# Gradient-based Optimization Methods in Deep Learning

Zhang Jinxiong

December 22, 2017

**Abstract**

Stochastic gradient descent is to optimize the finite sum objective functions based on gradient descent. Stochastic gradient descent (SGD) is the popular optimization algorithm in deep learning. It is easy to implement and of much advantages in differentiable optimization within high dimension space.

## 1 Introduction

Gradient descent or steepest descent is the fundamental optimization only using the gradient of the object function. It is simple and especially efficient in convex optimization problem. However, there are many cases that it is difficult or expensive to compute the gradient in high dimension space. To void that, we just grasp some "local" or "partial" information instead of the full information in high dimension space.

For example, the coordinate descent algorithm is to optimize the multi-variable function in one specific variable or coordinate while the rest are fixed sequentially.

It is natural for us to take the advantage of "partial" gradient information in place of the full gradient information with affordable computation cost. This paper is aimed to review some stochastic gradient descent methods.

## 2 Determinant Gradient Method

The stochastic gradient descent is rooted in classical or determinant method.

### 2.1 Gradient Descent

Gradient descent is based on the local information - it is the negative gradient that is most speedy to decrease for a function in a samll region around a given point. Let $f(x)$ be the objective function and $\nabla f(x) = g(x)$.

Note: $\alpha$ must be small enough to ensure that it is descent: $f(x_{k+1}) < f(x_k)$.

The step is key factor in the convergence proof of the gradient descent. One common setting is diminishing but not summable:

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

**Algorithm 1** Gradient Descent

1. Give the initial value: $x_0$;
2. Update the parameters with small step:

$$x_{k+1} = x_k - \alpha g(x_k).$$

## 2.2 The Momentum Methods

Momentum is form the physics. It is used to accelerate the speed of convergence.

**Algorithm 2** Gradient Descent with momentum

1. Give the initial value: $x_0$;
2. Compute the momentum: $v_k = \gamma v_{k-1} - \alpha g(\theta_k)$;
3. Update the parameters with small step:

$$\theta_{k+1} = \theta_k + v_k.$$

Nesterov's momentum is different from the classical momentum in the position where it is computed.

$$v_k = \gamma v_{k-1} - \alpha g(\theta_k + v_{k-1})$$
$$\theta_{k+1} = \theta_k + v_k$$

## 2.3 The Dynamics in Optimization

It is simple to prove the convergence rate of the above methods in a unified framework.

## 2.4 Some Examples

Some examples will show the details of the methods. And the step length and profitable direction are the focus of the gradient-based optimization methods.

### 2.4.1 The Projected Gradient Methods

This example is from Bingsheng He. It shows that the steps of the steepest descent method affect the speed of convergence dramatically.

### 2.4.2 Newton's Method for Quadratic Optimization

It takes one step in the direction of Newton's method for quadratic optimization. In theory, Newton's method just take one step to get the optimum.However, the gradient descent may take more steps.

# 3 Stochastic Gradient Descent

Stochastic methods are not to compute the exact gradient but to estimate the gradient by sampling. It is always applied to the optimization in the finite sum form. For example, it is to solve the mean square error of least square method in large scale.

The object function is called empirical risk function with the form $f(x_i;\theta) = \sum_{i=1}^{n} f_i(\theta)$, where $x_i$ is constant parameters given by training sample and $f_i$s have the same independent variables $\theta$.

Let $\nabla f_i(x) = g_i(x)$. The stochastic gradient descent with constant step is shown below.

**Algorithm 3** Primary Stochastic Gradient Descent

1. Give the initial value: $x_0$;
2. Randomly choose term index $k_i$ in the $k$th iteration from $1, 2, \cdots, n$;
3. Update the parameters:
$$\theta_{k+1} = \theta_k - \alpha g_{k_i}(\theta_k).$$

Note the difference with gradient descent: the index $k_i$ is a random variable! We can randomly choose a minibatch of training set to compute the gradient instead of the single sample: $\theta_{k+1} = \theta_k - \alpha \frac{1}{m} \sum_{i=1}^{m} g_{k_i}(\theta_k)$.

