Session 13: Stochastic gradient descent

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Contents

1. Gradient descent
2. Convergence analysis for convex functions
3. Improvements

Stochastic gradient descent

Setting

In machine learning, one often has to minimize functions of the form

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x).$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$.

Setting

Stochastic gradient descent

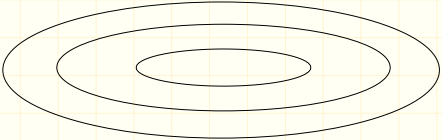
$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x).$$

Starting at some $x_0 \in \mathbb{R}^n$, perform the updates:

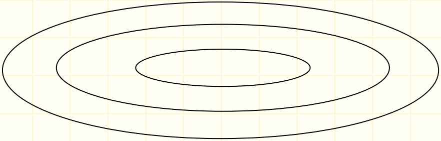
$$\begin{array}{lll} \text{Pick} & i & \text{uniformly at random in} \quad \{1, \ldots, N\}, \\ \text{Update} & \multicolumn{2}{l}{x_{t+1} = x_t - \alpha_t \nabla f_i(x_t),} \end{array}$$

Tradeoffs in SGD

Rapidly decaying step sizes



Slowly decaying step sizes



Convergence analysis

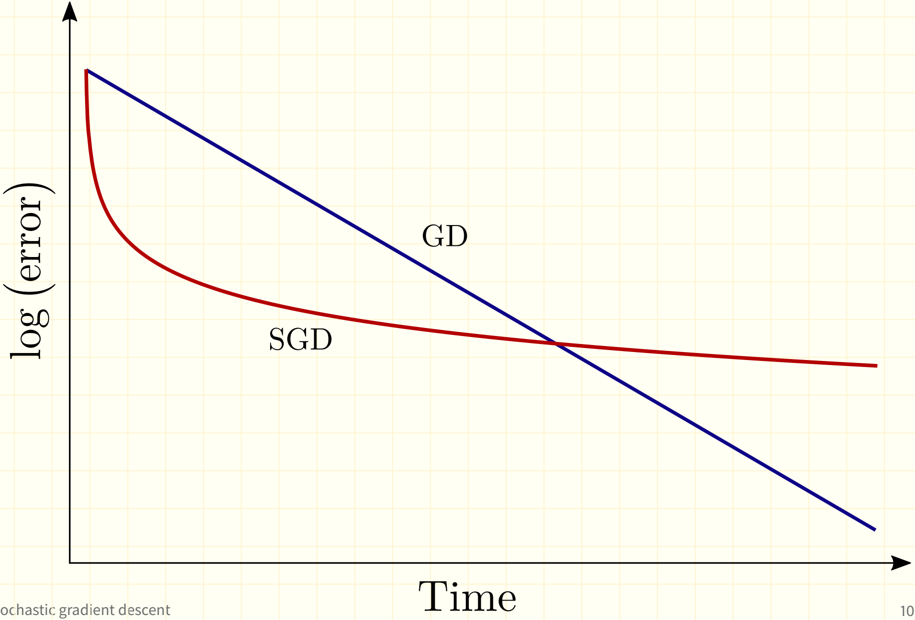
Convergence analysis

GD vs SGD

Gradient descent

Stochastic gradient descent

GD vs SGD



Questions?

Questions?