Computational Statistics and Optimisation

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Overline

Introduction

Least squares, quadratic objective functions

Global/local minima

Gradient descent

Forward-backward analysis

Forward-backward accelerated

Duality gap and stopping criterion

Overline

Motivation

Many problems in **Statistics / Machine Learning** have an **optimization** formulation, usually coming from a frequentist modeling

Rem: Bayesian methods would need other tools : approximations of integral instead of function minimization

Among many examples :

- Linear regression/least square (the most common problem)

Motivation

Many problems in **Statistics / Machine Learning** have an **optimization** formulation, usually coming from a frequentist modeling

Rem: Bayesian methods would need other tools : approximations of integral instead of function minimization

Among many examples :

- Linear regression/least square (the most common problem)
- regularized least square : ridge/Tikhonov, Lasso/basis pursuit and variants

Motivation

Many problems in **Statistics / Machine Learning** have an **optimization** formulation, usually coming from a frequentist modeling

Rem: Bayesian methods would need other tools : approximations of integral instead of function minimization

Among many examples :

- Linear regression/least square (the most common problem)
- regularized least square : ridge/Tikhonov, Lasso/basis pursuit and variants
- Logistic regression (w/o regularization)

Motivation

Many problems in **Statistics / Machine Learning** have an **optimization** formulation, usually coming from a frequentist modeling

Rem: Bayesian methods would need other tools : approximations of integral instead of function minimization

Among many examples :

- Linear regression/least square (the most common problem)
- regularized least square : ridge/Tikhonov, Lasso/basis pursuit and variants
- Logistic regression (w/o regularization)
- PCA, Sparse PCA

Motivation

Many problems in **Statistics / Machine Learning** have an **optimization** formulation, usually coming from a frequentist modeling

Rem: Bayesian methods would need other tools : approximations of integral instead of function minimization

Among many examples :

- Linear regression/least square (the most common problem)
- regularized least square : ridge/Tikhonov, Lasso/basis pursuit and variants
- Logistic regression (w/o regularization)
- PCA, Sparse PCA
- Matrix completion (*e.g.*, using trace norm regularization)

Motivation

Many problems in **Statistics / Machine Learning** have an **optimization** formulation, usually coming from a frequentist modeling

Rem: Bayesian methods would need other tools : approximations of integral instead of function minimization

Among many examples :

- Linear regression/least square (the most common problem)
- regularized least square : ridge/Tikhonov, Lasso/basis pursuit and variants
- Logistic regression (w/o regularization)
- PCA, Sparse PCA
- Matrix completion (*e.g.*, using trace norm regularization)

Classical regression / least square model

- p variables / features
- n observations

Simple linear model

$$y_i = + \sum_{j=1}^{p} \theta_j^* x_{i,j} + \varepsilon_i$$

$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \ldots, n$$

$$\mathbb{E}(\varepsilon) = 0$$

$$\text{System formulation} \begin{cases} y_1 & = \sum_{j=1}^{p} \theta_j^* x_{1,j} + \varepsilon_1 \\ & \vdots \\ y_n & = \sum_{j=1}^{p} \theta_j^* x_{n,j} + \varepsilon_n \end{cases}$$

Dimension p

Matrix model

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \theta_1^* \\ \vdots \\ \theta_p^* \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix}, \theta^* = \begin{pmatrix} \theta_1^* \\ \vdots \\ \theta_p^* \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boxed{\mathbf{y} = X\theta^* + \boldsymbol{\varepsilon}}$ or $y_i = \langle X_{i,:}, \theta^* \rangle + \varepsilon_i$ for $i = 1, \dots, n$

Rem: Notation $X = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ – features are columnwise

Matrix / vector formulation

$$\boxed{\mathbf{y} = X\theta^* + \boldsymbol{\varepsilon}}$$

- $\mathbf{y} \in \mathbb{R}^n$: observations
- $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$: features
- $\theta^* \in \mathbb{R}^p$: (true) model parameter - target
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$: noise

Overline

Introduction

Least squares, quadratic objective functions

Global/local minima

Gradient descent

Forward-backward analysis

Forward-backward accelerated

Duality gap and stopping criterion

Least square / Ridge estimator

Least square optimization problem :

$$\hat{\theta}^{\mathrm{LS}} \in \underset{\theta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|\mathbf{y} - X\theta\|_2^2 \right)$$

$$\hat{\theta}^{\mathrm{LS}} \in \underset{\theta \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} \left[y_i - \left(\sum_{j=1}^{p} \theta_j x_{i,j} \right) \right]^2$$

Ridge regression optimization problem (with parameter $\lambda > 0$)

$$\hat{\theta}_{\lambda}^{\mathrm{Ridge}} \in \underset{\theta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|\mathbf{y} - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right)$$

<u>Rem</u>: Later we will see the Lasso (ℓ_1 regularization), but it is not a smooth function

Quadratic function in dimension two

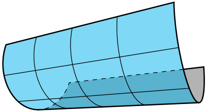
$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2b x_1 x_2 + c x_2^2 \end{cases}$$

A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

Quadratic function in dimension two

$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2 b x_1 x_2 + c x_2^2 \end{cases}$$

A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

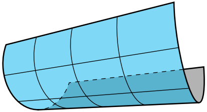


$(y_1, y_2) \mapsto y_1^2$

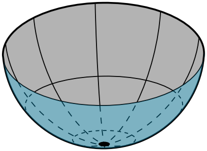
Quadratic function in dimension two

$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2b x_1 x_2 + c x_2^2 \end{cases}$$

A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



$(y_1, y_2) \mapsto y_1^2$

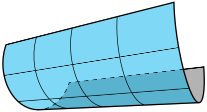


$(y_1, y_2) \mapsto y_1^2 + y_2^2$

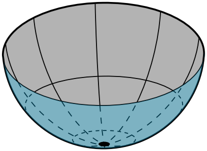
Quadratic function in dimension two

$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2b x_1 x_2 + c x_2^2 \end{cases}$$

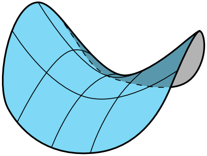
A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



$(y_1, y_2) \mapsto y_1^2$



$(y_1, y_2) \mapsto y_1^2 + y_2^2$

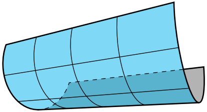


$(y_1, y_2) \mapsto y_1^2 - y_2^2$

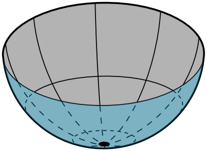
Quadratic function in dimension two

$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2 b x_1 x_2 + c x_2^2 \end{cases}$$

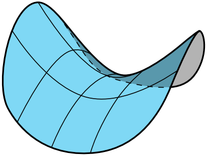
A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



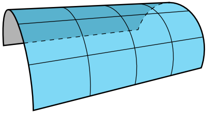
$(y_1, y_2) \mapsto y_1^2$



$(y_1, y_2) \mapsto y_1^2 + y_2^2$



$(y_1, y_2) \mapsto y_1^2 - y_2^2$

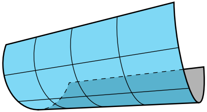


$(y_1, y_2) \mapsto -y_1^2$

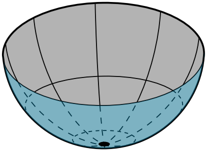
Quadratic function in dimension two

$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2 b x_1 x_2 + c x_2^2 \end{cases}$$

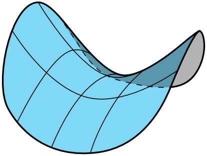
A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



