Introduction to Accelerated gradient method

Yuxuan Zhou
FMT
February 8th 2018

Contents

Introduction

Why first-order methods?

- Fast and suitable for large scale data optimization with relatively low accuracy requirement, e.g. Machine Learning,Signal Processing......
- example

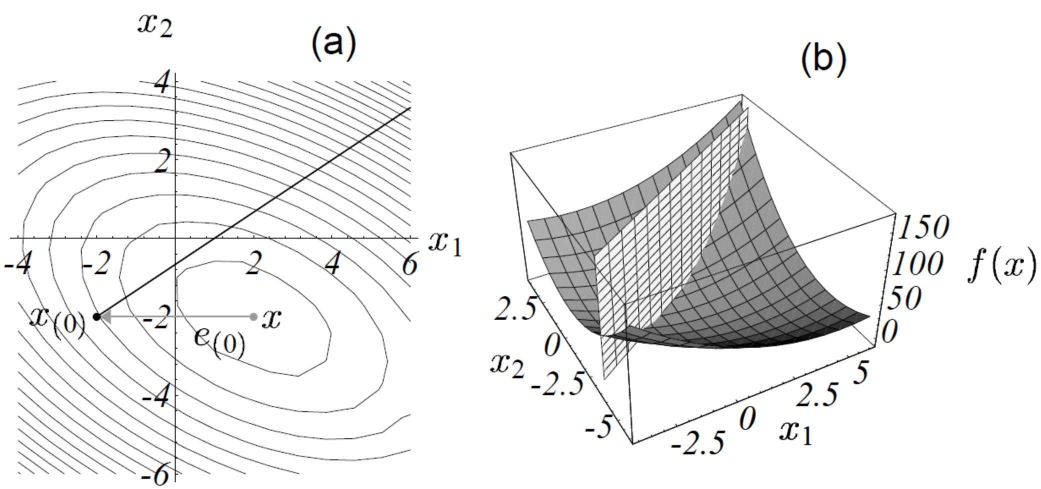
$$f(x) = x^T Ax + bx + c$$

$$\nabla f(x) = Ax + b$$

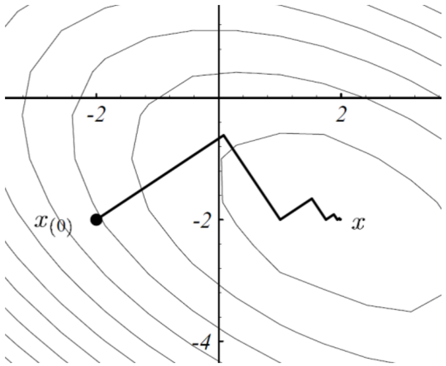
$$\nabla^2 f(x) = A$$

Gradient Descent Method
Algorithm

given *a starting point* $x_0 \in$ ***dom*** f
repeat
step $:\Delta x_k = -\nabla f(x_k).$
Line search. Choose a step size t.
Update. $x_{k+1} = x_k - t\nabla f(x)$
until *Stopping criterion is satisfied.*



