Why Does Stochastic Gradient Descent Slow Down in Low-Precision Training?

Vincent-Daniel Yun JUYOUNG.YUN@USC.EDU
University of Southern California, USA
Open Neural Network Research Lab, MODULABS, Republic of Korea

Abstract

Low-precision training has become crucial for reducing the computational and memory costs of large-scale deep learning. However, quantizing gradients introduces magnitude shrinkage, which can change how stochastic gradient descent (SGD) converges. In this study, we explore SGD convergence under a gradient shrinkage model, where each stochastic gradient is scaled by a factor $q_k \in (0, 1]$. We show that this shrinkage affect the usual stepsize μ_k with an effective stepsize $\mu_k q_k$, slowing convergence when $q_{\min} < 1$. With typical smoothness and bounded-variance assumptions, we prove that low-precision SGD still converges, but at a slower pace set by $q_{\min}$, and with a higher steady error level due to quantization effects. We analyze theoretically how lower numerical precision slows training by treating it as gradient shrinkage within the standard SGD convergence setup.

1. Introduction

Deep learning models [13] have grown rapidly in size, while the amount of training data has increased exponentially with the development of the Internet [20]. Training such large-scale models requires substantial GPU memory and computational resources [22], motivating the widespread adoption of low-precision and mixed-precision training. Modern accelerators typically compute gradients in low precision (such as FP16, FP8, or FP4) [3, 15, 23, 27, 29] for speed and memory efficiency, while maintaining model parameters and accumulators in full precision (FP32). This hybrid approach is now standard practice in large-scale training pipelines but introduces an important numerical side effect. Low-precision gradients do not preserve their original magnitudes faithfully. Quantization maps many small values to coarse levels or zero, causing what we refer to as a *systematic shrinkage* of gradient magnitudes during backpropagation. If g denotes the true full-precision gradient, the quantized gradient produced by FP16/FP8/FP4 arithmetic can be written as $\tilde{g} = q\,g + \varepsilon$, where $q \in (0, 1]$ is a shrinkage factor and ε represents quantization noise. In practice, FP16 introduces only mild shrinkage, whereas lower precisions such as FP8 or FP4 exhibit substantially larger contraction of gradient magnitudes because of their limited dynamic range. Since optimizers apply the update $w \leftarrow w - \mu\tilde{g}$, this shrinkage reduces the effective stepsize from μ to $\mu_{\text{eff}} = \mu q$, thereby slowing optimization and increasing the steady-state error compared to FP32 training. Although convergence of low-precision SGD has been established in several works [2, 7, 30], the direct mathematical effect of shrinkage on the convergence rate has not been clearly isolated.

In this work, we explicitly incorporate the shrinkage factor q into the classical SGD convergence framework [1, 12, 18]. By tracking how q modifies the descent inequality, the second-moment bounds, and the resulting convergence rate, we show that low-precision arithmetic effec-

© V.-D. Yun.

tively rescales the stepsize and slows down training. Our analysis provides a simple theoretical explanation for the empirically observed slowdown in FP16/FP8/FP4 training and suggests how stepsize scheduling should be adapted in such low-precision regimes. A comprehensive overview of low-precision techniques is provided in Appendix 2.

Our contribution. We analyze mixed-precision training under a gradient-shrinkage model and show that quantized gradients induce an effective stepsize μq that slows convergence and enlarges the error floor. To highlight this central implication of our analysis, we formally state the following key observation:

"Low-precision training reduces the effective stepsize to μq, thereby requiring a larger nominal stepsize to attain faster convergence."

By incorporating the shrinkage factor directly into the standard SGD convergence proof, we derive explicit bounds quantifying this slowdown and support the theory with short numerical experiments demonstrating the shrinkage in commonly used low-precision formats.

2. Related Works

Large neural network model has led to remarkable performance improvements [20], but also raised concerns over computational cost, energy efficiency, and accessibility. To address these, various compression and acceleration techniques have been proposed, including quantization [5, 17], pruning [11, 14], and knowledge distillation [16]. These approaches reduce model size, memory usage, and inference cost, often with minimal accuracy loss.

Reducing the precision of weights, activations, and gradients is an effective way to cut memory and computation requirements [6, 9]. Low-precision formats can be floating-point (e.g., FP16, FP8, FP4) [3, 15, 22–24, 27–29] or fixed-point [4]. While fixed-point offers speed and memory benefits, it often suffers from limited dynamic range, especially for complex tasks [21]. Mixed-precision training [22] combines low-precision computations with high-precision accumulations to maintain accuracy, and has been widely adopted in modern hardware, including NVIDIA GPUs and Google TPUs [19]. Specialized accelerators such as BitFusion [26] and FPGA-based solutions [25] further optimize low-precision execution.

Low-Precision Training Error. Prior work has attributed the accuracy degradation in low-precision training to several numerical effects. Earlier analyses [8] show that quantization introduces additional stochastic noise that limits statistical accuracy when left uncompensated. Subsequent studies [10] report that coarse quantization perturbs activation distributions, leading to bias shifts that are especially harmful in architectures with sensitive activation statistics. Other investigations [31] observe that low-bitwidth networks are more prone to suboptimal minima, suggesting that reduced precision inherently complicates optimization.

