FTML practical session 16

14 juin 2025

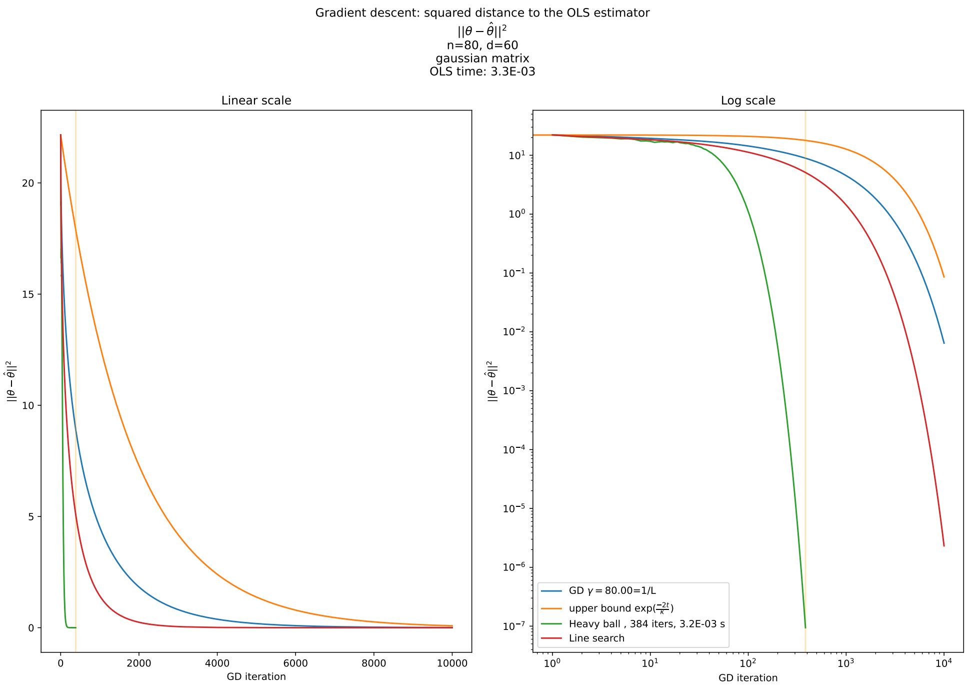

TABLE DES MATIÈRES

1 THE HEAVY-BALL METHOD

1.1 Convergence rates of GD for convex functions

We consider the optimization of a convex function $f : \theta \to f(\theta)$ using a gradient descent (GD). In particular, we consider the **convergence speed** of GD. This speed can be expressed in several manners. For instance, as the distance between the iterate θ_t and a minimizer θ^* (of course assuming that this minimizer exists and is unique), as a function of the iteration number t. It if possible to show the following results for two-times differentiable convex functions :

— if H is invertible ($\mu > 0$), we have a convergence rate in $\exp(-\frac{2t}{\kappa})$.

— if H is not invertible ($\mu = 0$), we have a convergence rate in $\mathcal{O}(\frac{1}{t})$ (probably one of the exercises of the project).

These rates are speed **upper bounds**, meaning that the convergence is at least as fast as those. Also note that these rates of convergence require the use of specific values of the learning rate γ, like for instance $\frac{1}{L}$ (but other values might also be used, depending on the context). We also see that these rates depend on the **condition number** of the Hessian H. If we note μ the smallest eigenvalue of the Hessian H, and L the largest, and if this Hessian is for instance symetric and definite positive, then

$$\kappa = \frac{L}{\mu} \tag{1}$$

However, the condition number might be defined also for general matrices, and even functions.
https://en.wikipedia.org/wiki/Condition_number

1.1.1 *Large condition numbers*

Hence, when κ is very large ($>> 1$), the convergence to the optimum might be very slow. Note that matrices with large condition numbers are not rare in large-scale machine learning and scientific computing applications. If the smallest eigen value μ of a given matrix H is very small, or even 0 (which will happen as soon as H is not full rank), κ will be very large as soon as the largest eigenvalue L is not also very small.

1.1.2 *Inertial methods*

When κ is large, some methods still exist in order to speed the convergence of gradient descent, such as **Heavy-ball.** This method consists in adding a **momentum term** to the gradient update term, such as the iteration now writes

$$\theta_{t+1} = \theta_t - \gamma \nabla_\theta f(\theta_t) + \beta(\theta_t - \theta_{t-1}) \tag{2}$$

where β and γ are real constants that should be tuned. The update $\theta_{t+1} - \theta_t$ is then a combination of the gradient $\nabla_\theta f(\theta_t)$ and of the previous update $\theta_t - \theta_{t-1}$. The goal of this method it to balance the effet of oscillations in the gradient. The heavy-ball method is called an **inertial method.** When f is a general convex function (not necessary quadratic), some generalizations exist, such as **Nesterov acceleration.** Many of the most famous variations of SGD, like RMSProp and Adam, optionally include such a momentum term.

1.2 Simulation

Using **exercise_1_heavy_ball/main.py**, try to manually find γ and β values such that the convergence is faster with the Heavy-Ball method than with a standard GD.