Zigzag Phenomenon

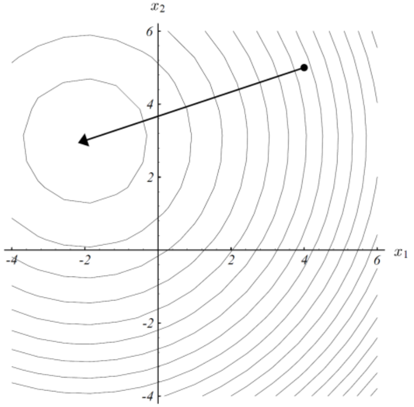


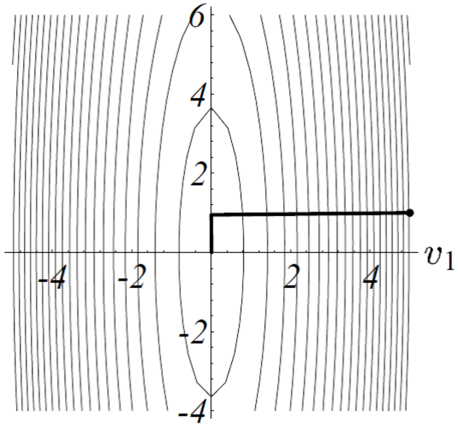
Deciding Factors

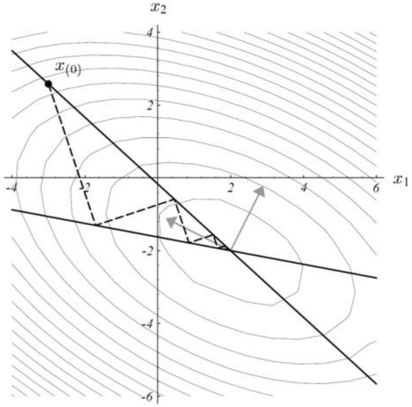
- Condition number:

$$\kappa = \frac{\sigma_{max}}{\sigma_{min}}$$

- Starting point



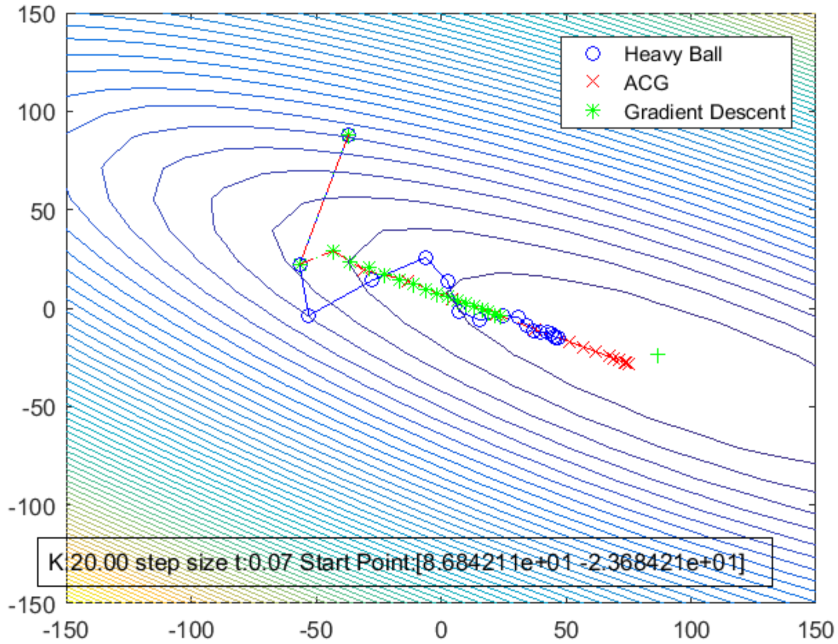




Introduction

Momentum Methods(Heavy Ball and Accelerated Gradient Descent)

- A momentum term is introduced, in order to help escaping from the ‚Valley'
- Momentum Methods are no longer descent methods, since the function values are not monotonically decreasing