While these studies primarily focus on noise amplification, activation distortion, or architectural sensitivity, our work offers a complementary and distinct perspective. We show that low-precision arithmetic induces a *systematic shrinkage* of gradient magnitudes, which directly reduces the effective stepsize during optimization. By incorporating this shrinkage into a standard SGD convergence framework, we provide a clean mathematical explanation for the slower convergence and elevated error floor observed in low-precision training.

3. Problem Setup

Notation. We follow the standard SGD convergence proof [1]. The expectation over all sources of randomness (data sampling and quantization) is written $\mathbb{E}[\cdot]$, while the conditional expectation given the σ-algebra $\mathcal{F}_k$ of all randomness up to iteration k is $\mathbb{E}[\cdot \mid \mathcal{F}_k]$. At iteration k, the stochastic gradient is $g(w_k, \xi_k) = \nabla F(w_k; \xi_k)$ and the low-precision gradient is $\tilde{g}(w_k, \xi_k) = q_k\, g(w_k, \xi_k) + \varepsilon_k$, where the shrinkage factor $q_k \in [q_{\min}, q_{\max}] \subset (0, 1]$ and the quantization noise ε_k satisfies $\mathbb{E}[\varepsilon_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\varepsilon_k\|_2^2] \le \sigma_\varepsilon^2$. If the nominal stepsize is $\mu_k > 0$, the effective stepsize is $\mu_k q_k$.

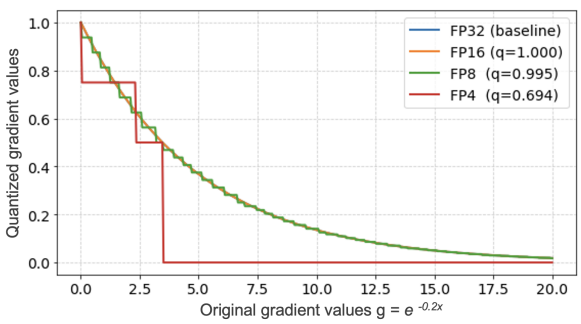

Figure 1: Quantization effect on a slowly decaying gradient-like function $g = e^{-0.2x}$ without AMP or loss scaling.

Problem Setup. We minimize the expected loss $F(w) = \mathbb{E}_\xi[\ell(\xi, w)]$ over $w \in \mathbb{R}^d$, where ξ is drawn from data distribution $\mathcal{D}$. The optimization goal is to find $w_* \in \mathbb{R}^d$ such that $F(w_*) = \min_w F(w)$. In the full-precision setting, SGD updates parameters via $w_{k+1} = w_k - \mu_k\, g(w_k, \xi_k)$, where $g(w_k, \xi_k) = \nabla F(w_k; \xi_k)$ and $\xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{D}$. In low-precision formats (e.g., FP16, FP8, FP4), the update becomes $w_{k+1} = w_k - \mu_k q_k\, g(w_k, \xi_k) - \mu_k\, \varepsilon_k$,.

Figure 1 illustrates this phenomenon for a smoothly decaying gradient-like signal g, showing how quantization maps many small values to zero or coarse levels, thereby reducing the overall magnitude. From FP16 to FP4, the shrinkage factor q decreases noticeably, and the gradient curve deviates more from the FP32 baseline. The q values were computed by measuring the ratio $\|\tilde{g}\|_2 / \|g\|_2$ after quantizing g to each format without AMP or loss scaling.

Under standard convergence assumptions [1] but with the low-precision modifications above, the descent inequality effectively replaces μ_k by $\mu_k q_{\min}$, leading to slower convergence when $q_{\min} < 1$, while the noise term ε_k adds extra variance to the error floor. In the next theoretical analysis section, we show that low-precision SGD still converges under these conditions by adapting a basic proof of SGD convergence [1], and highlight how the stepsize shrinkage impacts the convergence speed.

4. Theoretical Analysis

We prove SGD convergence, showing that low-precision SGD converges more slowly, using two key ingredients: (i) smoothness of the objective and (ii) bounds on the first/second moments of the stochastic gradients $\{\tilde{g}(w_k, \xi_k)\}$ under the standard SGD proof of convergence [1]. Here, for notational simplicity, we denote $\mathbb{E}_{\xi_k, \mu_k, \varepsilon_k}[\cdot]$ by $\mathbb{E}_{\xi_k}[\cdot]$.

Assumption 1 (Lipschitz-continuous objective gradients) *The objective $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and its gradient ∇F is L-Lipschitz continuous: $\|\nabla F(w) - \nabla F(\bar{w})\|_2 \le L\|w - \bar{w}\|_2$ where $\forall \{w, \bar{w}\} \subset \mathbb{R}^d$. This condition ensures that the gradient does not vary too rapidly*

with respect to w, a standard requirement for convergence analysis. A direct consequence is

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^\top (w - \bar{w}) + \frac{1}{2} L \|w - \bar{w}\|_2^2, \quad \forall \{w, \bar{w}\} \subset \mathbb{R}^d. \tag{1}$$

