PTML 4: 15/04/2022

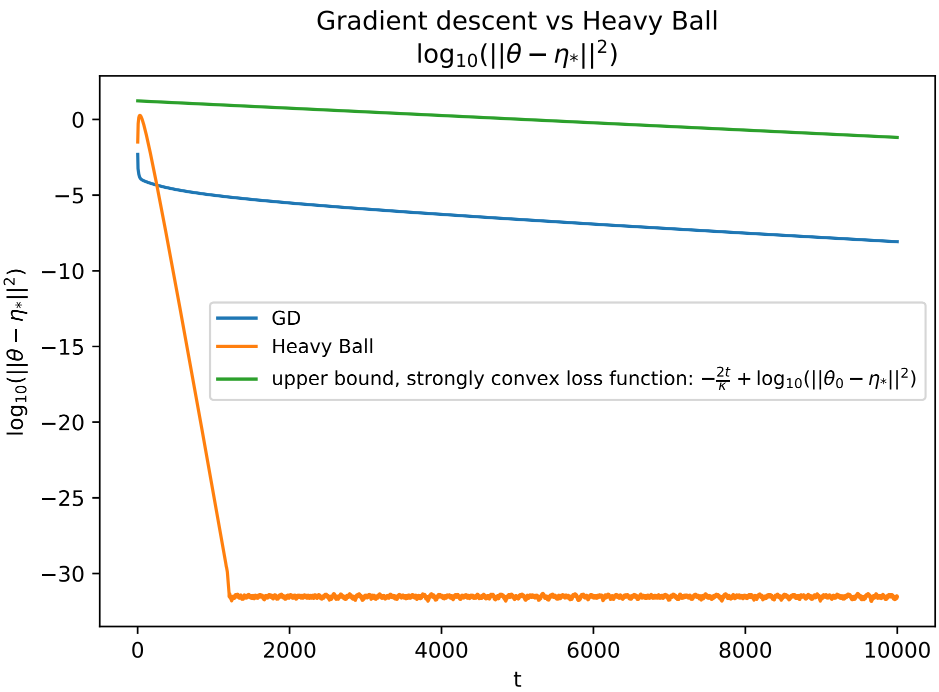

TABLE DES MATIÈRES

1 GRADIENT DESCENT ON A LEAST-SQUARES PROBLEM

In this section the setting is the same as in the first section of TP3.

1.1 The heavy-ball method

When κ is very large, the convergence might become very slow. Some methods exist in order to speed it up, such as **Heavy-ball.** This method consists in adding a **momentum term** to the gradient update term, such as the iteration now writes

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

The update $\theta_{t+1} - \theta_t$ is then a combination of the gradient $\nabla_{\theta_t} f$ and of the previous update $\theta_t - \theta_{t-1}$. The goal of this method it might balance the effet of oscillations in the gradient.

We will use these parameters :

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

and

$$\beta = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2 \tag{3}$$

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.**

1.1.1 *Impact on convergence rate*

Assuming $\mu > 0$, 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

$$a_t = \langle \theta_t - \eta^*, u_\lambda \rangle \tag{4}$$

Exercice 1 : Show that

$$a_{t+1} = (1 - \gamma\lambda + \beta) a_t - \beta a_{t-1} \tag{5}$$

Exercice 2 : Compute the constant-recursive sequence a_t, and show that there exists a constant C_λ that depends on the initial conditions (trough A and B, and a_0), such that

$$\forall t, a_t \leqslant (\sqrt{\beta})^t C_\lambda \tag{6}$$

https://en.wikipedia.org/wiki/Constant-recursive_sequence

If u_i is a basis of orthogonal normed vectors with eigenvalues λ_i, we have that

$$\begin{aligned} ||\theta_t - \eta^*||^2 &= \sum_{i=1}^{d} (\langle \theta_t - \eta^*, u_i \rangle)^2 \\ &\leqslant \sum_{i=1}^{d} (\sqrt{\beta})^{2t} C_{\lambda_i} \\ &= (\sqrt{\beta})^{2t} D \end{aligned} \tag{7}$$

with

$$D = \sum_{i=1}^{d} C_{\lambda_i} \tag{8}$$

We can now remark that

$$\begin{aligned}\sqrt{\beta} &= \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} \\ &= \frac{1-\sqrt{\frac{\mu}{L}}}{1+\sqrt{\frac{\mu}{L}}} \\ &\leqslant 1-\sqrt{\frac{\mu}{L}} \\ &= 1-\frac{1}{\sqrt{\kappa}}\end{aligned} \tag{9}$$

Finally, as $1-\frac{1}{\sqrt{\kappa}} \leqslant \exp(-\frac{1}{\sqrt{\kappa}})$,

$$||\theta_t - \eta^*||^2 = \mathcal{O}(\exp(-\frac{2t}{\sqrt{\kappa}})) \tag{10}$$

Conclusion : with the heavy-ball momentum term, we changed the convergence rate of $\mathcal{O}(\exp(-\frac{2t}{\kappa}))$ to a convergence rate of $\mathcal{O}(\exp(-\frac{2t}{\sqrt{\kappa}}))$. This means that characteristic convergence time went from κ to $\sqrt{\kappa}$. If κ is large, which is the case we are interested in, this can be a great improvement.

