STAT 593
Robust statistics:
Gradient Descent

Joseph Salmon

http://josephsalmon.eu

Télécom Paristech, Institut Mines-Télécom

&

University of Washington, Department of Statistics

(Visiting Assistant Professor)

Outline

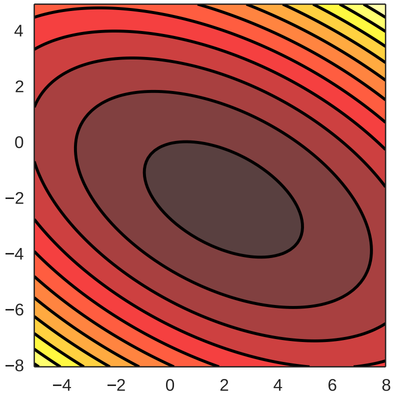
Reminder

Convexity for optimization

Gradient descent

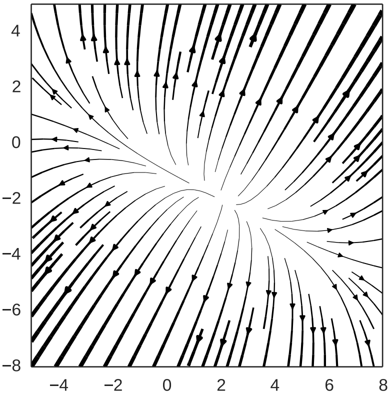
Level lines / gradient flow

Level set of a (quadratic) function



Level lines / gradient flow

Gradient flow of the same function



Level lines / gradient flow

Level set and gradient flow of the same function

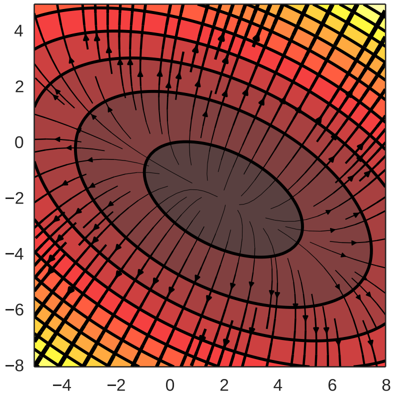


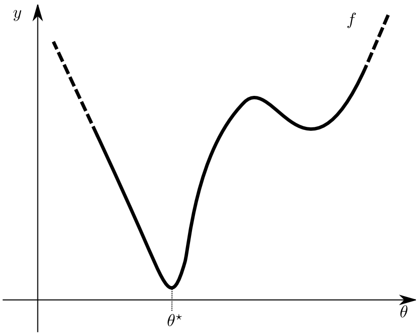
Table of Contents

Existence of a minimum

Theorem

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

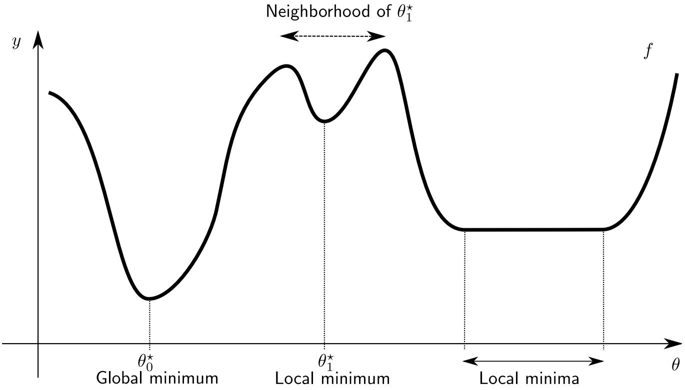
$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$$



Local vs global minima

Definition: local minimum

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



Rem: a global minimum is also a local minimum

Convex case: local = global

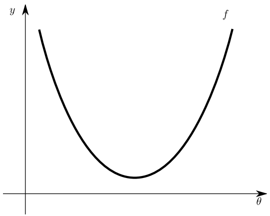
Theorem

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

Convex case: local = global

Theorem

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

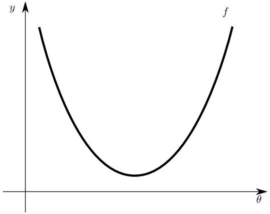


Convex: 1 global minimum

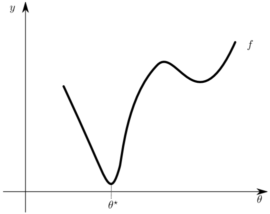
Convex case: local = global

Theorem

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



Convex: 1 global minimum



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

Table of Contents

Convex functions and tangents

Theorem

For a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then for any $(\theta^\star, \theta) \in \mathbb{R}^d \times \mathbb{R}^d$, the following holds:

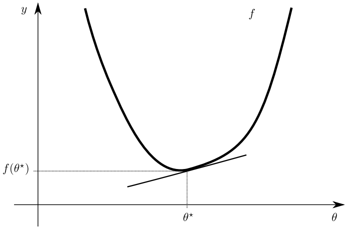
$$f(\theta) \geq f(\theta^\star) + \langle \nabla f(\theta^\star), \theta - \theta^\star \rangle$$

Convex functions and tangents

Theorem

For a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then for any $(\theta^\star, \theta) \in \mathbb{R}^d \times \mathbb{R}^d$, the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle \nabla f(\theta^\star), \theta - \theta^\star \rangle$$

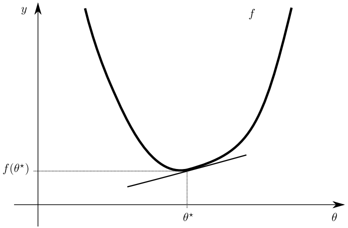


Convex functions and tangents

Theorem

For a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then for any $(\theta^\star, \theta) \in \mathbb{R}^d \times \mathbb{R}^d$, the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle \nabla f(\theta^\star), \theta - \theta^\star \rangle$$



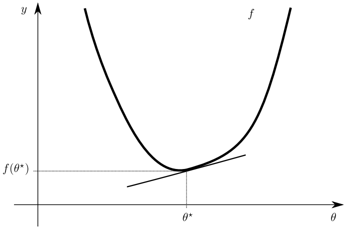
Proof: for $\theta^\star, \theta$ and α, define
$\theta_\alpha = \alpha\theta^\star + (1-\alpha)\theta$

Convex functions and tangents

Theorem

For a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then for any $(\theta^\star, \theta) \in \mathbb{R}^d \times \mathbb{R}^d$, the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle \nabla f(\theta^\star), \theta - \theta^\star \rangle$$



Proof: for $\theta^\star, \theta$ and α, define

$\theta_\alpha = \alpha\theta^\star + (1-\alpha)\theta$

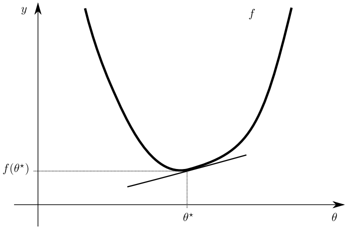
$f(\theta) \geq \frac{1}{1-\alpha}[f(\theta_\alpha) - \alpha f(\theta^\star)]$

Convex functions and tangents

Theorem

For a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then for any $(\theta^\star, \theta) \in \mathbb{R}^d \times \mathbb{R}^d$, the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle \nabla f(\theta^\star), \theta - \theta^\star \rangle$$



Proof: for $\theta^\star, \theta$ and α, define

$$\theta_\alpha = \alpha\theta^\star + (1-\alpha)\theta$$

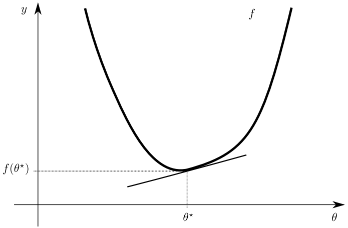
$$\begin{aligned} f(\theta) &\geq \tfrac{1}{1-\alpha}[f(\theta_\alpha) - \alpha f(\theta^\star)] \\ &= f(\theta^\star) + \tfrac{1}{1-\alpha}[f(\theta_\alpha) - f(\theta^\star)] \end{aligned}$$

Convex functions and tangents

Theorem

For a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then for any $(\theta^\star, \theta) \in \mathbb{R}^d \times \mathbb{R}^d$, the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle \nabla f(\theta^\star), \theta - \theta^\star \rangle$$



Proof: for $\theta^\star, \theta$ and α, define

$$\theta_\alpha = \alpha\theta^\star + (1-\alpha)\theta$$

$$\begin{aligned} f(\theta) &\geq \tfrac{1}{1-\alpha}[f(\theta_\alpha) - \alpha f(\theta^\star)] \\ &= f(\theta^\star) + \tfrac{1}{1-\alpha}[f(\theta_\alpha) - f(\theta^\star)] \\ &\underset{\alpha \to 1}{\to} f(\theta^\star) + \langle f'(\theta^\star), \theta - \theta^\star \rangle \end{aligned}$$