Assumption 2 (First and second moment limits with quantization) *The objective function and SGD satisfy the following:*
(a) *The sequence of iterates $\{w_k\}$ is contained in an open set over which F is bounded below by a scalar $F_{\inf}$. This requires F to be bounded below in the region of iterates.*
(b) *There exist scalars $\mu_G \geq \mu > 0$ and $q_{\min} > 0$ such that, for all $k \in \mathbb{N}$. This ensures $-\tilde{g}(w_k, \xi_k)$ is a sufficient descent direction with magnitude comparable to $\nabla F(w_k)$ but reduced by $q_{\min}$,*

$$\nabla F(w_k)^\top \mathbb{E}_{\xi_k}[\tilde{g}(w_k, \xi_k)] \geq q_{\min} \mu \|\nabla F(w_k)\|_2^2, \tag{2}$$

$$\|\mathbb{E}_{\xi_k}[\tilde{g}(w_k, \xi_k)]\|_2 \leq q_{\max} \mu_G \|\nabla F(w_k)\|_2. \tag{3}$$

(c) *There exist scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $k \in \mathbb{N}$, where $\tilde{M} := q_{\max}^2 M + M_\varepsilon$ and $\tilde{M}_V := q_{\max}^2 M_V + M_{\varepsilon,V}$ account for quantization noise. This bounds the variance of $\tilde{g}(w_k, \xi_k)$, allowing it to be nonzero at stationary points and grow quadratically for convex quadratics*

$$\mathbb{V}_{\xi_k, q_k, \varepsilon_k} \|\tilde{g}(w_k, \xi_k)\|_2 \leq \tilde{M} + \tilde{M}_V \|\nabla F(w_k)\|_2^2, \tag{4}$$

Assumption 3 (Strong convexity) *The objective $F : \mathbb{R}^d \to \mathbb{R}$ is strongly convex: there exists $c > 0$ such that*

$$F(\overline{w}) \geq F(w) + \nabla F(w)^\top (\overline{w} - w) + \frac{1}{2} c \|\overline{w} - w\|_2^2, \tag{5}$$

for all $(\overline{w}, w) \in \mathbb{R}^d \times \mathbb{R}^d$. This implies F has a unique minimizer $w_* \in \mathbb{R}^d$ with $F_* := F(w_*)$, and

$$2c\,(F(w) - F_*) \leq \|\nabla F(w)\|_2^2 \quad \text{for all } w \in \mathbb{R}^d. \tag{6}$$

Lemma 1 *Under Assumption 1, the iterates of SGD with low-precision gradient $\tilde{g}(w_k, \xi_k)$ satisfy, for all $k \in \mathbb{N}$,*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^\top \mathbb{E}_{\xi_k}[\tilde{g}(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 L\, \mathbb{E}_{\xi_k}[\|\tilde{g}(w_k, \xi_k)\|_2^2]. \tag{7}$$

Lemma 2 *Under Assumptions 1 and 2, the iterates of SGD satisfy, for all $k \in \mathbb{N}$,*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq \underbrace{-q_{\min} \mu \alpha_k \|\nabla F(w_k)}_{\textit{stepsize shrinkage}} \|_2^2 + \frac{1}{2} \alpha_k^2 L\, \mathbb{E}_{\xi_k}[\|\tilde{g}(w_k, \xi_k)\|_2^2], \tag{8}$$

$$\leq -\underbrace{\left(q_{\min} \mu - \frac{1}{2} \alpha_k L \tilde{M}_G\right)}_{\textit{stepsize shrinkage}} \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L \tilde{M}. \tag{9}$$

Theorem 3 (Strongly Convex Objective, Fixed Stepsize with Quantization) *Under Assumptions 1, 2, and 3 with $F_{\inf} = F_*$, we suppose SGD uses the low-precision gradient $\tilde{g}(w_k, \xi_k)$ satisfying $0 < \bar{\alpha} \leq \frac{\mu_q}{L \tilde{M}_G}$, where $\mu_q := q_{\min} \mu$. Then, for all $k \in \mathbb{N}$,*

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha} L \tilde{M}}{2c\mu_q} + \underbrace{(1 - \bar{\alpha} c \mu_q)^{k-1}}_{\textit{stepsize shrinkage}} \left(F(w_1) - F_* - \frac{\bar{\alpha} L \tilde{M}}{2c\mu_q} \right), \tag{10}$$

$$\xrightarrow[k \to \infty]{} \quad \frac{\bar{\alpha} L \tilde{M}}{2c\mu_q} = \frac{1}{q_{min}} \frac{\bar{\alpha} L \tilde{M}}{2c\mu} \tag{11}$$

Note that when $q_{\min} < 1$, the reduced factor $\mu_q = q_{\min}\mu$ makes $(1 - \bar{\alpha}c\mu_q)^{k-1}$ decay at a slower rate, thereby reducing the convergence speed. Specifically, since μ_q appears in the denominator of the limit term $\frac{\bar{\alpha}L\tilde{M}}{2c\mu_q}$, a smaller μ_q increases this term, leading to a larger asymptotic error bound.

For stepsizes $\alpha_r = \alpha_1 2^{-r}$, the bound is $\mathbb{E}[F(w_{k_r+1}) - F_*] \leq 2F_{\alpha_r}$ and $F_{\alpha_r} := \frac{\alpha_r L\tilde{M}}{2c\mu_q}$ requiring $k_{r+1} - k_r \approx \frac{\log 3}{\alpha_r c\mu_q} = \mathcal{O}(2^r)$, with the Robbins–Monro condition $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$.