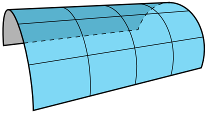
$(y_1, y_2) \mapsto y_1^2$



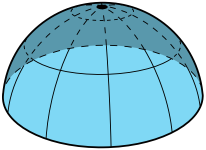
$(y_1, y_2) \mapsto y_1^2 + y_2^2$



$(y_1, y_2) \mapsto y_1^2 - y_2^2$



$(y_1, y_2) \mapsto -y_1^2$

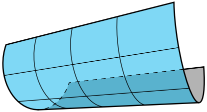


$(y_1, y_2) \mapsto -(y_1^2 + y_2^2)$

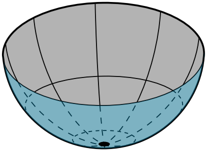
Quadratic function in dimension two

$$\begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x_1, x_2) & \mapsto x^\top A x = a x_1^2 + 2 b x_1 x_2 + c x_2^2 \end{cases}$$

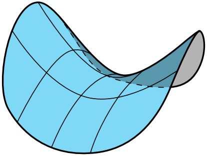
A symmetric real matrix : $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



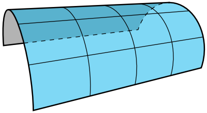
$(y_1, y_2) \mapsto y_1^2$



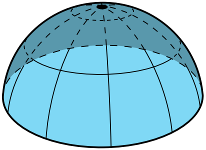
$(y_1, y_2) \mapsto y_1^2 + y_2^2$



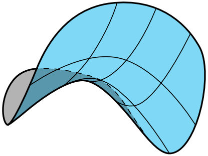
$(y_1, y_2) \mapsto y_1^2 - y_2^2$



$(y_1, y_2) \mapsto -y_1^2$



$(y_1, y_2) \mapsto -(y_1^2 + y_2^2)$



$(y_1, y_2) \mapsto y_2^2 - y_1^2$

Quadratic function / least square / solving linear system

For a matrix $A \in \mathbb{R}^{p\times p}$ and $b \in \mathbb{R}^p$ the following are equivalent :

- Solving in x a system $Ax = b$
- Minimizing w.r.t to x the function $f(x) = \frac{1}{2}x^\top A^\top Ax - b^\top Ax$

Example :

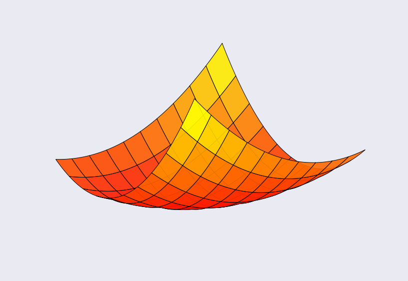
$$f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 6x_2^2 + 4x_1x_2) - 2x_1 + 8x_2$$

with

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \text{ and } b = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Optimization in $\mathbb{R}^p$

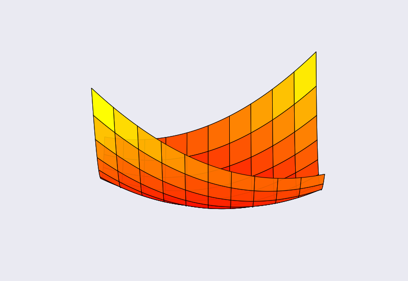
Quadratic function (Positive)



Example: $f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 6x_2^2 + 4x_1x_2) - 2x_1 + 8x_2$

Optimization in $\mathbb{R}^p$

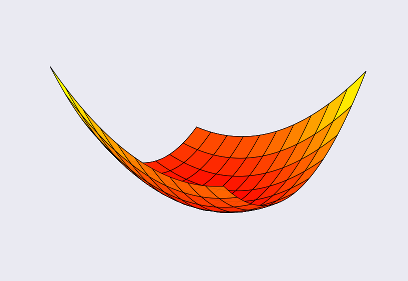
Quadratic function (Positive)



Example: $f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 6x_2^2 + 4x_1x_2) - 2x_1 + 8x_2$

Optimization in $\mathbb{R}^p$

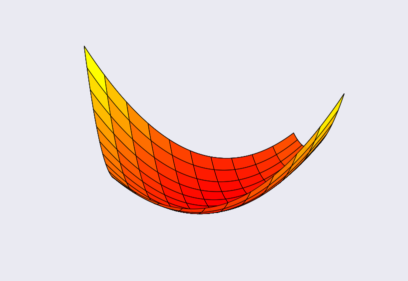
Quadratic function (Positive)



Example: $f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 6x_2^2 + 4x_1x_2) - 2x_1 + 8x_2$

Optimization in $\mathbb{R}^p$

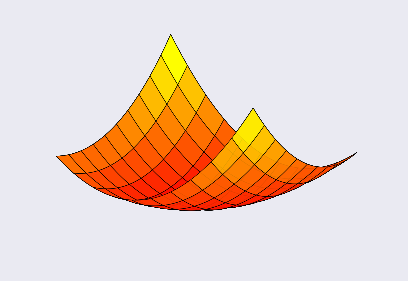
Quadratic function (Positive)



Example: $f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 6x_2^2 + 4x_1x_2) - 2x_1 + 8x_2$

Optimization in $\mathbb{R}^p$

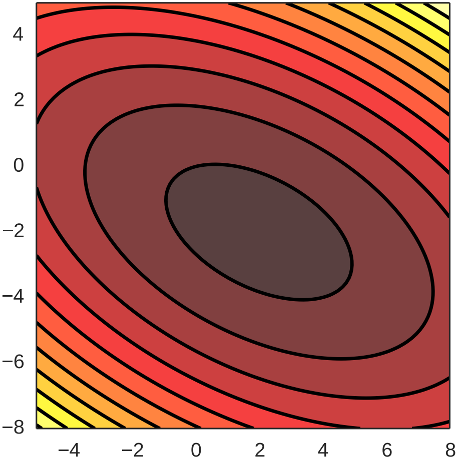
Quadratic function (Positive)



Example: $f(x_1, x_2) = \frac{1}{2}(3x_1^2 + 6x_2^2 + 4x_1x_2) - 2x_1 + 8x_2$

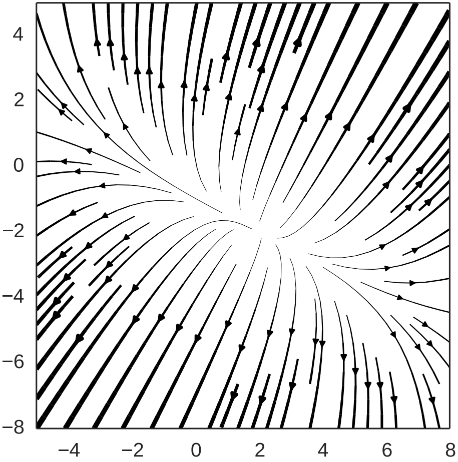
Level lines / gradient flow

Level set of the same function



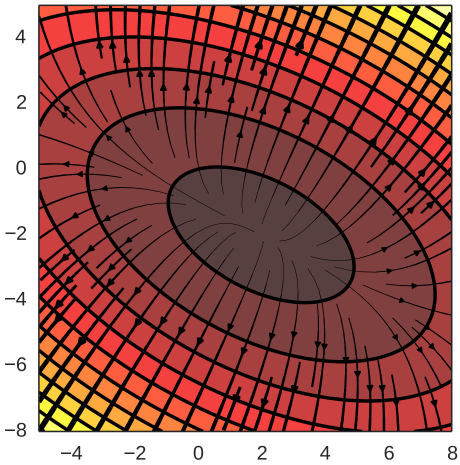
Level lines / gradient flow

Gradient flow of the same function



Level lines / gradient flow

Level set and gradient flow of the same function



Least square case

Canonical problem :

$$\hat{\theta}^{\mathrm{LS}} \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} \left(\frac{1}{2} \|\mathbf{y} - X\theta\|_2^2 \right)$$

Note that $f(\theta) = \frac{1}{2}\|\mathbf{y} - X\theta\|_2^2 = \frac{1}{2}\theta^\top X^\top X\theta - \langle \theta, X^\top \mathbf{y} \rangle + \frac{1}{2}\|y\|_2^2$
Hence the problem is quadratic.