Sub-gradients / sub-differential

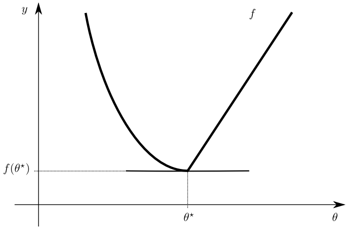
Definition

For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, $u \in \mathbb{R}^d$ is a **sub-gradient** of f at $\theta^\star$, if for any $\theta \in \mathbb{R}^d$ the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle$$

The **sub-differential** is the set of all sub-gradients:

$$\partial f(\theta^\star) = \{u \in \mathbb{R}^d : \forall \theta \in \mathbb{R}^d, f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle\}$$



Sub-gradients / sub-differential

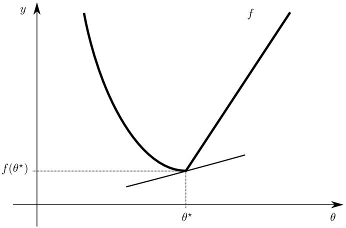
Definition

For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, $u \in \mathbb{R}^d$ is a **sub-gradient** of f at $\theta^\star$, if for any $\theta \in \mathbb{R}^d$ the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle$$

The **sub-differential** is the set of all sub-gradients:

$$\partial f(\theta^\star) = \{u \in \mathbb{R}^d : \forall \theta \in \mathbb{R}^d, f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle\}$$



Sub-gradients / sub-differential

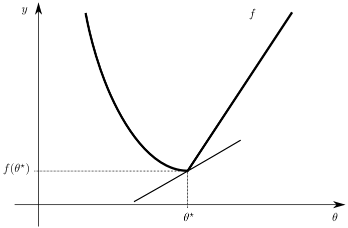
Definition

For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, $u \in \mathbb{R}^d$ is a **sub-gradient** of f at $\theta^\star$, if for any $\theta \in \mathbb{R}^d$ the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle$$

The **sub-differential** is the set of all sub-gradients:

$$\partial f(\theta^\star) = \{u \in \mathbb{R}^d : \forall \theta \in \mathbb{R}^d, f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle\}$$



Sub-gradients / sub-differential

Definition

For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, $u \in \mathbb{R}^d$ is a **sub-gradient** of f at $\theta^\star$, if for any $\theta \in \mathbb{R}^d$ the following holds:

$$f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle$$

The **sub-differential** is the <u>set</u> of all sub-gradients:

$$\partial f(\theta^\star) = \{u \in \mathbb{R}^d : \forall \theta \in \mathbb{R}^d, f(\theta) \geq f(\theta^\star) + \langle u, \theta - \theta^\star \rangle\}$$

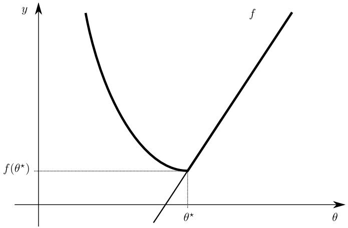


Table of Contents

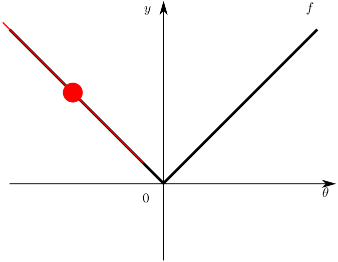
Sub-differential for the absolute value

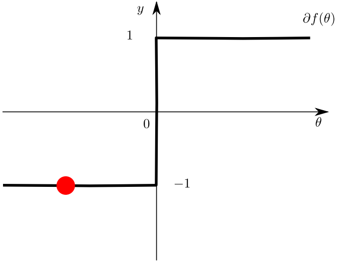
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$

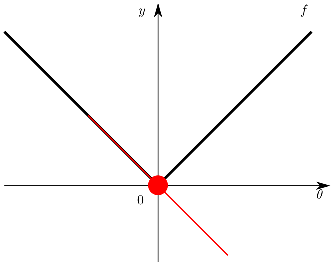




Sub-differential for the absolute value

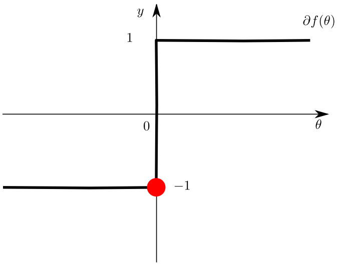
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$



Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$



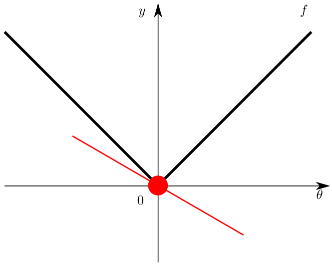
Sub-differential for the absolute value

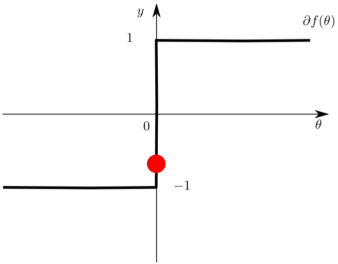
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$





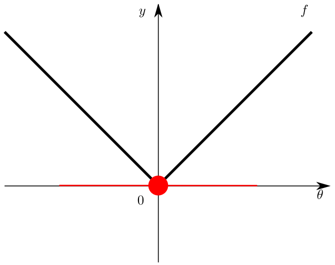
Sub-differential for the absolute value

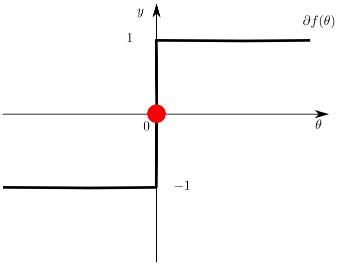
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$





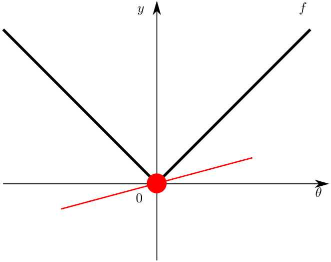
Sub-differential for the absolute value

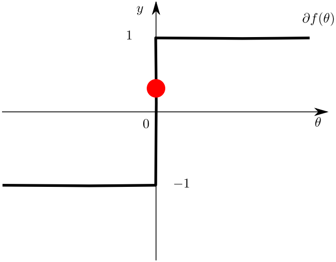
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$





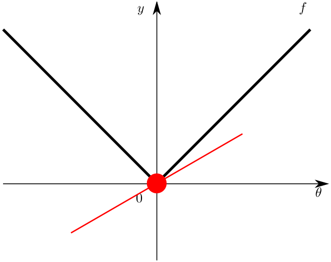
Sub-differential for the absolute value

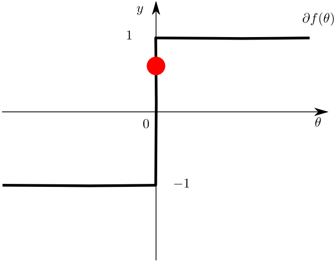
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$





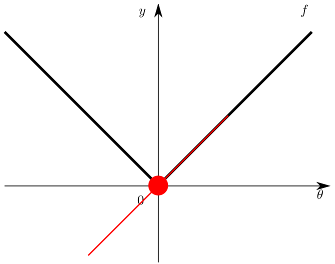
Sub-differential for the absolute value

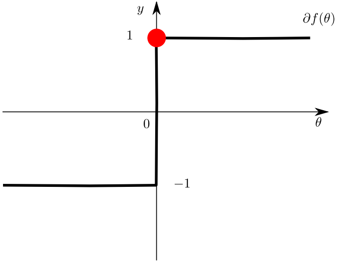
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$





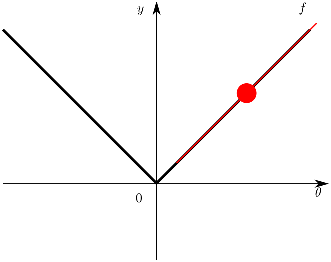
Sub-differential for the absolute value

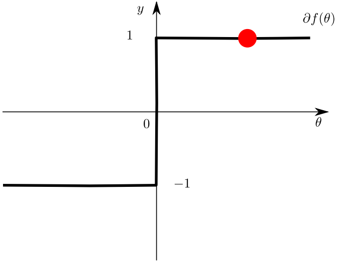
Function: abs

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ \theta & \mapsto |\theta| \end{cases}$$

Sub-differential: sign

$$\partial f(\theta) = \begin{cases} \{-1\} & \text{if } \theta \in]-\infty, 0[\\ \{1\} & \text{if } \theta \in]0, +\infty[\\ [-1, 1] & \text{if } \theta = 0 \end{cases}$$





Properties

- when the function f has a gradient at θ, its sub-differential is a singleton reduced to the standard gradient, *i.e.,*:

$$\partial f(\theta) = \{\nabla f(\theta)\}$$

- **Separable function**: for $f(x_1, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j)$,

$$\partial f(x_1, \ldots, x_p) = \partial f_1(x_1) \times \cdots \times \partial f_p(x_p)$$

- existence can be tricky but is ok for standard convex (continuous) function[1].