Theorem 4 (Strongly Convex Objective, Diminishing Stepsizes with Quantization) *Under Assumptions 1, 2, and 3 (with $F_{\inf} = F_*$), suppose that SGD with the low-precision gradient $\tilde{g}(w_k, \xi_k)$ is run with a stepsize sequence $\alpha_k = \frac{\beta}{\gamma+k}, \beta > \frac{1}{c\mu_q}, \gamma > 0$, such that $\alpha_1 \leq \frac{\mu_q}{L\tilde{M}_G}$, where $\mu_q := q_{\min}\mu$. Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies*

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu_q}{\gamma + k}, \tag{12}$$

where $\nu_q := \max\left\{\frac{\beta^2 L\tilde{M}}{2(\beta c\mu_q - 1)},\ (\gamma+1)(F(w_1) - F_*)\right\}$. Since μ_q appears in the denominator of $\frac{\beta^2 L\tilde{M}}{2(\beta c\mu_q - 1)}$, a smaller μ_q reduces the denominator and thus increases the bound.

Based on Theorem 4, when $q_{\min} < 1$, the reduced effective coefficient $\mu_q = q_{\min}\mu$ makes the $O(1/k)$ convergence rate slower and increases the constant factor in the bound.

Remark 5 (Convergence Under Quantization) *Even though quantization introduces the shrinkage factor $q_{\min} < 1$ and reduces the effective coefficient $\mu_q = q_{\min}\mu$, which slows the $O(1/k)$ convergence rate, Theorem 4 still guarantees convergence. Since the bound $\nu_q/(\gamma + k)$ tends to zero as $k \to \infty$, the optimality gap satisfies $\mathbb{E}[F(w_k) - F_*] \to 0$. Thus, low-precision SGD continues to converge to the optimum, with the shrinkage factor influencing only the speed of convergence.*

Comparison with full precision. In the full-precision case ($q_k \equiv 1$, $\varepsilon_k \equiv 0$), the same argument yields $\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu}{\gamma+k}$ and $\nu := \max\left\{\frac{\beta^2 LM}{2(\beta c\mu - 1)},\ (\gamma+1)(F(w_1) - F_*)\right\}$. Relating the noise-dominated terms, we define the inflation factor $\rho := \frac{\tilde{M}}{M} \cdot \frac{\beta c\mu - 1}{\beta c\mu_q - 1} (> 0)$, so that is $\frac{\beta^2 L\tilde{M}}{2(\beta c\mu_q - 1)} = \rho \cdot \frac{\beta^2 LM}{2(\beta c\mu - 1)}$. Consequently where $A := \frac{\beta^2 LM}{2(\beta c\mu - 1)}$, $B := (\gamma+1)(F(w_1) - F_*)$,

$$\nu_q = \max\{\rho A,\ B\}, \nu = \max\{A,\ B\}, \tag{13}$$

In particular, if $\rho \geq 1$ (e.g., when $q_{\min} < 1$ and $\tilde{M}$ is sufficiently large relative to M so that $\tilde{M}/M \geq \frac{\beta c\mu_q - 1}{\beta c\mu - 1}$), then $\nu_q \geq \nu$ and thus $\frac{\nu}{\gamma+k} \leq \frac{\nu_q}{\gamma+k}$ for all $k \in \mathbb{N}$. This makes explicit that quantization (via μ_q and $\tilde{M}$) weakens the bound compared to full precision.

5. Experimental Results

We trained a ResNet-50 model on CIFAR-10 for 100 epochs with a batch size of 256 and compared the behavior of SGD under different gradient precisions. Figures 2 (A) and (B) use a stepsize of 1×10^{-4} and show that FP32, FP16, and FP8 produce nearly identical training dynamics, while FP4 exhibits noticeably slower loss reduction and lower test accuracy, reflecting the effect of gradient shrinkage on the effective stepsize.

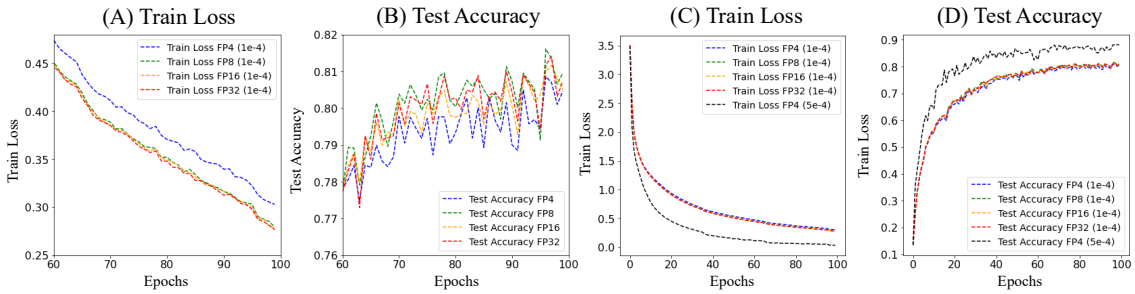

Figure 2: Training curves for ResNet-50 on CIFAR-10 over 100 epochs. (A) Train loss using FP4, FP8, FP16, and FP32 quantized gradients with a stepsize of 1×10^{-4}. (B) Test accuracy under the same setting. (C) Train loss using FP4 gradients with an increased stepsize of 5×10^{-4}. (D) Test accuracy for FP4 with the larger stepsize.

