov's method

Theorem

Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ be an L -smooth convex function with minimizers, and let (x_n, y_n) be generated by Nesterov's method with convenient α_n, β_n .

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

Break



The stochastic gradient method

Context and definition

Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$, and let Ξ be a probability space.

The **stochastic gradient method** is defined by

$$x_{n+1} = x_n - \alpha_n g(x_n, \xi_n),$$

where $\alpha_n > 0$, (ξ_n) is an i.i.d. sequence of random variables in Ξ , and $g : \mathbb{R}^N \times \Xi$ is intended to approximate ∇f in the sense that

$$\mathbb{E}_{\xi}(g(x, \xi)) \sim \nabla f(x)$$

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

Examples

- ❶ **Noisy Gradients:** $g(x, \xi) = \nabla f(x) + \xi$, with $\mathbb{E}(\xi) = 0$.
- ❷ **Incremental Gradient Method:** For $f = \frac{1}{M} \sum_{m=1}^M f_m$, at iteration n , we select $j_n \in \{1, \dots, M\}$ and compute $x_{n+1} = x_n - \alpha_n \nabla f_{j_n}(x_n)$.
- ❸ **Empirical Risk Minimization:** The empirical risk is defined by
$$R(\phi) = \frac{1}{M} \sum_{m=1}^M \ell(\phi(x_m), y_m).$$

Key assumption for convergence

We suppose that f is convex, \hat{x} is a minimizer of f , and there exist $L, B \geq 0$ such that

$$\mathbb{E}_{\xi} [\|g(x, \xi)\|^2] \leq L^2 \|x - \hat{x}\|^2 + B^2.$$

Example

In the context of incremental gradient, this holds if each f_m is L_m -Lipschitz, attains its minimum at some \hat{x}_m , and we set

$$L^2 = \frac{2}{M} \sum_{m=1}^M L_m^2 \quad \text{and} \quad B^2 = \frac{2}{M} \sum_{m=1}^M L_m^2 \|\hat{x}_m - \hat{x}\|^2.$$

Convergence results I: $L = 0$

Set

$$\sigma_n = \sum_{k=0}^n \alpha_k, \quad \tau_n = \sum_{k=0}^n \alpha_k^2, \quad \text{and} \quad \bar{x}_n = \frac{1}{\sigma_n} \sum_{k=0}^n \alpha_k x_k.$$

Then, for each $n \geq 1$, we have

$$\mathbb{E}[f(\bar{x}_n) - \min(f)] \leq \frac{D_0^2 + \tau_n B}{2\sigma_n},$$

where $D_0 = \text{dist}(x_0, S)$.

Convergence results II: L possibly nonzero


Assume f is μ -strongly convex and $\alpha_n \equiv \alpha$. For each $n \geq 1$, we have

$$\mathbb{E}[\|x_n - x^*\|^2] \leq D_0^2 (1 - 2\alpha\mu + \alpha^2 L^2)^n + \frac{\alpha B^2}{2\mu - \alpha L^2}.$$


Questions

- Can we obtain a convergence rate for $\mathbb{E}(f(x_n) - \min(f))$?
- If $B = 0$, we obtain linear convergence. What is the best possible rate? How does this compare with the deterministic gradient method?
- If $B \neq 0$, can we obtain convergence by using vanishing step sizes?

Variants

- Batching for incremental gradient. 
- Nesterov's Acceleration (similar with heavy ball)

$$\begin{cases} y_n &= x_n + \beta_n(x_n - x_{n-1}) \\ x_{n+1} &= SG(y_n) \end{cases}$$

- Parameter selection
 - Step sizes (learning rates) prescribed *a priori*: Epochs 
 - Adaptive: Adam
- Variance reduction
 - SVRG, SAG, SAGA

Adam: Adaptive Moment Estimation (2015)

- The **direction** is updated by

$$d_n = [\beta_1 d_{n-1} + (1 - \beta_1) g_n] (1 - \beta_1^n)^{-1}.$$

- The **second order moment** is estimated by

$$v_n^{(i)} = \left[\beta_2 v_{n-1}^{(i)} + (1 - \beta_2) \left(g_n^{(i)} \right)^2 \right] (1 - \beta_2^n)^{-1}$$