[1] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. New York: Springer, 2011, pp. xvi+468.

More properties[2]

For a convex function, the following holds true:

$$\partial f(\theta) = \left\{ s \in \mathbb{R}^d : \langle s, u \rangle \leq \lim_{t \to 0_+} \frac{f(\theta + tu) - f(\theta)}{t}, \text{ for all } u \in \mathbb{R}^d \right\}$$

- (Positive combinations) For any $t_1, t_2 > 0$, any function f_1, f_2, and any $\theta \in \mathbb{R}^d$, then
$$\partial(t_1 f_1 + t_2 f_2)(\theta) = t_1 \partial f_1(\theta) + t_2 \partial f_2(\theta)$$
- (Linear pre-composition) For any matrix A, any $\theta \in \mathbb{R}^d$ and any function f, then
$$\partial(f \circ A)(\theta) = A^\top \partial f(A\theta)$$

[2] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. I*. Vol. 305. Berlin: Springer-Verlag, 1993.

ℓ_1-prox : soft thresholding

$$x^* \in \arg\min_{x\in\mathbb{R}} f_{\lambda,z}(x) \iff 0 \in \partial f_{\lambda,z}(x^*) \text{ for } f_{\lambda,z}(x) = \frac{(z-x)^2}{2} + \lambda|x|$$

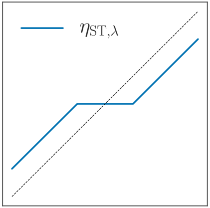
$$\iff 0 \in \partial f_{\lambda,z}(x^*) = x^* - z + \lambda \underbrace{\partial|\cdot|(x^*)}_{\mathrm{sign}(x^*)}$$

$$\iff x^* \in z - \lambda\,\mathrm{sign}(x^*)$$

Considering the cases $x^* > 0, x^* = 0, x^* < 0$, this leads to

$$x^* = \begin{cases} 0 & \text{si } |z| \leq \lambda \\ z - \lambda & \text{si } \ z \ \geq \lambda \\ z + \lambda & \text{si } \ z \ \leq -\lambda \end{cases}$$

$$\boxed{x^* := \eta_{\mathrm{ST},\lambda}(z) = \mathrm{sign}(z)(|z| - \lambda)_+}$$



Soft thresholding through sub-gradients (vector case)

$x^* \in \underset{x \in \mathbb{R}^p}{\arg\min} f_{\lambda,z}(x)$ for $f_{\lambda,z}(x) = \frac{\|z-x\|^2}{2} + \lambda \|x\|_1$ can be written:

$$\boxed{x^* := \eta_{\mathrm{ST},\lambda}(z)}$$

i.e., on can apply the soft-thresholding component wise:

$$\forall j \in [p], \quad x_j^* = \eta_{\mathrm{ST},\lambda}(z_j) := \mathrm{sign}(z_j)(|z_j| - \lambda)_+$$

proof: use separability of $\|x\|_1 = \sum_{j=1}^{p} |x_j|$

Median with sub-gradients

Definition

Median : $$\mathrm{Med}_n(\mathbf{x}) \in \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} |\theta - x_i| := \|\theta \mathbf{1}_n - \mathbf{x}\|_1 = f(\theta)$$

In practice, one can use the following:

Property

Any median $\mathrm{Med}_n(\mathbf{x})$ satisfies:

$$\#\{i \in [n] : x_i < \mathrm{Med}_n(\mathbf{x})\} \leq \#\{i \in [n] : x_i \geq \mathrm{Med}_n(\mathbf{x})\}$$
$$\#\{i \in [n] : x_i > \mathrm{Med}_n(\mathbf{x})\} \leq \#\{i \in [n] : x_i \leq \mathrm{Med}_n(\mathbf{x})\}$$

Proof

$$
\begin{aligned}
0 \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}} \partial f \cdot (\theta) &\iff 0 \in \mathbf{1}_n^\top \partial \left\|\cdot\right\|_1 (\theta \mathbf{1}_n - \mathbf{x}) \\
&\iff 0 \in \sum_{\theta > x_i} 1 - \sum_{\theta < x_i} 1 + \sum_{x_i = \theta} \operatorname{sign}(\theta - x_i) \\
&\iff 0 \in \#\{\theta > x_i\} - \#\{\theta < x_i\} \\
&\qquad + \sum_{x_i = \theta} \operatorname{sign}(\theta - x_i)
\end{aligned}
$$

Proof

$$
\begin{aligned}
0 \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}} \partial f \cdot(\theta) &\iff 0 \in \mathbf{1}_n^\top \partial \left\|\cdot\right\|_1 (\theta \mathbf{1}_n - \mathbf{x}) \\
&\iff 0 \in \sum_{\theta > x_i} 1 - \sum_{\theta < x_i} 1 + \sum_{x_i = \theta} \operatorname{sign}(\theta - x_i) \\
&\iff 0 \in \#\{\theta > x_i\} - \#\{\theta < x_i\} \\
&\qquad + \sum_{x_i=\theta} \operatorname{sign}(\theta - x_i)
\end{aligned}
$$

Hence, by upper bounding sign by 1:

$$
\#\{\theta < x_i\} \leq \#\{\theta > x_i\} + \sum_{x_i=\theta} 1 = \#\{\theta \geq x_i\}.
$$

Proof

$$
\begin{aligned}
0 \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}} \partial f \cdot(\theta) &\iff 0 \in \mathbf{1}_n^\top \partial \left\|\cdot\right\|_1 (\theta \mathbf{1}_n - \mathbf{x}) \\
&\iff 0 \in \sum_{\theta > x_i} 1 - \sum_{\theta < x_i} 1 + \sum_{x_i = \theta} \operatorname{sign}(\theta - x_i) \\
&\iff 0 \in \#\{\theta > x_i\} - \#\{\theta < x_i\} \\
&\qquad + \sum_{x_i = \theta} \operatorname{sign}(\theta - x_i)
\end{aligned}
$$

Hence, by upper bounding sign by 1:

$$\#\{\theta < x_i\} \leq \#\{\theta > x_i\} + \sum_{x_i = \theta} 1 = \#\{\theta \geq x_i\}.$$

Similarly, by lower bounding sign by -1:

$$\#\{\theta < x_i\} \leq \#\{\theta < x_i\} + \sum_{x_i = \theta} 1 = \#\{\theta \leq x_i\}$$

Table of Contents

Fermat's rule

Theorem

A point $\theta^\star$ minimizes a convex function $f : \mathbb{R}^d \to \mathbb{R}$ iff $0 \in \partial f(\theta^\star)$

Proof: use the sub-gradient definition:

- $0 \in \partial f(\theta^\star)$ iff $\forall \theta \in \mathbb{R}^d, f(\theta) \geq f(\theta^\star) + \langle 0, \theta - \theta^\star \rangle = f(\theta^\star)$

Fermat's rule

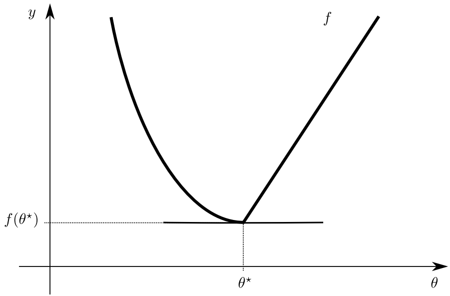
Theorem

A point $\theta^\star$ minimizes a convex function $f : \mathbb{R}^d \to \mathbb{R}$ iff $0 \in \partial f(\theta^\star)$

Proof: use the sub-gradient definition:

- $0 \in \partial f(\theta^\star)$ iff $\forall \theta \in \mathbb{R}^d, f(\theta) \geq f(\theta^\star) + \langle 0, \theta - \theta^\star \rangle = f(\theta^\star)$

Rem: visually, a horizontal tangent is admissible



Example: Fermat's rule for the Lasso

$$\beta^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \quad \left(\frac{1}{2}\|y - X\beta\|_2^2 \quad + \lambda \|\beta\|_1 \right)$$

Necessary and sufficient optimality conditions (Fermat's Rule):

$$\boxed{\forall j \in [\![1, p]\!],\ \mathbf{x}_j^\top \left(\frac{y - X\beta^{(\lambda)}}{\lambda} \right) \in \begin{cases} \{\mathrm{sign}(\beta^{(\lambda)})_j\} & \text{if } (\beta^{(\lambda)})_j \neq 0, \\ [-1, 1] & \text{if } (\beta^{(\lambda)})_j = 0. \end{cases}}$$