Figures 2 (C) and (D) repeat the FP4 experiment with a larger stepsize of 5×10^{-4}. In this setting, FP4 recovers much faster descent and substantially improved accuracy. These results empirically confirm that the degradation observed in low-precision training arises primarily from stepsize shrinkage and can be mitigated by increasing the nominal stepsize.

6. Conclusion

Our analysis establishes that the slowdown in low-precision SGD arises fundamentally from the gradient shrinkage factor $q_{\min} < 1$. By incorporating this factor into the descent inequality and the moment bounds, we show that the effective stepsize becomes $\alpha_k \mu_q$ with $\mu_q = q_{\min}\mu$, directly scaling down the rate at which the iterates approach the optimum. This modification yields slower convergence in both the fixed-stepsize and diminishing-stepsize regimes and increases the asymptotic error floor.

Our proof follows the classical SGD convergence framework and isolates a simple and explicit mechanism, namely stepsize shrinkage, that explains why lower numerical precision leads to reduced optimization speed. The accompanying experiments further validate this effect: FP4 training converges more slowly under the same nominal stepsize, while increasing the stepsize compensates for shrinkage and restores behavior comparable to higher-precision training. Together, these theoretical and empirical findings clarify the role of gradient shrinkage in low-precision optimization and offer practical guidance for selecting stepsizes in mixed-precision training.

7. Acknowledgement

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

[1] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

[2] Matteo Cacciola, Antonio Frangioni, Masoud Asgharian, Alireza Ghaffari, and Vahid Partovi Nia. On the convergence of stochastic gradient descent in low-precision number formats. *arXiv preprint arXiv:2301.01651*, 2023. doi: 10.48550/arXiv.2301.01651.

[3] Léopold Cambier, Anahita Bhiwandiwalla, Ting Gong, et al. Shifted and squeezed 8-bit floating point format for low-precision training of deep neural networks. *arXiv preprint arXiv:2001.05674*, 2020.

[4] Xi Chen, Xiaolin Hu, Hucheng Zhou, and Ningyi Xu. FxpNet: Training a deep convolutional neural network in fixed-point representation. In *Proceedings of the International Joint Conference on Neural Networks*, 2017.

[5] Jungwook Choi, Zhiwei Wang, Swagath Venkataramani, Puneet Chuang, Vijayalakshmi Srinivasa, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. In *International Conference on Learning Representations*, 2018.

[6] Christopher De Sa, Matthew Feldman, Christopher Ré, and Kunle Olukotun. Understanding and optimizing asynchronous low-precision stochastic gradient descent. In *Proceedings of International Symposium on Computer Architecture*, 2017.

[7] Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R. Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018. doi: 10.48550/arXiv.1803.03383.

[8] Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018.

[9] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2024.

[10] Alexander Finkelstein, Uri Almog, and Mark Grobman. Fighting quantization bias with bias. *arXiv preprint arXiv:1906.03193*, 2019.

[11] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. In *International Conference on Learning Representations*, 2019.

[12] Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2024. doi: 10.48550/arXiv.2301.11235.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[14] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015.

[15] Zhiwei Hao, Jianyuan Guo, Li Shen, Yong Luo, Han Hu, Guoxia Wang, Dianhai Yu, Yonggang Wen, and Dacheng Tao. Low-precision training of large language models: Methods, challenges, and opportunities. *arXiv preprint arXiv:2505.01043*, 2025. doi: 10.48550/arXiv.2505.01043.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

[17] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] Arnulf Jentzen and Adrian Riekert. A proof of convergence for stochastic gradient descent in the training of artificial neural networks with relu activation for constant target functions. *Zeitschrift für angewandte Mathematik und Physik*, 73(5):188, 2022. doi: 10.1007/s00033-022-01716-w.

[19] Ulrich Koster et al. Bf16: Revisiting bf16 training. *Proceedings of the International Conference on Machine Learning*, 2020.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[21] Darryl D. Lin, Sachin S. Talathi, and V. Sreekanth Annapureddy. Fixed point quantization of deep convolutional networks. In *Proceedings of the International Conference on Machine Learning*, 2016.

[22] Paulius Micikevicius, Sharan Narang, Jonah Alben, Greg Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

[23] Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. FP8 Formats for Deep Learning. *arXiv preprint arXiv:2209.05433*, 2022. doi: 10.48550/arXiv.2209.05433.

[24] Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, et al. Fp8-lm: Training fp8 large language models. *arXiv preprint arXiv:2310.18313*, 2023.

[25] Sascha Ristov, Erez Malkin, and Zeljko Zilic. Efficient deep learning inference on embedded systems using fixed-point arithmetic on fpgas. *Journal of Signal Processing Systems*, 91(1): 1–13, 2019.

[26] Hardik Sharma, Jongse Park, Naveen Suda, Liangzhen Lai, Benson Chau, Joon Kyung Kim, Vikas Chandra, and Hadi Esmaeilzadeh. Bit Fusion: Bit-level dynamically composable architecture for accelerating deep neural networks. In *Proceedings of International Symposium on Computer Architecture*, 2017.