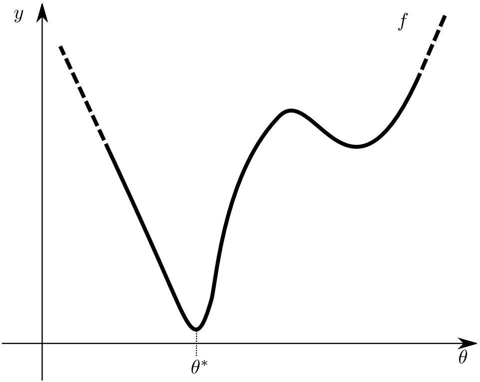
Rem: the (Gram) matrix $X^\top X$ is **positive-semidefinite**
Rem: Uniqueness is not always guaranteed, since one needs
$\ker(X^\top X) = \ker(X) \neq \{0\}$

Overline

Existence of a minimum

Coercive functions

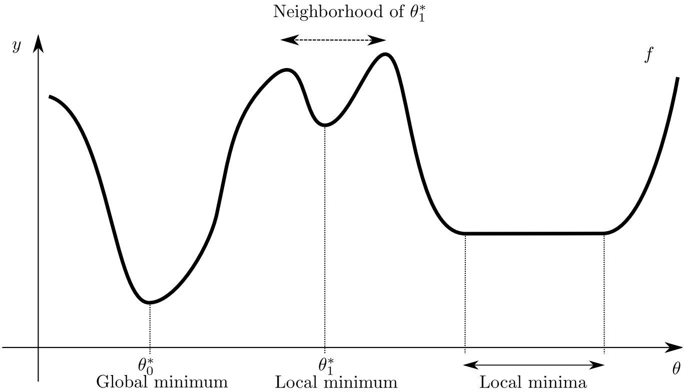
Let a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ be continuous satisfying $\lim_{\|x\|\to\infty} f(\theta) = +\infty$ (*i.e.*, **coercive**) then, there exists a point $\theta^\star$ where the minimum is reached : $\theta^\star \in \underset{\theta\in\mathbb{R}^p}{\arg\min}\, f(\theta)$



Local vs global minima

Definition : local minimum

$f : \mathbb{R}^p \mapsto \mathbb{R}$ has **local minimum** at $\theta^\star$ if $\theta^\star$ is a minimum of f restricted to a neighborhood of $\theta^\star$



<u>Rem</u>: : a global minimum is also a local minimum

Convex case : local = global

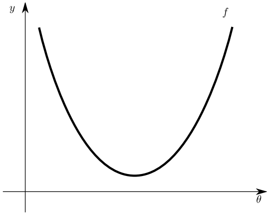
Theorem : equivalence local/global in the convex case

If a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is convex, then any local minimum of f also a global minimum of f.

Convex case : local = global

Theorem : equivalence local/global in the convex case

If a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is convex, then any local minimum of f also a global minimum of f.

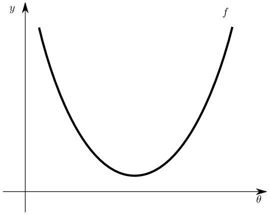


Convex : 1 global minimum

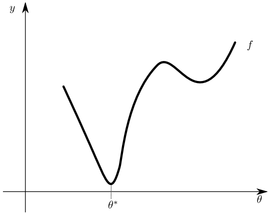
Convex case : local = global

Theorem : equivalence local/global in the convex case

If a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is convex, then any local minimum of f also a global minimum of f.



Convex : 1 global minimum

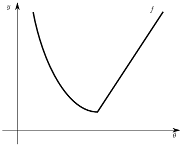


Non-convex : 2 local min. & 1 global min.

Convexity and minimum

Various types of behavior for convex functions

- global minimum *e.g.*, quadratic, etc.

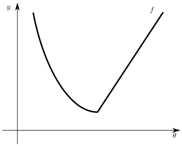


global minimum

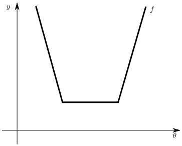
Convexity and minimum

Various types of behavior for convex functions

- global minimum *e.g.*, quadratic, etc.
- several minima *e.g.*, piecewise-affine (quadratic possible too !)



global minimum

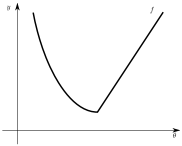


interval of minima

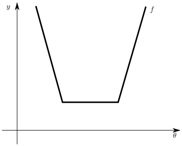
Convexity and minimum

Various types of behavior for convex functions

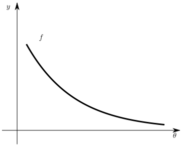
- global minimum *e.g.*, quadratic, etc.
- several minima *e.g.*, piecewise-affine (quadratic possible too !)
- no minimum, lower bounded *e.g.*, exponential function



global minimum



interval of minima

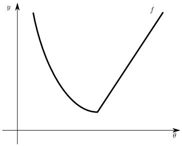


lower bounded

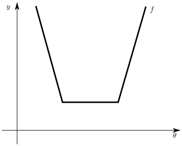
Convexity and minimum

Various types of behavior for convex functions

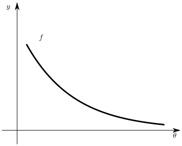
- global minimum *e.g.*, quadratic, etc.
- several minima *e.g.*, piecewise-affine (quadratic possible too !)
- no minimum, lower bounded *e.g.*, exponential function
- no minimum, lower bound is $-\infty$ *e.g.*, affine or $-\log(\cdot)$



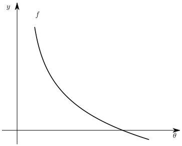
global minimum



interval of minima



lower bounded



not lower bounded

Overline

Gradient descent : intuition

- General formulation : minimize f (in $\mathbb{R}^p$) by finding iteratively a new point for which f has decreased the most
- First order approximation :

$$f(\theta) \approx f(\theta^0) + \langle \nabla f(\theta^0), \theta - \theta^0 \rangle$$

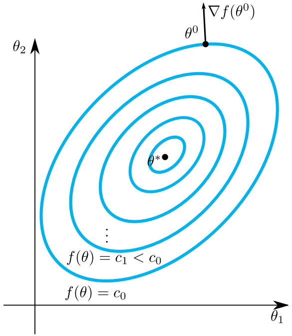
- Solution to decrease the most the function f around θ_0 (Cauchy-Schwartz) : "align" with the opposite direction to the gradient $\theta - \theta^0 = -\alpha \nabla f(\theta^0)$
- $\alpha > 0$ controls the "speed" with which one progresses in that direction.This parameter is called the **step size**

Gradient descent : algorithm

Data: initialization θ^0, max. iterations T, stopping criterion ε, step α
Result: for θ^T "close" to a minimum of f
for $1 \leqslant t \leqslant T$ **do**
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
  STOP if stopping criterion is smaller than ε
end

Possible stopping criterion :

- $\|\nabla f(\theta^t)\| \leqslant \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leqslant \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leqslant \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leqslant \varepsilon$
- duality gap (when easy to compute)

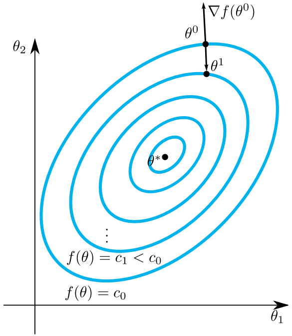


Gradient descent : algorithm

Data: initialization θ^0, max. iterations T, stopping criterion ε, step α
Result: for θ^T "close" to a minimum of f
for $1 \leqslant t \leqslant T$ **do**
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
  STOP if stopping criterion is smaller than ε
end

Possible stopping criterion :

- $\|\nabla f(\theta^t)\| \leqslant \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leqslant \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leqslant \varepsilon$ or $\frac{\|\theta^{t+1}-\theta^t\|}{\|\theta^t\|} \leqslant \varepsilon$
- duality gap (when easy to compute)