Remember that $\kappa = \frac{L}{\mu}$, and that μ may be very small when n or d is large. For instance, in the case of Ridge regression, we have seen in the previous session that for instance, μ can be of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ (see the computation of the optimal regularisation parameter). Hence, κ may be of order $\sqrt{n}$ or higher.

1.1.2 *Simulation*

Exercice 3 : Use the file **TP_3_GD_strongly_convex _heavy_ball.py** to implement the Heavy-ball method and compare the convergence speed results to that of GD. You should obtain something like figures 1 and 2.

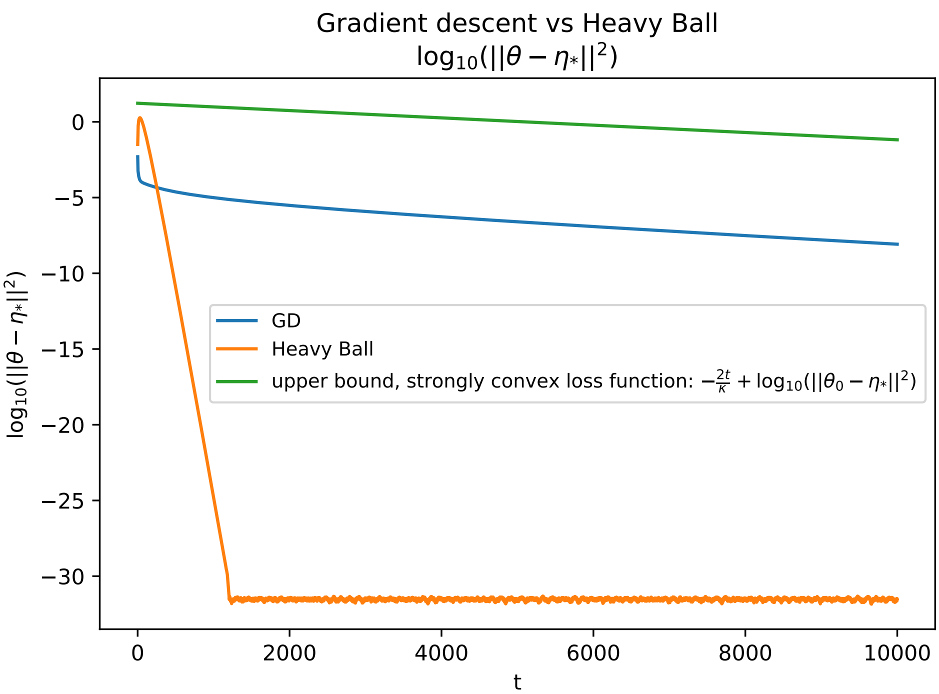

FIGURE 1 – Heavy-ball vs GD

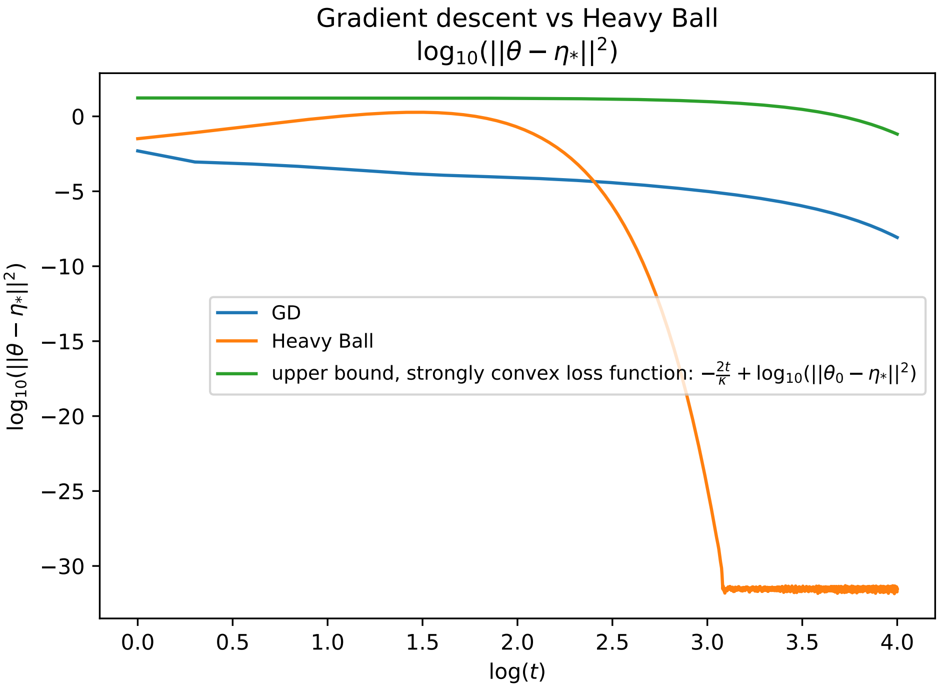

Figure 2 – Heavball vs GD, logarithmic scale

Références