- The **step size (learning rate)** is set at $\alpha_n^{(i)} = \frac{\alpha}{\sqrt{v_n^{(i)} + \varepsilon}}.$

- Finally, the next iterate is computed by $x_{n+1} = x_n - \alpha_n d_n.$

Variance Reduction

- Motivation:

- Let \mathcal{X}, \mathcal{Y} be two random variables and set $\mathcal{Z} = \mathcal{X} - (\mathcal{Y} - \mathbb{E}(\mathcal{Y}))$.
- Then, $\mathbb{E}(\mathcal{Z}) \sim \mathbb{E}(\mathcal{X})$ and $\mathbb{V}(\mathcal{Z}) = \mathbb{V}(\mathcal{X}) - 2\text{Cov}(\mathcal{X}, \mathcal{Y}) + \mathbb{V}(\mathcal{Y})$.
- If \mathcal{X}, \mathcal{Y} are **highly correlated**, then $\mathbb{V}(\mathcal{Z})$ is small.

- SAG (Stochastic Average Gradient):

- Strongly convex case (2012).
- Convex case (2014).
- SAGA (2014): Unbiased, suitable for nonsmooth and non-strongly convex functions.

SAGA

We have $f(x) = \frac{1}{M} \sum_{m=1}^M f_m(x)$

If $j \in \{1, \dots, M\}$ is picked uniformly at random, then

$$\mathbb{E}_j(\nabla f_j(x)) = \frac{1}{M} \sum_{m=1}^M \nabla f_m(x).$$

In the notation introduced above, we set

$$\mathcal{X} = \nabla f_{j_n}(x_{n+1})$$

$$\mathcal{Y} = \nabla f_{j_n}(x_n)$$

$$\mathcal{Z} = \nabla f_{j_n}(x_{n+1}) - \left[\nabla f_{j_n}(x_n) - \frac{1}{M} \sum_{m=1}^M \nabla f_m(x_n) \right]$$

SAGA

We have $f(x) = \frac{1}{M} \sum_{m=1}^M f_m(x)$

Instead of

$$\mathcal{Z} = \nabla f_{j_n}(x_{n+1}) - \left[\nabla f_{j_n}(x_n) - \frac{1}{M} \sum_{m=1}^M \nabla f_m(x_n) \right],$$

which would be costly, we define

$$\mathcal{Z}' = \nabla f_{j_n}(x_{n+1}) - \left[\nabla f_{j_n}(x_n) - \frac{1}{M} \sum_{m=1}^M \mathbf{g}_m^{(n)} \right],$$

where we have $\mathbf{g}_m^{(n)}$ in storage for $m = 1, \dots, M$.

SAGA in practice

Begin with $x_0 \in \mathbb{R}^N$, compute $g_m^{(0)} = \nabla f_m(x_0)$ for $m = 1, \dots, M$, and construct a matrix $\mathcal{G}_0 = \begin{bmatrix} g_1^{(0)} & \dots & g_M^{(0)} \end{bmatrix}$.

After iteration n , we have a point x_n and a matrix \mathcal{G}_n .

Pick $j_n \in \{1, \dots, M\}$ uniformly at random, compute $g_{j_n}^{(n+1)} = \nabla f_{j_n}(x_n)$, and update \mathcal{G}_n to \mathcal{G}_{n+1} by replacing only the j_n -th column by $g_{j_n}^{(n+1)}$.

Finally, $x_{n+1} = x_n - \alpha_n \left[g_{j_n}^{(n+1)} - g_{j_n}^{(n)} + \frac{1}{M} \sum_{m=1}^M g_m^{(n)} \right]$, with $\alpha_n > 0$.




Introduction to Optimization

**Lecture 09: Simple constraints and projected gradient.
Nonsmooth convex functions and subgradient method.**



A quick reminder

Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex. For each $x \in \mathbb{R}^N$, the **projection** of x onto C , denoted by $P_C(x)$, is the point in C , which is the closest to x .

