THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 385: Stat Models for Big Data

Lecture 4: GD with momentum.

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

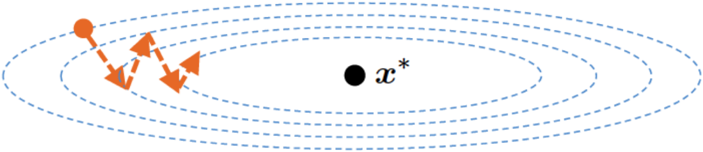
https://psarkar.github.io/teaching

Polyak's heavy ball method

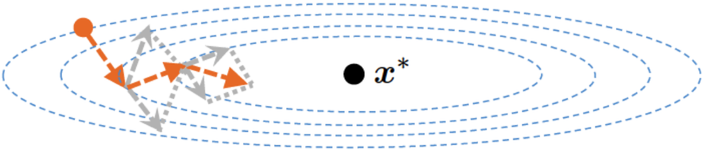
Figure 1: B. Polyak

- $$\beta_{t+1} = \beta_t - \alpha \nabla f(\beta_t) + \underbrace{\theta(\beta_t - \beta_{t-1})}_{\text{momentum term}}$$

Momentum



gradient descent



heavy-ball method

Recall GD?

- For a L smooth and μ convex optimization problem, i.e. $\mu I \preceq \|H\| \preceq LI$,

$$\|\beta_t - \beta^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\beta_0 - \beta^*\|$$

  where $\kappa = L/\mu$ i.e. the condition number of the Hessian.
- For the same problem, using Polyak's method we can show that,

$$\left\| \begin{bmatrix} \beta_{t+1} - \beta^* \\ \beta_t - \beta^* \end{bmatrix} \right\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^t \left\| \begin{bmatrix} \beta_1 - \beta^* \\ \beta_0 - \beta^* \end{bmatrix} \right\|$$

Momentum method

- Recall we have:

$$\begin{aligned}\beta_{t+1} - \beta^* &= (1+\theta)(\beta_t - \beta^*) - \alpha \nabla f(\beta_t) - \theta(\beta_{t-1} - \beta^*) \\ &= ((1+\theta)I - \alpha \nabla^2 f(z_t))(\beta_t - \beta^*) - \theta(\beta_{t-1} - \beta^*)\end{aligned}$$

- This gives the dynamic system:

$$\begin{bmatrix} \beta_{t+1} - \beta^* \\ \beta_t - \beta^* \end{bmatrix} \leq \begin{bmatrix} (1+\theta)I - \alpha \nabla^2 f(z_t) & -\theta I \\ I & 0 \end{bmatrix} \begin{bmatrix} \beta_t - \beta^* \\ \beta_{t-1} - \beta^* \end{bmatrix}$$

Momentum method

- We need to upper bound the norm of
  $M := \begin{bmatrix} (1+\theta)I - \alpha\nabla^2 f(z_t) & -\theta I \\ I & 0 \end{bmatrix}$
- It can be shown that:

$$\begin{aligned} \|M\| &= \left\| \begin{bmatrix} (1+\theta) - \alpha\Lambda & -\theta I \\ I & 0 \end{bmatrix} \right\| \\ &= \max_i \left\| \begin{bmatrix} (1+\theta) - \alpha\lambda_i & -\theta \\ 1 & 0 \end{bmatrix} \right\| \end{aligned}$$

- Eigenvalues of the 2×2 matrix can be written as a solution of the following quadratic:

$$\sigma^2 - \sigma((1+\theta) - \alpha\lambda_i) + \theta = 0$$

Momentum method

- If $((1+\theta)-\alpha\lambda_i)^2 \leq 4\theta$, the roots are imaginary and the magnitude is $\sqrt{\theta}$
- This is satisfied if

$$\theta \in [(1-\sqrt{\alpha\lambda_i})^2, (1+\sqrt{\alpha\lambda_i})^2]$$

- But recall that $\lambda_i \in [\mu, L]$.
- If we set $1-\sqrt{\alpha L} = -(1-\sqrt{\alpha\mu})$, then we have

$$\alpha = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2 \qquad \theta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$$

- So the new contraction factor becomes $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

Nesterov's Accelerated Gradient

- If we only assume that $\|\nabla^2 f(x)\| \leq L$ and not strong convexity, then in your homework you will prove that

$$f(\beta_t) - f(\beta^*) \leq c_L \frac{\|\beta_0 - \beta^*\|^2}{t}$$

- Note that this is much weaker than the linear convergence we saw before.
- Question is can we do better?

Nesterov's Accelerated Gradient

Figure 2: Y. Nesterov

- Keep track of two vectors x_t and y_t
- $x_{t+1} = y_t - \alpha_t \nabla f(y_t)$
- $y_{t+1} = x_{t+1} + \underbrace{\frac{t}{t+3}}_{\mu_{t+1}}(x_{t+1} - x_t)$

Nesterov's Accelerated Gradient

- Can be re-written as:

$$x_{t+1} = x_t + \mu(x_t - x_{t-1}) - \alpha_t \nabla f(x_t + \mu_t(x_t - x_{t-1}))$$

- Very much like the momentum method, but computes the derivative at a future step.

Nesterov's Accelerated Gradient

Nesterov's Accelerated Gradient

Nesterov's Accelerated Gradient

Momentum method

Momentum method

Acknowledgment

Y. Chen's large scale Optimization class at Princeton.