Analyzing Stochastic Gradient Descent Method

Tushar Agarwal

1 Introduction

21st century is indeed the age of Big-Data, i.e., data-sets so large that the existing computing resources are not enough to process them within practical time-frames. While there have been rapid improvements on the hardware side, the software implementations are still trying to catch up. One of the obvious challenges that arises when utilizing these huge data-sets to train a model, is the optimization involved, which is typical for any sort of parameter estimation. Due to the shear size of data-sets, the advanced algorithms like Newton's method, which have several advantages regarding convergence, are unusable due to their computational complexities and practical resource limitations. As a result, practitioners have resorted to simpler methods among which the most widely and effectively used one is the Gradient Descent (GD) Method.

The objective of this paper is to analyze the Stochastic Gradient Descent (SGD) method. It is an attempt to answer questions like when and why SGD method would converge to a desirable solution. It is an expectation that the insights from the analysis presented in this paper may assist practitioners to come-up with more effective heuristics. Moreover, theoreticians may be able to extend the analysis to more advanced variants of SGD like the ones with adaptive step-sizes, which are not covered in this paper. The analysis is largely based on section 4 of the detailed paper on the same topic by Bottou et al. [1].

In this section the typical optimization problem in large-scale machine learning is presented along-with the variants of gradient descent algorithm used to solve it. Some pros and cons of these variants are discussed in order to motivate the use of Stochastic Gradient Descent methods. Then, a very generic form of the SGD algorithm is presented that encompasses many of its present-day variants. This is done to make the further analysis of the algorithm, in section 2, applicable to several cases simultaneously. The analysis in section 2 first covers the case of strongly convex objective function with a fixed step-size (Theorem 3). Then the more general non-convex case is analyzed for both, a fixed step-size (Theorem 4) and a diminishing step size (Theorem 5). Furthermore, a convergence in probability result is established for l_2 norm of the gradient (Theorem 6). Both, theorems 3 and 4 help gain insight into a commonly used heuristic by practitioners. Theorems 5 and 6 bring out the importance of a diminishing step-size which helps converge even when the estimates of the expected gradient are noisy. Finally the insights are summarized in section 3.

1.1 Typical Problem Formulation

In this sub-section, the typical optimization problem in large-scale machine learning is considered. However, the results can be easily extended to other optimization problems as well. Consider a given data-set, $D = \{(x_i, y_i)\}_i$ where (x_i, y_i) pairs are drawn from an unknown joint distribution, $P(x, y)$. Here, notation x, y denotes inputs and outputs of the system under consideration respectively and let $x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}$. Let $i \in [N]$ where $N \in \mathbb{N}$ are total number of samples in D and notation $[n]$ denotes the enumeration of natural numbers upto and including $n \in \mathbb{N}$. Typically a fixed form of prediction function, $h(\cdot; w) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ is assumed parameterized by a real vector, $w \in \mathbb{R}^p$. The problem is to estimate this w using data-set D. This is typically done by minimizing an appropriate loss function, $l(\cdot, \cdot, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^p \to \mathbb{R}$. The loss, $l(x, y, w)$ corresponds to error in model fitting when $h(x; w)$ and y are the predicted and true outputs respectively.

Now, given data D, the target usually is to minimize $l(x, y, w)$ with respect to w. Ideally, it is desired to minimize the expected loss, $F(w) = \mathbb{E}_{x,y}[l(x, y, w)]$ where the subscript of the expectation denotes the random variables over which it is taken. But the joint distribution, $P(x, y)$ is not explicitly known. So, an approximation to the expected loss is used, typically the empirical mean, i.e., $F(w) = \frac{1}{N}\sum_{i=1}^{N}(l(x_i, y_i, w))$. Hence, it is desired to reach a stationary point, i.e., a point where the gradient ∇F (or its estimate) is zero, while minimizing the objective function, $F(w)$. This is usually done by iteratively moving in the direction of steepest descent, $-\nabla F(w_k)$, where k is the iteration number. This algorithm is known as Gradient Descent (GD).