- ① The positive orthant \mathbb{R}_+^N . 
- ② Balls $\bar{B}(x_0, r)$ and boxes $\prod [a_i, b_i]$. 
- ③ Affine spaces, such as $\{x_0\} + V$, or $\{x \in \mathbb{R}^N : Ax = b\}$. 

The projected gradient method

Consider the problem

$$\min_{x \in C} f(x),$$

where

- $f : \mathbb{R}^N \rightarrow \mathbb{R}$, and
- $C \subset \mathbb{R}^N$ is nonempty, closed, convex and easy to project onto.

The set of solutions is denoted by S , and the optimal value by f^* .

Idea: At each iteration, perform a gradient step and then project onto C .

$$x_{n+1} = P_C(x_n - \alpha_n \nabla f(x_n)).$$

Convergence of the projected gradient method

Theorem

Let f be convex and L -smooth, and suppose $S \neq \emptyset$. Iterate $x_{n+1} = P_C(x_n - \alpha \nabla f(x_n))$ with $0 < \alpha \leq \frac{1}{L}$.

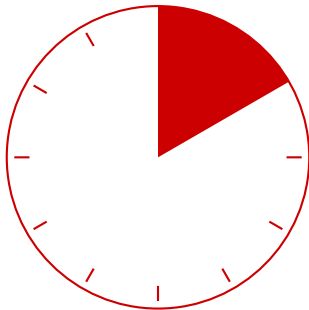
① x_n converges to a point in S , and

$$f(x_n) - f^* \leq \frac{\text{dist}(x_0, S)^2}{2\alpha n}.$$

② If f is μ -strongly convex, then

$$f(x_n) - f^* \leq \frac{f(x_0) - f^*}{(1 + \alpha\mu)^n}.$$

Break



Nonsmooth convex functions

Let $f : A \subset \mathbb{R}^N \rightarrow \mathbb{R}$ be convex. If f is differentiable at x , then

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$$

for all $y \in A$.

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 $y \in A$. The **subdifferential** of f at x , denoted by $\partial f(x)$, is the set of all the subgradients of f at x .

A few simple but important examples

Example (A big ice cream cone)

Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ be defined by $f(x) = \|x\|$.

Example (The ℓ^1 -norm)

Let $\|\cdot\|_1 : \mathbb{R}^N \rightarrow \mathbb{R}$ be defined by $\|x\|_1 = |x_1| + \cdots + |x_N|$.

Example (The indicator function)

Let $C \subset \mathbb{R}^N$ be nonempty, closed and convex. The **indicator function** of C is the function $\iota_C : \mathbb{R}^N \rightarrow \mathbb{R}$, defined as $\iota_C(x) = 0$ for all $x \in C$.

Introduction to Optimization

Lecture 10: Subgradient descent. The proximal-gradient algorithm.



Reminder

A vector $v \in \mathbb{R}^N$ is a **subgradient** of f a convex function $f : \text{dom}(f) \subset \mathbb{R}^N \rightarrow \mathbb{R}$ at the point x if

$$f(y) \geq f(x) + v \cdot y - x$$

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

Proposition

If $f : \text{dom}(f) \subset \mathbb{R}^N \rightarrow \mathbb{R}$ is convex, for each $x \in \text{int}(\text{dom}(f))$, there exist $L_x, r_x > 0$ such that

$$|f(z) - f(y)| \leq L_x \|z - y\|$$

for all $z, y \in B(x, r_x)$. Moreover, $\emptyset \neq \partial f(x) \subseteq \bar{B}(0, L_x)$.

The subgradient method

$$x_{n+1} = x_n - \alpha v_n \text{ with } v_n \in \partial f(x_n)$$

Proposition

Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ be convex and Lipschitz-continuous with constant M ($|f(x) - f(y)| \leq M\|x - y\|$) with minimizers, and let (x_n) be defined by the subgradient method. Set $\bar{x}_n = \frac{1}{n+1} \sum_{k=0}^n x_k$. Then,

$$\min_{k=1, \dots, n} (f(x_k) - \min(f)) \leq f(\bar{x}_n) - \min(f) \leq \frac{\alpha M^2}{2} + \frac{\text{dist}(x_0, S)^2}{2\alpha(n+1)}.$$

Question

Given $\varepsilon > 0$, after how many iterations can we be sure to have found a point \hat{x} such that $f(\hat{x}) - \min(f) \leq \varepsilon$?

