

Difficult to Be Partial to Stability or Speed in Traditional SGD: How Can SGD Be Modified to Balance Stability and Speed in Image Classification?

Guo Chen

Computer Science and technology, Beijing Jiao Tong University, Beijing, China

Abstract: In deep learning, training generative models generally requires large-scale datasets. The stochastic gradient descent (SGD) algorithm is widely adopted for optimizing such models. However, traditional SGD method have several limitations. For example, the unavoidable noise in large datasets can lead to unstable parameter updates and poor gradient convergence. Various modified SGD algorithms have been proposed in order to address these issues. This paper compares several representative variants of SGD from the perspectives of stability and convergence speed, and analyses their applicability in image classification tasks. The paper aims to introduce a comprehensive overview of the balance involved and to guide the selection of appropriate optimizers in practice.

Keywords: Deep Learning; Stochastic Gradient Descent; Gradient Convergence; Image Classification

1. Introduction

In recent years, machine learning and deep learning have been increasingly deployed in everyday intelligent services, including speech recognition and image classification [1]. Deep learning often relies on generative models, which are typically trained using the stochastic gradient descent (SGD) algorithm.

SGD is one of the fundamental optimization algorithms in deep learning [9]. It processes input training datasets in mini-batches and computes the gradient of the loss function with respect to the model parameters. The parameters are iteratively updated based on the result of the loss function. It is widely recognized that this mechanism is effective in minimizing the loss function and improving the predictive accuracy of neural networks. However, traditional SGD is not without drawbacks. For instance, as deep learning models are often trained on ever-larger datasets to achieve higher accuracy, the increased volume of

data inevitably introduces more noise into the gradient estimates [13]. This noise can cause erratic and unstable updates, which leads to slow speed and poor convergence [4]. Additionally, when using a small batch size, the updating process becomes relatively slow and inefficient [3]. However, using a large batch size may reduce the stochasticity and lead to poor generalization. The behavior of SGD in such noisy and high-dimensional settings can result in error amplification and convergence to suboptimal points [7].

Researchers have proposed many modifications to the basic SGD algorithm for stability and speed. These variants aim to strike a better balance between convergence stability and training efficiency. This paper focuses on the scenario of image classification and compare several representative modified SGD algorithms and evaluates them in terms of gradient convergence behaviour and processing speed. The paper summarizes the main advantages and disadvantages of each algorithm in a concise comparison table, which can serve as a practical reference for professionals and researchers.

2. Challenges of SGD in Image Classification

Training deep networks for image classification involves optimizing a highly non-convex loss function over millions of parameters. The gradient estimated from a mini-batch is an unbiased but noisy approximation of the full-batch gradient [2]. This noise can cause the optimization trajectory to oscillate, especially in regions of high curvature, which significantly slows down the convergence.

The heterogeneity of gradient magnitudes across parameters is another challenge. In deep networks, different layers exhibit vastly different gradient scales, which leads to a not optimal single global learning rate [5]. Adaptive methods attempt to address this by adjusting step sizes per coordinate. Furthermore, the amount of information in the training samples also varies. Some samples are

easy to classify and contribute little to learning, while hard samples provide more useful gradient information. Standard SGD treats all samples equally, missing opportunities to accelerate learning by focusing on more critical information [13]. Additionally, training on sensitive and privacy data requires gradient clipping and noise injection, which can further affect stability and convergence [11].

3. Representative Modified SGD Algorithms

This section reviews major categories of SGD variants and analyzes their mechanisms for improving stability or speed.

3.1 Momentum-Based Methods

Momentum methods accumulate past gradients to smooth the update direction. The classical momentum, also known as the heavy-ball method, introduces a velocity term that helps accelerate convergence along consistent directions and dampens oscillations [1]. Nesterov accelerated gradient (NAG) evaluates the gradient at a front position, which provides faster convergence in convex settings [6]. In deep learning, momentum coefficients around 0.9 are commonly used. These methods effectively reduce gradient noise and improve stability without sacrificing speed, though careful tuning of the learning rate is generally required.