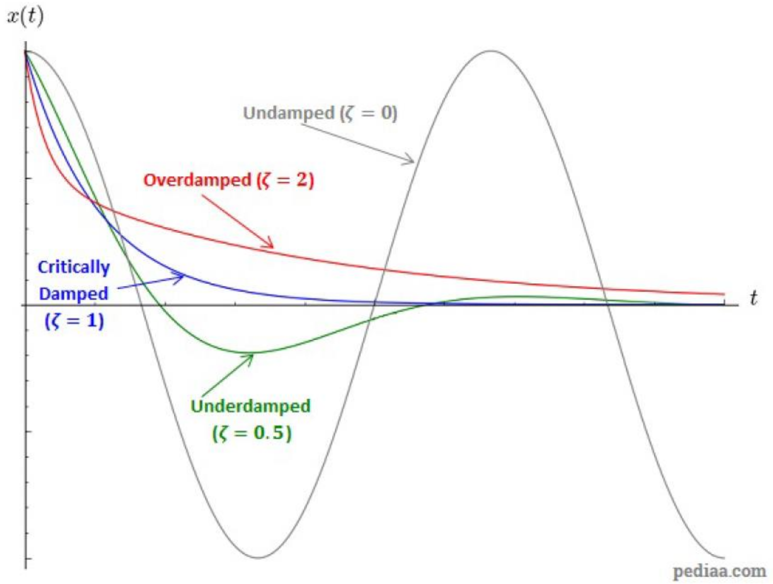
Math Concepts

Heavy Ball Method

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})$$

Comparing to second order ODE in continuous time

$$\beta \ddot{x} + a\dot{x} + b\nabla f(x) = 0$$



$$c_c = 2\sqrt{km}$$

$$\zeta = \frac{c}{c_c},$$

$$\zeta = \frac{\text{actual damping}}{\text{critical damping}},$$

Math Concepts

Convergence of iterative methods

Consider the case 2 by 2 matrix:

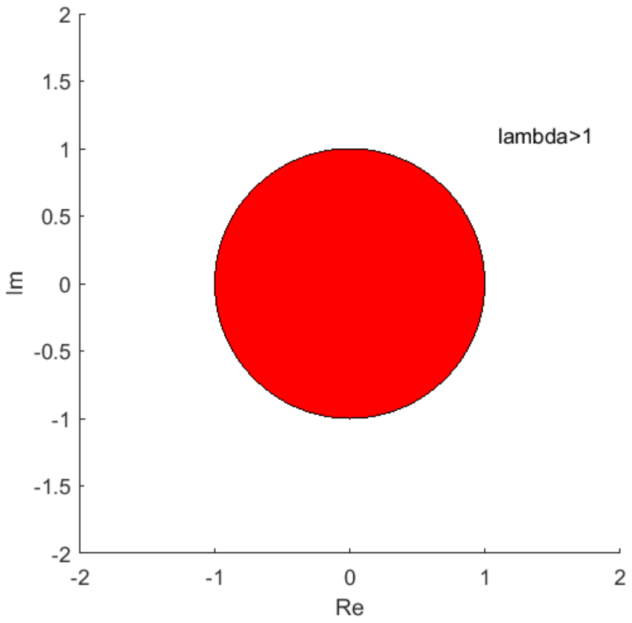
$$x_{k+1} = Bx_k$$

$$x_0 = av_1 + bv_2$$

After k iterations:

$$x_k = B^k x_0$$

$$B^k = a||\lambda_1||^k v_1 + b||\lambda_2||^k v_2$$



Math Concepts

-Lipschitz continuity of the gradient function:

$$||\nabla f(y) - \nabla f(x)||_2 \leq L||y - x||_2$$

-Strong convexity:

$$f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}||y - x||^2 \leq f(y)$$

$$\Big\downarrow$$

$$\mu I \leq \nabla^2 f(x) \leq LI \quad = \quad \begin{array}{c} \sigma_{min}(\nabla^2 f(x) \geq \mu \\ \sigma_{max}(\nabla^2 f(x)) \leq L \end{array}$$

Math Concepts

Heavy Ball Methods

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})$$

It can be transformed into a matrix form:

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ x_k - x^* \end{bmatrix} \right\|_2 =$$

$$\left\| \begin{bmatrix} x_k + \beta(x_k - x_{k-1}) - x^* \\ x_k - x^* \end{bmatrix} - \alpha \begin{bmatrix} \nabla f(x) \\ 0 \end{bmatrix} \right\|_2$$

