Evolution of Iterative Optimisation Algorithms till Adam Optimiser

1. Gradient Descent

It is a first order iterative method for finding the minimum of a function. It is based on the fact that the function value decreases in the opposite direction of the gradient at a particular point, if it is not a local minimizer.

UPDATE RULE:

Challenges and Drawbacks –

* 1. Choosing a proper step size- if too small, slow convergence; if too large, may cause function to fluctuate around minimum or even diverge.
  2. Same step size for update of all components of X.
  3. Same alpha for all iterations, unable to adapt to dataset characteristics.

1. Newton’s Method

It is a second order iterative method for finding the minimum of a function. In this method of optimization, the objective function is approximated by a quadratic function

around , and then a step is taken towards the minimum of that quadratic function.

UPDATE RULE:

Challenges and Drawbacks-

* 1. Newtons method is infeasible to compute in high dimensional data sets, as computation of the Hessian matrix is expensive.
  2. All the drawbacks of gradient descent method hold here as well.

1. Adding Momentum

Gradient descent has trouble navigating ravines, i.e. areas where surface curves more steeply in one dimension than in the other. In these scenarios, GD oscillates across the slopes of the ravine while only making hesitant progress along the bottom towards the local optimum.

Adding momentum accelerates GD in the relevant direction and dampens oscillation. It is done by adding a fraction of the update vector of the past time step to the current update vector.

UPDATE RULE:

The momentum term increases for dimensions whose gradients point in the same directions and reduces updates for dimensions whose gradients change directions. As a result, we gain faster convergence and reduced oscillation.

Challenges and Drawbacks-

* 1. In this method, we first compute the gradient at current location, then take a big jump in the direction of the updated accumulated gradient. So, it does not know when to slow down, and can overshoot the minimum.

1. Nesterov accelerated gradient

Nesterov momentum has a kind of prescience. It has a notion of where it is going and knows to slow down before the hill slopes up again.

First, it makes a big jump in the direction of the previous accumulated gradient. Then it measures the gradient where it ends up, and makes a correction.

This corrective update prevents from going too fast and overshooting the minimum.

UPDATE RULE:

Challenges and Drawbacks-

* 1. Up until now, we have adapted the direction to the slope of the objective function and sped up the descent. However, we would also like to adapt our step size to each individual component .

1. Adagrad

It adapts step size rate to the components, performing larger updates for components in whose direction, the gradient is less steep, and smaller updates for those components in whose direction, the gradient is steeper. This prevents unnecessary oscillations in the steeper direction, and moves faster in the less steep direction.

It also uses different step size for every component Xi.

UPDATE RULE:

Challenges and Drawbacks-

* 1. Adagrad’s main weakness is its accumulation of the squared gradients in the denominator. Since every added term is positive, the accumulated sum keeps growing during training. This in turn causes the step size to shrink and eventually become infinitesimally small, at which point the algorithm is no longer able to acquire additional knowledge.

1. RMSprop

Restricts window of accumulated past gradients to some fixed size, by using exponential averaging.

By the above formula, R approximately becomes the average of last gradients.

UPDATE RULE:

1. Adam

This algorithm also computes adaptive step sizes for each . In addition to dividing the step size by the decaying average of past square gradients like RMSprop, Adam also replaces the simple gradient term by an exponentially decaying average of past gradients, thus incorporating momentum.

As and are initialized to 0 vector, their subsequent values are biased towards 0, especially during the initial time steps, and especially when the decay rates are small.

( and are close to 1). Performing bias correction -

UPDATE RULE:

Adam works well in practice and compares favorably to other adaptive optimisation algorithms.