<u>Rem</u>: for OLS the **normal equation** are $\mathbf{x}_j^\top \left(y - X\beta^{(\lambda)} \right) = 0$

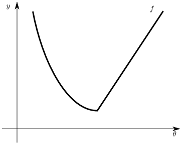
<u>Rem</u>: There exists a **critical** value $\lambda_{\max} = \max_{j \in [\![1,p]\!]} |\langle \mathbf{x}_j, y \rangle|$ s.t.

$$\forall \lambda > \lambda_{\max},\ \beta^{(\lambda)} = 0$$

Convexity and minimum

Various types of behavior for convex functions

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

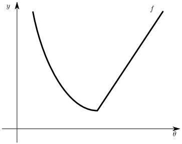


global minimum

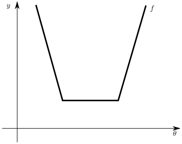
Convexity and minimum

Various types of behavior for convex functions

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



global minimum

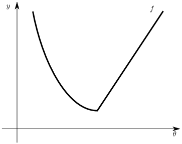


several minima

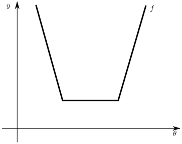
Convexity and minimum

Various types of behavior for convex functions

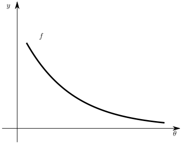
- global minimum *e.g.*, quadratic, etc.
- several minima *e.g.*, piecewise-affine (quadratic possible too!)
- no minimum, lower bounded *e.g.*, exponential function



global minimum



several minima

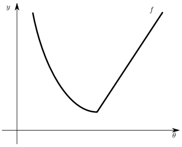


lower bounded

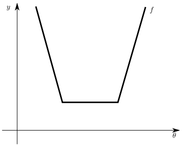
Convexity and minimum

Various types of behavior for convex functions

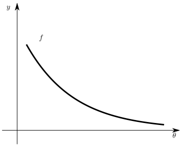
- global minimum *e.g.*, quadratic, etc.
- several minima *e.g.*, piecewise-affine (quadratic possible too!)
- no minimum, lower bounded *e.g.*, exponential function
- no minimum, lower bound is $-\infty$ *e.g.*, affine or $-\log(\cdot)$



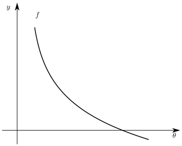
global minimum



several minima



lower bounded



not lower bounded

Gradient descent : intuition

- General formulation: minimize f by finding iteratively a new point for which f has decreased the most

Gradient descent : intuition

- General formulation: minimize f by finding iteratively a new point for which f has decreased the most

- First order approximation:

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

Gradient descent : intuition

- General formulation: minimize f 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 $\theta - \theta^0 = -\alpha \nabla f(\theta^0)$

Gradient descent : intuition

- General formulation: minimize f 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**

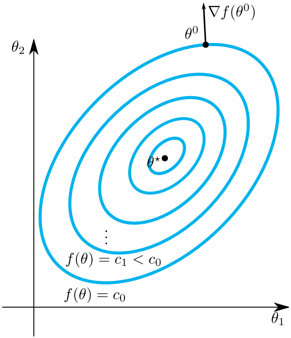
Gradient descent: algorithm

Algorithm: GRADIENT DESCENT

input : step size α, max. iterations $t_{\max}$, stopping criterion ε
init : θ^0
for $1 \leq t \leq t_{\max}$ **do**
  Break if stopping criterion smaller than ε
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
return $\theta^{t_{\max}}$ *"close" to a minimum of f*

Possible stopping criterion:

- $\|\nabla f(\theta^t)\| \leq \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leq \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ or $\frac{\|\theta^{t+1}-\theta^t\|}{\|\theta^t\|} \leq \varepsilon$
- duality gap (when available)



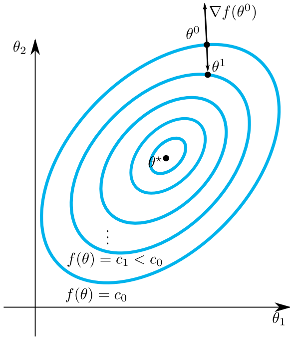
Gradient descent: algorithm

Algorithm: GRADIENT DESCENT

input : step size α, max. iterations $t_{\max}$, stopping criterion ε
init : θ^0
for $1 \leq t \leq t_{\max}$ **do**
  Break if stopping criterion smaller than ε
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
return $\theta^{t_{\max}}$ *"close" to a minimum of f*

Possible stopping criterion:

- $\|\nabla f(\theta^t)\| \leq \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leq \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leq \varepsilon$
- duality gap (when available)



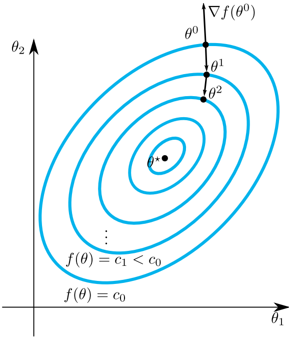
Gradient descent: algorithm

Algorithm: GRADIENT DESCENT

input : step size α, max. iterations $t_{\max}$, stopping criterion ε
init : θ^0
for $1 \leq t \leq t_{\max}$ **do**
  Break if stopping criterion smaller than ε
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
return $\theta^{t_{\max}}$ *"close" to a minimum of f*

Possible stopping criterion:

- $\|\nabla f(\theta^t)\| \leq \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leq \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leq \varepsilon$
- duality gap (when available)



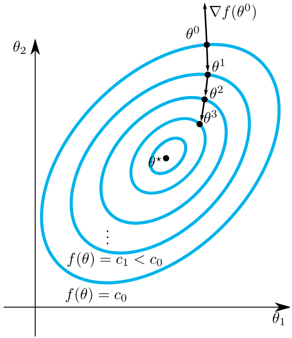
Gradient descent: algorithm

Algorithm: GRADIENT DESCENT

input : step size α, max. iterations $t_{\max}$, stopping criterion ε
init : θ^0
for $1 \leq t \leq t_{\max}$ **do**
    Break if stopping criterion smaller than ε
    $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
return $\theta^{t_{\max}}$ *"close" to a minimum of f*

Possible stopping criterion:

- $\|\nabla f(\theta^t)\| \leq \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leq \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ or $\frac{\|\theta^{t+1}-\theta^t\|}{\|\theta^t\|} \leq \varepsilon$
- duality gap (when available)



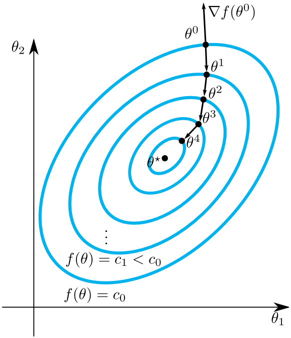
Gradient descent: algorithm

Algorithm: GRADIENT DESCENT

input : step size α, max. iterations $t_{\max}$, stopping criterion ε
init : θ^0
for $1 \leq t \leq t_{\max}$ **do**
  Break if stopping criterion smaller than ε
  $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
return $\theta^{t_{\max}}$ *"close" to a minimum of f*

Possible stopping criterion:

- $\|\nabla f(\theta^t)\| \leq \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leq \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leq \varepsilon$
- duality gap (when available)



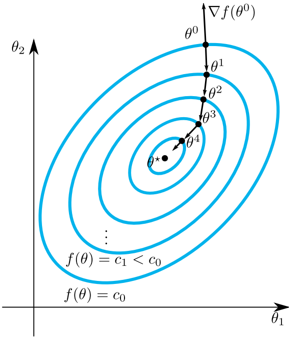
Gradient descent: algorithm

Algorithm: GRADIENT DESCENT

input : step size α, max. iterations $t_{\max}$, stopping criterion ε
init : θ^0
for $1 \leq t \leq t_{\max}$ **do**
    Break if stopping criterion smaller than ε
    $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla f(\theta^t)$
return $\theta^{t_{\max}}$ *"close" to a minimum of f*

Possible stopping criterion:

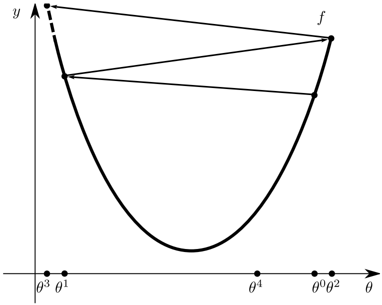
- $\|\nabla f(\theta^t)\| \leq \varepsilon$
- $f(\theta^{t+1}) - f(\theta^t) \leq \varepsilon$
- $\|\theta^{t+1} - \theta^t\| \leq \varepsilon$ or $\frac{\|\theta^{t+1} - \theta^t\|}{\|\theta^t\|} \leq \varepsilon$
- duality gap (when available)