It is important to note that ∇F is a gradient of an expectation. In most cases of interest, it can be replaced by an expectation of the gradient. In case of a finite sum, it is straight-forward and in case of an integral sign, Leibniz integral rule can be used. Leibniz integral rule may in-turn be proved through Dominated Convergence Theorem under some mild regularity conditions, including boundedness of $l(x, y, w)$. So, an estimate of expected value of the gradient ∇F, over the samples (x_i, y_i), is needed. If this estimate is computed using the entire data-set D before taking the step, it is known as Batch Gradient Descent (BGD) but if estimate of gradient is computed by randomly selecting 1 sample from D, it is known as Stochastic Gradient Descent (SGD). Both approaches have their advantages and disadvantages but in practice, what is widely used is something in-between the two, i.e., Mini-Batch Gradient Descent (MBGD) where gradient is estimated using a small subset of samples of D. Frequently, in machine learning literature, MBGD is also referred to as the Stochastic Gradient Descent as the gradient estimates are obtained from random sampling and hence are noisy. The convergence of the three variants is qualitatively shown in figure 1 over a contour plot of a two dimensional quadratic objective function with minimum at the center.

1.2 Motivations for SGD

There are several motivations for using SGD over BGD in practice. First and foremost would be that even if the gradient estimates for the i^{th} sample $\nabla F = \nabla_w l(x_i, y_i, w)$ are unbiased estimator of the true gradient, the standard error in approximating expectation from empirical mean reduces at the rate proportional to $\frac{1}{\sqrt{N}}$, where N are

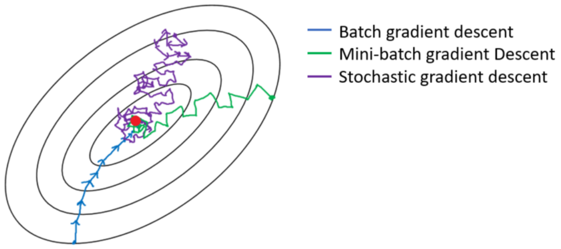

Figure 1: Qualitative convergence of the three variants of GD.
(Image from Towards Data Science)

total number of samples used in approximating the expectation. This is a direct consequence of Weak Law of Large Numbers in probability theory. So, these less than linear returns in accuracy of the estimate of ∇F motivate use of fewer samples. For eg., compare two hypothetical estimates of the expected gradient, one based on 100 examples and another based on 10,000 examples. The latter requires 100 times more computation than the former but reduces the standard error of the expectation only by a factor of 10.

Another motivation is the presence of redundancy, especially in large data-sets. It is very likely that several samples make very similar contributions to the gradient. This makes it useless to utilize the entire data-set for estimating the expected gradient. Hence, SGD is a more efficient use of resources in this regard.

But SGD has its cons as well, the primary one being noise in the expected gradient estimates. As it will be shown in the following section, if this noise in the estimates, is very high, the asymptotic optimality gap would be higher as well. Another disadvantage is, SGD doesn't allow exploiting parallelization to speed-up the training which is a crucial practical consideration. Fortunately, both of these flaws can be mitigated by MBGD and hence it is the most widely used algorithm. Taking a random subset of samples to estimate expected gradient, reduces the variance of the estimate as well as allows parallelization during iterative optimization updates. For the sake of this paper and further analysis, all these variants are referred as SG method as they are encompassed into a very generic Algorithm 1.

1.3 The Algorithm

The generalized stochastic gradient descent algorithm is presented in Algorithm 1. It presumes existence of 3 computational tools as follows:

1. A way to generate realizations of random variable $d_k = \{(x_i, y_i)\}_{i\in[N]}$, i.e., a set of random samples from the data distribution at iteration k.

2. A way to compute stochastic vector, $g(w_k, d_k)$ given both, d_k and w_k.

3. A method for selecting the step size, α_k at epoch k.

Indeed, the form of the algorithm presented is very general in the sense that the gradient vector $g(w_k, d_k)$ can be estimated in any way as outlined in the subsection

1.1. Apart from that, $g(w_k, d_k)$ can also be estimated using stochastic-Newton and Quasi-Newton directions and the following analysis will still be valid.

Algorithm 1 SG Method

```
Choose an initial iterate, $w_1$, of parameters.
for $k \leftarrow 1$ to $K$ do                                  ▷ K refers to no. of iterations
    Generate a realization of the random variable $d_k$
    Calculate the stochastic vector $g(w_k, d_k)$
    Select a step size $\alpha_k$ based on a desired criteria.
    Evaluate next iterate, $w_{k+1} = w_k - \alpha_k g(w_k, d_k)$
end for
```

2 Analysis of SG Method

This section will make an attempt at getting some insight into the convergence properties of SG method. The essential assumptions required for all the following analysis are stated first. The assumptions for specific cases will be mentioned later, as needed.

