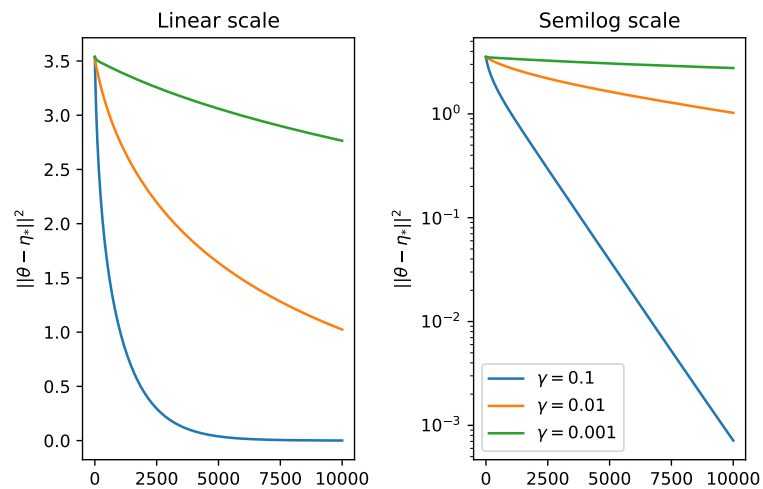


# FTML practical session 5

Gradient descent: squared distance to the OLS estimator  
 $\|\theta - \theta^*\|^2$



## TABLE DES MATIÈRES

1	Convexity	2
2	Application of unsupervised learning to classification	2
2.1	Meteorological data : dimensionality reduction and visualization . . .	2
3	Gradient descent and line search for a least squares problem	2
3.1	Setting . . . . .	2
3.2	Vanilla gradient descent . . . . .	3
3.3	Line search : A nested optimization problem . . . . .	4

## 1 CONVEXITY

- Show that  $x \rightarrow x^2$  is convex on  $\mathbb{R}$ , using the definition.
- Show that  $x \rightarrow \sin(x)$  is non-convex on  $\mathbb{R}$ , using the definition. Note that for non-convex functions, it is possible to have a 0 gradient or derivative at a point  $x$ , without  $x$  being a local minimum or even a local extremum.
- Find a differentiable function from  $\mathbb{R}$  to  $\mathbb{R}$  that is non-convex on  $\mathbb{R}$ , but the restriction to  $\mathbb{R}_+$  is convex.

## 2 APPLICATION OF UNSUPERVISED LEARNING TO CLASSIFICATION

In this section, we build classifiers based on an unsupervised preprocessing of the data. We hence combine supervised and unsupervised learning in order to solve a problem.

### 2.1 Meteorological data : dimensionality reduction and visualization

A meteorological station has gathered 1600 data samples in dimension 6, thanks to 6 sensors, that represent various physical measurements (such as wind speed, humidity, temperature, etc). The operators of the station would like to predict the risk of a tempest the next day, but first, they need to reduce the dimensionality of the data, in order to apply a supervised learning algorithm on the reduced data.

The data are stored in the `exercice_1_dimensionality_reduction/data/` folder.

Find a dimensionality reduction method and a dimension (2 or 3), that seems to allow to predict the label based on the projected components only, first by making scatter plots of the projected data, and by coloring the data according to their label. Verify this by training a classifier that learns to predict the labels based on the projections only, with a very good accuracy.

Template files in the folder :

— `main.py`

[https://scikit-learn.org/stable/modules/unsupervised\\_reduction.html](https://scikit-learn.org/stable/modules/unsupervised_reduction.html)

Later in the course, we will also study nonlinear dimensionality reduction methods. Note that for this particular problem, it was also possible to solve it without the first unsupervised learning stage.

## 3 GRADIENT DESCENT AND LINE SEARCH FOR A LEAST SQUARES PROBLEM

We come back to a least squares problem, that we will use to show an example of gradient descent algorithm, with a non-constant learning rate.

