Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.6065
Published by: Zeus Press

Stochastic Gradient Descent and the Law of Large Numbers: A Probabilistic Analysis of Convergence

Haoting Chai*

School of Mathematical Sciences, Dalian University of Technology, Dalian, China

**Corresponding author: Haoting Chai.*

Abstract

Stochastic Gradient Descent (SGD) is one of the most fundamental optimization algorithms in modern large-scale machine learning. However, by relying on only a minuscule number of random samples to estimate gradients at each iteration, it inevitably introduces stochastic noise. This paper aims to delve into the mathematical essence of how this highly randomized algorithm achieves stable convergence from a probabilistic perspective. The paper first rigorously reviews the theoretical foundations and mathematical proofs of the Weak and Strong Laws of Large Numbers. Subsequently, it constructs a probabilistic model for stochastic gradients and thoroughly analyzes the update mechanism of the SGD algorithm. The study demonstrates that the stochastic gradients in single iterations of SGD form a sequence of independent and identically distributed random vectors, whose mathematical expectation is the true full-batch gradient. Based on the Law of Large Numbers, the sample mean of these stochastic gradients converges asymptotically to the true gradient in probability (or almost surely), thereby effectively averaging out and dissipating the random noise during long-term iterations. The conclusion indicates that the effectiveness of SGD is mathematically equivalent to "unbiased estimation combined with the Law of Large Numbers," and its overall optimization behavior asymptotically approximates standard, noise-free gradient descent.

Keywords

stochastic gradient descent, law of large numbers, convergence analysis, unbiased estimation, machine learning optimization

1. Introduction

With the rapid development of artificial intelligence and machine learning technologies, particularly within the realm of deep learning, the essence of model training has fundamentally evolved into solving complex, high-dimensional, non-linear optimization problems. In this optimization process, gradient-based algorithms play an indispensable role [1]. Traditionally, Batch Gradient Descent (BGD) guides parameter updates by computing the average gradient across the entire training dataset. Its mathematical properties are highly stable, guaranteeing convergence to a local or global minimum. However, when confronted with the massive scale of modern data, BGD requires traversing the entire dataset for every single iteration, resulting in prohibitively high computational costs that struggle to meet the efficiency demands of practical engineering.

To overcome this computational bottleneck, Stochastic Gradient Descent (SGD) emerged and rapidly became the industry standard [1,2]. Unlike BGD, SGD and its variants (such as mini-batch SGD) rely on estimating gradients from a randomly selected, very small subset of samples at each iteration. While this sampling mechanism drastically improves computational efficiency, it also introduces non-negligible random noise, rendering the parameter update trajectory highly volatile. This naturally leads to a profound question: why can an algorithm laden with random error ultimately achieve stable convergence and locate optimal solutions?

The core key to unraveling the underlying mechanism of this phenomenon lies in the cornerstone of probability theory: the Law of Large Numbers (LLN) [3]. The LLN mathematically formalizes the objective principle that the average result of a large number of random events inevitably tends toward stability. In the context of optimization algorithms, it provides a solid theoretical guarantee for the unbiased estimation of stochastic gradients and the subsequent dissipation of noise over long-term iterations.

2. Theoretical Foundations of the Law of Large Numbers

2.1 Weak and Strong Laws of Large Numbers

Theorem 2.1 (Weak Law of Large Numbers) [4]

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables satisfying: $E[X_i] = \mu, Var(X_i) = \sigma^2 < \infty$, Define the sample mean as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, Then, for any given $\varepsilon > 0$ it follows that":

$$\lim_{n\to\infty} P(|\ \bar{X}_n - \mu\ | > \varepsilon) = 0 \tag{1}$$

which implies $\bar{X}_n \xrightarrow{P} \mu$.

Proof:

First, one needs to calculate the expectation of the sample mean: $E[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$. Next, its variance is determined. Due to the independence of the random variables, obtaining : $Var(\bar{X}_n) = Var$, $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$.

According to Chebyshev's inequality, it is found that $P(|\ \bar{X}_n - \mu\ | > \varepsilon) \le \frac{Var(\bar{X}_n)}{\varepsilon^2}$. Substituting the variance into the inequality yields:

$$P(|\ \bar{X}_n - \mu\ | > \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2} \tag{2}$$

Since $\frac{\sigma^2}{n\varepsilon^2} \to 0 (n \to \infty)$; it follows that:

$$\lim_{n\to\infty} P(|\ \bar{X}_n - \mu\ | > \varepsilon) = 0 \tag{3}$$

That is, $\bar{X}_n \xrightarrow{P} \mu$, This completes the proof.