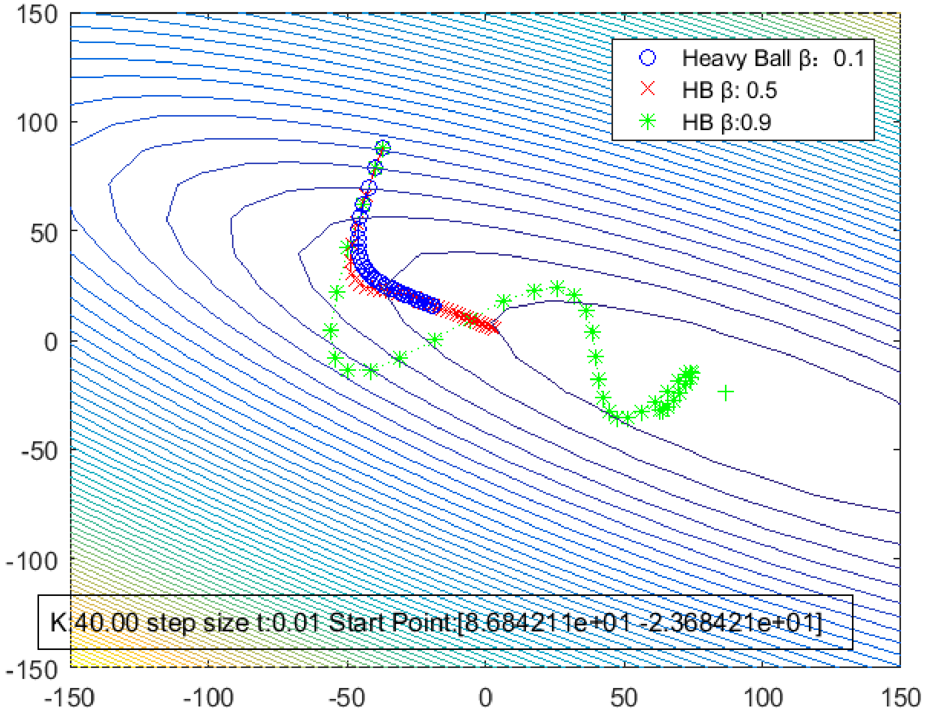
After linearlisation:

$$= \left\| \begin{bmatrix} (1+\beta)I - \alpha \nabla^2 f(z_k) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \end{bmatrix} \right\|$$

Heavy Ball Methods

Influence of the momentum term

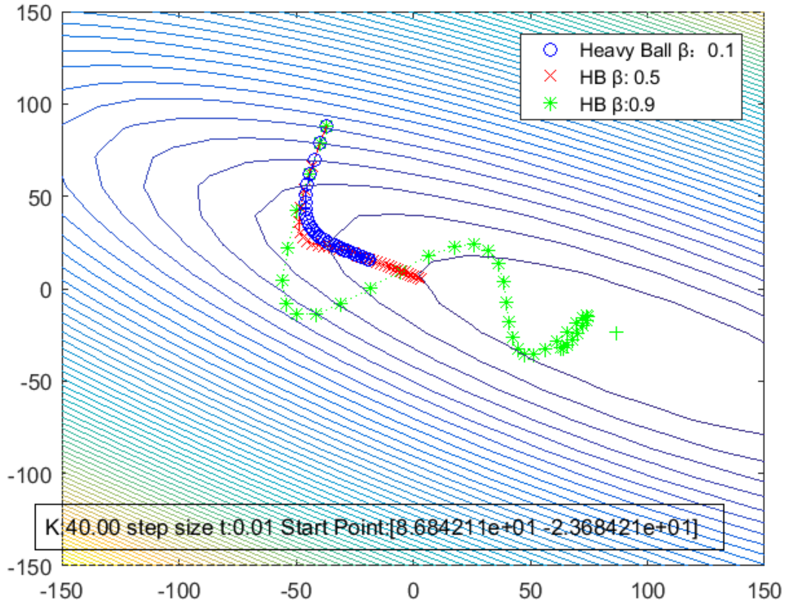
- Momentum increases the convergence rate while causing the overshooting.