3.2 Adaptive Learning Rate Methods

Adaptive methods adjust the learning rate for each parameter based on historical gradient magnitudes. SGDF (SGD with Wiener Filter) introduces adaptive filtering that can dynamically balance the historical momentum and current noise based on the gradient variance, which can significantly improve the stability of convergence. SGDF can also minimize MSE of gradient estimation via optimal time-varying gain, which can obtain more precise gradient directions and reduce the number of iterations to increase the speed of convergence [5]. DIVEBATCH is another adaptive method, which dynamically adjusts the batch size based on gradient diversity [8]. DIVEBATCH uses small batches initially and increases the batch size as the training progresses. This dynamic adjustment can improve the convergence speed only at the cost of a slight sacrifice of final accuracy.

3.3 Communication-Efficient Distributed Methods

In distributed training, scalability is often restricted by communication overhead. Synchronous SGD (SSGD) aggregates gradients from all workers at each iteration, which can ensure consistent convergence but lead to high latency. Asynchronous SGD (ASGD) removes the synchronization barrier, however, introduces weight staleness that can degrade accuracy.

LR-SGD reduces communication by randomly selecting complete layers for synchronization instead of compressing gradients [12]. This layer-based random sparsification introduces little extra computation and achieves competitive accuracy. Crossover-SGD further improves upon gossip-based methods by using overlapped segment-wise communication and random network topology with fair peer selection [10]. Its hierarchical variant limits the number of working units in gossip communication, which significantly enhances scalability and also maintains high validation accuracy.

3.4 Sampling and Privacy-Preserving Methods

Focal Sampling biases SGD towards samples that are found to be more important after a few training epochs. The sample importance is computed from the prediction probability and the sampling distribution is updated only once, which can reduce the probability of the online computational overhead. This method not only increases the convergence speed, but also balances the correct class importance. It is found that Focal Sampling biases SGD often outperforms standard SGD on imbalanced datasets. The result of a study also demonstrates that balancing by class importance can be more effective than balancing by sample count [13].

Differentially private SGD (DP-SGD) adds gradient clipping and noise to protect individual training samples [11]. In order to adapt to the changing gradient scales during training, Moving DP-SGD introduces progressive decay of both the clipping threshold and noise magnitude. This dynamic adjustment improves stability and provides strong privacy compared to fixed-parameter DP-SGD.

4. Trade-Offs Between Stability and Speed

The reviewed algorithms navigate the balance between the stability and speed in different ways. Momentum methods provide a good balance with little overhead, making them a reliable baseline. Adaptive methods often achieve faster early convergence but may sacrifice final

generalization, especially in image classification scenario [12]. Distributed methods like Crossover-SGD [10] significantly reduce training time in multi-worker setups, however their performance depends on hyperparameters such as the delay steps or network topology.

Sampling-based methods like Focal Sampling improve both speed and accuracy when the dataset has a difficulty imbalance, however their benefits depend on models and tasks [13]. Privacy-preserving methods may inevitably lead

to some utility loss, however, adaptive variants like Moving DP-SGD can mitigate this by adjusting the noise schedule [11].

After analyzing the mechanism of each modified SGD variant, it is found that no single optimizer dominates across all scenarios. The optimal choice depends on the size of dataset, model architecture, computational resources, and whether generalization or training speed is prioritized. Table 1 summarizes the main characteristics of each category.

Table 1. Comparison of SGD Variants for Image Classification

Category	Representative Methods	Stability	Speed	Main Advantage	Main Limitation
Momentum	Momentum, NAG	High	Moderate	Simple, effective	Requires tuning
Adaptive	SGDF, DIVEBATCH	High	High	Per-parameter scaling, robust	Potential generalization gap
Distributed	LR-SGD, Crossover-SGD	High	Very high	Reduced communication	Hyperparameter sensitivity
Sampling	Focal Sampling	High	Moderate	Focus on information samples	Model-dependent
Privacy	DP-SGD, Moving DP-SGD	Moderate	Low	Privacy guarantee	Utility trade-off

5. Discussion and Recommendations

5.1 Practical Guidelines

For image classification tasks, SGD with momentum remains an optimal scheme due to its simplicity and outstanding generalization [9]. When faster initial convergence is desired, SGDF [5] or DIVEBATCH [8] are recommended. However, it is necessary to carefully verify the final accuracy. LR-SGD [12] and Crossover-SGD [10] can effectively reduce time while preserving accuracy for distributed training with many workers. If the dataset is imbalanced, Focal Sampling provides a computationally cheap way to improve performance [13]. For applications with strict privacy requirements, Moving DP-SGD provides a more stable alternative to standard DP-SGD [11].