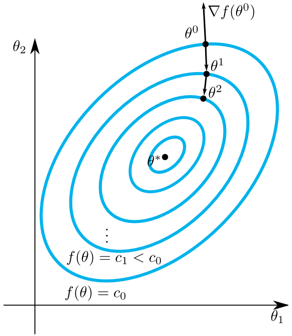


Gradient descent : algorithm

Data: initialization θ^0, max. iterations T, stopping criterion ε, step α
Result: for θ^T "close" to a minimum of f
for $1 \leqslant t \leqslant T$ **do**
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
  STOP if stopping criterion is smaller than ε
end

Possible stopping criterion :

- $\|\nabla f(\theta^t)\| \leqslant \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leqslant \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leqslant \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leqslant \varepsilon$
- duality gap (when easy to compute)

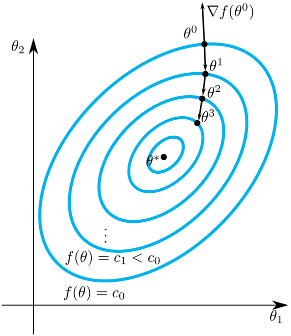


Gradient descent : algorithm

Data: initialization θ^0, max. iterations T, stopping criterion ε, step α
Result: for θ^T "close" to a minimum of f
for $1 \leqslant t \leqslant T$ **do**
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
  STOP if stopping criterion is smaller than ε
end

Possible stopping criterion :

- $\|\nabla f(\theta^t)\| \leqslant \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leqslant \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leqslant \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leqslant \varepsilon$
- duality gap (when easy to compute)

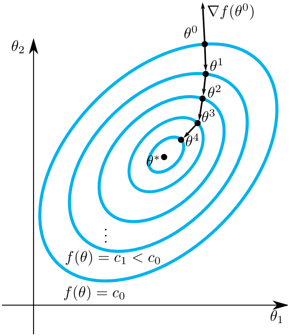


Gradient descent : algorithm

Data: initialization θ^0, max. iterations T, stopping criterion ε, step α
Result: for θ^T "close" to a minimum of f
for $1 \leqslant t \leqslant T$ **do**
    $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
    STOP if stopping criterion is smaller than ε
end

Possible stopping criterion :

- $\|\nabla f(\theta^t)\| \leqslant \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leqslant \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leqslant \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leqslant \varepsilon$
- duality gap (when easy to compute)

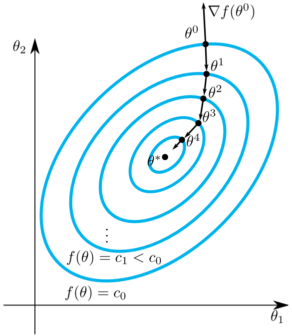


Gradient descent : algorithm

Data: initialization θ^0, max. iterations T, stopping criterion ε, step α
Result: for θ^T "close" to a minimum of f
for $1 \leqslant t \leqslant T$ **do**
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
  STOP if stopping criterion is smaller than ε
end

Possible stopping criterion :

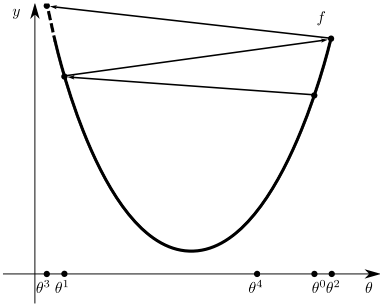
- $\|\nabla f(\theta^t)\| \leqslant \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leqslant \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leqslant \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leqslant \varepsilon$
- duality gap (when easy to compute)



Mind the step...size (1D case)

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}$$

α : crucial parameter to insure convergence toward a minimum

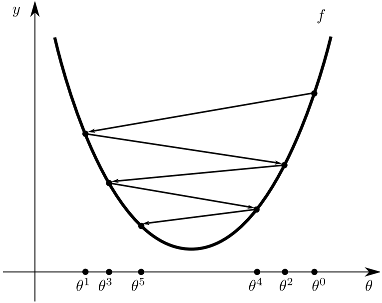


Divergence : really too large step size

Mind the step...size (1D case)

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}$$

α : crucial parameter to insure convergence toward a minimum

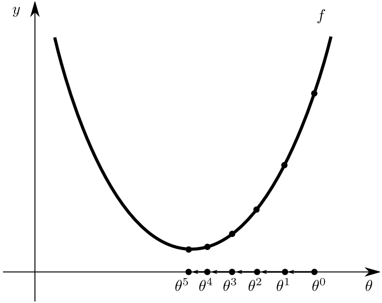


Slow convergence : still too large step size

Mind the step...size (1D case)

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}$$

α : crucial parameter to insure convergence toward a minimum

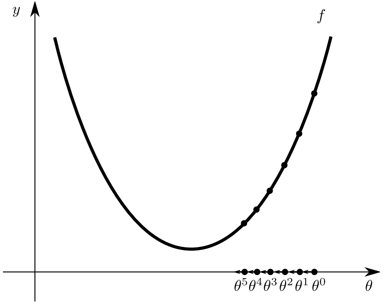


Fast convergence : good step size

Mind the step...size (1D case)

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}$$

α : crucial parameter to insure convergence toward a minimum

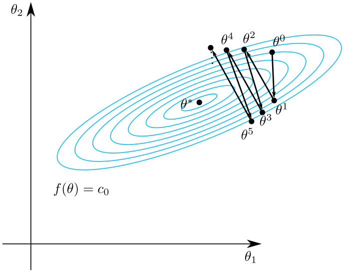


Slow convergence : too small step size

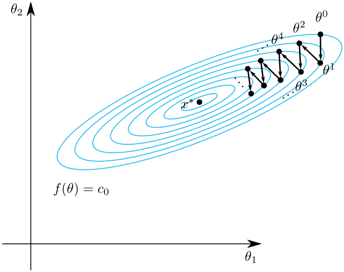
Mind the step...size (2D case)

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}$$

α : crucial parameter to insure convergence toward a minimum



Too large step



Too small step

Convergence : Lipschitz gradient

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}$$

Convergence rate for fixed step size

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (x, y), \quad \|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\|$$

Result : for any minimum $\theta^\star$ of f, if $\alpha \leqslant \frac{1}{L}$ then θ^T satisfies

$$f(\theta^T) - f(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha T}$$

- Faster : for better initialization, larger α, more steps !

Rem: if f is twice differentiable $\nabla^2 f(x) \preceq L \cdot Id$

Convergence : proof

Point 1 : gradient L-Lipschitz implies quadratic upper bound

$$\forall(\theta, \theta') \quad f(\theta) \leqslant f(\theta') + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{L}{2}\|\theta' - \theta\|^2$$

Point 2 : remind $\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)$ with Point 1

$$f(\theta^{t+1}) \leqslant f(\theta^t) - (1 - \frac{L\alpha}{2})\alpha\|\nabla f(\theta^t)\|^2$$

Point 3 : use convexity, $0 < \alpha \leqslant \frac{1}{L}$, $ab = (a^2 + b^2 - (a-b)^2)/2$ and the defintion of θ^{t+1}

$$\begin{aligned} f(\theta^{t+1}) \leqslant& f(\theta^\star) + \nabla f(\theta^t)^\top(\theta^t - \theta^\star) - \frac{\alpha}{2}\|\nabla f(\theta^t)\|^2 \\ =& f(\theta^\star) + \frac{1}{2\alpha}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2) \end{aligned}$$

Convergence proof (bis)

Point 4 : Telescopic sums

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\big(f(\theta^{t+1})-f(\theta^\star)\big) &\leqslant \frac{1}{T}\frac{1}{2\alpha}(\|\theta^0-\theta^\star\|^2-\|x^T-\theta^\star\|^2)\\ &\leqslant \frac{1}{2\alpha T}\|\theta^0-\theta^\star\|^2\end{aligned}$$

From Point 2, $f(\theta^{t+1}) \leqslant f(\theta^t)$, hence

$$f(\theta^{t+1})-f(x^\star) \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\big(f(\theta^{t+1})-f(\theta^\star)\big) \leqslant \frac{1}{2\alpha T}\|\theta^0-\theta^\star\|^2$$

Limits of convergence

- The convergence holds for $\alpha < 2/L$ (*cf.* Nesterov (2004) [p. 69])
- One needs to know the constant L, to find a correct (scaling) step size. It is not always known by the practitioner.
- A small constant step size is not the solution : it would lead to (very) slow convergence...