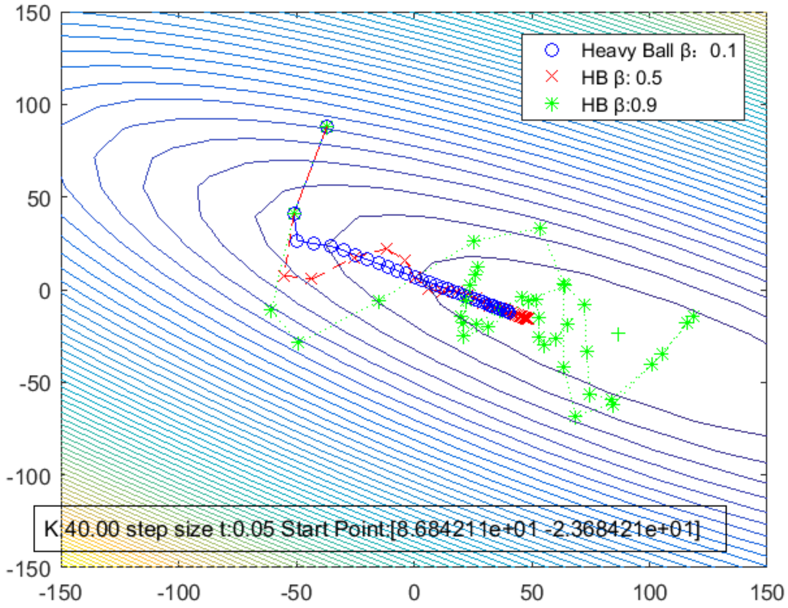


Heavy Ball Methods

Optimal value of β *and* α *are not easy to estimate*

- trade-off between step size and momentum
- By larger step size, it diverges first with a higher momentum value.





Accelerated Gradient Method

Algorithm 2(basic ACG)

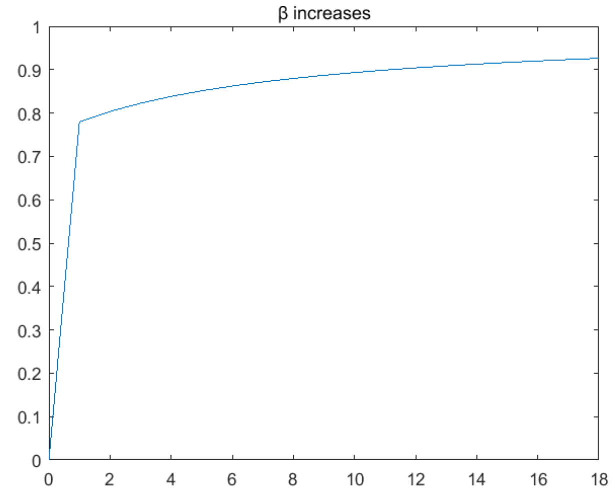
$$Set\ \gamma_0 = 1, x_1 = x_0 - t\,\nabla f(x_0)$$

$$y_{k+1} = x_k + \beta_k(x_k - x_{k-1})$$

$$x_{k+1} = y_{k+1} + -t_k \nabla f(y_k)$$

$$\gamma_k = \frac{1}{2}(4\gamma_{k-1}^2 + \gamma_{k-1}^4)^{\frac{1}{2}} - \gamma_{k-1}^2$$