Main Assumptions:

1. The objective function $F : \mathbb{R}^p \to \mathbb{R}$ is continously differentiable and its gradient, $\nabla F : \mathbb{R}^p \to \mathbb{R}^p$ is Lipschitz continuous with Lipschitz constant $L > 0$, i.e.,

$$\|\nabla F(w) - \nabla F(\bar{w})\|_2 \leq L\|w - \bar{w}\|_2 \quad \forall w, \bar{w} \in \mathbb{R}^p \tag{1}$$

2. The sequence of iterates w_k is contained in an open set over which F is bounded below by a scalar F_{inf}.

3. $\exists$ scalars $\mu_G \geq \mu > 0$ such that, $\forall k \in \mathbb{N}$,

$$\nabla F(w_k)^T \mathbb{E}_{d_k}[g(w_k, d_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \tag{2}$$

$$\|\mathbb{E}_{d_k}[g(w_k, d_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2 \tag{3}$$

4. $\exists$ scalars $M > 0$ and $M_V > 0$ such that $\forall k \in \mathbb{N}$,

$$\mathbb{V}_{d_k}[g(w_k, d_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2 \tag{4}$$

$\mathbb{E}_{d_k}[\cdot]$ and $\mathbb{V}_{d_k}[\cdot]$ denote expected value and variance taken with respect to the distribution of the random variable d_k given w_k.

First assumption ensures that the gradient of ∇F doesn't change arbitrarily quickly because otherwise it won't be a good indicator of how far to move to decrease F. Second assumption merely wants objective to be lower bounded over the parameter space of w. This is usually the case, for eg: squared loss is lower bounded by 0. Third assumption ensures that in expectation, the function $-g(w_k, d_k)$, is a direction of sufficient descent with a norm comparable to the norm of ∇F. This also occurs commonly in practice . For eg, if $g(w_k, d_k)$ is an unbiased estimator of ∇F, the conditions hold trivially with equality and for $\mu_G = \mu = 1$. Fourth assumption restricts the variance of the estimate $g(w_k, d_k)$ but in a relatively minor manner. For eg, if F is a complex quadratic function, the variance is allowed to be non-zero at any stationary point and can increase quadratically in any direction.

Under the first assumption, the following lemma can be proved.

Lemma 1. *The iterates of SG satisfy the following inequality* $\forall k \in \mathbb{N}$

$$\mathbb{E}_{d_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{d_k}[g(w_k, d_k)] + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{d_k}[\|g(w_k, d_k)\|_2^2] \quad (5)$$

Proof. From assumption 1, it can be shown that (refer to Appendix B of [1]),

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \frac{1}{2} L \|w - \bar{w}\|_2^2 \qquad \forall w, \bar{w} \in \mathbb{R}^p \quad (6)$$

Putting $w = w_{k+1}$ and $\bar{w} = w_k$ and rearranging gives,

$$\begin{aligned} F(w_{k+1}) - F(w_k) &\leq \nabla F(w_k)^T (w_{k+1} - w_k) + \frac{1}{2} L \|w_{k+1} - w_k\|_2^2 \\ &\leq -\alpha_k \nabla F(w_k)^T g(w_k, d_k) + \frac{1}{2}\alpha_k^2 L \|g(w_k, d_k)\|_2^2 \end{aligned}$$

Taking expectation in this inequality with respect to d_k and noting that w_k doesn't depend on d_k, the result follows. □

The third and fourth assumptions basically put further restrictions on first and second moment of the gradient estimates $g(w_k, d_k)$. Under all 4 assumptions, the following lemma can be proved.

