On Gradient Descent
and its Variants

Ozaner Hansha

March 27, 2021

Abstract

We examine the gradient descent (GD) algorithm, with respect to arbitrary real functions $f : \mathbb{R}^n \to \mathbb{R}$. In particular, we elaborate on its formulation, complexity, and convergence conditions. We then introduce and do the same with two variants of GD, namely stochastic gradient descent (SGD) and Adam.

Background

In the wake of the huge popularity afforded to machine learning... blah blah the combination of GD and backpropagation has become the 'workhorse of machine learning.' (find that quote). Or, more accurately, SGD and backpropagation. Or even *more* accurately, Adam and backpropagation. Then blah blah get into how these are modifications of GD to be more efficient in some dimension or other for ML applications.

Gradient Descent

Overview

Here I'll give an intuitive overview of the algorithm as basically descending a (high-dimensional) 'hill'. Use pictures (cant put gifs in a pdf) of an arrow traveling down some 3d objective function. Make a passing note that to find this steepest decline (ie. gradient), our objective function must be (sub)differentiable.

Formulation

Here I'll actually outline the preconditions of the algorithm, give the steps in a sort of math-ish, psuedocode, then provide an implementation in, say, python.[3] [1]

Complexity

Here we'll analyze its computational complexity, both spatial and temporal. If I can, I'll highlight how slow computing the gradient over all sample points is in a model fitting example which I can use to lead into SGD which only uses a random sample to calculate the gradient.

Convergence

Deal with convergence conditions here. From what I can tell, the only sure-fire convergence condition there is is that the objective function is convex. Talk about subgradients and box the final convergence results for both fixed and variable step sizes. [2]

Stochastic Gradient Descent

Overview

Intro what's changed and why it was changed. Maybe a nice graphic.

Same sections are before, but shorted since less needs to be introduced and mostly focuses on what's changed.[3] [1]

Adam

Overview

Intro what's changed and why it was changed. Maybe a nice graphic.

Same sections are before, but shorted since even less needs to be introduced and mostly focuses on what's changed.[3] [1]

References

[1] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv e-prints*, page arXiv:1609.04747, September 2016.

[2] Ryan Tibshirani. 10-725: Optimization, Sep 2013.

[3] Jiawei Zhang. Gradient Descent based Optimization Algorithms for Deep Learning Models Training. *arXiv e-prints*, page arXiv:1903.03614, March 2019.