5.2 Open Challenges

Although having achieved these advances, several challenges persist. Hyperparameter tuning is significant for many algorithms, and transferable tuning strategies are still under development. The theoretical understanding of why adaptive methods sometimes generalize worse than SGD is incomplete. Moreover, the interaction between optimizers and other training techniques, for instance, data augmentation, learning rate schedules, and weight decay, is not fully involved [13]. Future research should focus on developing algorithms that can automatically adjust data characteristics and model

architectures, in order to reduce the burden of manual tuning.

6. Conclusion

This paper reviewed modified SGD algorithms for image classification, focusing on their ability to balance stability and convergence speed. Momentum methods improve stability, adaptive methods accelerate convergence, distributed methods reduce communication overhead, and sampling methods focus on information sample. Each category has different advantages and limitations, and the optimal choice depends on the specific requirements. The comparative summary provided here can serve as a practical reference for professionals and researchers. Continued development of adaptive and principled optimization techniques will further promote the efficiency and effectiveness of deep learning in image classification and beyond.

References

- [1] Fucheng, D., Wanjie, W., Ao, G., Xiaoqi, W., & Fan, W. (2025). Gradient Descent Algorithm Survey. <https://doi.org/10.48550/arxiv.2511.20725>
- [2] Ziyin, L., Liu, K., Mori, T., & Ueda, M. (2021). Strength of Minibatch Noise in SGD. <https://doi.org/10.48550/arxiv.2102.05375>
- [3] Ziyin, L., Li, B., Galanti, T., & Ueda, M. (2023). Type-II Saddles and Probabilistic Stability of Stochastic Gradient Descent. <https://doi.org/10.48550/arxiv.2303.13093>
- [4] Block, A., Foster, D. J., Krishnamurthy, A.,

- Simchowit, M., & Zhang, C. (2023). Butterfly Effects of SGD Noise: Error Amplification in Behavior Cloning and Autoregression. <https://doi.org/10.48550/arxiv.2310.11428>
- [5] Yao, Z., Yu, R., Chang, G., Li, Y., Zhang, Y., & Li, D. (2023). Signal Processing Meets SGD: From Momentum to Filter. <https://doi.org/10.48550/arxiv.2311.02818>
- [6] Hermant, J., Renaud, M., Aujol, J.-F., Dossal, C., & Rondepierre, A. (2026). Gradient correlation is a key ingredient to accelerate SGD with momentum. <https://doi.org/10.48550/arxiv.2410.07870>
- [7] Do, T., Hannibal, S., & Jentzen, A. (2024). Non-convergence to global minimizers in data driven supervised deep learning: Adam and stochastic gradient descent optimization provably fail to converge to global minimizers in the training of deep neural networks with ReLU activation. <https://doi.org/10.48550/arxiv.2410.10533>
- [8] Chen, Y., Wang, Y., & Sundaram, H. (2025). DIVEBATCH: Accelerating Model Training Through Gradient-Diversity Aware Batch Size Adaptation. <https://doi.org/10.48550/arxiv.2509.16173>
- [9] Nagwekar, A. (2025). Towards Guided Descent: Optimization Algorithms for Training Neural Networks At Scale. <https://doi.org/10.48550/arxiv.2512.18373>
- [10] Yeo, S., Bae, M., Jeong, M., Kwon, O., & Oh, S. (2023). Crossover-SGD: A gossip-based communication in distributed deep learning for alleviating large mini-batch problem and enhancing scalability. *Concurrency and Computation*, 35(15), Article e7508. <https://doi.org/10.1002/cpe.7508>
- [11] Liu, E., Chu, Z., & Zhang, X. (2025). Wasserstein GAN for moving differential privacy protection. *Scientific Reports*, 15(1), Article 19634. <https://doi.org/10.1038/s41598-025-03178-2>
- [12] Zhang, Z., Hu, Y., & Ye, Q. (2022). LR-SGD: Layer-based Random SGD For Distributed Deep Learning. *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 6–11. <https://doi.org/10.1145/3532213.3532215>
- [13] Quercia, A., Nader, F., Morrison, A., Scharr, H., & Assent, I. (2025). Focal Sampling: SGD biased towards early important samples for efficient image classification with augmentation selection: Focal Sampling: SGD biased towards early important samples. *Knowledge and Information Systems*, 67(11), 11161–11191. <https://doi.org/10.1007/s10115-025-02563-7>