Example: $\theta \mapsto \frac{\|X\theta - y\|_2^2}{2}$ then $L = \lambda_{\max}(X^\top X)$ (spectral radius)

Line search

For faster convergence, it might be recommended to "optimize" the step size at each iteration, *i.e.*, α^t might evolve with iterations. Denote by $d^t = -\nabla f(\theta^t)$ the current (gradient) descent direction

Full line search optimization

Minimization of the amplitude, by solving the following 1D problem :

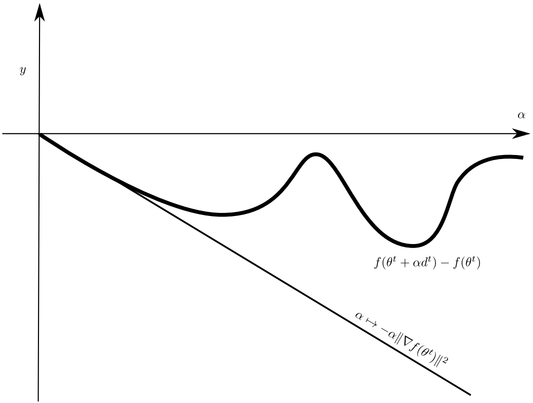
$$f(\theta^t + \alpha^t d^t) = \min_{\alpha \geqslant 0} f(\theta^t + \alpha d^t)$$

Rem: Need the 1D problem to be simple to solve.

Line search II

Armijo rule (or geometric backtracking)

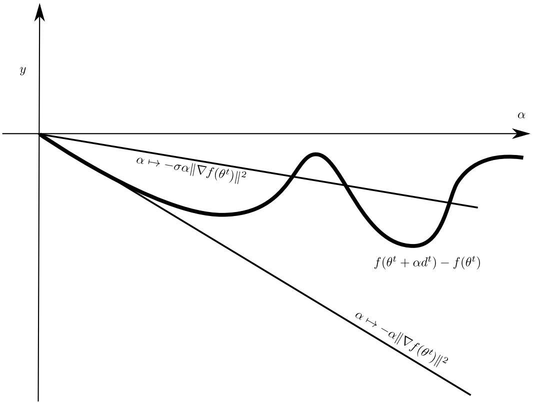
Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$: where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma \beta^m s \langle \nabla f(\theta^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(\theta^t)\|^2$



Line search II

Armijo rule (or geometric backtracking)

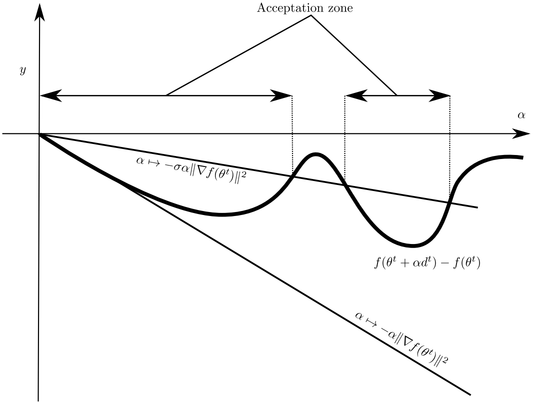
Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$: where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma \beta^m s \langle \nabla f(\theta^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(\theta^t)\|^2$



Line search II

Armijo rule (or geometric backtracking)

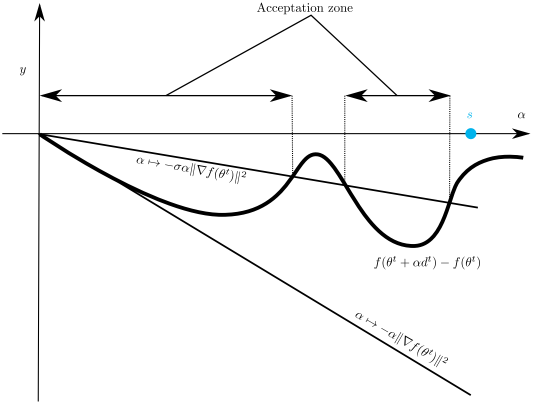
Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$: where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma \beta^m s \langle \nabla f(\theta^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(\theta^t)\|^2$



Line search II

Armijo rule (or geometric backtracking)

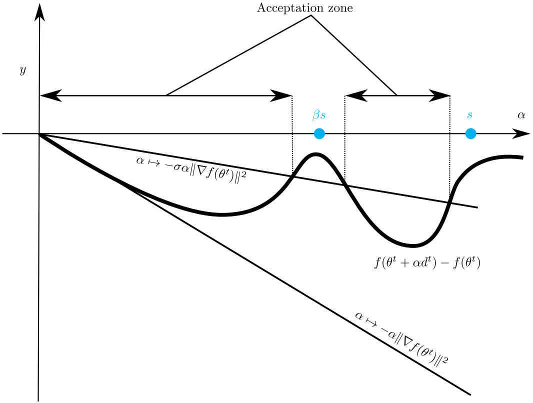
Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$:
where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma \beta^m s \langle \nabla f(\theta^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(\theta^t)\|^2$



Line search II

Armijo rule (or geometric backtracking)

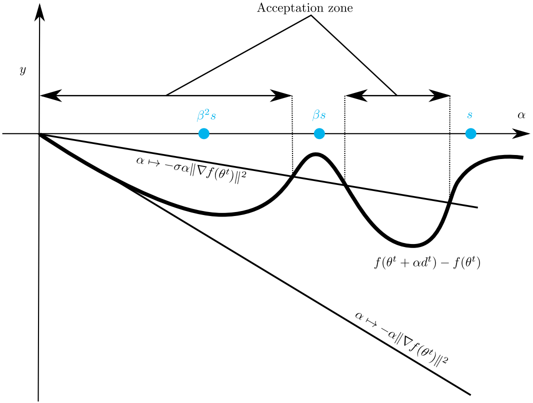
Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$:
where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma \beta^m s \langle \nabla f(\theta^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(\theta^t)\|^2$



Line search II

Armijo rule (or geometric backtracking)

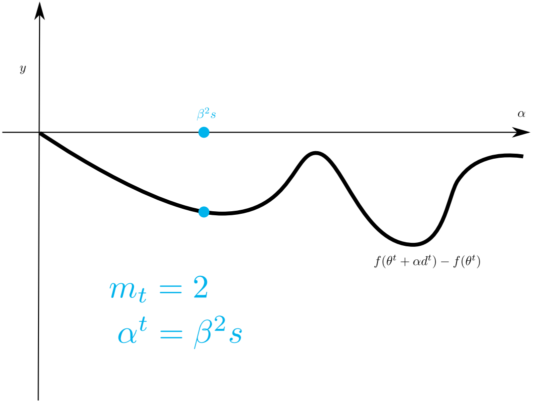
Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$:
where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma \beta^m s \langle \nabla f(\theta^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(\theta^t)\|^2$



Line search II

Armijo rule (or geometric backtracking)

Fix $s > 0$, $\sigma \in]0,1[$, and $\beta \in]0,1[$, need to choose $\alpha^t = \beta^{m_t} s$: where m_t is the first integer such that
$f(\theta^t + \beta^m s d^t) - f(\theta^t) \leqslant \sigma\beta^m s\langle\nabla f(\theta^t), d_t\rangle = -\sigma\beta^m s\|\nabla f(\theta^t)\|^2$