Lemma 2. *The iterates of SG satisfy the following inequality* $\forall k \in \mathbb{N}$

$$\begin{aligned} \mathbb{E}_{d_k}[F(w_{k+1})] - F(w_k) &\leq -\mu \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{d_k}[\|g(w_k, d_k)\|_2^2] \quad (7) \\ &\leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right) \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M \quad (8) \end{aligned}$$

where $M_G = M_V + \mu_G^2 \geq \mu^2 > 0$

Proof. From lemma 1 and (2), result (7) follows as,

$$\begin{aligned} \mathbb{E}_{d_k}[F(w_{k+1})] - F(w_k) &\leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{d_k}[g(w_k, d_k)] + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{d_k}[\|g(w_k, d_k)\|_2^2] \\ &\leq -\mu \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{d_k}[\|g(w_k, d_k)\|_2^2] \end{aligned}$$

The definition of variance is,

$$\mathbb{V}_{d_k}[g(w_k, d_k)] = \mathbb{E}_{d_k}[\|g(w_k, d_k)\|_2^2] + \|\mathbb{E}_{d_k}[g(w_k, d_k)]\|_2^2 \quad (9)$$

Using (3) and (4) with this definition gives,

$$\mathbb{E}_{d_k}[\|g(w_k, d_k)\|_2^2] \leq M + M_G \|\nabla F(w_k)\|_2^2 \quad (10)$$

Using this result in second term of right-hand-side (RHS) of (7) yields (8). □

Both lemmas 1 and 2 reveal that regardless of how the method arrived at the iterate w_k, the optimization process and convergence continues in a Markovian manner, i.e., depends only on the iterate w_k, d_k and α_k and not on any past iterates. This is obvious from (8) where the expected difference between the two values of the objective function is upper bounded by a deterministic quantity. First term in (8) is strictly negative for small α_k. However, the second term could be large enough to allow the objective value to increase. Balancing these two terms is critical in the design of SG method.

2.1 SG Method for Strongly Convex Objective

In this subsection, the analysis is carried further with an additional assumption of strong convexity of the objective function, $F(w)$, i.e., $\exists$ a scalar $c > 0$ such that

$$F(w) \geq F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \frac{1}{2} L \|w - \bar{w}\|_2^2 \quad \forall w, \bar{w} \in \mathbb{R}^p \tag{11}$$

Therefore, $F(w)$ has a unique minimizer, say $w_* \in \mathbb{R}^p$ and let $F_* = F(w_*)$. Whenever a function is strongly convex, a fact (without proof) from convex analysis can be used, which is,

$$2c(F(w) - F_*) \leq \|\nabla F(w)\|_2^2 \quad \forall w \in \mathbb{R}^p \tag{12}$$

It is also to be noted that from (6) and (11), $c \leq L$. The first theorem of convergence analysis can now be stated and proved.

Theorem 3. *Under four main assumptions and additional strong convexity assumption in (11), if the SG method is run with a fixed step size, $\alpha_k = \alpha \quad \forall k \in \mathbb{N}$ such that*

$$0 < \alpha < \frac{\mu}{LM_G} \tag{13}$$

Then the expected optimality gap, $\forall k \in \mathbb{N}$ can be upper bounded as

$$\begin{aligned} \mathbb{E}[F(w_k) - F_*] &\leq \frac{\alpha LM}{2c\mu} + (1 - \alpha c\mu)^{k-1} \left(F(w_1) - F_* - \frac{\alpha LM}{2c\mu} \right) \\ &\xrightarrow[k \to \infty]{} \frac{\alpha LM}{2c\mu} \end{aligned} \tag{14}$$

where the total expectation, $\mathbb{E}[F(w_k)] = \mathbb{E}_{d_1}\mathbb{E}_{d_2}...\mathbb{E}_{d_{k-1}}[F(w_k)]$

Proof. Using (8) from Lemma 2 with (12) and (13) gives,

$$\begin{aligned} \mathbb{E}_{d_k}[F(w_{k+1})] - F(w_k) &\leq -\left(\mu - \frac{1}{2}\alpha LM_G \right) \alpha \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\ &\leq -\frac{1}{2}\alpha\mu \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\ &\leq -\alpha c\mu (F(w_k) - F_*) + \frac{1}{2}\alpha^2 LM \end{aligned}$$

Subtracting F_* both sides and rearranging gives,

$$\mathbb{E}_{d_k}[F(w_{k+1})] - F_* \leq (1 - \alpha c\mu)(F(w_k) - F_*) + \frac{1}{2}\alpha^2 LM$$

Taking total expectation and subtracting constant $\frac{\alpha LM}{2c\mu}$ both sides yields,

$$\begin{aligned}\mathbb{E}[F(w_{k+1}) - F_*] - \frac{\alpha LM}{2c\mu} &\leq (1 - \alpha c\mu)\mathbb{E}[F(w_k) - F_*] + \frac{1}{2}\alpha^2 LM - \frac{\alpha LM}{2c\mu} \\ &= (1 - \alpha c\mu)\left(\mathbb{E}[F(w_k) - F_*] - \frac{\alpha LM}{2c\mu}\right)\end{aligned} \tag{15}$$

