Assignment#3

Name: Ahmed Hani Ibrahim

Subject: CS624

ADAM Optimizer vs. Gradient Descent Optimizer

Before we get starting with the differences between Gradient Descent and the newly developed approach ADAM that was presented in ICLR 2015 conference, we need to understand some concepts in machine learning. We will illustrate the meaning of Adaptive Learning that is used in ADAM.

**Machine Learning Generations**

Machine Learning has passed through multiple generations; each generation has its own metrics regarding the approach to learn a machine. We can conclude them in 4 generations, Rule-based systems, Simple Machine Learning, Deep Learning, Adaptive Learning. We will talk briefly about each one of them

**Rule-based systems**

A very naïve approach for learning that needs the human efforts to teach the machine some simple rules, and using these rules, the machine can reach to some other rules. For example, if we say that “Ahmed” has a bother called “Mohamed”, and we add that “Mohamed” has a brother called “Ibrahim”. Using inference and basic operations, the machine can easily say that “Ahmed” has a brother called “Ibrahim”. We can formulate the problem in the following.

* Given that hasBrother(“Ahmed”, “Mohamed”), hasBrother(“Mohamed”, “Ibrahim”), what is the relation between “Ahmed” and “Ibrahim”?
* Using inference, hasBrother(“Ahmed”, Brother(“Mohamed”))

This is a simple way to make the machine add something new in its knowledge. Regularly, this is represented in Ontology.  
  
The problem with this approach is that it is very naïve, and not applicable in today’s tasks. Applying such approach in systems will definitely causes poor accuracy.

**Simple Machine Learning**

A huge step towards modern machine learning. Unlike Rule-based systems, the machine can learn from the given data and increases its ability to predict a new given data. The data is represented as a bag of features that are fed into the classifier, using the features and its label, the classifier can update its parameter to increase its ability of predicting the new data.

This approach is named “Supervised Learning”, in which you have some labeled data and features, you train the classifier by them, then if there are new data, the machine can predict the label.

A simple example for that, when you say a Cat, you know that it is a Cat, because you have saw many cats before and know that the cats have specific shape and features, such as eye color, hair color, and other things you saw before in cats.

**Deep Learning**

A dominating field in machine learning in this era. Instead of getting the features from the data, you let the machine decides the features and extracts them from the data without human interaction.

An example for that, assume that you have a large data of cats images, and you don’t know what are the features that are very important to recognize that cats, so, you use deep learning to make the machine extracts the features by itself by feeding the data of cats, then the machine will get the features that heavily effects on cat’s recognition.

This approach is commonly used now because it saves the time of features engineering by humans.

**Adaptive Learning**

Although the great successes of Deep Learning, they have discovered that human’s interaction is needed to improve the accuracy of the learning approach. Instead of making the interaction is just input the data to the classifier, we need to see during the learning some parameters that if we manipulate them in the running time, we could obtain better results.

You can imagine it as a combination between Deep Learning and Reinforcement learning. For each step when training the data, there are some choices that needed to be taken with some parameters values, the choices have some rewards, and we force the machine to choose the best decision based on some heuristics techniques that increase its reward. One of the parameters that we can treat it as adaptively is the *Learning Rate.*

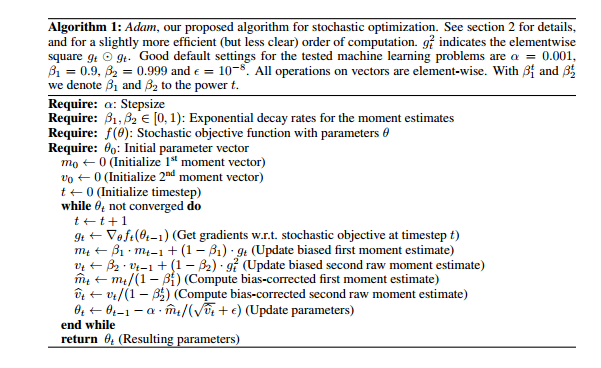
**ADAM Optimizer vs. Gradient Descent Optimizer**

ADAM optimizer is a good example of using adaptive learning to estimate its parameters. Let’s talk about the learning rate in both, the Gradient Descent and ADAM.

In Gradient Descent, we always begin with some value for learning rate that doesn’t change for every iteration, we can manually update the learning rate in each iteration, but we don’t use a specific criterion for changing the learning rate value.

The problems with are commonly known as “Under-fitting” and “Over-fitting”. “Under-fitting” means that the training wasn’t good enough to reach the global minima of the problem, “Over-fitting” means that the training was over-estimated that made the machine “Memorize” the data not “Learn” from data that will make the machine suffer from “Generalization” problem.

ADAM is derived from “Adaptive Moment Estimation”, the figure shows the algorithm pseudo-code.



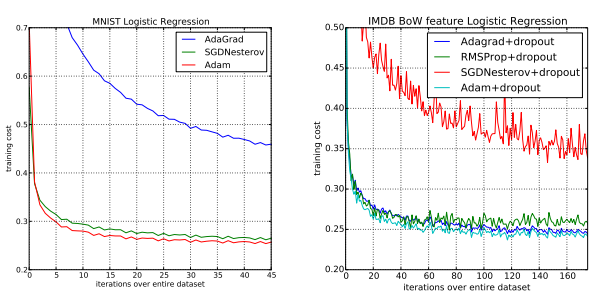
ADAM deals with first and second of the gradient moments (means), in other words, it depends on 2 parameters,  andwhere is an exponential decay rates for the moment estimates, it has a value between [0, 1) and  is a step size. Our interest is in the update parameters step. To compare between the values of the parameters, we need to get the delta of the parameters values in t and t – 1, so, the update formula is |∆t| = α · (1 − β1)/ √ 1 − β2. Here, the best case of the learning rate is when (1 − β1) > √ 1 − β2, because it will effect when multiply it with the step size. This case happens when gradient results has been zeros before the current time t.

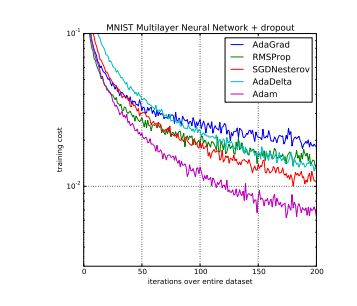
If the case when (1 − β1) = √ 1 − β2 , then the effeteness will be smaller as the update will only depends on the step size.

These cases happened because of the values of the decay rates where it is estimated under some heuristic values.

**Experiments**

These are some results when using Gradient Descent and ADAM optimizer. It shows that the huge advantage of using ADAM over Gradient.





**References**

1. [**http://arxiv.org/pdf/1412.6980v8.pdf**](http://arxiv.org/pdf/1412.6980v8.pdf)
2. [**http://arxiv.org/pdf/1212.5701v1.pdf**](http://arxiv.org/pdf/1212.5701v1.pdf)
3. [**http://idibon.com/the-fourth-generation-of-machine-learning-adaptive-learning/**](http://idibon.com/the-fourth-generation-of-machine-learning-adaptive-learning/)
4. [**http://sebastianruder.com/optimizing-gradient-descent/**](http://sebastianruder.com/optimizing-gradient-descent/)