Variance reduction for stochastic gradient methods

Axel Böhm

September 10, 2021

1 Introduction

A common Task in (supervised) machine learning:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{\text{loss for } i\text{-th sample}} + \underbrace{\psi(x)}_{\text{regularizer}}$$

where the i-th sample is (a_i, y_i).

- linear regression: $f_i(x) = (a_i^T x - y_i)^2$, and $\psi = 0$
- logistic regression: $f_i(x) = \log(1 + e^{-y_i a_i^T x})$, and $\psi = 0$ "sigmoid function" and logistic loss.
- Lasso: f_i as for linear regression but $\psi(x) = \|x\|_i$
- SVM: $f_i(x) = \max\{0, 1 - y_i a_i^T x\}$ and $\psi(x) = \|x\|^2$

Stochastic gradient descent

We already noticed that:

- large stepsizes fail to supress variability