Noting that by (13),

$$0 < \alpha c\mu < \frac{c\mu^2}{LM_G} \leq \frac{c\mu^2}{L\mu^2} = \frac{c}{L} \leq 1$$

So, (15) is a contraction inequality. Hence, the desired result follows by applying (15) repeatedly over all iterations $k \in \mathbb{N}$. □

If there is no noise in the gradient computation or if the noise decays with $\|\nabla F(w_k)\|_2^2$, i.e., if $M = 0$, then a geometric (R-linear) convergence to the optimal value is achieved. This is a standard result for BGD method with a sufficiently small positive α.

If the gradient computation is noisy, a fixed α_k can still be used and the expected objective values will converge geometrically (or R-linearly) to a neighborhood of the optimal value. But, after some point, the noise in the gradient estimates will prevent further progress. This limitation can also be seen from (8). As the solution is approached, the first term on RHS becomes smaller (since $\nabla F(w_k) \to 0$) but the second term remains constant. From (14), selecting a smaller α_k worsens the contraction constant but allows to arrive closer to the optimal value. This provides important insight into effectiveness of a heuristic strategy often employed in practice by which SG method is run with a fixed α_k and, if progress appears to stall, a smaller α_k is selected and the process is repeated.

2.2 SG Methods for Non-Convex Objective

In this subsection, the analysis is extended to the most widely used and practical case, i.e. when the objective $F(w)$ is non-convex and the step-size, α_k is diminishing. Before that, a result for fixed α_k is derived which has similar conclusions as from the previous theorem.

Theorem 4. *Under the four main assumptions, if the SG method is run with a fixed step size, $\alpha_k = \alpha \quad \forall k \in \mathbb{N}$ such that*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{16}$$

then $\forall K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{K\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu\alpha} \tag{17}$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K\mu\alpha} \xrightarrow[K\to\infty]{} 0 \tag{18}$$

Proof. From the condition in (16) and taking total expectation in (8) gives,

$$\begin{aligned}\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)] &\leq -(\mu - \frac{1}{2}\alpha L M_G)\alpha\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha^2 LM \\ &\leq -\frac{1}{2}\mu\alpha\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha^2 LM\end{aligned}$$

Summing both sides for $k \in [K]$ and using main assumption 2 gives,

$$F_{inf} - \mathbb{E}[F(w_1)] \leq \mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_1)] \leq -\frac{1}{2}\mu\alpha\sum_{k=1}^{K}\mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}LMK\alpha^2$$

Rearranging terms gives (17) and further multiplying both sides by $\frac{1}{K}$ yields (18) □

Very similar to the conclusions of the previous theorem, if there is no noise in the gradient computation or if the noise decays with $\|\nabla F(w_k)\|_2^2$, i.e., if $M = 0$, then sum of squared gradients remains finite as seen from (17) which implies $\{\|\nabla F(w_k)\|_2\} \to 0$ as $K \to \infty$. This is again a standard result for BGD method with a sufficiently small positive α.

In case of noisy gradient estimates, while one cannot bound the expected optimality gap as in the convex case, it is possible to bound the average norm of the gradient of the objective function observed on w_k visited during the first K iterations. This quantity gets smaller when K increases, indicating that the SG method spends increasingly more time in regions where the objective function has a (relatively) small gradient. Similar to the conclusions of the previous theorem, the asymptotic result of (18) illustrates that the noise eventually stops the progress towards the optimum. The average norm of the gradients can be made arbitrarily small by selecting a smaller α_k, but doing so also reduces the speed at which the average norm of the gradient approaches its limiting value.

In the following theorem, the convergence in case of a diminishing step-size is analyzed.