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

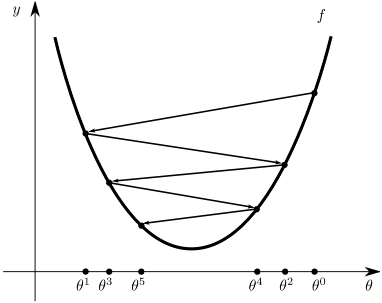


Divergence: really too large step size

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

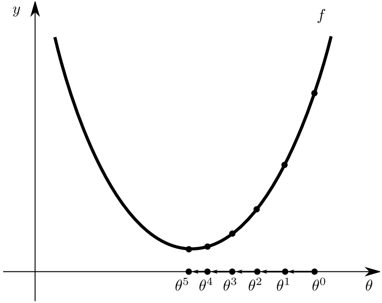


Slow convergence : still too large step size

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

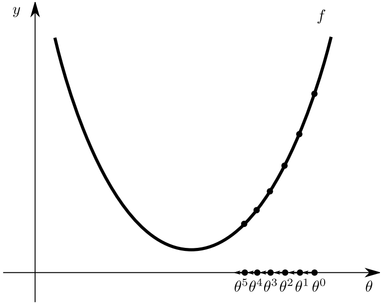


Fast convergence : good step size

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

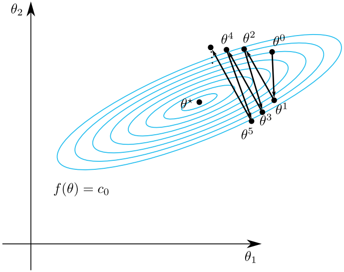


Slow convergence : too small step size

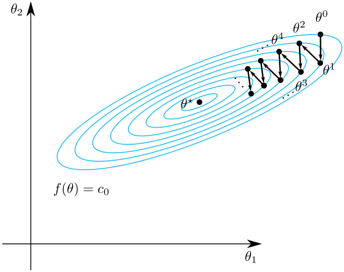
Mind the step...size (2D case)

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

α: crucial parameter to insure convergence toward a minimum



Too large step



Too small step

Smooth function: gradient Lipschitz

Quadratic majorization

If f is convex, differentiable with gradient L-Lipschitz, *i.e.,*

$$\forall(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|$$

then the following holds: $\quad \forall(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d,$

$$0 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq \frac{L}{2} \left\|\theta' - \theta\right\|^2$$

Smooth function: gradient Lipschitz

Quadratic majorization

If f is convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|$$

then the following holds: $\quad \forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d,$

$$0 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq \frac{L}{2} \left\|\theta' - \theta\right\|^2$$

Rem: positivity is a consequence of convexity. The second inequality will be proved later on.

Smooth function: gradient Lipschitz

Quadratic majorization

If f is convex, differentiable with gradient L-Lipschitz, *i.e.*,

$$\forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\|$$

then the following holds: $\quad \forall (\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$0 \leq f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle \leq \frac{L}{2} \left\| \theta' - \theta \right\|^2$$

Rem: positivity is a consequence of convexity. The second inequality will be proved later on.

Rem: if f is twice differentiable $\nabla^2 f \preceq L \cdot \mathrm{Id}_d$ in the sense that $L \cdot \mathrm{Id}_d - \nabla^2 f$ is semi-definite positive, then ∇f is L-Lipschitz

Majorization/minimization

Fix θ^0, and assume the previous inequality holds for any $\theta \in \mathbb{R}^d$:

$$f(\theta) - f(\theta^0) - \langle \nabla f(\theta^0), \theta - \theta^0 \rangle \leq \frac{L}{2} \left\| \theta^0 - \theta \right\|^2$$

yields

$$\begin{aligned} f(\theta) \leq & f(\theta^0) + \langle \nabla f(\theta^0), \theta - \theta^0 \rangle + \frac{L}{2} \|\theta^0 - \theta\|^2 \\ = & \tfrac{L}{2} \left\| \theta^0 - \tfrac{1}{L} \nabla f(\theta^0) - \theta \right\|^2 + f(\theta^0) - \tfrac{1}{2L} \left\| \nabla f(\theta^0) \right\|^2 := Q_L(\theta^0, \theta) \end{aligned}$$

Hence : $\quad \forall \theta \in \mathbb{R}^d, \quad \begin{cases} Q_L(\theta^0, \theta^0) = f(\theta^0) \\ f(\theta) \leq Q_L(\theta^0, \theta) \end{cases}$. This leads to a tight upper bound that can be simply minimized, since

$$\operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} Q_L(\theta^0, \theta) = \theta^0 - \frac{1}{L} \nabla f(\theta^0)$$

Example on a simple case: Huber function

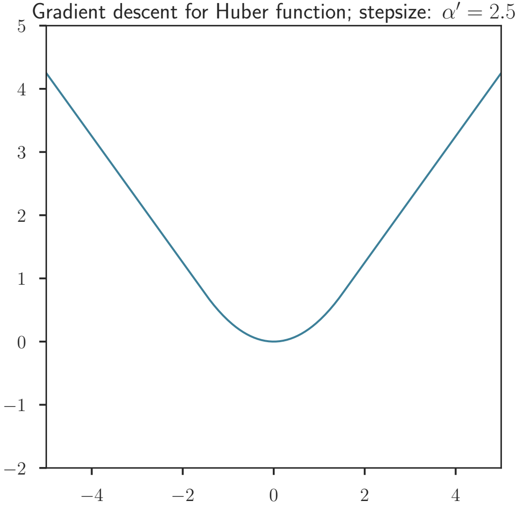
Remind that

$$\rho_\alpha = \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

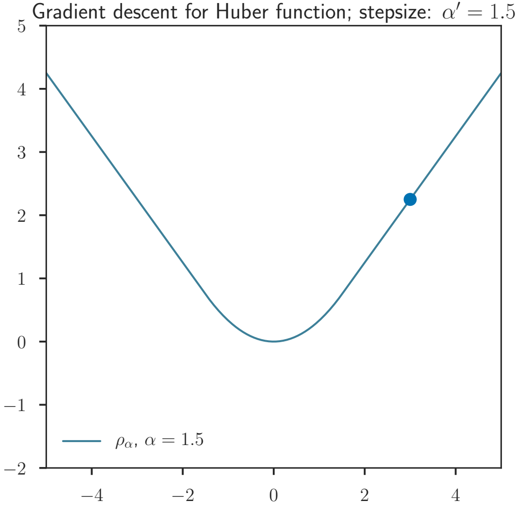
Then, one can show[3] that this is a convex function with gradient L-Lipschitz for $L = \frac{1}{\alpha}$.

[3] A. Beck and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.