## 3.1 The Difference with Full Gradient Descent

The sufficient difference in SGD and GD is estimated gradient in SGD is far from the full original gradient in GD. The noise of the estimated gradient makes the convergence more slow. Leon Bottou have claimed that the convergence speed of stochastic gradient descent is in fact limited by the noisy approximation of the true gradient.

# 4 Adaptive Learning Rate

There are some strategies of choosing the step sizes.

### 4.0.1 Adagrad

It is based on the experience that infrequent parameters can give more information than the frequent parameters. Thus it is designed to perform larger updates for infrequent and smaller updates for frequent parameters.

It individually adapts the learning rates of all model parameters by scaling them inversely proportional to the square root of the sum of all the historical squared values of the gradient.

### 4.0.2 RMSprop

RMSProp uses an **exponentially decaying average**to discard history from the extreme past.

**Algorithm 4** Adagrad Algorithm

1. Sample a minibatch of m examples from the training set $\{x_{k_1}, x_{k_2}, \ldots, x_{k_m}\}$ with corresponding targets $y_{k_i}$;
2. Compute the gradient:$g_k = \frac{1}{m} \sum_{i=1}^{m} g_{k_i}(\theta)$;
3. Accumulate squared gradient: $r_{k+1} = r_k + g_{k+1} \odot g_{k+1}$, where $\odot$ is multiplication applied element-wise;
4. Update: $\theta_{k+1} = \theta_k - \frac{\varepsilon}{\delta + \sqrt{r_{k+1}}} \odot g_k$ (Division and square root applied element-wise).

**Algorithm 5** RMSprop Algorithm

1. Sample a minibatch of m examples from the training set $\{x_{k_1}, x_{k_2}, \ldots, x_{k_m}\}$ with corresponding targets $y_{k_i}$;
2. Compute the gradient:$g_k = \frac{1}{m} \sum_{i=1}^{m} g_{k_i}(\theta)$;
3. Accumulate squared gradient: $r_{k+1} = \rho r_k + (1-\rho) g_{k+1} \odot g_{k+1}$, where $\odot$ is multiplication applied element-wise;
4. Update: $\theta_{k+1} = \theta_k - \frac{\varepsilon}{\delta + \sqrt{r_{k+1}}} \odot g_k$ (Division and square root applied element-wise).

#### 4.0.3 Adam

Adaptive Moment Estimation (Adam) is another method that computes adaptive learning rates for each parameter. It can be seen as a variant on the combination of RMSProp and momentum.

**Algorithm 6** Adam

Sample a minibatch of m examples from the training set $\{x_{k_1}, x_{k_2}, \ldots, x_{k_m}\}$ with corresponding targets $y_{k_i}$;
Compute the gradient:$g_k = \frac{1}{m} \sum_{i=1}^{m} g_{k_i}(\theta)$;
Update biased first moment estimate: $s_{k+1} = \rho_1 s_k + (1-\rho_1) g_k$;
Update biased second moment estimate: $r_{k+1} = \rho_2 r_k + (1-\rho_2) g_k \odot g_k$;
Correct bias in first moment: $\hat{s}_{k+1} = \frac{s_{k+1}}{1-\rho_1^k}$;
Correct bias in second moment: $\hat{r}_{k+1} = \frac{r_{k+1}}{1-\rho_2^k}$;
Update: $\theta_{k+1} = \theta_k - \varepsilon \frac{\hat{s}_{k+1}}{\sqrt{\hat{r}_{k+1}} + \delta}$ .

Adam is most popular in deep learning community.

## 5 Variance Reduction

The stochastic gradient descent is not always descent due to its inherent variance.

**Algorithm 7** SVRG

**Parameters:**s update frequency m and learning rate $\alpha$.
**Initialize:**$\tilde{\theta}$
**Iterate** for $s = 1, 2, \cdots$ $\tilde{\theta} = \tilde{\theta}_{s-1}$
$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} g_i$
$\theta_0 = \tilde{\theta}$
**Iterate** Randomly pick $i_k \in \{1, 2, \cdots, m\}$ and update weight:

$$\theta_k = \theta_{k-1} - \alpha(g_{i_k}(\theta_{k-1}) - g_{i_k}(\tilde{\theta}) + \tilde{\mu}).$$

end
**Option I:** set $\tilde{\theta}_{s-1} = \theta_m$
**Option II:**set $\tilde{\theta}_{s-1} = \theta_t$ for randomly chosen $t \in \{1, 2, \cdots, m-1\}$

#### 5.0.4 Stochastic Average Gradient

Like stochastic gradient (SG) methods, the SAG method's iteration cost is independent of the number of terms in the sum. However, by incorporating a memory of previous gradient values the SAG method achieves a faster convergence rate than black-box SG methods.

