FTML practical session 14

14 juin 2025

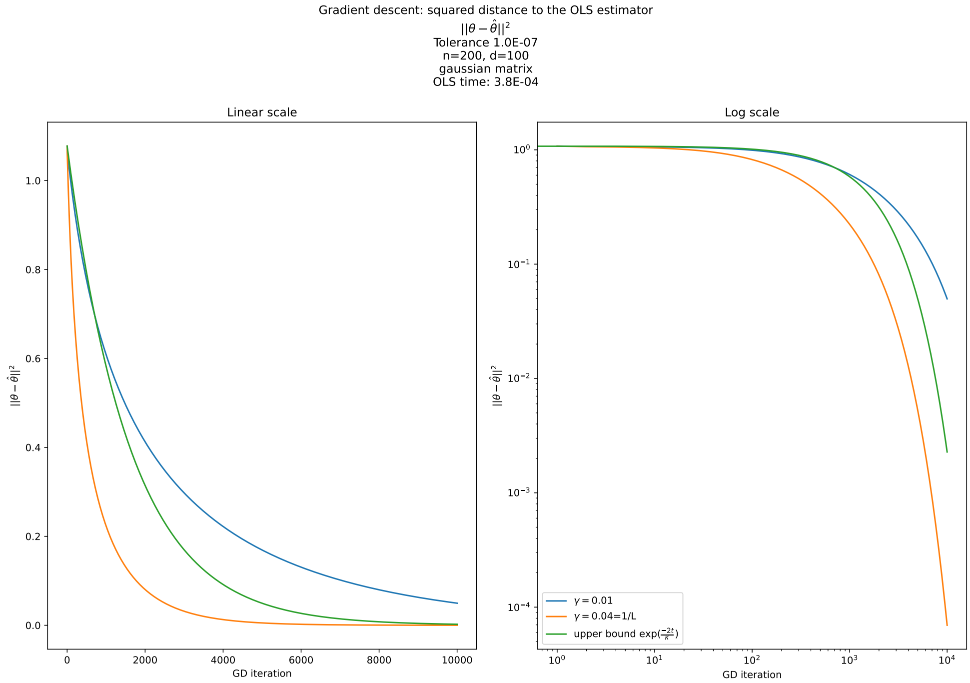

TABLE DES MATIÈRES

1 Convergence speed of gradient descent 2

1 CONVERGENCE SPEED OF GRADIENT DESCENT

1.1 Setting

We want to study the speed of convergence of the minimization of a convex function f defined over $\mathbb{R}^d$, with gradient descent.

$$\forall t, \ \theta_{t+1} \leftarrow \theta_t - \gamma \nabla_\theta f(\theta_t) \tag{1}$$

where t is the iteration index.

We will study the specific case of linear regression (OLS), but the results parially generalize to general convex functions. We use the usual objects :

— design matrix $X \in \mathbb{R}^{n,d}$

— label vector $y \in \mathbb{R}^n$.

— loss function

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

— $||.||$ is the usual euclidean norm.

The gradient and the Hessian write :

$$\nabla_\theta f(\theta) = \frac{1}{n} X^T (X\theta - y) \tag{3}$$

$$H = \frac{1}{n} X^T X \tag{4}$$

We note θ^* the minimizers of f. All minimizers verify that

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

or

$$H\theta^* = \frac{1}{n} X^T y \tag{6}$$

If H is not invertible, they might be not unique, but all have the same function value $f(\theta^*)$.

— H is symmetric, positive semi-definite.

— H is invertible if and only if its smallest eigenvalue μ is > 0, in which case f is strongly convex (see section 2.3 in https://github.com/nlehir/FTML/blob/master/lecture_notes/lecture%20notes.pdf)

1.2 Convergence speed of gradient descent

We assume that $\mu > 0$, meaning that H is invertible. Let us study the convergence speed of GD towards θ^* (that exsits and is unique).