### 3.1 Setting

#### 3.1.1 Notations

We recall the setting and the notations here :

—  $\mathcal{X} = \mathbb{R}^d$

- $y = \mathbb{R}$
- $\theta = \mathbb{R}^d$  defines the estimator, which is the function  $x \rightarrow x^T \theta$ .
- design matrix :  $X \in \mathbb{R}^{n,d}$
- vector of outputs (labels) :  $y \in \mathbb{R}^n$ .

We want to minimize the function  $f$  representing the empirical risk :

$$f(\theta) = \frac{1}{2n} \|X\theta - y\|^2 \quad (1)$$

### 3.1.2 Gradient descent

We recall that the gradient of  $f$  and its Hessian write :

$$\begin{aligned} \nabla_{\theta} f(\theta) &= \frac{1}{n} X^T (X\theta - y) \\ &= H\theta - \frac{1}{n} X^T y \end{aligned} \quad (2)$$

$$H = \frac{1}{n} X^T X \quad (3)$$

An **iteration** of the gradient algorithm writes :

$$\theta_{t+1} = \theta_t - \gamma \nabla_{\theta} f(\theta_t) \quad (4)$$

### 3.1.3 Global minimizers

We consider the global minimizers of  $f$ , noted  $\eta^*$ . They all necessary verify

$$\nabla_{\theta} f(\eta^*) = 0 \quad (5)$$

Which means that

$$H\eta^* = \frac{1}{n} X^T y \quad (6)$$

As  $\theta \rightarrow f(\theta)$  is convex, we know that this condition is also sufficient : any vector that cancels the gradient is a global minimum.

If  $H$  is not invertible, there might be several minimizes. However, if  $f$  is strongly convex, then  $H$  is invertible and  $\eta^*$  is unique, we note it  $\theta^*$ . We will consider such a case here. In general,  $H$  is a positive semi-definite matrix, and here we hence consider a case where it is definite positive.

## 3.2 Vanilla gradient descent

Implement a gradient descent (GD) on the data contained in [exercice\\_2\\_line\\_search/data/](#), and experiment the learning rate in order to observe convergence or divergence.

Template files in the folder :

- **main.py**
- **algorithms.py** (to edit)

You should observe something like figure 1 in the cases of convergence.

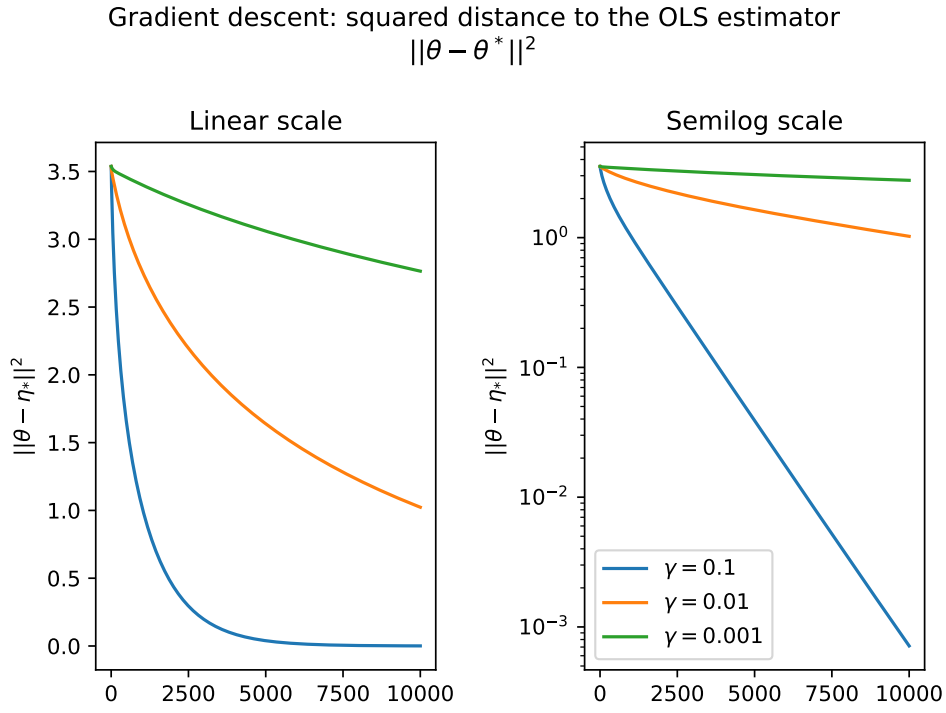


FIGURE 1 – Constant step size gradient descent and for several learning rates.