$$\beta_k =_k (1 - \gamma_{k-1}^{-1})$$

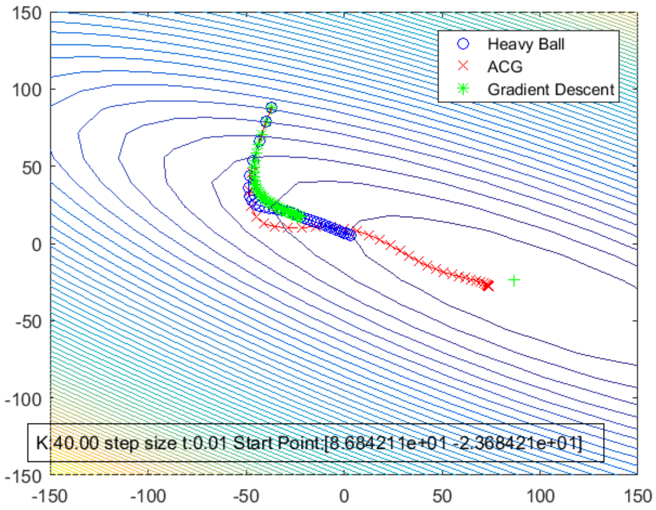


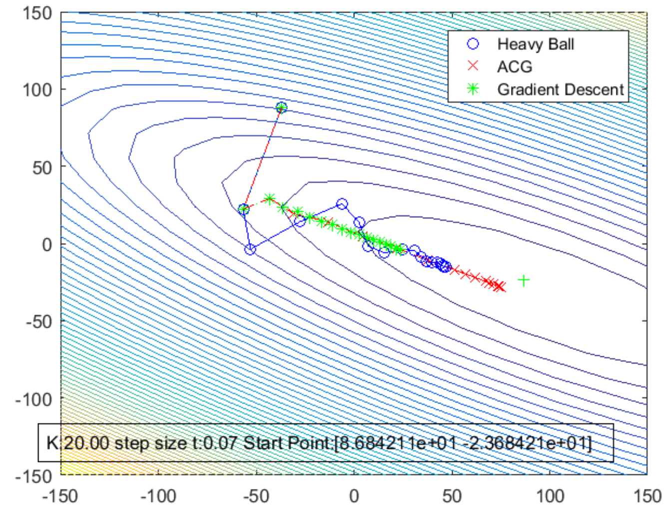
- The coefficient of the momentum term is a function instead of a constant
- The gradient is evaluated at the intermediate point y_k

Accelerated Gradient Mehtod

Comparison of the convergence rate

- With a small enough step size, Nesterov's method converges fastest, followed by Heavy Ball method and Gradient Desent, and its performance is more stable at difference step sizes

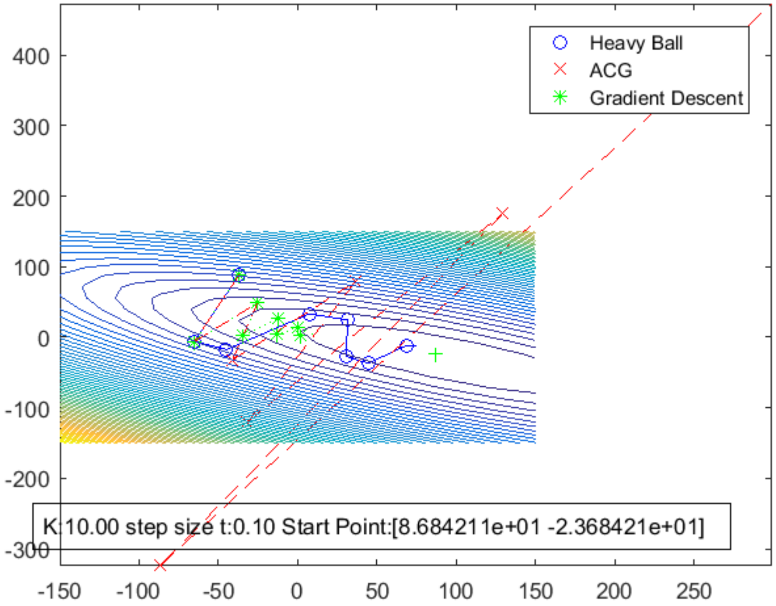




Accelerated Gradient Method

Step size choice

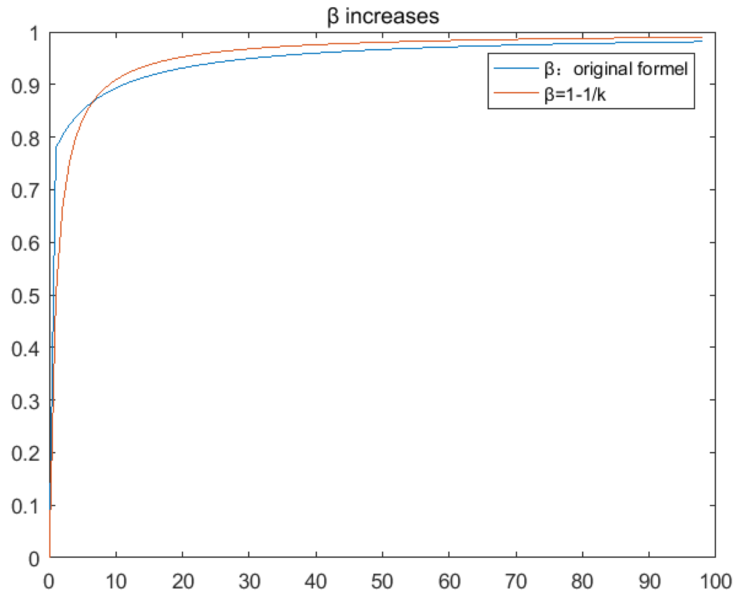
- Notably, ACG and HB with high β is more sensitive to the step size. When the step size is larger, it begins to diverge, whereas the other two methods stays stable.(β of Heavy Ball Method is set to a high value:0.7)