Extended real-valued functions, I

We extend the codomain \mathbb{R} to $\mathbb{R} \cup \{+\infty\}$, with the conventions that $+\infty > \gamma$ for all $\gamma \in \mathbb{R}$, and some algebraic operations are allowed:

- $+\infty + \gamma = +\infty$ for all $\gamma \in \mathbb{R}$,
- $\gamma(+\infty) = +\infty$ for all $\gamma > 0$, and $0(+\infty) = 0$.

The **(effective) domain** and **epigraph** of a function $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is

$$\begin{aligned}\text{dom}(f) &= \{x \in \mathbb{R}^N : f(x) < +\infty\} \\ \text{epi}(f) &= \{(x, z) \in \mathbb{R}^{N+1} : f(x) \leq z\},\end{aligned}$$

respectively. We will **always** assume that $\text{dom}(f) \neq \emptyset$.

Notice that $\text{dom}(\lambda f + g) = \text{dom}(f) \cap \text{dom}(g)$.

Extended real-valued functions, II

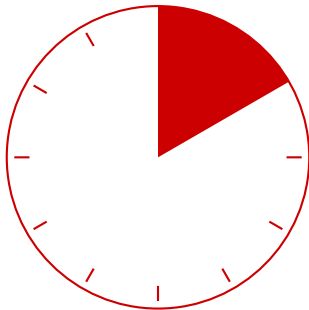
A function $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is **closed** if $\text{epi}(f)$ is closed.

Examples

- 1 Every continuous function is closed.
- 2 The **indicator function** of $C \subset \mathbb{R}^N$ ($C \neq \emptyset$) is $\iota_C : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$, defined as $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = +\infty$ if $x \notin C$.
Here, $\text{dom}(\iota_C) = C$ and $\text{epi}(\iota_C) = C \times [0, +\infty)$.
This function is closed if C is closed, and convex if C is convex.

If $f : \mathbb{R}^N \rightarrow \mathbb{R}$, $\min\{f(x) : x \in C\} = \min\{f(x) + \iota_C(x) : x \in \mathbb{R}^N\}$.

Break



Closedness and proximity operator

If $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is closed and convex, and $y \in \mathbb{R}^N$, the function

$$f_y(x) = f(x) + \frac{1}{2}\|x - y\|^2$$

is closed and strongly convex. Its subdifferential is given by

$$\partial f_y(x) = \partial f(x) + x - y,$$

for each $x \in \text{dom}(f)$. The unique minimizer of f_y is denoted by $\text{prox}_f(y)$, and is characterized by

$$y - \text{prox}_f(y) \in \partial f(\text{prox}_f(y)).$$

The proximal method

If $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is closed and convex, and fix $\alpha > 0$. From an initial point $x_0 \in \mathbb{R}^N$, define a sequence inductively by

$$x_{n+1} = \text{prox}_{\alpha f}(x_n).$$

Exercise

If $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is closed and convex, and $S \neq \emptyset$, then x_n converges to a point in S . Moreover,

$$f(x_n) - \min(f) \leq \frac{\text{dist}(x_0, S)^2}{2\alpha n}, \quad n \geq 1.$$

Introduction to Optimization

Lecture 11: Convergence of the proximal-gradient algorithm. Conjugate functions.



university of
 groningen

Proximal-gradient algorithm

Suppose we want to find the minima of $f = g + h$, where $g : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is closed and convex, and $h : \mathbb{R}^N \rightarrow \mathbb{R}$ is L -smooth and convex.