Theorem 2.2 (Kolmogorov's Strong Law of Large Numbers) [4,5]

Let $X_1, X_2, \ldots$be a sequence of i.i.d. random variables satisfying： $E[X_i] = \mu, Var(X_i) < \infty$， Then： $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$， which means that

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1 \tag{4}$$

Proof: Let $Y_i = X_i - \mu$,which yields $E[Y_i] = 0$.

By defining the partial sum $S_n = \sum_{i=1}^{n} Y_i$, it follows that: $\bar{X}_n - \mu = \frac{S_n}{n}$. Thus, the objective is equivalent to proving that $\frac{S_n}{n} \to 0\ \ a.s.$

Step 1: Constructing the Variance Series. Given $Var(Y_i) = \sigma^2$, consider the following series: $\sum_{n=1}^{\infty} \frac{Var(Y_n)}{n^2} = \sigma^2 \sum_{n=1}^{\infty} \frac{1}{n^2}$. Since the harmonic series limits $\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$, it follows that

$$\sum_{n=1}^{\infty} \frac{Var(Y_n)}{n^2} < \infty \tag{5}$$

Step 2: Applying Kolmogorov's Inequality. For the partial sum $S_n = \sum_{i=1}^{n} Y_i$, Kolmogorov's inequality states:

$$P\left(\max_{k \le n} \mid S_k \mid \ge \lambda \right) \le \frac{Var(S_n)}{\lambda^2} \tag{6}$$

Where $Var(S_n) = n\sigma^2$.

Step 3: Constructing a Subsequence. Consider the subsequence $n = 2^k$. Thus:

$$P\left(\mid S_{2^k} \mid \ge \varepsilon 2^k \right) \le \frac{Var(S_{2^k})}{\varepsilon^2 2^{2k}} = \frac{2^k \sigma^2}{\varepsilon^2 2^{2k}} = \frac{\sigma^2}{\varepsilon^2 2^k} \tag{7}$$

Consequently, the series of these probabilities converges:

$$\sum_{k=1}^{\infty} P(\mid S_{2^k} \mid \ge \varepsilon 2^k) < \infty \tag{8}$$

Step 4: Applying the Borel-Cantelli Lemma. According to the Borel-Cantelli lemma, if $\sum P(A_k) < \infty$,then $P(A_k \text{ i.o.}) = 0$. Therefore, $\frac{S_{2^k}}{2^k} \to 0$.

Step 5: Generalization to all n. For any given n,there exists an integer k such that $2^k \le n < 2^{k+1}$.In this case:

$$\frac{|S_n|}{n} \le \frac{|S_{2^k}|}{2^k} + \frac{|S_n - S_{2^k}|}{2^k} \tag{9}$$

By utilizing the independence of the random variables and variance estimation, it can be proven that the second term almost surely converges to 0 as $k \to \infty$. Thus,$\frac{S_n}{n} \to 0\ a.s.$ which implies $\bar{X}_n \to \mu\ a.s.$ This completes the proof.

2.2 Statistical Significance of the Law of Large Numbers

The core concept of the Law of Large Numbers can be summarized as follows: random errors gradually cancel each other out during the averaging process. If the random variables possess a finite variance, the variance of the sample mean is given by:

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} \tag{10}$$

This demonstrates that as the sample size increases, the fluctuation of the average decreases. This theoretical property serves as a crucial mathematical foundation for the stable operation and convergence of numerous randomized algorithms in machine learning, such as Stochastic Gradient Descent.

3. The Relationship Between the Law of Large Numbers and Stochastic Gradient Descent

3.1 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is one of the most fundamental optimization algorithms [1,2] in machine learning, primarily used to minimize loss functions by iteratively updating model parameters. Unlike Batch Gradient Descent, which computes gradients using the entire dataset, the core idea of SGD is to randomly select a single training sample at each iteration to estimate the gradient and immediately update the model parameters [5,6].

The advantages of SGD are as following.

The first is Handling Large-Scale Data. When dealing with massive datasets (e.g., millions or billions of samples), Batch Gradient Descent requires traversing the entire dataset for a single iteration, resulting in massive computational overhead. In contrast, SGD computes updates based on a single sample, significantly reducing computational complexity and drastically accelerating iteration speed.