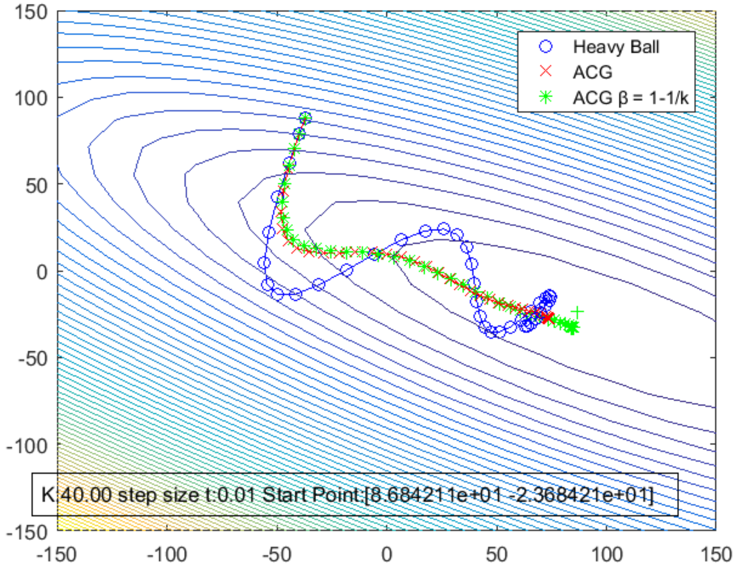


Accelerated Gradient Method

Function of β

- Substitute the original formel with $\beta = 1 - 1/k$

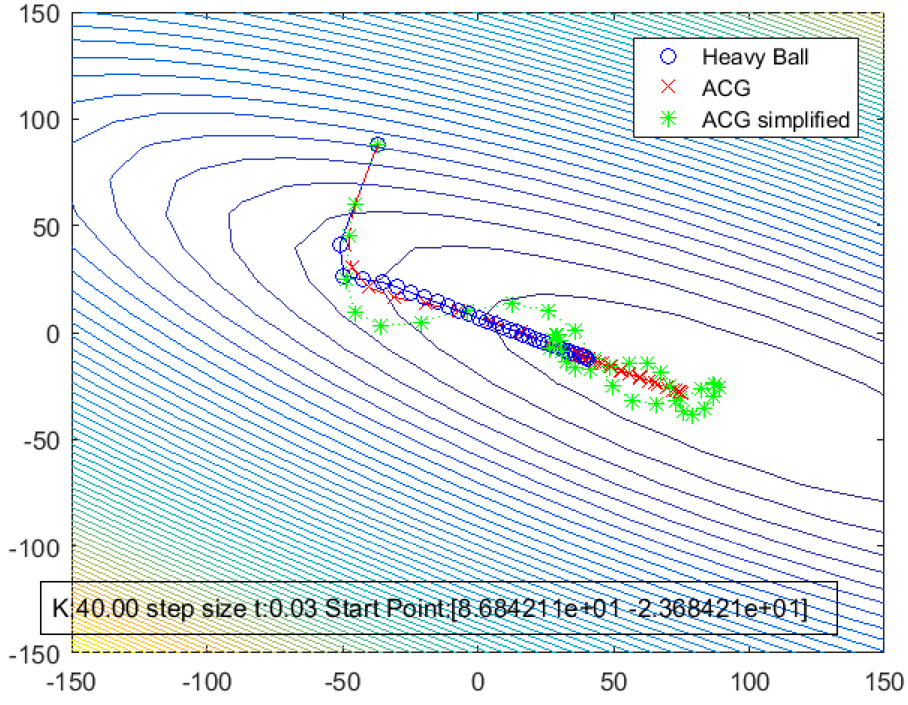




Accelerated Gradient Method

Gradient of the intermediate point

- Calculating the gradient at the intermediate point seems to guarantee the convergence



End

Thanks for attention!