Lecture 23 : Accelerating Gradient Descent (Use momentum)

- Gradient descent $\quad x_{k+1} = x_k - s\nabla f(x_k)$
- Accelerated gradient descent
  1. momentum added
  2. Nesterov formula
- Stochastic gradient descent

Momentum

$$\boxed{\begin{aligned} x_{k+1} &= x_k - s z_k \\ z_k &= \nabla f_k + \beta z_{k-1} \end{aligned}} \quad \rightarrow \quad \boxed{\begin{aligned} x_{k+1} &= x_k - s z_k \\ z_{k+1} - S x_{k+1} &= \beta z_k \end{aligned}}$$

$f = \frac{1}{2} x^T S x$

$\nabla f = Sx$

$$\begin{bmatrix} 1 & 0 \\ -S & 1 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}_k$$

$Sq = \lambda q$

Suppose $x_k = c_k q$

x_k is tracking eigenvector

$z_k = d_k q$

$S x_k = c_k \lambda q$

$\downarrow$

∇f_k

$x_k \qquad\qquad x_{k+1}$

$$\begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix} \begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} \cancel{q} = \begin{bmatrix} -s & 1 \\ \beta & 0 \end{bmatrix} \begin{bmatrix} d_k \\ c_k \end{bmatrix} \cancel{q}$$

$c_{k+1} q = c_k q - s\, d_k q$

$d_{k+1} q - \lambda c_{k+1} q = \beta d_k q$

$$\begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix} \begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

$$\begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -s \\ \lambda & \beta - \lambda s \end{bmatrix} \begin{bmatrix} c_k \\ d_k \end{bmatrix} = R \begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

$\begin{bmatrix} c_k \\ d_k \end{bmatrix}$ is multiplied by R every step

Simple, beautiful steps come from tracking one eigenvector, which makes the whole problem scalar.

$$[x \;\; y]\begin{bmatrix} 5 & 2 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

$$[x \;\; y]\begin{bmatrix} 5x+2y \\ 2x+3y \end{bmatrix} = 5x^2+4xy+3y^2$$

Make R as small as possible for fast convergence

How to choose λ ?

$$0 < m \leq \lambda \leq M$$

eigenvalue of S

$$\frac{M}{m} = K = \text{condition \# of } S$$

e_1, e_2 = eigenvalues of R

depends on eigenvalues of S

* Choose s and β to minimize $\max[\,|e_1(\lambda)|, |e_2(\lambda)|\,]$

for $\lambda_{min}(S) \leq \lambda \leq \lambda_{max}(S)$

$$m \leq \lambda \leq M$$

It turned out that:

$$s_{optimal} = \left(\frac{2}{\sqrt{M}+\sqrt{m}}\right)^2 = \left(\frac{2}{1+\sqrt{b}}\right)^2$$

$$\beta_{optimal} = \left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^2 = \left(\frac{1-\sqrt{b}}{1+\sqrt{b}}\right)^2$$

for 2x2 example $S=\begin{bmatrix} 1 & 0 \\ 0 & b \end{bmatrix}$

$$\left|\text{eigenvalues of } R\right| < \left(\frac{1-\sqrt{b}}{1+\sqrt{b}}\right)^2$$