Example

A typical example in image and signal processing, statistics, ML, is

$$f(x) = \frac{1}{2} \|Ax - b\|^2 + \rho \|x\|_1$$

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

Proximal-gradient algorithm

The **proximal-gradient** method consists in applying proximal iterations while linearizing the smooth function:

$$\begin{aligned}x_{n+1} &= \text{prox}_{\alpha g}(x_n - \alpha \nabla h(x_n)) \\ &= \text{Argmin} \left\{ g(x) + \frac{1}{2\alpha} \|x - (x_n - \alpha \nabla h(x_n))\|^2 \right\}.\end{aligned}$$

This subproblem has a unique solution characterized by

$$0 \in \partial g(x_{n+1}) + \nabla h(x_n) + \frac{1}{\alpha}(x_{n+1} - x_n).$$

Convergence of proximal-gradient sequences

Theorem

Let $f = g + h$, where $g : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is closed and convex, and $h : \mathbb{R}^N \rightarrow \mathbb{R}$ is L -smooth and convex. Take $\alpha \in (0, 1/L]$ and define (x_n) by

$$x_{n+1} = \text{prox}_{\alpha g}(x_n - \alpha \nabla h(x_n)), \quad n \geq 0.$$

If $S \neq \emptyset$, x_n converges to an $\hat{x} \in S$, and

$$f(x_n) - \min(f) \leq \frac{\text{dist}(x_0, S)^2}{2\alpha n}, \quad n \geq 1.$$

Moreover, $\lim_{n \rightarrow \infty} n(f(x_n) - \min(f)) = 0$.

Sketch of the proof

$f(x_n) + g(x_n)$ is nonincreasing

$$\begin{cases} f(x_{n+1}) & \leq f(x_n) + \nabla f(x_n) \cdot (x_{n+1} - x_n) + \frac{L}{2} \|x_{n+1} - x_n\|^2 \\ g(x_{n+1}) & \leq g(x_n) + \left(\frac{x_n - x_{n+1}}{\alpha} - \nabla f(x_n) \right) \cdot (x_{n+1} - x_n). \end{cases}$$

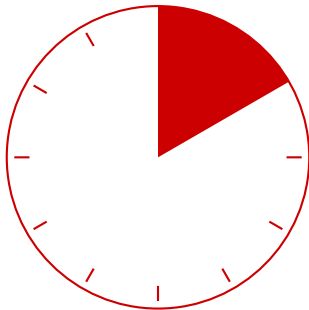
Compatibility

$$\|x_{n+1} - x_n - \alpha(\nabla f(x_{n+1}) - \nabla f(x_n))\|^2 \leq \|x_{n+1} - x_n\|^2.$$

Convergence rate

$$\begin{cases} f(x_n) & \leq f(p) + \nabla f(x_n) \cdot (x_n - p) \\ g(x_{n+1}) & \leq g(p) + \left(\frac{x_n - x_{n+1}}{\alpha} - \nabla f(x_n) \right) \cdot (x_{n+1} - p). \end{cases}$$

Break



The Fenchel conjugate

The **Fenchel conjugate** of a closed convex function $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is the function $f^* : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$f^*(x^*) = \sup_{x \in \mathbb{R}^N} \{x^* \cdot x - f(x)\}.$$

Examples

- In \mathbb{R} : (1) $f(x) = x$; (2) $f(x) = \frac{1}{2}x^2$; (3) $f(x) = e^x$.
- In \mathbb{R}^N : (4) $f(x) = c \cdot x + \alpha$; (5) $f(x) = \phi(\|x\|)$.

Basic properties

Proposition

- ❶ The function f^* is closed, convex and not identically $+\infty$.
- ❷ If $f \leq g$, then $f^* \geq g^*$.
- ❸ **Fenchel-Young Inequality:** $f(x) + f^*(x^*) \geq x^* \cdot x$.
There is equality if, and only if, $x^* \in \partial f(x)$.
- ❹ $f^{**} = f$.
- ❺ **Legendre-Fenchel Reciprocity Formula:** $x^* \in \partial f(x)$ if, and only if, $x \in \partial f^*(x^*)$.
- ❻ Let $\mu\ell = 1$. Then, f is μ -strongly convex if, and only if, f^* is ℓ -smooth.

Introduction to Optimization

Lecture 12: Duality and algorithms.