Theorem 5. *Under the four main assumptions, if the SG method is run with a step size sequence satisfying the classical Robbins-Monro conditions [2], i.e.,*

$$\sum_{k=1}^{\infty}\alpha_k = \infty \quad \textit{and} \quad \sum_{k=1}^{\infty}\alpha_k^2 < \infty \tag{19}$$

then $\forall K \in \mathbb{N}$

$$\lim_{K\to\infty}\mathbb{E}\left[\sum_{k=1}^{K}\|\nabla F(w_k)\|_2^2\right] < \infty \tag{20}$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K}\|\nabla F(w_k)\|_2^2\right] \xrightarrow[K\to\infty]{} 0 \tag{21}$$

where, $A_K = \sum_{k=1}^{K}\alpha_k$

Proof. From the second condition in (19), it is clear that $\{\alpha_k^2\} \to 0 \Rightarrow \{\alpha_k\} \to 0$. Therefore, without loss of generality (WLOG), it can be assumed that $\alpha_k L M_G \leq \mu \;\; \forall k \in \mathbb{N}$ (i.e., the first $k_0 \in \mathbb{N}$ for which this is true, can be considered as the first iterate). Now, taking total expectation in (8) gives,

$$\begin{aligned}\mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_k)] &\leq -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 LM \\ &\leq -\frac{1}{2}\mu\alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 LM\end{aligned}$$

Summing both sides for $k \in [K]$ gives,

$$F_{inf} - \mathbb{E}[F(w_1)] \leq \mathbb{E}[F(w_{k+1})] - \mathbb{E}[F(w_1)] \leq -\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}LM \sum_{k=1}^{K} \alpha_k^2$$

Multiplying both sides by $\frac{2}{\mu}$ and rearranging the terms yields,

$$\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leq \frac{2(\mathbb{E}[F(w_1)] - F_{inf})}{\mu} + \frac{LM}{\mu}\sum_{k=1}^{K}\alpha_k^2$$

Note that by the second condition of (19), the RHS converges to a finite value as $K \to \infty$. This proves (20). Furthermore, by first condition of (19), $K \to \infty \Rightarrow A_K \to \infty$. Hence, (21) follows. □

This theorem establishes results about a weighted sum-of-squares and a weighted average of squared gradients of F similar to those in the previous one. However, unlike the previous theorem, the result (21) states that the weighted average norm of the squared gradients converges to zero even if the gradients are noisy, i.e., if $M > 0$. This is due to the fact that the α_k need not be fixed anymore and just need to satisfy (19).

The following theorem will show a stronger and important convergence result for l_2 norm of the gradient of F.

Theorem 6. *For any $K \in \mathbb{N}$ let $k(K) \in [K]$ be a random index chosen with probability proportional to $\{\alpha_k\}_{k=1}^{K}$. If the conditions of Theorem 5 hold, then,*

$$\lim_{K\to\infty} \|\nabla F(w_{k(K)})\|_2 \xrightarrow[in\;\; probability]{} 0 \tag{22}$$

Proof. For any $\epsilon > 0$, using (20) and Markov's inequality gives,

$$\mathbb{P}(\|\nabla F(w_k)\|_2 \geq \epsilon) = \mathbb{P}(\|\nabla F(w_k)\|_2^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[\mathbb{E}_k[\|\nabla F(w_k)\|_2^2]]}{\epsilon^2} \xrightarrow[K\to\infty]{} 0$$

which proves the desired result by definition of convergence in probability. □

The result of this theorem is a stronger conclusion than the previous theorem but is only valid for gradient of F at a randomly selected iterate.

3 Conclusion

In this paper, an analysis of stochastic gradient method is presented. First, a very generic form of the Stochastic Gradient Descent algorithm is presented that encompasses many of its present-day variants. This is done to make the analysis of the algorithm applicable to several cases. The analysis is based on 4 main assumptions. It first covers the case of a strongly convex objective function with a fixed step-size (Theorem 3). Then the more general non-convex case is analyzed for both, a fixed step-size (Theorem 4) and a diminishing step size (Theorem 5). Both, Theorem 3 and 4 help gain insight into the effectiveness of a popularly used heuristic by practitioners, i.e., starting with a fixed step-size, waiting for the decrease in objective value to stall followed by a reduction in the step-size and repeating the process. Theorem 5 and 6 bring out the importance of another widely used practice, i.e., a diminishing step-size schedule. It helps converge close to a local-minimum even when the estimates of the expected gradient are noisy. It is anticipated that the analysis and insights presented in this paper can assist practitioners to come up with more effective heuristics. Moreover, theoreticians may be able to extend the analysis to more advanced variants of SGD like the ones with adaptive step-sizes.

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[2] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.