1.2.1 *Step 1*

Show that

$$\forall \theta \in \mathbb{R}^d, f(\theta) - f(\theta^*) = \frac{1}{2} (\theta - \theta^*)^T H (\theta - \theta^*) \tag{7}$$

1.2.2 *Step 2*

Show that

$$\forall t \in \mathbb{N}, \theta_t = \theta_{t-1} - \gamma H(\theta_{t-1} - \theta^*) \tag{8}$$

1.2.3 Step 3

Deduce that :

$$\theta_t - \theta^* = (I - \gamma H)(\theta_{t-1} - \theta^*) \tag{9}$$

and that

$$\theta_t - \theta^* = (I - \gamma H)^t(\theta_0 - \theta^*) \tag{10}$$

where θ_0 is the initial value of θ.

1.2.4 Step 4

We can use two measures of performance of the gradient algorithm. Using the previous results, they write :

— Distance to minimizer :

$$||\theta_t - \theta^*||^2 = (\theta_0 - \theta^*)^T(I - \gamma H)^{2t}(\theta_0 - \theta^*) \tag{11}$$

— Convergence in function values :

$$f(\theta_t) - f(\theta^*) = \frac{1}{2}(\theta_0 - \theta^*)^T(I - \gamma H)^{2t}H(\theta_0 - \theta^*) \tag{12}$$

We introduce the **condition number** $\kappa = \frac{L}{\mu}$ where L is the largest eigenvalue of H. By convention, if $\mu = 0$, $L = +\infty$. Show that with a good choice of γ, we obtain an **exponential convergence**

$$||\theta_t - \theta^*||^2 \leqslant (1 - \frac{1}{\kappa})^{2t}||\theta_0 - \theta^*||^2 \tag{13}$$

We note that

$$(1 - \frac{1}{\kappa})^{2t} \leqslant \exp(-\frac{1}{\kappa})^{2t} = \exp(-\frac{2t}{\kappa}) \tag{14}$$

1.2.5 Simulation

Run a simulation that plots both the upper bound and the convergence speed of GD on a least squares problem, like seen on Figure 1. You can adapt the code from **tp_05_line_search**.

1.2.6 Non strongly convex functions

If $\mu = 0$, we do not have an exponential convergence guarantee, but rather a convergence rate in $\mathcal{O}(\frac{1}{t})$ (the proof is different).

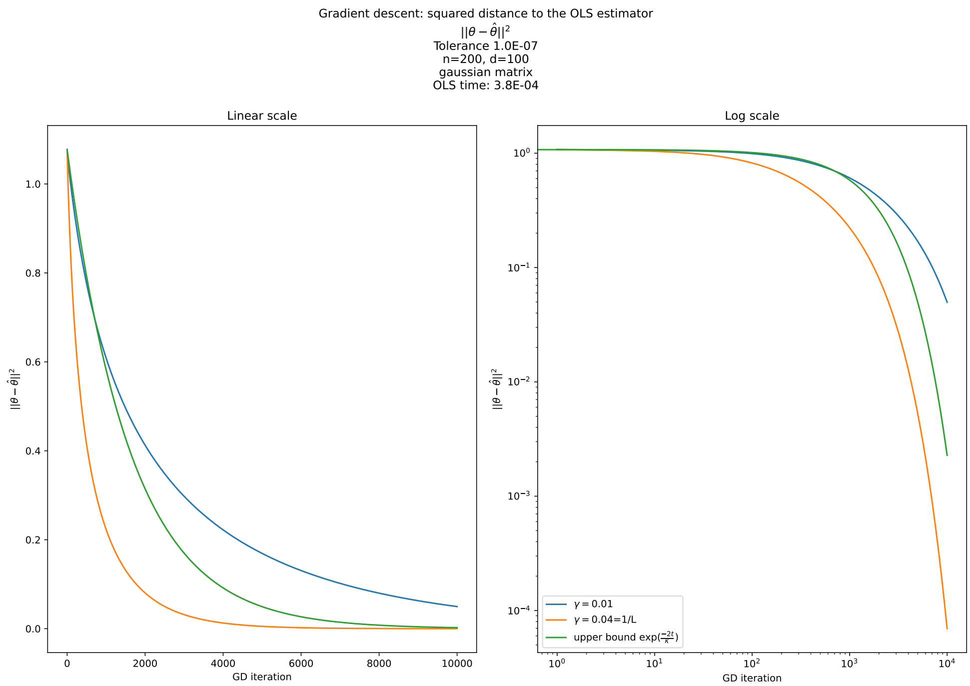

Figure 1 – Comparison of the upper bound and of the actual convergence speed.