[27] Xiao Sun, Jungwook Choi, Chia-Yu Chen, Naigang Wang, Swagath Venkataramani, Vijayalakshmi (Viji) Srinivasan, Xiaodong Cui, Wei Zhang, and Kailash Gopalakrishnan. Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Xiao Sun, Naigang Wang, Chia-Yu Chen, Jiamin Ni, Ankur Agrawal, Xiaodong Cui, Swagath Venkataramani, Kaoutar El Maghraoui, Vijayalakshmi (Viji) Srinivasan, and Kailash Gopalakrishnan. Ultra-low precision 4-bit training of deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1796–1807. Curran Associates, Inc., 2020.

[29] Juyoung Yun, Sol Choi, Francois Rameau, Byungkon Kang, and Zhoulai Fu. Revisiting 16-bit neural network training: A practical approach for resource-limited learning. *arXiv preprint arXiv:2305.10947*, 2025. doi: 10.48550/arXiv.2305.10947.

[30] Ruqi Zhang, Andrew Gordon Wilson, and Christopher De Sa. Low-precision stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2022. doi: 10.48550/arXiv.2206.09909.

[31] Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, and Ian Reid. Towards effective low-bitwidth convolutional neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Appendix

Assumption 1 (Lipschitz-continuous objective gradients) *The objective $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and its gradient ∇F is L-Lipschitz continuous: $\|\nabla F(w) - \nabla F(\bar{w})\|_2 \leq L\|w - \bar{w}\|_2$ where $\forall\{w, \bar{w}\} \subset \mathbb{R}^d$. This condition ensures that the gradient does not vary too rapidly with respect to w, a standard requirement for convergence analysis. A direct consequence is*

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^\top (w - \bar{w}) + \frac{1}{2} L\|w - \bar{w}\|_2^2, \quad \forall\{w, \bar{w}\} \subset \mathbb{R}^d. \tag{14}$$

Assumption 2 (First and second moment limits with quantization) *The objective function and SGD satisfy the following:*
(a) *The sequence of iterates $\{w_k\}$ is contained in an open set over which F is bounded below by a scalar $F_{\inf}$. This requires F to be bounded below in the region of iterates.*
(b) *There exist scalars $\mu_G \geq \mu > 0$ and $q_{\min} > 0$ such that, for all $k \in \mathbb{N}$. This ensures $-\tilde{g}(w_k, \xi_k)$ is a sufficient descent direction with magnitude comparable to $\nabla F(w_k)$ but reduced by $q_{\min}$,*

$$\nabla F(w_k)^\top \mathbb{E}_{\xi_k}[\tilde{g}(w_k, \xi_k)] \geq q_{\min}\mu\|\nabla F(w_k)\|_2^2, \tag{15}$$

$$\|\mathbb{E}_{\xi_k}[\tilde{g}(w_k, \xi_k)]\|_2 \leq q_{\max}\mu_G\|\nabla F(w_k)\|_2. \tag{16}$$

(c) *There exist scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $k \in \mathbb{N}$, where $\tilde{M} := q_{\max}^2 M + M_\varepsilon$ and $\tilde{M}_V := q_{\max}^2 M_V + M_{\varepsilon,V}$ account for quantization noise. This bounds the variance of $\tilde{g}(w_k, \xi_k)$, allowing it to be nonzero at stationary points and grow quadratically for convex quadratics*

$$\mathbb{V}_{\xi_k, q_k, \varepsilon_k}\|\tilde{g}(w_k, \xi_k)\|_2 \leq \tilde{M} + \tilde{M}_V\|\nabla F(w_k)\|_2^2, \tag{17}$$

Assumption 3 (Strong convexity) *The objective $F : \mathbb{R}^d \to \mathbb{R}$ is strongly convex: there exists $c > 0$ such that*

$$F(\overline{w}) \geq F(w) + \nabla F(w)^\top (\overline{w} - w) + \frac{1}{2} c\|\overline{w} - w\|_2^2, \tag{18}$$

for all $(\overline{w}, w) \in \mathbb{R}^d \times \mathbb{R}^d$. This implies F has a unique minimizer $w_* \in \mathbb{R}^d$ with $F_* := F(w_*)$, and

$$2c\,(F(w) - F_*) \leq \|\nabla F(w)\|_2^2 \quad \textit{for all } w \in \mathbb{R}^d. \tag{19}$$

Lemma 1 *Under Assumption 1, the iterates of SGD with low-precision gradient $\tilde{g}(w_k, \xi_k)$ satisfy, for all $k \in \mathbb{N}$,*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^\top \mathbb{E}_{\xi_k}[\tilde{g}(w_k, \xi_k)] + \frac{1}{2}\alpha_k^2 L\, \mathbb{E}_{\xi_k}[\|\tilde{g}(w_k, \xi_k)\|_2^2]. \tag{20}$$

Proof From Assumption 1,

$$F(w_{k+1}) - F(w_k) \leq \nabla F(w_k)^\top (w_{k+1} - w_k) + \frac{1}{2} L\|w_{k+1} - w_k\|_2^2 \tag{21}$$

$$\leq -\alpha_k \nabla F(w_k)^\top \big(\underbrace{q_k\, g(w_k, \xi_k) + \varepsilon_k}_{\text{stepsize shrinkage}}\big) + \frac{1}{2}\alpha_k^2 L\, \|\underbrace{q_k\, g(w_k, \xi_k) + \varepsilon_k}_{\text{stepsize shrinkage}}\|_2^2. \tag{22}$$

Taking expectations over $(\xi_k, q_k, \varepsilon_k)$, with w_k fixed, yields (20). ■