Line search III

Armijo's rule or geometric backtracking

In practice : *cf.* Bertsekas (1999)

- $s = 1$
- $\beta = 1/2$ or $\beta = 1/10$
- $\sigma \in [10^{-5}, 10^{-1}]$

Analysis of line search (L-Lipschitz gradient)

Properties of the Armijo rule

$$\alpha^t = s \text{ or } \alpha^t \in [2\beta(1-\sigma)/L, 2(1-\sigma)/L]$$

and so

$$\alpha_t \geqslant \min(s, 2\beta(1-\sigma)/L)$$

Proof : reminding Point 2, with $\theta^{t+1} = \theta^t - \alpha^t \nabla f(\theta^t)$:

$$f(\theta^{t+1}) \leqslant f(\theta^t) - (1 - \frac{L\alpha^t}{2})\alpha^t \|\nabla f(\theta^t)\|^2$$

so if $\alpha^t \leqslant 2(1-\sigma)/L$ then $f(\theta^{t+1}) \leqslant f(\theta^t) - \sigma\alpha^t \|\nabla f(\theta^t)\|^2$ and any value smaller than $2(1-\sigma)/L$ would be Armijo admissible.
By definition, the iteration is accepted if the previous was not : so $\beta^{m-1}s > 2(1-\sigma)/L$ and $\beta^m s \leqslant 2(1-\sigma)/L$

Rem: The Armijo prevent the step size to be too small

Convergence for the Armijo rule

$$\boxed{\theta^{t+1} = \theta^t - \alpha^t \nabla f(\theta^t)}$$

<u>Rem</u>: Choosing $\sigma \leqslant 1/2$, $\quad f(\theta^{t+1}) \leqslant f(\theta^t) - \sigma\alpha \|\nabla f(\theta^t)\|^2$ and the same proof works

Convergence rate

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta'), \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leqslant L \|\theta - \theta'\|$$

Result : for any minimum $\theta^\star$ of f then θ^T satisfies

$$f(\theta^T) - f(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2 \min(s, 2\beta(1-\sigma)/L)\, T}$$

<u>Rem</u>: Trade-off between more restricted zone (large β, small σ) and more computations (*i.e.*, more function evaluations)

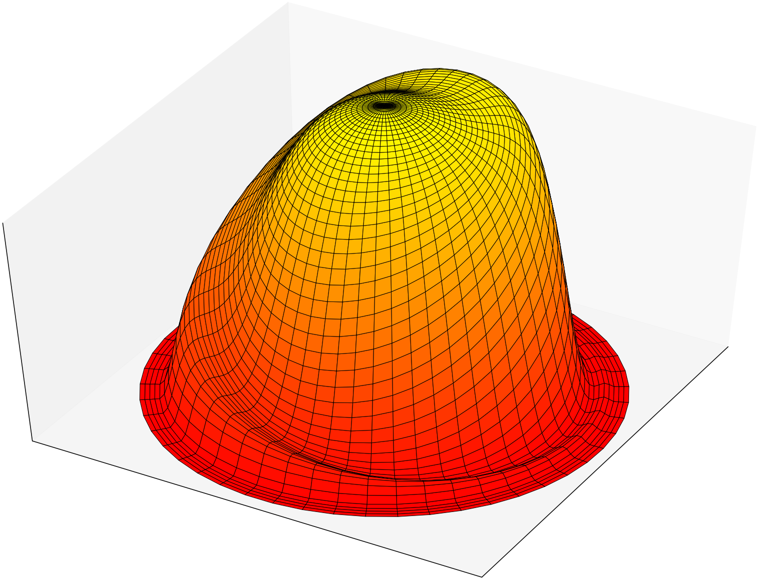
Convergence of the iterates

- The convergence of the iterates is not guaranteed for all smooth functions
- more convergence difficulties in infinite dimension spaces...
- One needs convexity for iterates convergence, otherwise counter-example Bertsekas (1999) or Absil *et al.* 2005 even for $\mathcal{C}^\infty$ functions

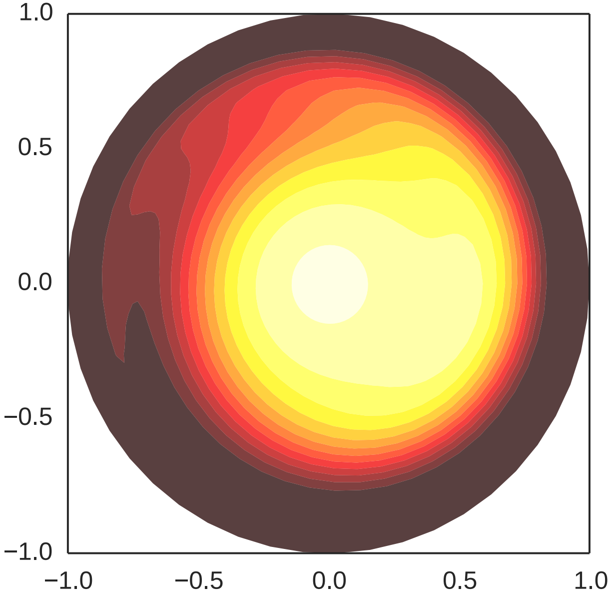
  Example: Mexican hat (in polar equation)

$$f(r,\theta) = \begin{cases} e^{-\frac{1}{1-r^2}}\left(1 - \frac{4r^4}{4r^4+(1-r^2)^2}\sin(\theta - \frac{1}{1-r^2})\right) & \text{if } r < 1 \\ 0 & \text{otherwise} \end{cases}$$

Counter example : spiraling toward zero



Counter example : spiraling toward zero



Analysis with strong-convexity

The following definition is not standard, but is taken from Hiriart-Urruty and Lemaréchal (1993), p. 280

Definition : strongly convex function

A convex function f is called μ-strongly convex if for all $\theta, \theta' \in \mathbb{R}^d$ the following (quadratic lower bound) holds true :

$$f(\theta) \geqslant f(\theta') + \langle s, \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|_2^2, \quad \forall s \in \partial f(\theta')$$

Rem: The standard definition is that $f - 1/2\mu |\cdot|^2$ is convex
Rem: if f is twice differentiable $\nabla^2 f(\theta) \succeq \mu \cdot Id$
Example: $\theta \mapsto \frac{\|X\theta - y\|_2^2}{2}$ then $\mu = \lambda_{\min}(X^\top X)$, and $\lambda_{\min}(X^\top X)/\lambda_{\max}(X^\top X)$ is the condition number of the matrix X

Strong-convexity + gradient Lipschitz

Property

Assume that f is closed, μ-strongly convex and has gradient L-Lipschitz, then f has a unique minimizer $\theta^\star$ satisfying :

$$\frac{\mu}{2}\|\theta - \theta^*\|_2^2 \leqslant f(\theta) - f(\theta^\star)$$

Corollary : control of gradient descent iterates

Under the same assumption with $\alpha \leqslant 1/L$, θ^T satisfies

$$\|\theta^T - \theta^\star\|_2^2 \leqslant \frac{1}{\alpha\mu T}\|\theta^0 - \theta^\star\|_2^2$$

<u>Rem</u>: if $\alpha = 1/L$ the constant factor is L/μ (condition number)
<u>Rem</u>: Even geometric convergence rate Nesterov (2004) [p.70] :

$$\|\theta^T - \theta^\star\|_2^2 \leqslant \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^T \|\theta^0 - \theta^\star\|_2^2 \quad (\text{ for } \alpha = \frac{2}{\mu + L})$$

Overline

Composite minimization

One aims at minimizing :

$$F = f + g$$

- f smooth : ∇f is L-Lipschitz
- g proximable (prox-capable) : prox_g can be "efficiently" computed, where

$$\boxed{\mathrm{prox}_g(y) = \arg\min_{z \in \mathbb{R}^d} \left(\frac{1}{2}\|z - y\|_2^2 + g(z) \right)}$$