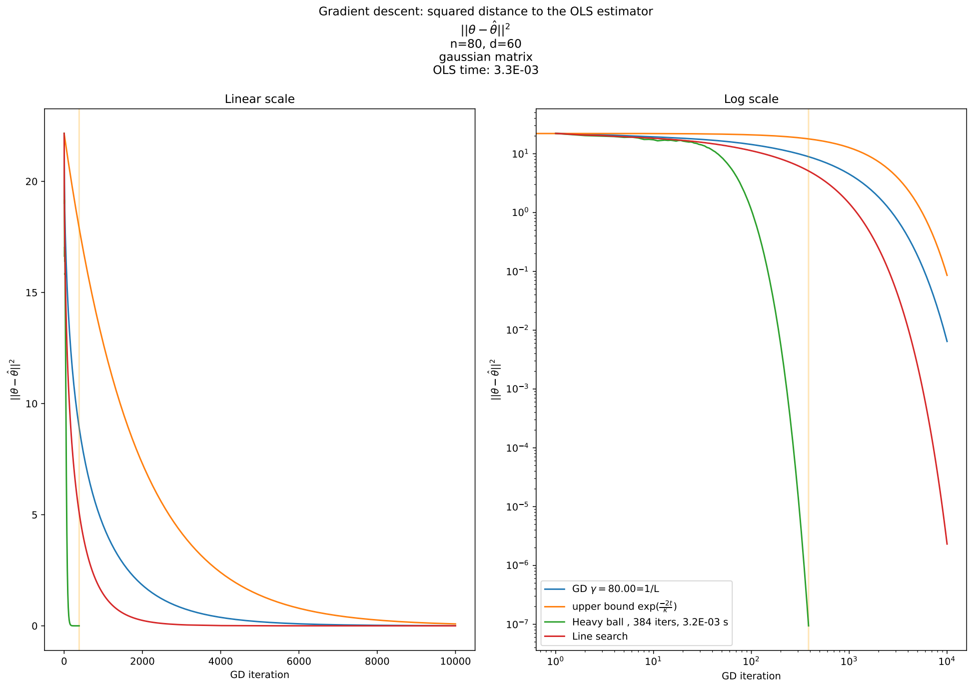

FIGURE 1 – Heavy ball vs GD, semilog scale

1.3 Analytical proof of γ and β values

Assuming $\mu > 0$, in a least squares problem, it is possible to show that the characteristic convergence time with the heavy-ball momentum term is $\sqrt{\kappa}$ instead of κ, **if** β **and** γ **are tuned well**. With the heavy-ball momentum term, we can change the convergence (upper bound) from $\mathcal{O}(\exp(-\frac{2t}{\kappa}))$ to $\mathcal{O}(\exp(-\frac{2t}{\sqrt{\kappa}}))$. If κ is large, which is the case we are interested in, this can be a significant improvement.

The update $\theta_{t+1} - \theta_t$ is then a combination of the gradient $\nabla_\theta f(\theta_t)$ and of the previous update $\theta_t - \theta_{t-1}$. This method might balance the effet of oscillations in the gradient. We will use these parameters :

$$\gamma = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} \tag{3}$$

and

$$\beta = \Big(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\Big)^2 \tag{4}$$

We keep the same notations as in the former practical sessions dedicated to the least squares problem. μ is the smallest eigenvalue of the Hessian H and L is the largest. Assuming $\mu > 0$ (strongly convex function), we will show that the characteristic convergence time with the heavy-ball momentum term is $\sqrt{\kappa}$ instead of κ.

Let λ be an eigenvalue of H and u_λ a eigenvector for thie eigenvalue. We are interested in the evolution of $\langle \theta_t - \eta^*, u_\lambda \rangle$.

We note

$$a_t = \langle \theta_t - \eta^*, u_\lambda \rangle \tag{5}$$

Exercice 1 : Show that

$$a_{t+1} = (1 - \gamma\lambda + \beta)a_t - \beta a_{t-1} \tag{6}$$

Exercice 2 : Compute the constant-recursive sequence a_t, and show that there exists a constant C_λ that depends on the initial conditions, such that

$$\forall t, a_t \leqslant (\sqrt{\beta})^t C_\lambda \tag{7}$$

https://en.wikipedia.org/wiki/Constant-recursive_sequence
If u_i is a basis of orthonormal vectors with eigenvalues λ_i, we have that

$$\begin{aligned} \|\theta_t - \eta^*\|^2 &= \sum_{i=1}^{d} (\langle \theta_t - \eta^*, u_i \rangle)^2 \\ &\leqslant \sum_{i=1}^{d} (\sqrt{\beta})^{2t} C_{\lambda_i} \\ &= (\sqrt{\beta})^{2t} D \end{aligned} \tag{8}$$

with

$$D = \sum_{i=1}^{d} C_{\lambda_i} \tag{9}$$

We can now remark that

$$\begin{aligned} \sqrt{\beta} &= \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \\ &= \frac{1 - \sqrt{\frac{\mu}{L}}}{1 + \sqrt{\frac{\mu}{L}}} \\ &\leqslant 1 - \sqrt{\frac{\mu}{L}} \\ &= 1 - \frac{1}{\sqrt{\kappa}} \end{aligned} \tag{10}$$

Exercice 3 : Conclude