Variations on Stochastic Gradient Descent

Computational Statistics

Johan Larsson

Department of Mathematical Sciences, University of Copenhagen

October 15, 2024

Stochastic Gradient Descent

Last Time
Introduced stochastic gradient descent (SGD) and mini-batch version thereof.

Algorithm 1: Mini-Batch SGD

Data: $\gamma_0 > 0$

repeat

$A_k \leftarrow$ random mini-batch of m samples;

$x_k \leftarrow x_{k-1} - \frac{\gamma_k}{|A_k|} \sum_{i \in A_k} \nabla f_i(x_{k-1})$;

until *convergence*;

Stochastic Gradient Descent

Last Time
Introduced stochastic gradient descent (SGD) and mini-batch version thereof.

Problems
We indicated that there were problems with vanilla SGD: poor convergence, erratic behavior.

Algorithm 1: Mini-Batch SGD

Data: $\gamma_0 > 0$

repeat

$A_k \leftarrow$ random mini-batch of m samples;

$x_k \leftarrow x_{k-1} - \frac{\gamma_k}{|A_k|} \sum_{i \in A_k} \nabla f_i(x_{k-1})$;

until *convergence*;

Today

How can we improve stochastic gradient descent?

Momentum
Base update on combination of gradient step and previous point.

Today

How can we improve stochastic gradient descent?

Momentum
Base update on combination of gradient step and previous point.

Adaptive Gradients
Adapt learning rate to particular feature.

Momentum

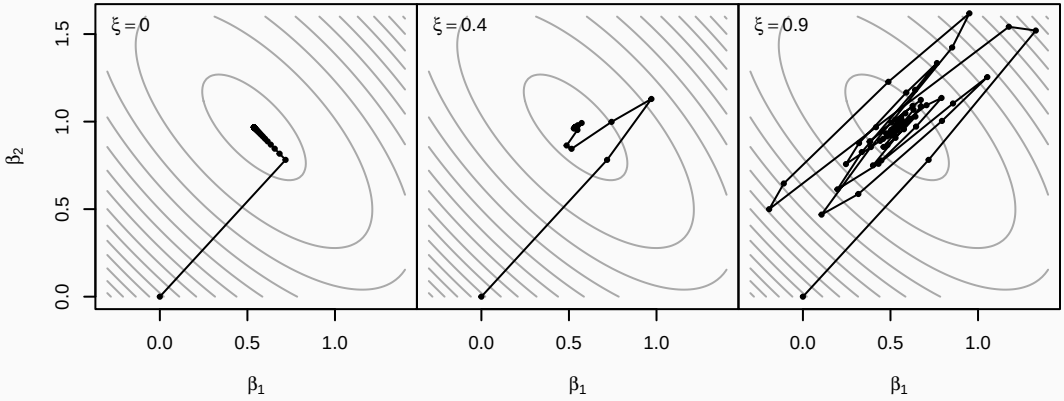

Figure 1: Trajectories of GD for different momentum values