The SAG iterations take the form:

$$\theta_{k+1} = \theta_k - \frac{\alpha_k}{n}\sum_{i=1}^{n} y_{i_k},$$

where at each iteration a random index $i_k$ is selected and we set:

$$y_{i_k} = \begin{cases} g_i(\theta_k), & \text{if } i = i_k; \\ y_{i_{k-1}}, & \text{otherwise.} \end{cases}$$

### 5.1 SVRG

Rie Johnson and Tong Zhang introduced an explicit variance reduction method for stochastic gradient descent. It is proved that this method enjoys the same fast convergence rate as those of stochastic dual coordinate ascent (SDCA) and Stochastic Average Gradient (SAG) for *smooth and strongly convex* functions.

Its aim is to reduce the variance of the gradient so that the process is more stable.

### 5.2 SVRG++

It is improved SVRG for non-strongly convex or sum-of-non-convex settings.The objective function is

$$F(x) = \sum_{i=1}^{n} f_i(x) + h(x),$$

where $h(x)$ is the regularization term. Let $g_i(x) = \nabla_x f_i(x)$.

**Algorithm 8** SVRG++

1. During the $s$th epoch, $\hat{\mu}_{s-1} = \frac{1}{m_{s-1}} \sum_{i=1}^{m} g_i(\hat{x}_{s-1})$;
2. $m_s = 2^s m_0$
3. For $k \in \{0, 1, 2, \cdots, m_s - 1\}$

Randomly pick $i_k \in \{1, 2, \cdots, n\}$ and update weight:

$$\hat{w} = g_{i_k}(x_k^s) - g_{i_k}(\hat{x}_{s-1}) + \hat{\mu}_s,$$

$$x_{k+1}^s = argmin_x\{h(x) + \frac{1}{2\gamma}\|x - x_k^s\| + < x, \hat{w} >\}$$

4. $\hat{x}_s = \frac{1}{m_s} \sum_{k=1}^{m_s} x_k^s$ ;
5. $x_0^{s+1} = x_{m_s}^s$

### 5.3 SCSD

As a member of the SVRG family of algorithms, SCSG makes use of gradient estimates at two scales, with the number of updates at the faster scale being governed by a geometric random variable. Unlike most existing algorithms in this family, both the computation cost and the communication cost of SCSG do not necessarily scale linearly with the sample size $n$; indeed, these costs are independent of $n$ when the target accuracy is low.

**Algorithm 9** SCSD

1.During the $s$th epoch and $s \in \{1, 2, \cdots, T\}$, $\hat{\theta} = \hat{\theta}_{s-1}$, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{m} g_i(\hat{\theta})$,$\theta_0 = \hat{\theta}$;
2.For $k \in \{1, 2, \cdots, M\}$, where $M \sim Geom(\frac{B}{B+1})$ Randomly pick $i_k \in \{1, 2, \cdots, n\}$ and update weight:

$$\theta_k = \theta_{k-1} - \alpha(g_{i_k}(\theta_{k-1}) - g_{i_k}(\hat{\theta}) + \hat{\mu}).$$

3.In this epoch:$\hat{\theta}_s = \theta_M$
4.(Strongly convex case)$\hat{\theta} = \theta_T$ or (Not strongly convex case)$\hat{\theta} = \frac{1}{T} \sum_{k=1}^{T} \theta_k$

## 6 The Stochastic Properties

The randomness occurs in the stochastic gradient-based optimization, including:

- Initial values: It is important to choose the initial values in practical aplications such as deep learning.
- The estimated gradient: The estimation is not based on reliable of statistical method due to the lack of the distribution of the original full gradient.
- The noise of input data: The raw data is always filled with noise.