Rem: g might be non-smooth in this formulation
More details on "prox" properties in Parikh and Boyd (2013)

Examples of proximity operators

$$\boxed{\operatorname{prox}_g(y) = \underset{z\in\mathbb{R}^d}{\arg\min}\left(\frac{1}{2}\|z-y\|_2^2 + g(z)\right)}$$

- Null function : if $g = 0$, then $\operatorname{prox}_g = \mathrm{Id}$
- Smooth function ∇g exists :

$$\operatorname{prox}_g(y) = (\mathrm{Id} + \nabla g)^{-1}(y)$$

- Indicator function : $g = \iota_C$ for a closed convex set $C \subset \mathbb{R}^p$,

$$\operatorname{prox}_g(y) = \pi_C(y), \quad \text{projection over the set } C$$

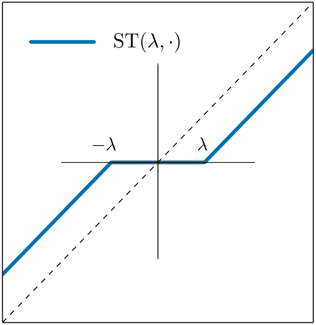
Examples of proximity operators (II) : Soft-Thresholding

Case where $g(x) = \lambda |x|$ (absolute value)

$$\begin{aligned}
\operatorname{prox}_g(y) &= \mathrm{ST}(\lambda, y) \\
&= \arg\min_{\beta \in \mathbb{R}} \Big(\frac{(y-\beta)^2}{2} + \lambda|\beta| \Big) \\
&= \operatorname{sign}(y) \cdot (|y| - \lambda)_+
\end{aligned}$$

with $(\cdot)_+ = \max(0, \cdot)$

Proof : use sub-gradients of $|\cdot|$ and Fermat condition



Rem: Any $|y| > \lambda$, is shrinked toward zero by a factor λ ; **bias !**

Forward-Backward algorithm

Notation : $\boxed{\phi_\alpha(\theta) := \operatorname{prox}_{\alpha g}\left(\theta - \alpha \nabla f(\theta)\right)}$

Forward-Backward algorithm (for minimizing $F = f + g$) :

Input: Initialization θ^0, step size α
Result: θ^T
while *not converged* **do**
$\quad \theta^{t+1} = \phi_\alpha(\theta^t)$
end

Rem: Link with majorization-minimization techniques

$$\phi_\alpha(\theta) = \arg\min_{\theta'} \left(f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{1}{2\alpha} \|\theta' - \theta\|^2 + g(\theta') \right)$$

Rem: Often referred to as "Iteratives Soft-Thresholding Algorithm"

Convergence : f gradient Lipschitz

$$\boxed{\theta^{t+1} = \phi_\alpha(\theta^t) = \mathrm{prox}_{\alpha g}(\theta^t - \alpha \nabla f(\theta^t))}$$

Convergence rate for fixed step size

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta'), \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leqslant L \|\theta - \theta'\|$$

Result : for any minimum $\theta^\star$ of F, if $\alpha \leqslant \frac{1}{L}$ then θ^T satisfies

$$F(\theta^T) - F(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha T}$$

Rem: same bound as in the case with $g \equiv 0$

Proof :

Point 1 : for $\alpha > 0$ and $\hat{x} = \phi_\alpha(\bar{x})$ then for all y :

$$\boxed{F(\hat{x}) + \frac{\|\hat{x} - y\|_2^2}{2\alpha} \leqslant F(y) + \frac{\|\bar{x} - y\|_2^2}{2\alpha}}$$

Proof : $H_\alpha(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + \frac{1}{2\alpha}\|y - \bar{x}\|^2 + g(y)$
H_α is $1/\alpha$-strongly convex and $H(\cdot) - 1/(2\alpha)\|\cdot\|_2^2$ is convex
(*cf.* page 280, Hiriart-Urruty and Lemaréchal (1993))

$$\hat{x} = \arg\min_y H_\alpha(y) \quad (\textit{cf. } \text{two slides up})$$

Remind that $0 \in \partial H_\alpha(\hat{x})$ and apply the definition of $1/\alpha$-strong convexity to y and $\hat{x}$:

$$\forall y, H_\alpha(\hat{x}) + 1/(2\alpha)\|\hat{x} - y\|_2^2 \leqslant H_\alpha(y)$$

Proof continued (Point 1)

This means :

$$g(\hat{x}) + f(\bar{x}) + \langle \nabla f(\bar{x}), \hat{x} - \bar{x} \rangle + \frac{1}{2\alpha}(\|\hat{x} - \bar{x}\|^2 + \|\hat{x} - y\|^2) \leqslant$$
$$g(y) + f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle + \frac{1}{2\alpha}\|y - \bar{x}\|^2$$

By convexity of f :

$$f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle \leqslant f(y)$$

and by the choice $\alpha \leqslant 1/L$ the following bound holds :

$$f(\hat{x}) \leqslant f(\bar{x}) + \langle \nabla f(\bar{x}), \hat{x} - \bar{x} \rangle + \frac{1}{2\alpha}\|\hat{x} - \bar{x}\|^2$$

So Point 1 holds : $\boxed{F(\hat{x}) + \frac{1}{2\alpha}\|\hat{x} - y\|^2 \leqslant F(y) + \frac{1}{2\alpha}\|y - \bar{x}\|^2}$

Last part of the proof

Point 1 states : $F(\hat{x}) + \frac{1}{2\alpha}\|\hat{x} - y\|^2 \leqslant F(y) + \frac{1}{2\alpha}\|y - \bar{x}\|^2$,

$$\text{Choosing :} \begin{cases} y = & \theta^\star \quad \text{(any minimizer of } F\text{)} \\ \bar{x} = & \theta^t \\ \hat{x} = & \theta^{t+1} \end{cases}$$

$$\text{Yields} \quad F(\theta^{t+1}) + \frac{1}{2\alpha}\|\theta^{t+1} - \theta^\star\|^2 \leqslant F(\theta^\star) + \frac{1}{2\alpha}\|\theta^\star - \theta^t\|^2$$

This leads to Point 3 of the smooth case :

$$F(\theta^{t+1}) \leqslant F(\theta^\star) + \frac{1}{2\alpha}(\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2)$$

and a telescopic argument provides to the desired bound.

Overline

Forward-Backward accelerated algorithm

Notation : $\boxed{\phi_\alpha(\theta) := \operatorname{prox}_{\alpha g}\left(\theta - \alpha \nabla f(\theta)\right)}$

Forward-Backward algorithm

Input: Initialization θ^0, step size α, a sequence $(\mu_t)_{t\in\mathbb{N}}$ satisfying :$\mu_1 = 1$ and $\mu_{t+1}^2 - \mu_{t+1} \leqslant \mu_t^2$

Result: θ^T

while *not converged* **do**

$$\theta^{t+1} = \phi_\alpha(z^t)$$
$$z^{t+1} = \theta^{t+1} + \frac{\mu_{t+1}-1}{\mu_{t+2}}(\theta^{t+1} - \theta^t)$$

end

Examples of admissible sequences :

- $\mu_{t+1} = \sqrt{\mu_t^2 + 1/4} + 1/2$ (*i.e.*, $\mu_{t+1}^2 - \mu_{t+1} = \mu_t^2$)
- $\mu_{t+1} = (t+1)/2$
- $\mu_{t+1} = (t+a-1)/a$, with $a > 2$

Convergence : Lipschitz gradient

$$\boxed{\theta^{t+1} = \phi_\alpha(\theta^t) = \mathrm{prox}_{\alpha g}(\theta^t - \alpha \nabla f(\theta^t))}$$

Convergence rate for fixed step size

Hypothesis : f convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta'), \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leqslant L \|\theta - \theta'\|$$