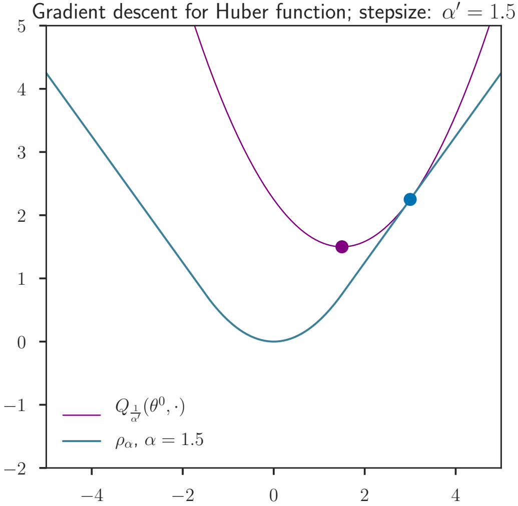
Maximization/minimization



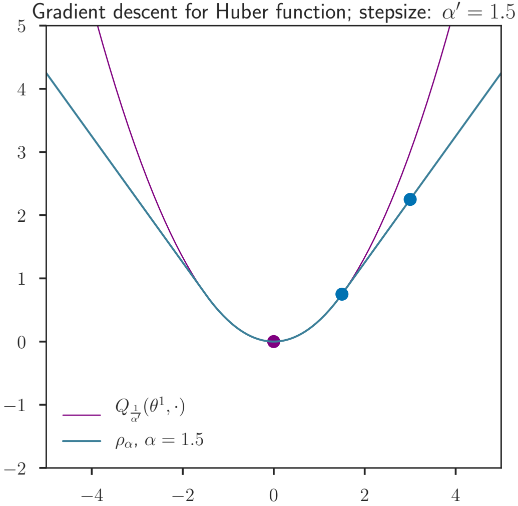
Maximization/minimization



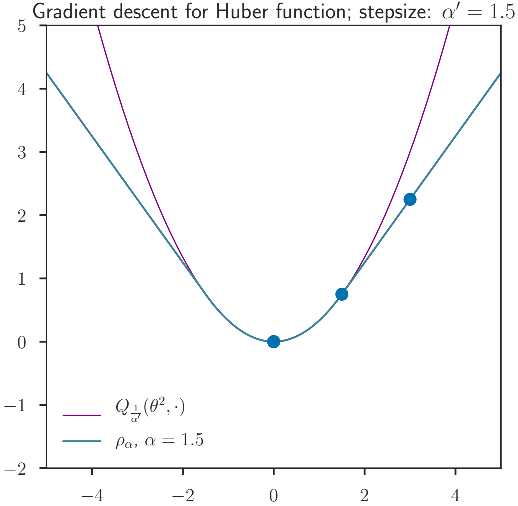
Maximization/minimization



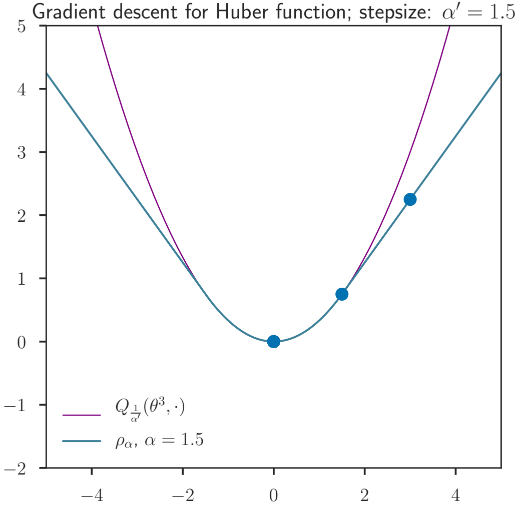
Maximization/minimization



Maximization/minimization



Maximization/minimization



Proof of majorization

Define $\phi(t) = f(t\theta + (1-t)\theta') = f(\theta' + t(\theta - \theta'))$. Then, ϕ is differentiable and $\phi'(t) = \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle$. Hence,

Proof of majorization

Define $\phi(t) = f(t\theta + (1-t)\theta') = f(\theta' + t(\theta - \theta'))$. Then, ϕ is differentiable and $\phi'(t) = \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle$. Hence,

$$f(\theta) - f(\theta') = \phi(1) - \phi(0) = \int_0^1 \phi'(t)dt$$

Proof of majorization

Define $\phi(t) = f(t\theta + (1-t)\theta') = f(\theta' + t(\theta - \theta'))$. Then, ϕ is differentiable and $\phi'(t) = \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle$. Hence,

$$\begin{aligned} f(\theta) - f(\theta') =& \phi(1) - \phi(0) = \int_0^1 \phi'(t) dt \\ =& \int_0^1 \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle \, dt \end{aligned}$$

Proof of majorization

Define $\phi(t) = f(t\theta + (1-t)\theta') = f(\theta' + t(\theta - \theta'))$. Then, ϕ is differentiable and $\phi'(t) = \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle$. Hence,

$$
\begin{aligned}
f(\theta) - f(\theta') =& \phi(1) - \phi(0) = \int_0^1 \phi'(t) dt \\
=& \int_0^1 \left\langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \right\rangle dt \\
=& \int_0^1 \left\langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') - \nabla f(\theta') \right\rangle dt \\
& + \langle \theta - \theta', \nabla f(\theta') \rangle
\end{aligned}
$$

Proof of majorization

Define $\phi(t) = f(t\theta + (1-t)\theta') = f(\theta' + t(\theta - \theta'))$. Then, ϕ is differentiable and $\phi'(t) = \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle$. Hence,

$$\begin{aligned} f(\theta) - f(\theta') =& \phi(1) - \phi(0) = \int_0^1 \phi'(t) dt \\ &= \int_0^1 \left\langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \right\rangle dt \\ &= \int_0^1 \left\langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') - \nabla f(\theta') \right\rangle dt \\ &\quad + \langle \theta - \theta', \nabla f(\theta') \rangle \end{aligned}$$

Use Cauchy-Schwartz Inequality and remind ∇f is L-Lipschitz:

$$\begin{aligned} & f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle \\ \leq & \int_0^1 \|\theta - \theta'\| \left\| \nabla f(t\theta + (1-t)\theta') - \nabla f(\theta') \right\| dt \end{aligned}$$

Proof of majorization

Define $\phi(t) = f(t\theta + (1-t)\theta') = f(\theta' + t(\theta - \theta'))$. Then, ϕ is differentiable and $\phi'(t) = \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle$. Hence,

$$
\begin{aligned}
f(\theta) - f(\theta') =& \phi(1) - \phi(0) = \int_0^1 \phi'(t) dt \\
=& \int_0^1 \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') \rangle \, dt \\
=& \int_0^1 \langle \theta - \theta', \nabla f(t\theta + (1-t)\theta') - \nabla f(\theta') \rangle \, dt \\
& + \langle \theta - \theta', \nabla f(\theta') \rangle
\end{aligned}
$$

Use Cauchy-Schwartz Inequality and remind ∇f is L-Lipschitz:

$$
\begin{aligned}
& f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle \\
\leq & \int_0^1 \|\theta - \theta'\| \, \|\nabla f(t\theta + (1-t)\theta') - \nabla f(\theta')\| \, dt \\
\leq & L \|\theta - \theta'\|^2 \int_0^1 t dt = \frac{L}{2} \|\theta - \theta'\|^2
\end{aligned}
$$

$\square$

Table of Contents

Convergence in the smooth case

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}, \text{ for } t = 1, \dots, t_{\max}$$

Convergence rate for fixed step size

If f is convex, differentiable, with L-Lipschitz gradient, for any minimum $\theta^\star$ of f, if $\alpha \leq \frac{1}{L}$ then $\theta^{t_{\max}}$ satisfies

$$f(\theta^{t_{\max}}) - f(\theta^\star) \leq \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha t_{\max}}$$

[4] Y. Nesterov. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

Convergence in the smooth case

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}, \text{ for } t = 1, \dots, t_{\max}$$

Convergence rate for fixed step size

If f is convex, differentiable, with L-Lipschitz gradient, for any minimum $\theta^\star$ of f, if $\alpha \leq \frac{1}{L}$ then $\theta^{t_{\max}}$ satisfies

$$f(\theta^{t_{\max}}) - f(\theta^\star) \leq \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha t_{\max}}$$

Rem: convergence is all the more fast that

[4] Y. Nesterov. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

Convergence in the smooth case

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}, \text{ for } t = 1, \dots, t_{\max}$$

Convergence rate for fixed step size

If f is convex, differentiable, with L-Lipschitz gradient, for any minimum $\theta^\star$ of f, if $\alpha \leq \frac{1}{L}$ then $\theta^{t_{\max}}$ satisfies

$$f(\theta^{t_{\max}}) - f(\theta^\star) \leq \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha t_{\max}}$$

Rem: convergence is all the more fast that

- a better initialization is found, *i.e.*, $\|\theta^0 - \theta^\star\|^2$ small

[4] Y. Nesterov. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

Convergence in the smooth case

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}, \text{ for } t = 1, \dots, t_{\max}$$

Convergence rate for fixed step size

If f is convex, differentiable, with L-Lipschitz gradient, for any minimum $\theta^\star$ of f, if $\alpha \leq \frac{1}{L}$ then $\theta^{t_{\max}}$ satisfies

$$f(\theta^{t_{\max}}) - f(\theta^\star) \leq \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha t_{\max}}$$

Rem: convergence is all the more fast that

- a better initialization is found, *i.e.*, $\|\theta^0 - \theta^\star\|^2$ small
- a larger step size α is used (best: $\alpha = 1/L$)

[4] Y. Nesterov. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

Convergence in the smooth case

$$\boxed{\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)}\,, \text{ for } t = 1, \ldots, t_{\max}$$

Convergence rate for fixed step size

If f is convex, differentiable, with L-Lipschitz gradient, for any minimum $\theta^\star$ of f, if $\alpha \leq \frac{1}{L}$ then $\theta^{t_{\max}}$ satisfies

$$f(\theta^{t_{\max}}) - f(\theta^\star) \leq \frac{\|\theta^0 - \theta^\star\|^2}{2\alpha t_{\max}}$$

Rem: convergence is all the more fast that

- a better initialization is found, *i.e.*, $\|\theta^0 - \theta^\star\|^2$ small
- a larger step size α is used (best: $\alpha = 1/L$)
- Nesterov accelerated variant[4], reaches $1/(t^{\max})^2$ (optimal)

[4] Y. Nesterov. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

Convergence: proof

Fact 1: gradient L-Lipschitz implies quadratic upper bound

$$\forall(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad f(\theta) \leq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{L}{2} \left\| \theta' - \theta \right\|^2$$