### 3.3 Line search : A nested optimization problem

Considering a fixed iteration step  $\theta_t$ , we note

$$\alpha(\gamma) = \theta_t - \gamma \nabla_{\theta} f(\theta_t) \in \mathbb{R}^d \quad (7)$$

The **exact line search** method attempts to find the optimal step  $\gamma^*$ , at each iteration. This means, given the position  $\theta_t$ , the parameter  $\gamma$  that minimizes the function defined by

$$\begin{aligned} g(\gamma) &= f(\theta_t - \gamma \nabla_{\theta} f(\theta_t)) \\ &= f(\alpha(\gamma)) \end{aligned} \quad (8)$$

Is  $g : \mathbb{R} \rightarrow \mathbb{R}$  a convex function ?

Find the value  $\gamma^*$  that minimizes  $\gamma \rightarrow g(\gamma)$  for a given  $\theta_t$ , and gradient descent where  $\gamma$  is set optimally thanks to this method (exact line search). Compare the convergence speeds by measuring the distance between the iterate and the OLS estimator  $\eta^*$  at each iteration.

You can uncomment the lines that call the `line_search()` function in the main file. You should observe something like figure 2

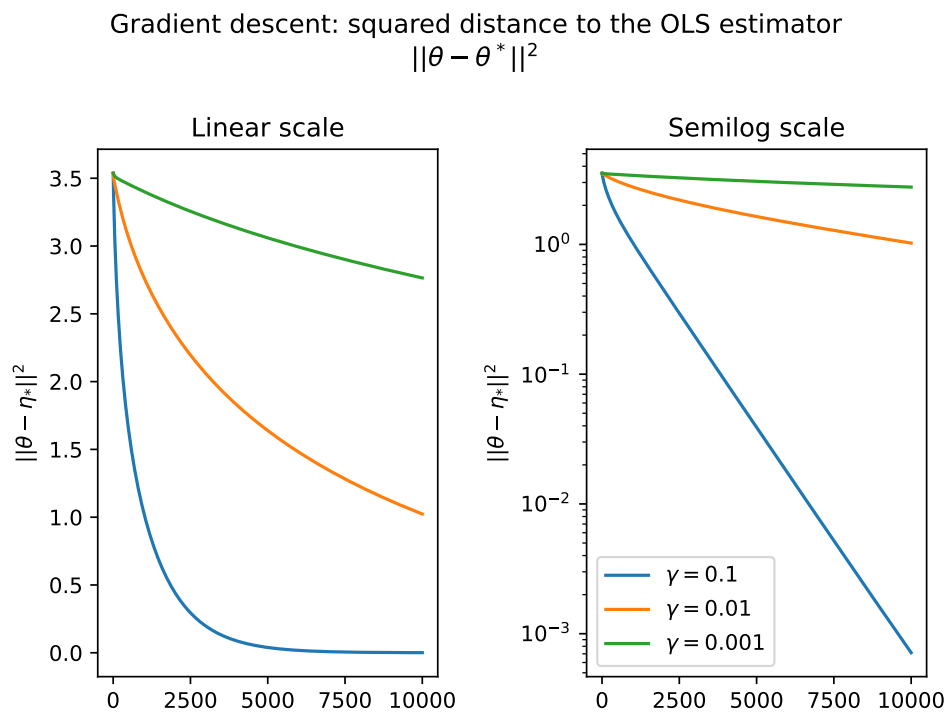


FIGURE 2 – Comparison between constant step size gradient descent and line search.