The Fenchel conjugate

The **Fenchel conjugate** of a closed convex function $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ is the closed convex function $f^* : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$f^*(x^*) = \sup_{x \in \mathbb{R}^N} \{x^* \cdot x - f(x)\}.$$

Fenchel-Young Inequality: $f(x) + f^*(x^*) \geq \langle x^*, x \rangle$.

There is equality if, and only if, $x^* \in \partial f(x)$.

Legendre-Fenchel Reciprocity Formula: $x^* \in \partial f(x) \iff x \in \partial f^*(x^*)$.

Fenchel-Rockafellar duality

Let $P \in \mathbb{R}^{M \times N}$, and let $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^M \rightarrow \mathbb{R} \cup \{+\infty\}$ be closed and convex.

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

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

Proposition

The **duality gap** $v + v^*$ is nonnegative.

Characterization of the primal-dual solutions

Theorem

The following statements concerning $\hat{x} \in \mathbb{R}^N$ and $\hat{y} \in \mathbb{R}^M$ are equivalent:

- i) $-P^T \hat{y} \in \partial f(\hat{x})$ and $\hat{y} \in \partial g(P\hat{x})$;
- ii) $f(\hat{x}) + f^*(-P^T \hat{y}) = \langle -P^T \hat{y}, \hat{x} \rangle$ and $g(P\hat{x}) + g^*(\hat{y}) = \langle \hat{y}, P\hat{x} \rangle$;
- iii) $f(\hat{x}) + g(P\hat{x}) + f^*(-P^T \hat{y}) + g^*(\hat{y}) = 0$; and
- iv) $\hat{x} \in S$ and $\hat{y} \in S^*$ and $v + v^* = 0$.

Moreover, if $\hat{x} \in S$ and g is continuous, there exists $\hat{y} \in \mathbb{R}^M$ such that all four statements hold.

Structured optimization problem

We consider the problem

$$\min \{f(x) + g(Px) + h(x)\},$$

where

- $P \in \mathbb{R}^{M \times N}$;
- $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^M \rightarrow \mathbb{R} \cup \{+\infty\}$ are closed and convex; and
- $h : \mathbb{R}^N \rightarrow \mathbb{R}$ is ℓ -smooth and convex.

Primal-dual algorithm

Chambolle-Pock (2011), Condat-Vũ, (2013):

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

with $\tau \sigma \|P\|^2 + \frac{\tau \ell}{2} \leq 1$.

Proposition

Limit points are solutions of the problem.

Remark (Implementation trick: Moreau's Identity)

$$\operatorname{prox}_{\sigma g^*}(y) = y - \sigma \operatorname{prox}_{\sigma^{-1}g}(\sigma^{-1}y).$$

TV Regularization

The **Total Variation Regularization Problem** is

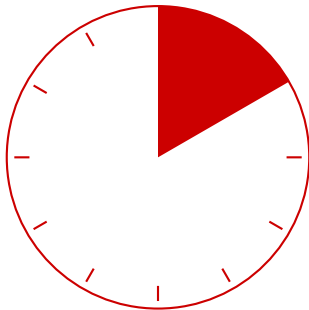
$$\min_{x \in \mathbb{R}^{N_1 \times N_2}} \left\{ \frac{1}{2} \|Fx - b\|^2 + \rho \|Dx\|_1 \right\},$$

where F models or approximates the process by which an image x has been deteriorated to produce b , and D is the **discrete gradient**.

Question

Can we apply the primal-dual algorithm to this problem?

Break



Linear programming

The **linear programming problem** is

$$(LP) \quad \min_{x \in \mathbb{R}^N} \{ c \cdot x : Ax \leq b \},$$

where $c \in \mathbb{R}^N$, A is a matrix of size $M \times N$, and $b \in \mathbb{R}^M$.

It is a primal problem with $f(x) = c \cdot x$ and $g(z) = \iota_{\mathbb{R}_+^M}(b - z)$.

Dual problem

$$(DLP) \quad \min_{y \in \mathbb{R}^M} \{ b \cdot y : A^T y + c = 0, \text{ and } y \geq 0 \}.$$

Exercise

Compute the dual of the dual.