Convergence: proof

Fact 1: gradient L-Lipschitz implies quadratic upper bound

$$\forall(\theta,\theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad f(\theta) \leq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{L}{2} \left\| \theta' - \theta \right\|^2$$

Fact 2: by convexity

$$f(\theta^t) \leq f(\theta^\star) - \langle \nabla f(\theta^t), \theta^\star - \theta^t \rangle = f(\theta^\star) + \langle \nabla f(\theta^t), \theta^t - \theta^\star \rangle$$

Convergence: proof

Fact 1: gradient L-Lipschitz implies quadratic upper bound

$$\forall(\theta,\theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad f(\theta) \leq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{L}{2} \left\|\theta' - \theta\right\|^2$$

Fact 2: by convexity

$$f(\theta^t) \leq f(\theta^\star) - \langle \nabla f(\theta^t), \theta^\star - \theta^t \rangle = f(\theta^\star) + \langle \nabla f(\theta^t), \theta^t - \theta^\star \rangle$$

Fact 3: as $\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)$ and $0 < \alpha \leq \frac{1}{L}$, with Fact 1

$$f(\theta^{t+1}) \leq f(\theta^t) - \alpha \left(1 - \tfrac{L\alpha}{2}\right) \|\nabla f(\theta^t)\|^2 \leq f(\theta^t) - \tfrac{\alpha}{2} \|\nabla f(\theta^t)\|^2$$

Convergence: proof

Fact 1: gradient L-Lipschitz implies quadratic upper bound

$$\forall(\theta,\theta')\in\mathbb{R}^d\times\mathbb{R}^d,\quad f(\theta)\leq f(\theta')+\langle\nabla f(\theta'),\theta-\theta'\rangle+\frac{L}{2}\left\|\theta'-\theta\right\|^2$$

Fact 2: by convexity

$$f(\theta^t)\leq f(\theta^\star)-\langle\nabla f(\theta^t),\theta^\star-\theta^t\rangle=f(\theta^\star)+\langle\nabla f(\theta^t),\theta^t-\theta^\star\rangle$$

Fact 3: as $\theta^{t+1}=\theta^t-\alpha\nabla f(\theta^t)$ and $0<\alpha\leq\frac{1}{L}$, with Fact 1

$$f(\theta^{t+1})\leq f(\theta^t)-\alpha\left(1-\tfrac{L\alpha}{2}\right)\|\nabla f(\theta^t)\|^2\leq f(\theta^t)-\tfrac{\alpha}{2}\|\nabla f(\theta^t)\|^2$$

Fact 4: using Fact 2 & 3, $ab=\frac{a^2+b^2-(a-b)^2}{2}$, $\theta^{t+1}=\theta^t-\alpha\nabla f(\theta^t)$:

$$f(\theta^{t+1})\leq f(\theta^\star)+\langle\nabla f(\theta^t),\theta^t-\theta^\star\rangle-\frac{\alpha}{2}\|\nabla f(\theta^t)\|^2$$

Convergence: proof

Fact 1: gradient L-Lipschitz implies quadratic upper bound

$$\forall(\theta,\theta') \in \mathbb{R}^d \times \mathbb{R}^d, \quad f(\theta) \leq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{L}{2} \|\theta' - \theta\|^2$$

Fact 2: by convexity

$$f(\theta^t) \leq f(\theta^\star) - \langle \nabla f(\theta^t), \theta^\star - \theta^t \rangle = f(\theta^\star) + \langle \nabla f(\theta^t), \theta^t - \theta^\star \rangle$$

Fact 3: as $\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)$ and $0 < \alpha \leq \frac{1}{L}$, with Fact 1

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

Fact 4: using Fact 2 & 3, $ab = \frac{a^2+b^2-(a-b)^2}{2}$, $\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t)$:

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

Convergence proof (bis)

Fact 4: Telescopic sum

$$\frac{1}{t_{\max}}\sum_{t=0}^{t_{\max}-1}\Big(f(\theta^{t+1})-f(\theta^\star)\Big) \leq \frac{1}{t_{\max}}\frac{1}{2\alpha}(\|\theta^0-\theta^\star\|^2-\|\theta^{t_{\max}}-\theta^\star\|^2)$$

$$\leq \frac{1}{2\alpha t_{\max}}\|\theta^0-\theta^\star\|^2$$

Convergence proof (bis)

Fact 4: Telescopic sum

$$\frac{1}{t_{\max}} \sum_{t=0}^{t_{\max}-1} \Big(f(\theta^{t+1}) - f(\theta^\star)\Big) \leq \frac{1}{t_{\max}} \frac{1}{2\alpha} (\|\theta^0 - \theta^\star\|^2 - \|\theta^{t_{\max}} - \theta^\star\|^2)$$

$$\leq \frac{1}{2\alpha t_{\max}} \|\theta^0 - \theta^\star\|^2$$

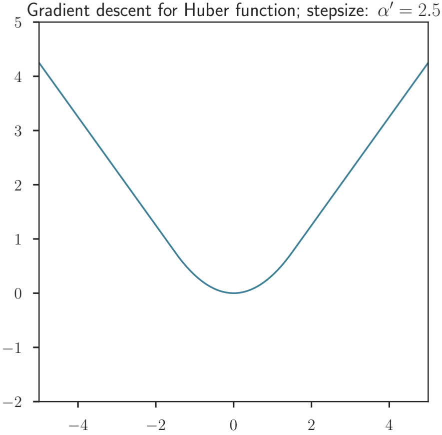
From Fact 3, for any $t \geq 0$, $f(\theta^{t+1}) \leq f(\theta^t)$, hence

$$f(\theta^{t_{\max}}) - f(\theta^\star) \leq \frac{1}{t_{\max}} \sum_{t=0}^{t_{\max}-1} \Big(f(\theta^{t+1}) - f(\theta^\star)\Big)$$

$$\leq \frac{1}{2\alpha t_{\max}} \|\theta^0 - \theta^\star\|^2$$

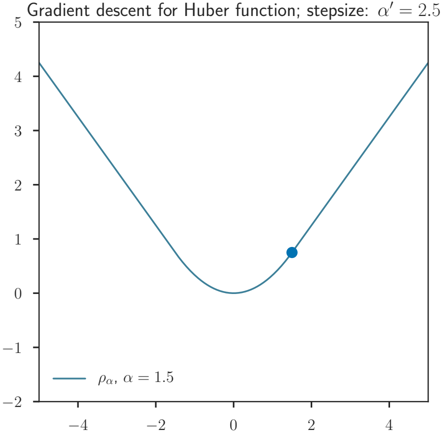
More on convergence

- A similar results holds for $\alpha < \frac{2}{L}$, *cf.* Nesterov (2004) [p. 69]



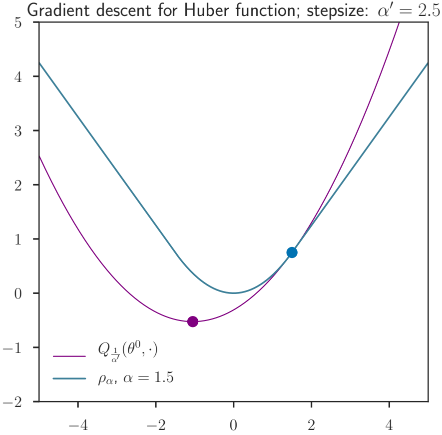
More on convergence

- A similar results holds for $\alpha < \frac{2}{L}$, *cf.* Nesterov (2004) [p. 69]



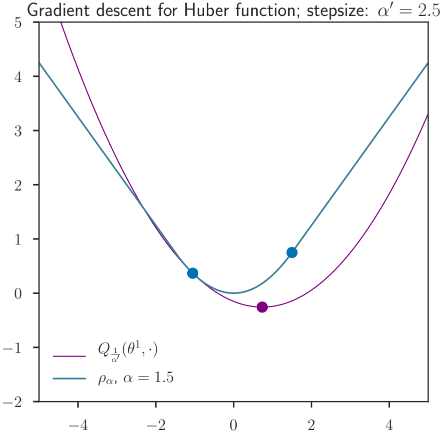
More on convergence

- A similar results holds for $\alpha < \frac{2}{L}$, *cf.* Nesterov (2004) [p. 69]



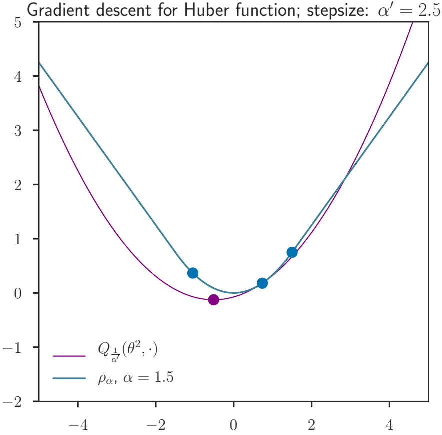
More on convergence

- A similar results holds for $\alpha < \frac{2}{L}$, *cf.* Nesterov (2004) [p. 69]