This bound expresses the expected one-step change as the sum of a descent term and a curvature-dependent penalty, both influenced by q_k and ε_k. If $\tilde{g}(w_k, \xi_k)$ is unbiased, then

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\, \mathbb{E}_{\xi_k}[\| \underbrace{q_k\, g(w_k, \xi_k) + \varepsilon_k}_{\text{stepsize shrinkage}} \|_2^2]. \tag{23}$$

We guarantee SGD convergence when the stochastic directions and stepsizes make the right-hand side of (20) bounded by a deterministic term that ensures sufficient descent in F. This requires constraints on the first and second moments of $\{\tilde{g}(w_k, \xi_k)\}$ to limit the effect of the last term in (23). We restrict the variance of $\tilde{g}$ as $\mathbb{V}_{\xi_k, q_k, \varepsilon_k}[\tilde{g}(w_k, \xi_k)] := \mathbb{E}\left[\|\tilde{g}(w_k, \xi_k)\|_2^2\right] - \|\mathbb{E}[\tilde{g}(w_k, \xi_k)]\|_2^2$.

Together with the variance of $\tilde{g}$, these give the second moment bound:

$$\mathbb{E}_{\xi_k}\left[\|\tilde{g}(w_k, \xi_k)\|_2^2\right] \leq \tilde{M} + \tilde{M}_G \|\nabla F(w_k)\|_2^2 \quad \tilde{M}_G := \tilde{M}_V + q_{\max}^2 \mu_G^2 \geq (q_{\min}\mu)^2 > 0. \tag{24}$$

The next lemma extends Lemma 1 under Assumption 2.

Lemma 2 *Under Assumptions 1 and 2, the iterates of SGD satisfy, for all $k \in \mathbb{N}$,*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq \underbrace{-q_{\min}\mu\alpha_k \|\nabla F(w_k)}_{\textit{stepsize shrinkage}} \|_2^2 + \frac{1}{2}\alpha_k^2 L\, \mathbb{E}_{\xi_k}[\|\tilde{g}(w_k, \xi_k)\|_2^2], \tag{25}$$

$$\leq -\underbrace{\left(q_{\min}\mu - \frac{1}{2}\alpha_k L \tilde{M}_G\right)}_{\textit{stepsize shrinkage}} \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \tilde{M}. \tag{26}$$

Proof From Lemma 1 with $\tilde{g}$, we have

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\alpha_k \nabla F(w_k)^\top \mathbb{E}_{\xi_k}\left[\tilde{g}(w_k, \xi_k)\right] + \frac{1}{2}\alpha_k^2 L\, \mathbb{E}_{\xi_k}\left[\|\tilde{g}(w_k, \xi_k)\|_2^2\right] \tag{27}$$

$$\leq \underbrace{-q_{\min}\mu\alpha_k \|\nabla F(w_k)\|_2^2}_{\text{stepsize shrinkage}} + \frac{1}{2}\alpha_k^2 L\, \mathbb{E}_{\xi_k}\left[\|\tilde{g}(w_k, \xi_k)\|_2^2\right], \tag{28}$$

which yields (25). Applying Assumption 2(c) and the bound in (24) gives (26). ■

Theorem 3 (Strongly Convex Objective, Fixed Stepsize with Quantization) *Under Assumptions 1, 2, and 3 with $F_{\inf} = F_*$, we suppose SGD uses the low-precision gradient $\tilde{g}(w_k, \xi_k)$ satisfying $0 < \bar{\alpha} \leq \frac{\mu_q}{L\tilde{M}_G}$, where $\mu_q := q_{\min}\mu$. Then, for all $k \in \mathbb{N}$,*

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha} L \tilde{M}}{2c\mu_q} + \underbrace{(1 - \bar{\alpha} c \mu_q)^{k-1}}_{\textit{stepsize shrinkage}} \left(F(w_1) - F_* - \frac{\bar{\alpha} L \tilde{M}}{2c\mu_q} \right), \tag{29}$$

$$\xrightarrow[k\to\infty]{} \quad \frac{\bar{\alpha} L \tilde{M}}{2c\mu_q} = \frac{1}{q_{min}} \frac{\bar{\alpha} L \tilde{M}}{2c\mu} \tag{30}$$

Note that when $q_{\min} < 1$, the reduced factor $\mu_q = q_{\min}\mu$ makes $(1 - \bar{\alpha}c\mu_q)^{k-1}$ decay at a slower rate, thereby reducing the convergence speed. Specifically, since μ_q appears in the denominator of the limit term $\frac{\bar{\alpha}L\tilde{M}}{2c\mu_q}$, a smaller μ_q increases this term, leading to a larger asymptotic error bound.

