Reviews of Stochastic Gradient Descent Variants and its Asychonous Version

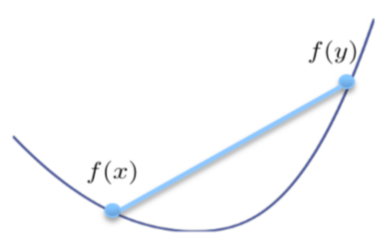
Introduction

In this review, we will introduce several at start-of-art gradient descent optimization algorthims.In full gradient descent optimization, we compute the cost gradient based on the complete training set. In case of very large datasets, using full gradient descent can be quite costly since we are only taking a single step for one pass overIn full gradient descent optimization, we compute the cost gradient based on the complete training set. In case of very large datasets, using full gradient descent can be quite costly since we are only taking a single step for one pass over

Notation

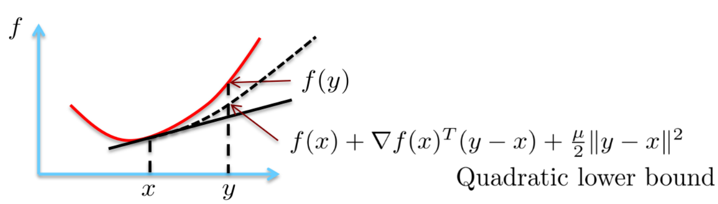
At first we introduce some notations about convex optimization.
Convex Function
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, if:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \forall x, y \in \mathbb{R}^n, \forall \alpha \in (0,1)$$



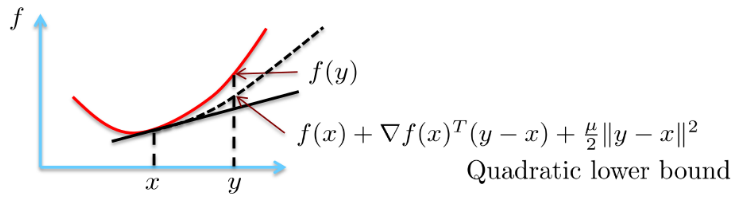
And it also means (Suppose f is differentiable)

$$f(y) \geq f(x) + \nabla f(x)^T(y-x), \forall x, y \in \mathbb{R}^n$$



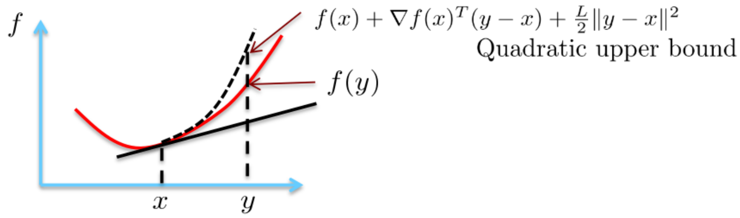
Strong Convexity
A differentiable function f is strongly convex if

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \forall x, y \in \mathbb{R}^n$$



Convex Function with Lipschitz Continuous
Let ∇f be Lipschitz continuous, ie., there exitst $L \geq 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \forall x, y \in \mathbb{R}^n$$



Gradient Descent

Update Rule
Convergence Rate

Stoastic Gradient Descent

Stoastic Average Gradient Descent(SAG)

Stoastic Gradient Descent with Predictive Variance Reduction(SVRG)