The second is Online Learning Support. SGD naturally accommodates streaming data processing, allowing the model to update parameters in real-time as new samples arrive.

The third is Escaping Local Minima. Because the gradient computed at each step is an estimate based on a single sample, it contains inherent random noise. This randomness helps the optimization algorithm escape sharp local minima during training, allowing it to find flatter minimum regions, which generally leads to better model generalization.

Suppose the objective loss function is $J(\theta)$, the model parameters are θ, and the learning rate is η, The specific steps of SGD are as follows. The first is to randomly initialize parameters θ. The second is to repeat the following steps until convergence or a maximum number of iterations is reached. One should randomly shuffle the training data. For each training sample (x_i, y_i), one need to compute the gradient of the loss function with respect to the parameters: $\nabla_\theta J(\theta; x_i, y_i)$. Finally, one should update parameters as $\theta := \theta - \eta \nabla_\theta J(\theta; x_i, y_i)$.

In practice, Mini-batch Stochastic Gradient Descent is most commonly used. This approach randomly selects a small batch of samples (e.g., 32 or 64) to compute the average gradient for parameter updates. It effectively balances the stability of batch updates with the computational efficiency of SGD and leverages hardware acceleration for matrix operations.

There are some challenges and improvements for this method.

Learning Rate Scheduling: SGD is highly sensitive to the learning rate. Common optimization strategies include introducing learning rate decay (e.g., exponential decay, cosine annealing) or employing adaptive learning rate mechanisms.

Oscillation Issues: To mitigate oscillations during parameter updates, a Momentum mechanism is frequently introduced to smooth the update direction and accelerate convergence.

Advanced Optimizers: By combining momentum and adaptive learning rates, advanced optimization algorithms such as Adam and RMSprop have been developed [7], becoming the standard choices for training deep learning models.

In summary, SGD is a simple yet powerful optimization framework that serves as the cornerstone of large-scale machine learning. Through the integration of mini-batch sampling, learning rate scheduling, and momentum, its inherent volatility is effectively controlled.

3.2 The Intrinsic Connection Between SGD and the Law of Large Numbers

Stochastic Gradient Descent (SGD) is one of the most fundamental optimization algorithms in machine learning, primarily used to minimize loss functions by iteratively updating model parameters. Unlike Batch Gradient Descent, which computes gradients using the entire dataset, the core idea of SGD is to randomly select a single training sample at each iteration to estimate the gradient and immediately update the model parameters.

Let the random index variable $i_t \sim Uniform\{1, \dots, N\}$ and be mutually independent. One can define the stochastic gradient at step t as

$$g_t(\theta) = \nabla \ell(\theta, z_{i_t}) \tag{11}$$

Here, $g_t(\theta)$ forms a sequence of independent and identically distributed (i.i.d.) random vectors, satisfying unbiasedness [8,9]:

$$E[g_t(\theta)] = \frac{1}{N} \sum_{i=1}^{N} \nabla \ell(\theta, z_i) = \nabla L(\theta) \tag{12}$$

There are a few core theorems [10,11], such as Average Convergence of Stochastic Gradients Theorem. For any fixed θ, let the average gradient over T iterations be $\bar{g}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g_t(\theta)$. Then

$$\bar{g}_T(\theta) \xrightarrow{P} \nabla L(\theta). \tag{13}$$

Furthermore, given the finite second moment, it follows that: $\bar{g}_T(\theta) \xrightarrow{a.s.} \nabla L(\theta)$.

Proof:

Let $X_t = g_t(\theta)$. Based on the previous assumptions, $X_t\ is\ i.i.d., with E[X_t] = \nabla L(\theta)$, and $E \parallel X_t \parallel^2 < \infty$. The Law of Large Numbers [3,4] is applied coordinate-wise for the vector case:

(1) Weak Convergence. For any coordinate k, let its component be $X_t^{(k)}$. Then

$$E[X_t^{(k)}] = (\nabla L(\theta))^{(k)} \tag{14}$$

By the Weak Law of Large Numbers,

$$\frac{1}{T}\sum_{t=1}^{T} X_t^{(k)} \xrightarrow{P} (\nabla L(\theta))^{(k)} \tag{15}$$

Consequently, convergence holds in vector form:

$$\bar{g}_T(\theta) \xrightarrow{P} \nabla L(\theta) \tag{16}$$

(2) Strong Convergence [4,5]. Given that $E \parallel X_t \parallel^2 < \infty$. Applying the Strong Law of Large Numbers coordinate-wise directly yields $\bar{g}_T(\theta) \xrightarrow{a.s.} \nabla L(\theta)$, This completes the proof.