Proof From Lemma 2, for all $k \in \mathbb{N}$ we have

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\left(\mu_q - \frac{1}{2}\bar{\alpha}L\tilde{M}_G\right)\bar{\alpha}\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 L\tilde{M} \tag{31}$$

$$\leq -\frac{1}{2}\bar{\alpha}\mu_q\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\bar{\alpha}^2 L\tilde{M} \tag{32}$$

$$\leq -\bar{\alpha}c\mu_q\left(F(w_k) - F_*\right) + \frac{1}{2}\bar{\alpha}^2 L\tilde{M}, \tag{33}$$

where (33) follows from the strong convexity bound (19). Subtracting F_* and taking expectations gives $\mathbb{E}[F(w_{k+1}) - F_*] \leq (1 - \bar{\alpha}c\mu_q)\,\mathbb{E}[F(w_k) - F_*] + \frac{1}{2}\bar{\alpha}^2 L\tilde{M}$. Subtracting $\frac{\bar{\alpha}L\tilde{M}}{2c\mu_q}$ from both sides yields

$$\mathbb{E}[F(w_{k+1}) - F_*] - \frac{\bar{\alpha}L\tilde{M}}{2c\mu_q} = (1 - \bar{\alpha}c\mu_q)\left(\mathbb{E}[F(w_k) - F_*] - \frac{\bar{\alpha}L\tilde{M}}{2c\mu_q}\right). \tag{34}$$

Since $0 < \bar{\alpha}c\mu_q \leq \frac{c\mu_q^2}{LM_G} \leq \frac{c\mu_q^2}{L\mu_k} = \frac{c}{L} \leq 1$, repeated application of (34) gives the bound in (29) and the claimed limit. ■

For stepsizes $\alpha_r = \alpha_1 2^{-r}$, the bound is $\mathbb{E}[F(w_{k_r+1}) - F_*] \leq 2F_{\alpha_r}$ and $F_{\alpha_r} := \frac{\alpha_r L\tilde{M}}{2c\mu_q}$ requiring $k_{r+1} - k_r \approx \frac{\log 3}{\alpha_r c\mu_q} = \mathcal{O}(2^r)$, with the Robbins–Monro condition $\sum_{k=1}^{\infty}\alpha_k = \infty$ and $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$.

Theorem 4 (Strongly Convex Objective, Diminishing Stepsizes with Quantization) *Under Assumptions 1, 2, and 3 (with $F_{\inf} = F_*$), suppose that SGD with the low-precision gradient $\tilde{g}(w_k, \xi_k)$ is run with a stepsize sequence $\alpha_k = \frac{\beta}{\gamma+k}, \beta > \frac{1}{c\mu_q}, \gamma > 0$, such that $\alpha_1 \leq \frac{\mu_q}{L\tilde{M}_G}$, where $\mu_q := q_{\min}\mu$. Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies*

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu_q}{\gamma + k}, \tag{35}$$

where $\nu_q := \max\left\{\frac{\beta^2 L\tilde{M}}{2(\beta c\mu_q - 1)},\ (\gamma+1)(F(w_1) - F_*)\right\}$. Since μ_q appears in the denominator of $\frac{\beta^2 L\tilde{M}}{2(\beta c\mu_q - 1)}$, a smaller μ_q reduces the denominator and thus increases the bound.

Proof From the choice of α_k, we have $\alpha_k L\tilde{M}_G \leq \alpha_1 L\tilde{M}_G \leq \mu_q$ for all $k \in \mathbb{N}$. Applying Lemma 2 and the strong convexity property, for all $k \in \mathbb{N}$:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\left(\mu_q - \frac{1}{2}\alpha_k L\tilde{M}_G\right)\alpha_k\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\tilde{M} \tag{36}$$

$$\leq -\frac{1}{2}\alpha_k\mu_q\|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\tilde{M} \tag{37}$$

$$\leq -\alpha_k c\mu_q\left(F(w_k) - F_*\right) + \frac{1}{2}\alpha_k^2 L\tilde{M}. \tag{38}$$

Subtracting F_* and taking expectations yields

$$\mathbb{E}[F(w_{k+1}) - F_*] \leq \underbrace{(1 - \alpha_k c\mu_q)}_{\text{stepsize shrinkage}} \mathbb{E}[F(w_k) - F_*] + \frac{1}{2}\alpha_k^2 L\tilde{M}. \tag{39}$$

We prove (35) by induction. For $k = 1$, the definition of ν_q guarantees (35). Assume (35) holds for some $k \geq 1$. Substitute into (39) with $\hat{k} := \gamma + k$:

$$\begin{aligned}
\mathbb{E}[F(w_{k+1}) - F_*] &\leq \left(1 - \frac{\beta c\mu_q}{\hat{k}}\right)\frac{\nu_q}{\hat{k}} + \frac{\beta^2 L\tilde{M}}{2\hat{k}^2} && (40)\\
&= \frac{\hat{k} - \beta c\mu_q}{\hat{k}^2}\nu_q + \frac{\beta^2 L\tilde{M}}{2\hat{k}^2} && (41)\\
&= \frac{\hat{k} - 1}{\hat{k}^2}\nu_q - \underbrace{\frac{\beta c\mu_q - 1}{\hat{k}^2}\nu_q + \frac{\beta^2 L\tilde{M}}{2\hat{k}^2}}_{\leq 0 \text{ by def. of } \nu_q} && (42)\\
&\leq \frac{\nu_q}{\hat{k} + 1}, && (43)
\end{aligned}$$

where the last inequality uses $\hat{k}^2 \geq (\hat{k} + 1)(\hat{k} - 1)$. Thus, (35) holds for $k + 1$, completing the induction. ■

Based on Theorem 4, when $q_{\min} < 1$, the reduced effective coefficient $\mu_q = q_{\min}\mu$ makes the $O(1/k)$ convergence rate slower and increases the constant factor in the bound.