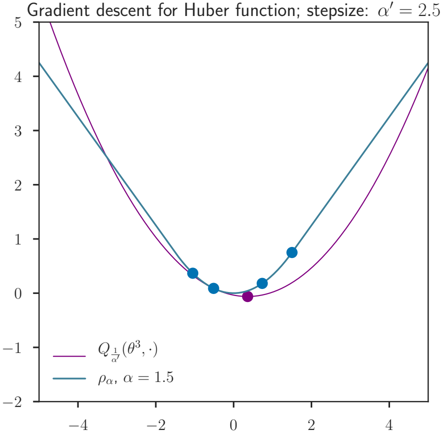
More on convergence

- A similar results holds for $\alpha < \frac{2}{L}$, *cf.* Nesterov (2004) [p. 69]



More on convergence

- A similar results holds for $\alpha < \frac{2}{L}$, *cf.* Nesterov (2004) [p. 69]



Convergence and limits

- One needs to know the constant L, to find a correct (scaling) step size. It is not always known by the practitioner.
  Example : $\theta \mapsto \frac{\|X\theta - y\|_2^2}{2}$ then $L = \lambda_{\max}(X^\top X)$; L is the **spectral radius**

- A small constant step size is not the solution : it would lead to (very) slow convergence...

Convergence of the iterates

- The iterates convergence is not guaranteed for all smooth functions, also more convergence difficulties in infinite dimension spaces

Convergence of the iterates

- The iterates convergence is not guaranteed for all smooth functions, also more convergence difficulties in infinite dimension spaces
- the iterates can be shown to converge for convex function with gradient L-Lipschitz and $\alpha < \frac{2}{L}$: there exists a solution $\theta^\star$ of the problem such that: $\theta^t \underset{t \to +\infty}{\to} \theta^\star$.

Convergence of the iterates

- The iterates convergence is not guaranteed for all smooth functions, also more convergence difficulties in infinite dimension spaces
- the iterates can be shown to converge for convex function with gradient L-Lipschitz and $\alpha < \frac{2}{L}$: there exists a solution $\theta^\star$ of the problem such that: $\theta^t \underset{t \to +\infty}{\longrightarrow} \theta^\star$.
- One needs convexity for iterates convergence, otherwise counter-example Bertsekas (1999) or Absil *et al.* 2005 even for $\mathcal{C}^\infty$ functions

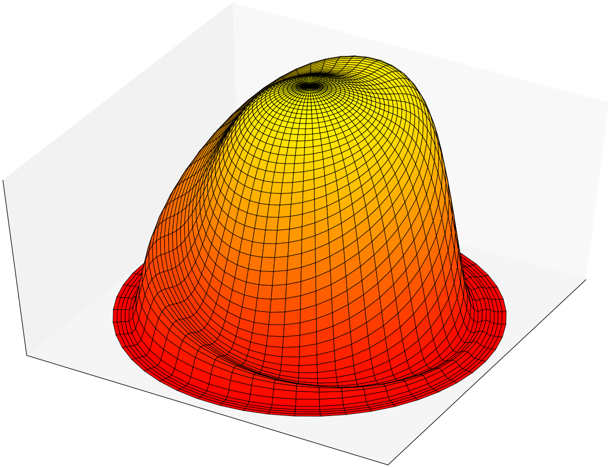
Convergence of the iterates

- The iterates convergence is not guaranteed for all smooth functions, also more convergence difficulties in infinite dimension spaces
- the iterates can be shown to converge for convex function with gradient L-Lipschitz and $\alpha < \frac{2}{L}$: there exists a solution $\theta^\star$ of the problem such that: $\theta^t \underset{t \to +\infty}{\longrightarrow} \theta^\star$.
- 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), see next slide

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

Counter example: spiraling toward zero



Counter example: spiraling toward zero

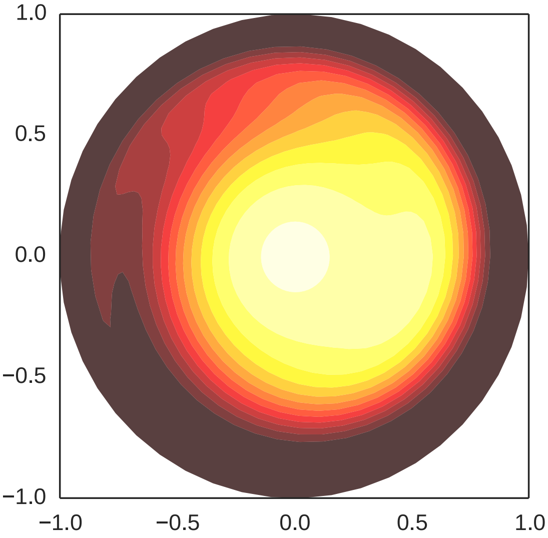


Table of Contents

Sub-gradient descent

Algorithm: SUB-GRADIENT DESCENT

input : max. iterations $t_{\max}$, step size $\alpha_t, t = 1, \dots, t_{\max}$,
stopping criterion ε

init : θ^0

for $1 \leq t \leq t_{\max}$ **do**

  Break if stopping criterion smaller than ε

  find $g_t \in \partial f(\theta^t)$

  $\theta^{t+1} \leftarrow \theta^t - \alpha_t g_t$

return $\theta^{t_{\max}}$ *"close" to a minimum of* f

Rem: theory[5] ensures convergence rate of $O(\log(t_{\max})/\sqrt{t_{\max}})$ when choosing $\alpha_t \propto 1/\sqrt{t}$

[5] Y. Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Applied Optimization. Boston, MA: Kluwer Academic Publishers, 2004.

Table of Contents

Analysis with strong-convexity

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

Definition

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:

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

Analysis with strong-convexity

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

Definition

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:

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

Rem: a more standard definition is that $f - 1/2\mu | \cdot |^2$ is convex

Analysis with strong-convexity

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

Definition

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:

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

Rem: a more standard definition is that $f - 1/2\mu|\cdot|^2$ is convex

Rem: if f is twice differentiable $\nabla^2 f(\theta) \succeq \mu \cdot Id$

Analysis with strong-convexity

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

Definition

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:

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

Rem: a more 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 $\lambda_{\max}(X^\top X)/\lambda_{\min}(X^\top X)$ is the (matrix) condition number of 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:

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

and the iterates converge provided $\alpha \leq \frac{1}{L}$

$$\|\theta^{t_{\max}} - \theta^\star\|_2^2 \leq \frac{1}{\alpha\mu t_{\max}}\|\theta^0 - \theta^\star\|_2^2$$

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:

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

and the iterates converge provided $\alpha \leq \frac{1}{L}$

$$\|\theta^{t_{\max}} - \theta^\star\|_2^2 \leq \frac{1}{\alpha\mu t_{\max}}\|\theta^0 - \theta^\star\|_2^2$$

Rem: if $\alpha = \frac{1}{L}$ the constant factor is $\frac{L}{\mu}$ (**condition number**)

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:

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

and the iterates converge provided $\alpha \leq \frac{1}{L}$

$$\|\theta^{t_{\max}} - \theta^\star\|_2^2 \leq \frac{1}{\alpha \mu t_{\max}} \|\theta^0 - \theta^\star\|_2^2$$

<u>Rem</u>: if $\alpha = \frac{1}{L}$ the constant factor is $\frac{L}{\mu}$ (**condition number**)

<u>Rem</u>: geometric convergence rate Nesterov (2004) [p.70]:

$$f(\theta) - f(\theta^\star) \leq \left(1 - \frac{\mu}{L}\right)^{t_{\max}} \|\theta^0 - \theta^\star\|_2^2 \quad (\text{for } \alpha = \tfrac{1}{L})$$

References I

- Absil, P., R. Mahony, and B. Andrews. "Convergence of the Iterates of Descent Methods for Analytic Cost Functions". In: *SIAM J. Optim.* 16.2 (2005), pp. 531–547.
- Bauschke, H. H. and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. New York: Springer, 2011, pp. xvi+468.
- Beck, A. and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.
- Bertsekas, D. P. *Nonlinear programming*. Athena Scientific, 1999.
- Hiriart-Urruty, J.-B. and C. Lemaréchal. *Convex analysis and minimization algorithms. I*. Vol. 305. Berlin: Springer-Verlag, 1993.
- Nesterov, Y. "A method for solving a convex programming problem with rate of convergence $O(1/k^2)$". In: *Soviet Math. Doklady* 269.3 (1983), pp. 543–547.

References II

- Nesterov, Y. *Introductory lectures on convex optimization*. Vol. 87. Applied Optimization. Boston, MA: Kluwer Academic Publishers, 2004.