Result : for any minimum $\theta^\star$ of F, if $\alpha \leqslant \frac{1}{L}$ then θ^T satisfies

$$F(\theta^T) - F(\theta^\star) \leqslant \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha\mu_T^2}$$

Rem: for common choices given above $\mu_t \approx t$, so the rate is $O(1/t^2)$, better than $O(1/t)$ (without acceleration)
Rem: define $F^* = F(\theta^\star)$ for the proof

Proof : rate for the Nesterov acceleration

Point 1 with $\hat{x} = \phi_\alpha(z^t)$, $\bar{x} = z^t$, $y = (1 - 1/\mu_{t+1})\theta^t + 1/\mu_{t+1} \cdot \theta^\star$

$$\boxed{F(\hat{x}) + \frac{\|\hat{x} - y\|_2^2}{2\alpha} \leqslant F(y) + \frac{\|\bar{x} - y\|_2^2}{2\alpha}}$$

with $u^{t+1} = \theta^t + \mu_{t+1}(\theta^{t+1} - \theta^t)$ and a little algebra gives :

$$\begin{aligned} F(\theta^{t+1}) + \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} &\leqslant F(y) + \frac{\|u^t - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} \\ F(\theta^{t+1}) - F^* - (1 - \frac{1}{\mu_{t+1}})(F(\theta^t) - F^*) &\leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha\mu_{t+1}^2} \\ \mu_{t+1}^2 \Delta F_{t+1}^* - (\mu_{t+1}^2 - \mu_{t+1})(\Delta F_t^*) &\leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha} \end{aligned}$$

(convexity of F and $\Delta F_{t+1}^* = F(\theta^{t+1}) - F^*$)

Proof continued

Define $\rho_{t+1} := \mu_{t+1} - \mu_{t+1}^2 + \mu_t^2 \geqslant 0$ so

$$\mu_{t+1}^2 \Delta F_{t+1}^* - (\mu_{t+1}^2 - \mu_{t+1})(\Delta F_t^*) \leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha}$$

$$\mu_{t+1}^2 \Delta F_{t+1}^* - \mu_t^2 \Delta F_t^* + \rho_{t+1} \Delta F_t^* \leqslant \frac{\|u^t - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^{t+1} - \theta^\star\|_2^2}{2\alpha}$$

Telescopic terms again (convention $\mu_0 = 0$ and $u_0 = x_0 = x_{-1}$)

$$\mu_T^2 \Delta F_T^* + \sum_{t=0}^{T} \rho_{t+1} \Delta F_t^* \leqslant \frac{\|u^0 - \theta^\star\|_2^2}{2\alpha} - \frac{\|u^T - \theta^\star\|_2^2}{2\alpha}$$

$$\mu_T^2 \Delta F_T^* \leqslant \frac{\|u^0 - \theta^\star\|_2^2}{2\alpha}$$

Convergence of the iterates

Very recent result : Chambolle and Dossal 2014 Proof out of the scope of this course
More reading on the previous theme :

- Nesterov (2004) for proofs, strong convexity, etc.
- Beck and Teboulle (2009) for ISTA/FISTA analysis
- Chambolle and Dossal (2014) for FISTA with larger choice of updating rules

Overline

Fenchel Duality for stopping criterion

F objective function, fix $\varepsilon > 0$ small, and stop when

$$\frac{F(\theta^{t+1}) - F(\theta_t)}{F(\theta^t)} \leqslant \varepsilon \text{ or } \nabla F(\theta^t) \leqslant \varepsilon$$

Alternative : leverage the **duality gap**

<u>Notation</u> : $\boxed{F(\theta) = f(X\theta) + g(\theta)}$ with $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^p \to \mathbb{R}$ and X : $n \times p$ matrix.

Fenchel-Duality

Consider the problem $\min_\theta F(\theta)$, then the following holds

$$\sup_u \{-f^*(u) - g^*(-X^\top u)\} \leqslant \inf_\theta \{f(X\theta) + g(\theta)\}$$

Moreover, if f and g are **convex**, then under mild assumptions, equality of both sides holds (**strong duality**, no **duality gap**)

proof : use Fenchel-Young inequality

Fenchel Duality

We denote by

- $\theta^\star$: primal optimal solution of $\inf_\theta\{f(X\theta) + g(\theta)\}$
- u^* : dual solution of $\sup_u\{-f^*(u) - g^*(-X^\top u)\}$

Define the **duality gap** by :

$$\Delta(\theta, u) = F(\theta) + f^*(u) + g^*(-X^\top u)$$

Property of the duality gap

$$\forall \theta, u, \quad \Delta(\theta, u) \geqslant F(\theta) - F(\theta^\star) \geqslant 0$$

proof : Fenchel-duality applied to a primal solution $\theta^\star$

Motivation for stopping criterion : $\boxed{\Delta(\theta, u) \leqslant \varepsilon \Rightarrow F(\theta) - F(\theta^\star) \leqslant \varepsilon}$

Example : Duality gap for the Lasso

Lasso objective : $\boxed{F(\theta) = \frac{1}{2}\|X\theta - y\|_2^2 + \lambda\|\theta\|_1}$

- $f(z) = \frac{1}{2}\|z - y\|_2^2; f^*(u) = \frac{1}{2}\|u\|_2^2 + \langle u, y\rangle$ (translation prop.)
- $g(\theta) = \lambda\|\theta\|_1; g^*(u) = \iota_{\{u, \|u\|_\infty \leqslant \lambda\}}$ (ℓ_∞ ball indicator)

- Duality gap : $$\begin{aligned}\Delta(\theta, u) =& F(\theta) + f^*(u) + g^*(-X^\top u)\\ =& F(\theta) + \frac{1}{2}\|u\|_2^2 + \langle u, y\rangle\end{aligned}$$

  as soon as $\|X^\top u\|_\infty \leqslant \lambda$, otherwise the bound is $+\infty$: useless

<u>Rem</u>: at optimum solutions and under mild assumptions
$\Delta(\theta^\star, u^*) = 0$

Example : Duality gap for the Lasso (II)

Possible choice :

- θ_t (current iterate of any iterative algorithm),
- $r_t = X\theta_k - y$ (minus current residuals)
- $u_t = \mu_t r_t$ with $\mu_t = \min(1, \lambda/\|X^\top r_t\|_\infty)$

Motivation for this choice : at optimum $u^* = \nabla f(X\theta^\star)$

Stopping criterion :

$$\begin{aligned} &\frac{1}{2}\|r_t\|_2^2 + \lambda\|\theta_t\|_1 + \frac{1}{2}\|u_t\|_2^2 + \langle u_t, y\rangle \leqslant \varepsilon \\ \Leftrightarrow\ &\frac{1}{2}(1+\mu_t^2)\|r_t\|_2^2 + \lambda\|\theta_t\|_1 + \mu_t\langle r_t, y\rangle \leqslant \varepsilon \end{aligned}$$

Références I

- P. Absil, R. Mahony, and B. Andrews.
  Convergence of the iterates of descent methods for analytic cost functions.
  SIAM Journal on Optimization, 16(2) :531–547, 2005.

- D. P. Bertsekas.
  Nonlinear programming.
  Athena Scientific, 1999.

- A. Beck and M. Teboulle.
  A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
  SIAM J. Imaging Sci., 2(1) :183–202, 2009.

- A. Chambolle and C. Dossal.
  How to make sure the iterates of fista converge.
  2014.

- J.-B. Hiriart-Urruty and C. Lemaréchal.
  Convex analysis and minimization algorithms. I, volume 305.
  Springer-Verlag, Berlin, 1993.

Références II

- Y. Nesterov.

  Introductory lectures on convex optimization, volume 87 of *Applied Optimization*.

  Kluwer Academic Publishers, Boston, MA, 2004.

- N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein.

  Proximal algorithms.

  Foundations and Trends in Machine Learning, 1(3) :1–108, 2013.