Define the gradient noise term at a single iteration as:

$$\xi_t = g_t(\theta_t) - \nabla L(\theta_t) \tag{17}$$

Given the current parameter θ_t , the expectation of this noise is zero, i.e., $E[\xi_t \mid \theta_t] = 0$. The SGD parameter update formula can be rewritten as:

$$\theta_{t+1} = \theta_t - \eta_t(\nabla L(\theta_t) + \xi_t) \tag{18}$$

The core role of the Law of Large Numbers [3,4,10] is demonstrated through the vanishing noise. Considering the average noise $\bar{\xi}_T = \frac{1}{T}\sum_{t=1}^{T} \xi_t$, at a fix point θ.Since$E[\xi_t] = 0$, and the variance is finite, by the Law of Large Numbers, $\bar{\xi}_T \to 0$.

This implies that the average stochastic gradient asymptotically approaches the true gradient:

$$\frac{1}{T}\sum_{t=1}^{T} g_t = \nabla L(\theta) + o(1) \tag{19}$$

Averaging the update equations yields:

$$\frac{1}{T}\sum_{t=1}^{T}(\theta_{t+1} - \theta_t) = -\frac{1}{T}\sum_{t=1}^{T} \eta_t\, g_t \tag{20}$$

If the learning rate satisfies appropriate conditions (e.g., bounded or decreasing) , combined with $\frac{1}{T}\sum_{t=1}^{T} g_t \to \nabla L(\theta)$, it follows that: The average update direction of SGD asymptotically converges to the negative direction of the true gradient,$-\nabla L(\theta)$.

4. Conclusion

In the Stochastic Gradient Descent algorithm, the stochastic gradient $g_t(\theta)$ obtained from a single sampling step constitutes a sequence of independent and identically distributed random vectors, whose mathematical expectation strictly equals the true gradient of the objective function, $\nabla L(\theta)$. According to the Law of Large Numbers, the sample average of these stochastic gradients converges to the true gradient in probability (or almost surely). Therefore, although a single parameter update contains random perturbations, from the macroscopic perspective of long-term iterations, these perturbations are gradually averaged out. This makes

the overall optimization behavior of SGD equivalent to a gradient descent process with asymptotically zero noise.

The logical relationship between stochastic gradient descent (SGD) and the law of large numbers can be distilled as follows: the essence of SGD is using the sample mean to estimate the expected gradient; the role of the law of large numbers is to ensure that the sample mean converges to the true expectation. Therefore, the final conclusion is that the effectiveness of SGD equals unbiased estimation plus the law of large numbers.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[2] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT (pp. 177–186).

[3] Durrett, R. Probability: Theory and Examples. 4th ed., Cambridge University Press, 2010.

[4] Billingsley, P. Probability and Measure. 3rd ed., Wiley, 1995.

[5] Tian, Y., Zhang, Y., & Zhang, H. (2023). Recent advances in stochastic gradient descent in deep learning. Mathematics, 11(3), 682.

[6] Sclocchi, A., & Wyart, M. (2024). On the different regimes of stochastic gradient descent. Proceedings of the National Academy of Sciences, 121(9), e2316301121.

[7] Li, T., Wang, B., Peng, C., & Yin, H. (2024). Stochastic gradient descent for kernel-based maximum correntropy criterion. Entropy, 26(12), 1104.

[8] Robbins, H., & Monro, S. (1951). A stochastic approximation method. Annals of Mathematical Statistics, 22(3), 400–407.

[9] Xia, L., Massei, S., & Hochstenbach, M. E. (2025). On the convergence of gradient descent with stochastic rounding errors under the Polyak–Łojasiewicz inequality. Computational Optimization and Applications, 90, 753–799.

[10] Lovas, A., & Rásonyi, M. (2023). Functional central limit theorem and strong law of large numbers for stochastic gradient Langevin dynamics. Applied Mathematics and Optimization, 88, 78.

[11] Nguegnang, G. M., Rauhut, H., & Terstiege, U. (2024). Convergence of gradient descent for learning linear neural networks. Advances in Continuous and Discrete Models, 2024, 23.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights