ANALYSIS OF QUANTIZED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Weight-quantized networks have small storage and fast inference, but training can still be time-consuming. This can be improved with distributed learning. To reduce the high communication cost due to worker-server synchronization, recently gradient quantization has also been proposed to train networks with full-precision weights. In this paper, we theoretically study how the combination of both weight and gradient quantization affects convergence. We show that (i) weight-quantized networks converge to an error related to the weight quantization resolution and weight dimension; (ii) quantizing gradients slows convergence by a factor related to the gradient quantization resolution and dimension; and (iii) clipping the gradient before quantization renders this factor dimension-free, thus allowing the use of fewer bits for gradient quantization. Empirical experiments confirm the theoretical convergence results, and demonstrate that quantized networks can speed up training and have comparable performance as full-precision networks.

1 INTRODUCTION

Deep neural networks are usually huge. Such a high demand in time and space can significantly limit the deployment on low-end devices. To alleviate this problem, many approaches have been recently proposed to compress deep networks. One direction is network quantization, which represents each network weight with a small number of bits. Besides significantly reducing the model size, it also accelerates network training and inference. Many weight quantization methods aim at approximating the full-precision weights in each iteration (Courbariaux et al., 2015; Lin et al., 2016; Rastegari et al., 2016; Li & Liu, 2016; Lin et al., 2017; Guo et al., 2017). Recently, loss-aware quantization minimizes the loss directly w.r.t. the quantized weights (Hou et al., 2017; Hou & Kwok, 2018; Leng et al., 2018), and often achieves better performance than approximation-based methods.

Distributed learning can further speed up training of weight-quantized networks (Dean et al., 2012). A key challenge is on reducing the expensive communication cost incurred during synchronization of the gradients and model parameters (Li et al., 2014a;b). Recently, algorithms that sparsify (Aji & Heafield, 2017; Wangni et al., 2017) or quantize the gradients (Seide et al., 2014; Wen et al., 2017; Alistarh et al., 2017; Bernstein et al., 2018) have been proposed.

In this paper, we consider quantization of both the weights and gradients in a distributed environment. Quantizing both weights and gradients has been explored in the DoReFa-Net (Zhou et al., 2016), QNN (Hubara et al.), WAGE (Wu et al., 2018) and ZipML (Zhang et al., 2017). We differ from them in two aspects. First, existing methods mainly consider learning with a single machine, and gradient quantization is used to reduce the computations in backpropagation. On the other hand, we consider a distributed environment, and use gradient quantization to reduce communication cost and accelerate distributed learning of weight-quantized networks. Weight update and quantization are performed at the server, and only quantized weights are transmitted over the network, leading to a reduction in communication cost from server to workers. Gradient quantization is performed by the workers. As only quantized gradients are transmitted, worker-to-server communication is also reduced. Moreover, instead of keeping both the full-precision and quantized weights as in a non-distributed environment, here the server only keeps the full-precision weights, while each worker only keeps the quantized weights. Thus, workers can be run on low-end devices with small storage.

Second, while DoReFa-Net, QNN and WAGE show impressive empirical results on the quantized network, theoretical guarantees are not provided. ZipML provides convergence analysis, but is limited to stochastic weight quantization, square loss with the linear model, and requires the stochastic gradients to be unbiased. This can be restrictive as most state-of-the-art weight quantization methods (Rastegari et al., 2016; Lin et al., 2016; Li & Liu, 2016; Guo et al., 2017; Hou et al., 2017; Hou & Kwok, 2018) are deterministic, and the resultant stochastic gradients are biased.

In this paper, we relax the restrictions on the loss function, and study in an online learning setting how biased stochastic gradients affect convergence. The main findings are:

1. With either full-precision or quantized gradients, the average regret of loss-aware weight quantization does not converge to zero, but to an error related to the weight quantization resolution Δ_w and dimension d. The smaller the Δ_w or d, the smaller is the error (Theorems 1 and 2).
2. With either full-precision or quantized gradients, the average regret converges with a $O(1/\sqrt{T})$ rate to the error, where T is the number of iterations. However, gradient quantization slows convergence (relative to using full-precision gradients) by a factor related to gradient quantization resolution Δ_g and d. The smaller the Δ_g or d, the faster is the convergence (Theorems 1 and 2).
3. Empirically, gradient clipping has been found effective in use with quantized gradient (Wen et al., 2017). We show theoretically that for gradients following the normal distribution, gradient clipping renders the speed degradation mentioned above dimension-free. However, an additional error is incurred. The convergence speedup and error are related to how aggressive clipping is performed. More aggressive clipping results in faster convergence, but a larger error (Theorem 3).
4. Empirical results verify the convergence analysis and show that quantizing gradients can significantly reduce communication cost, speed up training of weight-quantized networks, and obtain accuracy comparable with the use of full-precision gradients.

Notations. For a vector $\mathbf{x}$, $\sqrt{\mathbf{x}}$ is the element-wise square root, $\mathbf{x}^2$ is the element-wise square, Diag$(\mathbf{x})$ returns a diagonal matrix with $\mathbf{x}$ on the diagonal, and $\mathbf{x} \odot \mathbf{y}$ is the element-wise multiplication of vectors $\mathbf{x}$ and $\mathbf{y}$. For a matrix $\mathbf{Q}$, $\|\mathbf{x}\|_{\mathbf{Q}}^2 = \mathbf{x}^\top \mathbf{Q}\mathbf{x}$. For a matrix $\mathbf{X}$, $\sqrt{\mathbf{X}}$ is the element-wise square root, and diag$(\mathbf{X})$ returns a vector extracted from the diagonal elements of $\mathbf{X}$.

2 Preliminaries

2.1 Online Learning

Online learning continually adapts the model with a sequence of observations. It has been commonly used in the analysis of deep learning optimizers (Duchi et al., 2011; Kingma & Ba, 2015; Reddi et al., 2018). At time t, the algorithm picks a model with parameter $\mathbf{w}_t \in \mathcal{S}$, where $\mathcal{S}$ is a convex compact set. The algorithm then incurs a loss $f_t(\mathbf{w}_t)$. After T rounds, performance is evaluated by the regret $R(T) = \sum_{t=1}^T f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)$ and average regret $R(T)/T$, where $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{S}} \sum_{t=1}^T f_t(\mathbf{w})$ is the best model parameter in hindsight.

2.2 Weight Quantization

In BinaryConnect (Courbariaux et al., 2015), each weight is binarized using the sign function either deterministically or stochastically. In ternary-connect (Lin et al., 2016), each weight is stochastically quantized to $\{-1, 0, 1\}$. Stochastic weight quantization often suffers severe accuracy degradation, while deterministic weight quantization (as in the binary-weight-network (BWN) (Rastegari et al., 2016) and ternary weight network (TWN) (Li & Liu, 2016)) achieves much better performance.

In this paper, we will focus on loss-aware weight quantization, which further improves performance by considering the effect of weight quantization on the loss. Examples include loss-aware binarization (LAB) (Hou et al., 2017) and loss-aware quantization (LAQ) (Hou & Kwok, 2018). Let the full-precision weights from all L layers in the deep network be $\mathbf{w}$, and the corresponding quantized weight is denoted $Q_w(\mathbf{w}) = \hat{\mathbf{w}}$ where $Q_w(\cdot)$ is the weight quantization function.

At the tth iteration, the second-order Taylor expansion of $f_t(\hat{\mathbf{w}})$, i.e., $f_t(\hat{\mathbf{w}}_t) + \nabla f_t(\hat{\mathbf{w}}_t)^\top(\hat{\mathbf{w}} - \hat{\mathbf{w}}_t) + \frac{1}{2}(\hat{\mathbf{w}} - \hat{\mathbf{w}}_t)^\top \mathbf{H}_t(\hat{\mathbf{w}} - \hat{\mathbf{w}}_t)$ is minimized w.r.t. $\hat{\mathbf{w}}$, where $\mathbf{H}_t$ is the Hessian at $\hat{\mathbf{w}}_t$. $\mathbf{H}_t$ is expensive to compute. In practice, this is approximated by Diag$(\sqrt{\hat{\mathbf{v}}_t})$ where $\hat{\mathbf{v}}_t$ is the moving average: $\hat{\mathbf{v}}_t = \beta\hat{\mathbf{v}}_{t-1} + (1-\beta)\hat{\mathbf{g}}_t^2 = \sum_{j=1}^t (1-\beta)\beta^{t-j}\hat{\mathbf{g}}_j^2$ with $\beta \simeq 1$, and is readily available in popular deep network optimizers such as RMSProp and Adam. Moreover, Diag$(\sqrt{\hat{\mathbf{v}}_t})$ is also an estimate of $\text{Diag}(\sqrt{\text{diag}(\mathbf{H}_t^2)})$ (Dauphin et al., 2015). Computationally, the quantized weight can be obtained by first performing a preconditioned gradient descent $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\hat{\mathbf{v}}_t})^{-1}\hat{\mathbf{g}}_t$, followed by quantization via solving of the following optimization problem:[1]

$$\hat{\mathbf{w}}_{t+1} = Q_w(\mathbf{w}_{t+1}) = \arg\min_{\hat{\mathbf{w}}} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}\|^2_{\text{Diag}(\sqrt{\hat{\mathbf{v}}_t})} \quad \text{s.t.} \quad \hat{\mathbf{w}} = \alpha\mathbf{b},\ \alpha > 0,\ \mathbf{b} \in (\mathcal{S}_w)^d. \quad (1)$$

For binarization, $\mathcal{S}_w = \{-1, +1\}$, and a simple closed-form solution is obtained in (Hou et al., 2017). For m-bit linear quantization, $\mathcal{S}_w = \{-M_k, \dots, -M_1, M_0, M_1, \dots, M_k\}$, where $k = 2^{m-1} - 1$, $0 = M_0 < \cdots < M_k$ are uniformly spaced, with weight quantization resolution $\Delta_w = M_{r+1} - M_r$. An efficient approximate solution of (1) is obtained in (Hou & Kwok, 2018).

2.3 Stochastic Gradient Quantization

In a distributed learning environment, the main bottleneck is often on worker-to-server communication. This cost can be significantly reduced by quantizing the gradients (computed by the workers) before pushing to the server (Seide et al., 2014; Wen et al., 2017; Alistarh et al., 2017). For example, assuming that the full-precision gradient is 32-bit, the communication cost is reduced $32/m$ times when gradients are quantized to m bits. For the more advanced all-reduce communication model (Rabenseifner, 2004), the number of bits used for gradients increases by one at every aggregation step in the worst case[2]. However, even considering this effect, our proposed method still significantly reduces the network communication and speeds up training as will be shown in Figure 4 in the experiment section. In TernGrad (Wen et al., 2017), the gradients are ternarized. Here, we consider the more general m-bit linear quantization (Alistarh et al., 2017): $Q_g(\mathbf{g}_t) = s_t \cdot \text{sign}(\mathbf{g}_t) \odot \mathbf{q}_t$, where $\mathbf{g}_t$ is the stochastic gradient, $s_t = \|\mathbf{g}_t\|_\infty$, $\mathbf{q}_t \in (\mathcal{S}_g)^d$, and $\mathcal{S}_g = \{-B_k, \dots, -B_1, B_0, B_1, \dots, B_k\}$, with $k = 2^{m-1} - 1$, $0 = B_0 < B_1 < \cdots < B_k$ are uniformly spaced, and the gradient quantization resolution $\Delta_g = B_{r+1} - B_r$. The ith element $q_{t,i}$ in $\mathbf{q}_t$ is equal to B_{r+1} with probability $(|g_{t,i}|/s_t - B_r)/(B_{r+1} - B_r)$, and B_r otherwise. Here, r is an index satisfying $B_r \le |g_{t,i}|/s_t < B_{r+1}$. Note that $Q_g(\mathbf{g}_t)$ is an unbiased estimator of $\mathbf{g}_t$.

3 Properties of Quantized Network

In this section, we consider quantization of both weights and gradients in a distributed environment with N workers in data parallelism (Figure 1). At the tth iteration, worker $n \in \{1, 2, \dots, N\}$ computes the full-precision gradient $\hat{\mathbf{g}}_t^{(n)}$ w.r.t. the quantized weight and quantizes $\hat{\mathbf{g}}_t^{(n)}$ to $\tilde{\mathbf{g}}_t^{(n)} = Q_g(\hat{\mathbf{g}}_t^{(n)})$. The quantized gradients are then synchronized and averaged at the parameter server as: $\tilde{\mathbf{g}}_t = \frac{1}{N}\sum_{n=1}^N \tilde{\mathbf{g}}_t^{(n)}$. The server updates the second moment $\tilde{\mathbf{v}}_t$ based on $\tilde{\mathbf{g}}_t$, and also the full-precision weight as $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\tilde{\mathbf{v}}_t})^{-1}\tilde{\mathbf{g}}_t$. The weight is quantized using loss-aware weight quantization to produce $\hat{\mathbf{w}}_{t+1} = Q_w(\mathbf{w}_{t+1})$, which is then sent back to all the workers.

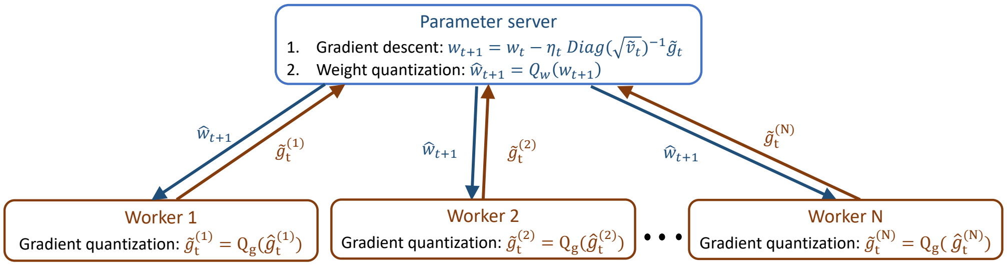

Parameter server
1. Gradient descent: $w_{t+1} = w_t - \eta_t\ Diag(\sqrt{\tilde{v}_t})^{-1}\tilde{g}_t$
2. Weight quantization: $\hat{w}_{t+1} = Q_w(w_{t+1})$
$\hat{w}_{t+1}$
$\tilde{g}_t^{(1)}$
$\hat{w}_{t+1}$
$\tilde{g}_t^{(2)}$
$\hat{w}_{t+1}$
$\tilde{g}_t^{(N)}$
Worker 1
Gradient quantization: $\tilde{g}_t^{(1)} = Q_g(\hat{g}_t^{(1)})$
Worker 2
Gradient quantization: $\tilde{g}_t^{(2)} = Q_g(\hat{g}_t^{(2)})$
• • •
Worker N
Gradient quantization: $\tilde{g}_t^{(N)} = Q_g(\hat{g}_t^{(N)})$

Figure 1: Distributed weight and gradient quantization with data parallelism.

[1] For simplicity of notations, we assume that the same scaling parameter α is used for all layers. Extension to layer-wise scaling is straightforward.

[2] Note that all the scaling parameter is shared for all workers.

3.1 Assumptions

Analysis on quantized networks has only been performed on models with (i) full-precision gradients and weights quantized by stochastic weight quantization (Li et al., 2017; De Sa et al., 2018), or simple deterministic weight quantization using the sign (Li et al., 2017); (ii) full-precision weights and quantized gradients (Alistarh et al., 2017; Wen et al., 2017; Bernstein et al., 2018); (iii) quantized weights and quantized gradients (Zhang et al., 2017), but limited to stochastic weight quantization, square loss on linear model (i.e., $f_t(\mathbf{w}_t) = (\mathbf{x}_t^\top \mathbf{w}_t - y_t)^2$) in Section 2.1), and unbiased gradient.

In this paper, we study the more advanced loss-aware weight quantization. As it is deterministic and has biased gradients, the above analysis do not apply. Moreover, we relax the assumptions on f_t as: (**A1**) f_t is convex; (**A2**) f_t is twice differentiable with Lipschitz-continuous gradient; and (**A3**) f_t has bounded gradient, i.e., $\|\nabla f_t(\mathbf{w})\| \leq G$ and $\|\nabla f_t(\mathbf{w})\|_\infty \leq G_\infty$ for all $\mathbf{w} \in \mathcal{S}$. These assumptions have been commonly used in convex online learning (Hazan, 2016; Duchi et al., 2011; Kingma & Ba, 2015) and quantized networks (Alistarh et al., 2017; Li et al., 2017). Obviously, the convexity assumption **A1** does not hold for deep networks. However, this facilitates analysis of deep learning models, and has been used in (Kingma & Ba, 2015; Reddi et al., 2018; Li et al., 2017; De Sa et al., 2018). Moreover, as will be seen, it helps to explain the empirical behavior in Section 4.

As in (Duchi et al., 2011; Kingma & Ba, 2015; Li et al., 2017), we assume that $\|\mathbf{w}_m - \mathbf{w}_n\| \leq D$ and $\|\mathbf{w}_m - \mathbf{w}_n\|_\infty \leq D_\infty$ for all $\mathbf{w}_m, \mathbf{w}_n \in \mathcal{S}$. Moreover, the learning rate η_t decays as $\eta/\sqrt{t}$, where η is a constant (Hazan, 2016; Duchi et al., 2011; Kingma & Ba, 2015; Li et al., 2017).

For simplicity of notations, we denote the full-precision gradient $\nabla f_t(\mathbf{w}_t)$ w.r.t. the full-precision weight by $\mathbf{g}_t$, and the full-precision gradient $\nabla f_t(Q_w(\mathbf{w}_t))$ w.r.t. the quantized weight by $\hat{\mathbf{g}}_t$. As f_t is twice differentiable (Assumption **A2**), using the mean value theorem, there exists $p \in (0,1)$ such that $\mathbf{g}_t - \hat{\mathbf{g}}_t = \nabla f_t(\mathbf{w}_t) - \nabla f_t(\hat{\mathbf{w}}_t) = \nabla^2 f_t\left(\hat{\mathbf{w}}_t + p(\mathbf{w}_t - \hat{\mathbf{w}}_t)\right)(\mathbf{w}_t - \hat{\mathbf{w}}_t)$. Let $\mathbf{H}'_t = \nabla^2 f_t(\hat{\mathbf{w}}_t + p(\mathbf{w}_t - \hat{\mathbf{w}}_t))$ be the Hessian at $\hat{\mathbf{w}}_t + p(\mathbf{w}_t - \hat{\mathbf{w}}_t)$. Moreover, let $\alpha = \max\{\alpha_1, \dots, \alpha_T\}$, where α_t is the scaling parameter in (1) at the tth iteration.

3.2 Weight Quantization with Full-Precision Gradient

When only weights are quantized, the update for loss-aware weight quantization is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\hat{\mathbf{v}}_t})^{-1}\hat{\mathbf{g}}_t,$$

where $\hat{\mathbf{v}}_t$ is the moving average of the (squared) gradients $\hat{\mathbf{g}}_t^2$ as discussed in Section 2.2.

Theorem 1. *For loss-aware weight quantization with full-precision gradients and $\eta_t = \eta/\sqrt{t}$,*

$$\begin{aligned} R(T) \;&\leq\; \frac{D_\infty^2\sqrt{dT}}{2\eta}\sqrt{\sum_{t=1}^{T}(1-\beta)\beta^{T-t}\|\hat{\mathbf{g}}_t\|^2} + \frac{\eta G_\infty\sqrt{d}}{\sqrt{1-\beta}}\sqrt{\sum_{t=1}^{T}\|\hat{\mathbf{g}}_t\|^2} \\ &\quad + \sqrt{L}D\sum_{t=1}^{T}\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t}}, \qquad (2) \\ \frac{R(T)}{T} \;&\leq\; O\left(\frac{d}{\sqrt{T}}\right) + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}. \qquad (3) \end{aligned}$$

For standard online gradient descent with the same learning rate scheme, $R(T)/T$ converges to zero at a $O(1/\sqrt{T})$ rate (Hazan, 2016). From Theorem 1, the average regret converges at the same rate, but only to a nonzero error $\sqrt{L}D\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}$ related to the weight quantization resolution Δ_w and dimension d.

3.3 Weight Quantization With Quantized Gradient

When both weights and gradients are quantized, the update for loss-aware weight quantization is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\tilde{\mathbf{v}}_t})^{-1}\tilde{\mathbf{g}}_t,$$

where $\tilde{\mathbf{g}}_t$ is the stochastically quantized gradient $Q_g(\nabla f_t(Q_w(\mathbf{w}_t)))$. The second moment $\tilde{\mathbf{v}}_t$ is the moving average of the (squared) quantized gradients $\tilde{\mathbf{g}}_t^2$. The following Proposition shows that gradient quantization significantly blows up the norm of the quantized gradient relative to its full-precision counterparts. Moreover, the difference increases with the gradient quantization resolution Δ_g and dimension d.

Proposition 1. $\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2) \leq (\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1)\|\hat{\mathbf{g}}_t\|^2$.

Theorem 2. *For loss-aware weight quantization with quantized gradients and* $\eta_t = \eta/\sqrt{t}$*,*

$$\begin{aligned}\mathbf{E}(R(T)) &\leq \frac{D_\infty^2\sqrt{dT}}{2\eta}\sqrt{\sum\nolimits_{t=1}^{T}(1-\beta)\beta^{T-t}\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2)} + \frac{\eta G_\infty\sqrt{d}}{\sqrt{1-\beta}}\sqrt{\sum\nolimits_{t=1}^{T}\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2)} \\ &\quad + \sqrt{L}D\sum\nolimits_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}_t'}}), \qquad (4)\\ \mathbf{E}\left(\frac{R(T)}{T}\right) &\leq O\left(\sqrt{\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1}\cdot\frac{d}{\sqrt{T}}\right) + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}. \qquad (5)\end{aligned}$$

The regrets in (2) and (4) are of the same form and differ only in the gradient used. Similarly, for the average regrets in (3) and (5), quantizing gradients slows convergence by a factor of $\sqrt{\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1}$, which is a direct consequence of the blowup in Proposition 1. These observations can be problematic as (i) deep networks typically have a large d; and (ii) distributed learning prefers using a small number of bits for the gradients, and thus a large Δ_g.

3.4 Weight Quantization with Quantized Clipped Gradients

To reduce convergence speed degradation caused by gradient quantization, gradient clipping has been proposed as an empirical solution (Wen et al., 2017). The gradient $\hat{\mathbf{g}}_t$ is clipped to Clip($\hat{\mathbf{g}}_t$), where Clip($\hat{g}_{t,i}$) $= \hat{g}_{t,i}$ if $|\hat{g}_{t,i}| \leq c\sigma$, and sign($\hat{g}_{t,i}$) $\cdot\, c\sigma$ otherwise. Here, c is a constant clipping factor, and σ is the standard deviation of elements in $\hat{\mathbf{g}}_t$. The update then becomes

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\text{Diag}(\sqrt{\check{\mathbf{v}}_t})^{-1}\check{\mathbf{g}}_t,$$

where $\check{\mathbf{g}}_t \equiv Q_g(\text{Clip}(\hat{\mathbf{g}}_t)) \equiv Q_g(\text{Clip}(\nabla f_t(Q_w(\mathbf{w}_t))))$ is the quantized clipped gradient. The second moment $\check{\mathbf{v}}_t$ is computed using the (squared) quantized clipped gradient $\check{\mathbf{g}}_t^2$.

As shown in Figure 2(a) of (Wen et al., 2017), the distribution of gradients before quantization is close to the normal distribution. Recall from Section 3.3 that the difference between $\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2)$ of the quantized gradient $\tilde{\mathbf{g}}_t$ and the full-precision gradient $\|\hat{\mathbf{g}}_t\|^2$ is related to the dimension d. The following Proposition shows that $\mathbf{E}(\|\check{\mathbf{g}}_t\|^2)/\mathbf{E}(\|\hat{\mathbf{g}}_t\|^2)$ becomes independent of d if $\hat{\mathbf{g}}_t$ follows the normal distribution and clipping is used.

Proposition 2. *Assume that* $\hat{\mathbf{g}}_t$ *follows* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$*, we have* $\mathbf{E}(\|\check{\mathbf{g}}_t\|^2) \leq ((2/\pi)^{\frac{1}{2}}c\Delta_g + 1)\mathbf{E}(\|\hat{\mathbf{g}}_t\|^2)$.

However, the quantized clipped gradient may now be biased (i.e., $\mathbf{E}(\check{\mathbf{g}}_t) = \text{Clip}(\hat{\mathbf{g}}_t) \neq \hat{\mathbf{g}}_t$). The following Proposition shows that the bias is related to the clipping factor c. A larger c (i.e., less severe gradient clipping) leads to smaller bias.

Proposition 3. *Assume that* $\hat{\mathbf{g}}_t$ *follows* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$*, we have* $\mathbf{E}(\|\mathit{Clip}(\hat{\mathbf{g}}_t) - \hat{\mathbf{g}}_t\|^2) \leq d\sigma^2(2/\pi)^{\frac{1}{2}}F(c)$*, where* $F(c) = -ce^{-\frac{c^2}{2}} + \sqrt{\frac{\pi}{2}}(1+c^2)(1 - \mathit{erf}(\frac{c}{\sqrt{2}}))$*, and* $\mathit{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,dt$ *is the error function.*

Theorem 3. *Assume that* $\hat{\mathbf{g}}_t$ *follows* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. *For loss-aware weight quantization with quantized clipped gradients and* $\eta_t = \eta/\sqrt{t}$*,*

$$\begin{aligned}\mathbf{E}(R(T)) &\leq \frac{D_\infty^2\sqrt{dT}}{2\eta}\sqrt{\sum\nolimits_{t=1}^{T}(1-\beta)\beta^{T-t}\mathbf{E}(\|\check{\mathbf{g}}_t\|^2)} + \frac{\eta G_\infty\sqrt{d}}{\sqrt{1-\beta}}\sqrt{\sum\nolimits_{t=1}^{T}\mathbf{E}(\|\check{\mathbf{g}}_t\|^2)} \\ &\quad + \sqrt{L}D\sum\nolimits_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}_t'}}) + D\sum\nolimits_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathit{Clip}(\hat{\mathbf{g}}_t) - \hat{\mathbf{g}}_t\|^2}), \qquad (6)\\ \mathbf{E}\left(\frac{R(T)}{T}\right) &\leq O\left(\sqrt{(2/\pi)^{\frac{1}{2}}c\Delta_g + 1}\frac{d}{\sqrt{T}}\right) + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}} + \sqrt{d}D\sigma(2/\pi)^{\frac{1}{4}}\sqrt{F(c)}. \qquad (7)\end{aligned}$$

Note that terms involving $\tilde{\mathbf{g}}_t$ in Theorem 2 are replaced by $\check{\mathbf{g}}_t$. Moreover, the regret has an additional term $D\sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\text{Clip}(\hat{\mathbf{g}}_t) - \hat{\mathbf{g}}_t\|^2})$ over that in Theorem 2. Comparing the average regrets in Theorems 1 and 3, gradient clipping before quantization slows convergence by a factor of $\sqrt{(2/\pi)^{\frac{1}{2}} c\Delta_g + 1}$, as compared to using full-precision gradients. This is independent of d as the increase in $\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2)$ is independent of d (Proposition 2). Hence, a Δ_g larger than the one in Theorem 2 can be used, and this reduces the communication cost in distributed learning.

A larger c (i.e., less severe gradient clipping) makes $F(c)$ smaller (Figure 2). Compared with (4) in Theorem 2, the extra error $\sqrt{d}D\sigma(2/\pi)^{\frac{1}{4}}\sqrt{F(c)}$ in (7) is thus smaller, but convergence is also slower. Hence, there is a trade-off between the two.

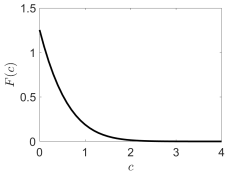


Figure 2: $F(c)$ vs the clipping factor c.

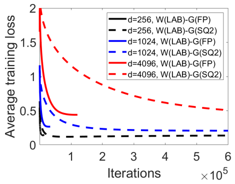
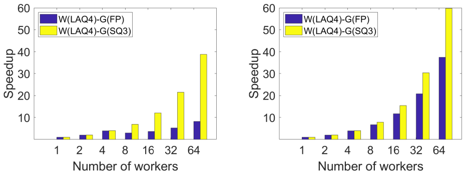


Figure 3: Convergence of LAB on the linear model with different d's.



(a) 1Gbps Ethernet. (b) 10Gbps Ethernet.

Figure 4: Speedup of ImageNet training on a 16-node GPU cluster. Each node has 4 1080ti GPUs with one PCI switch.

4 Experiments

4.1 Synthetic Data

In this section, we first study the effect of dimension d on the convergence speed and final error. As popular deep networks usually have hand-crafted architectures, we consider here a linear model with square loss as in (Zhang et al., 2017). Each entry of the model parameter is generated by uniform sampling from $[-0.5, 0.5]$. Samples $\mathbf{x}_i$'s are generated such that each entry of $\mathbf{x}_i$ is drawn uniformly from $[-0.5, 0.5]$, and the corresponding output y_i from $\mathcal{N}(\mathbf{x}_i^\top \mathbf{w}^*, (0.2)^2)$. At the tth iteration, a mini-batch of $B = 64$ samples are drawn to form $\mathbf{X}_t = [\mathbf{x}_1, \dots, \mathbf{x}_B]$ and $\mathbf{y}_t = [y_1, \dots, y_B]^\top$. The corresponding loss is $f_t(\mathbf{w}_t) = \|\mathbf{X}_t^\top \mathbf{w}_t - \mathbf{y}_t\|^2/2B$. The weights are quantized to 1 bit using LAB. The gradients are either full-precision (denoted FP) or stochastically quantized to 2 bits (denoted SQ2). The optimizer is RMSProp, and the learning rate is $\eta_t = \eta/\sqrt{t}$, where $\eta = 0.03$.

Figure 3 shows convergence of the average training loss $\sum_{t=1}^T f_t(\mathbf{w}_t)/T$, which differs from the average regret only by a constant.[3] As can be seen, for both full-precision and quantized gradients, a larger d leads to a larger loss upon convergence. Moreover, convergence is slower for larger d, particularly when the gradients are quantized. These agree with the results in Theorems 1 and 2.

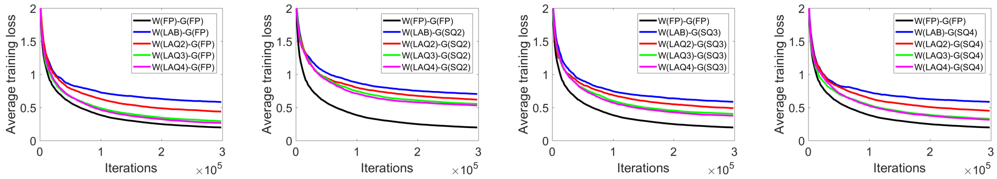
4.2 CIFAR-10

We follow (Wen et al., 2017) and use the same train/test split, Cifarnet model, data preprocessing, augmentation and distributed Tensorflow setup. Adam is used as the optimizer. The learning rate is decayed by a factor of 0.1 every 200 epochs.

4.2.1 Weight Quantization Resolution Δ_w

Weights are quantized to 1 bit (LAB), 2 bits (LAQ2), or m bits (LAQm). The gradients are full-precision (FP) or stochastically quantized to $m = \{2, 3, 4\}$ bits (SQm) without gradient clipping. Two workers are used in this experiment. Figure 5 shows convergence of the average training loss with different numbers of bits for the quantized weight. With full-precision or quantized gradients, weight-quantized networks have larger training losses than full-precision networks upon convergence. The more bits are used, the smaller is the final loss. This agrees with the results in Theorems 1 and 2. Table 1 shows the test set accuracies. Weight-quantized networks are less accurate than their full-precision counterparts, but the degradation is small when 3 or 4 bits are used.

[3] Legend "W(LAB)-G(FP)" means that weights are quantized using LAB and gradients are full-precision.



(a) G(FP). (b) G(SQ2). (c) G(SQ3). (d) G(SQ4).

Figure 5: Convergence with different numbers of bits for the weights on CIFAR-10. The gradient is full-precision (denoted G(FP)) or m-bit quantized (denoted G(SQm)) without gradient clipping.

Table 1: Testing accuracy (%) on CIFAR-10 with two workers.

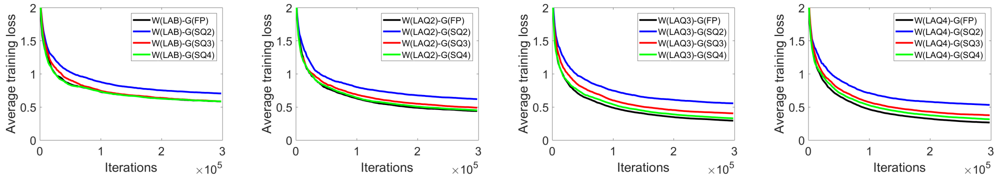
weight / gradient	FP	LAB	LAQ2	LAQ3	LAQ4
FP	83.74	80.37	82.11	83.14	83.35
SQ2 (no clipping)	81.40	78.67	80.27	81.27	81.38
SQ2 (clip,$c = 3$)	82.99	80.25	81.59	83.14	83.40
SQ3 (no clipping)	83.24	80.18	81.63	82.75	83.17
SQ3 (clip, $c = 3$)	83.89	80.13	81.77	82.97	83.43
SQ4 (no clipping)	83.64	80.44	81.88	83.13	83.47
SQ4 (clip, $c = 3$)	83.80	79.27	81.42	82.77	83.43

Table 2: Testing accuracy (%) on CIFAR-10 with varying number of workers (N).

weight	gradient	$N = 4$	$N = 8$	$N = 16$
FP	FP	83.28	83.38	83.76
LAQ3	FP	82.92	82.93	83.12
	SQ2 (no clipping)	81.53	81.08	81.30
	SQ2 (clip, $c = 3$)	83.01	82.94	82.73
	SQ3 (no clipping)	82.64	82.42	82.27
	SQ3 (clip, $c = 3$)	82.93	83.16	82.54
	SQ4 (no clipping)	83.03	82.53	82.61
	SQ4 (clip, $c = 3$)	82.51	83.07	82.52

4.2.2 Gradient Quantization Resolution Δ_g

Figure 6 shows convergence of the average training loss with different numbers of bits for the quantized gradients, again without gradient clipping. Using fewer bits yields a larger final error, and using 2- or 3-bit gradients yields larger training loss and worse accuracy than full-precision gradients (Figure 6 and Table 1). The fewer bits for the gradients, the larger the gap. The degradation is negligible when 4 bits are used. Indeed, 4-bit gradient sometimes has even better accuracy than full-precision gradient, as its inherent randomness encourages escape from poor sharp minima (Wen et al., 2017). Moreover, using a larger m results in faster convergence, which agrees with Theorem 2.

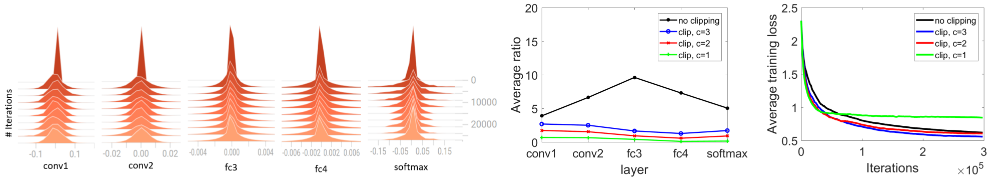


(a) W(LAB) (b) W(LAQ2). (c) W(LAQ3). (d) W(LAQ4).

Figure 6: Convergence with different numbers of bits for the gradients on CIFAR-10. The weight is binarized (denoted W(LAB)) or m-bit quantized (denoted W(LAQm)). Gradients are not clipped.

4.2.3 Gradient Clipping

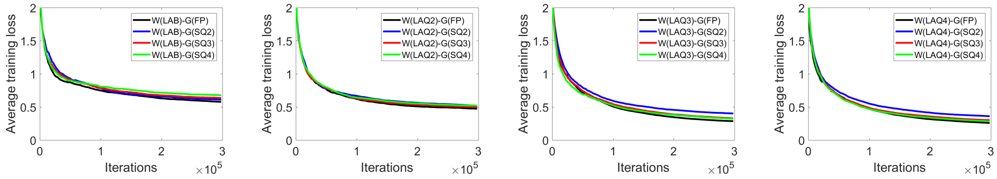
In this section, we perform experiments on gradient clipping, with clipping factor c in $\{1, 2, 3\}$. LAQ2 is used for weight quantization and SQ2 for gradient quantization. Figure 7(a) shows histograms of the full-precision gradients before clipping. As can be seen, the gradients at each layer before clipping roughly follow the normal distribution, which verifies the assumption in Section 3.4. Figure 7(b) shows the average $\|\tilde{\mathbf{g}}_t\|^2/\|\hat{\mathbf{g}}_t\|^2$ (for non-clipped gradients) and $\|\breve{\mathbf{g}}_t\|^2/\|\hat{\mathbf{g}}_t\|^2$ (for clipped gradients) over all iterations. The dimensionalities (d) of the various Cifarnet layers are "conv1": 1600, "conv2": 1600, "fc3": 884736, "fc4": 73728, "softmax": 1920. Layers with large d have large $\|\tilde{\mathbf{g}}_t\|^2/\|\hat{\mathbf{g}}_t\|^2$ values, which agrees with Proposition 1. With clipped gradients, $\|\breve{\mathbf{g}}_t\|^2/\|\hat{\mathbf{g}}_t\|^2$ is much smaller and does not depend on d, agreeing with Proposition 3. Figure 7(c) shows convergence of the average training loss. Using a smaller c (more aggressive clipping) leads to faster training (at the early stage of training) but larger final training loss, agreeing with Theorem 3.



(a) Histograms of gradients before clipping. (b) $\frac{\|\tilde{\mathbf{g}}_t\|^2}{\|\hat{\mathbf{g}}_t\|^2}$ or $\frac{\|\breve{\mathbf{g}}_t\|^2}{\|\hat{\mathbf{g}}_t\|^2}$. (c) Training curve.

Figure 7: Results for LAQ2 with SQ2 on CIFAR-10 with two workers. (a) Histograms of gradients at different Cifarnet layers before clipping (visualized by Tensorboard); (b) Average $\|\tilde{\mathbf{g}}_t\|^2/\|\hat{\mathbf{g}}_t\|^2$ (for non-clipped gradients) and $\|\breve{\mathbf{g}}_t\|^2/\|\hat{\mathbf{g}}_t\|^2$ (for clipped gradients); and (c) Training curves.

Figure 8 shows convergence of the average training loss with different numbers of bits for the quantized clipped gradient, with $c = 3$. By comparing[4] with Figure 6, gradient clipping achieves faster convergence, especially when the number of gradient bits is small. For example, 2-bit clipped gradient has comparable speed (Figure 8) and accuracy (Table 1) as full-precision gradient.



(a) W(LAB). (b) W(LAQ2). (c) W(LAQ3). (d) W(LAQ4).

Figure 8: Convergence with different numbers of bits for gradients (with $c = 3$) on CIFAR-10.

4.2.4 Varying the Number of Workers

We use 3-bit quantized weight (LAQ3), and gradients are full-precision or stochastically quantized to $m = \{2, 3, 4\}$ bits (SQm). The setup follows (Wen et al., 2017), and is in Appendix C. Table 2 shows the testing accuracies with varying number of workers N. Observations are similar to those in Section 4.2.3. 2-bit quantized clipped gradient has comparable performance as full-precision gradient, while the non-clipped counterpart requires 3 to 4 bits for comparable performance.

4.3 ImageNet

In this section, we train the AlexNet on ImageNet. We follow (Wen et al., 2017) and use the same data preprocessing, augmentation, learning rate, and mini-batch size. Quantization is not performed in the first and last layers, as is common in the literature (Zhou et al., 2016; Zhu et al., 2017; Polino et al., 2018; Wen et al., 2017). We use Adam as the optimizer. We experiment with 4-bit loss-aware weight quantization (LAQ4), and the gradients are either full-precision or quantized to 3 bits (SQ3).

Table 3 shows the accuracies with different numbers of workers. Weight-quantized networks have slightly worse accuracies than full-precision networks. Quantized clipped gradient outperforms the non-clipped counterpart, and achieves comparable accuracy as full-precision gradient.

Table 3: Top-1 and top-5 accuracies (%) on ImageNet.

weight	gradient	$N = 2$		$N = 4$		$N = 8$	
		top-1	top-5	top-1	top-5	top-1	top-5
FP	FP	55.08	78.33	55.45	78.57	55.40	78.69
LAQ4	FP	53.79	77.21	54.22	77.53	54.73	78.12
	SQ3 (no clipping)	52.48	75.97	52.87	76.40	53.18	76.62
	SQ3 (clip, $c = 3$)	54.13	77.27	54.23	77.55	54.34	78.07

Figure 4 shows the speedup in distributed training of a weight-quantized network with quantized/full-precision gradient compared to training with one worker using full-precision gra-

[4] Full-precision gradient curves are the same in Figures 6 and 8, and can be used as a common baseline.

dient. We use the performance model in (Wen et al., 2017), which combines lightweight profiling on a single node with analytical communication modeling. We also use the all-reduce communication model (Rabenseifner, 2004), in which each GPU communicates with its neighbor until all gradients are accumulated to a single GPU. We do not include the server's computation effort on weight quantization and the worker's effort on gradient clipping, which are negligible compared to the forward and backward propagations in the worker. As can be seen, when the bandwidth is small (Figure 4(a)), communication is the bottleneck, and using quantizing gradients is significantly faster than the use of full-precision gradients. With a larger bandwidth (Figure 4(b)), the difference in speedups is smaller. Moreover, note that on the 1Gbps Ethernet with quantized gradients, its speedup is similar to those on the 10Gbps Ethernet with full-precision gradients.

5 Conclusion

In this paper, we studied loss-aware weight-quantized networks with quantized gradient for efficient communication in a distributed environment. Convergence analysis is provided for weight-quantized networks with full-precision, quantized and quantized clipped gradients. Empirical experiments confirm the theoretical results, and demonstrate that quantized networks can speed up training and have comparable performance as full-precision networks.

References

A. F. Aji and K. Heafield. Sparse communication for distributed gradient descent. In *International Conference on Empirical Methods in Natural Language Processing*, pp. 440–445, 2017.

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pp. 1707–1718, 2017.

J. Bernstein, Y. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 2018.

M. Courbariaux, Y. Bengio, and J. P. David. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, pp. 3105–3113, 2015.

Y. Dauphin, H. de Vries, and Y. Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1504–1512, 2015.

C. De Sa, M. Leszczynski, J. Zhang, A. Marzoev, C. R. Aberger, K. Olukotun, and C. Ré. High-accuracy low-precision training. Preprint arXiv:1803.03383, 2018.

J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pp. 1223–1231, 2012.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Y. Guo, A. Yao, H. Zhao, and Y. Chen. Network sketching: Exploiting binary structure in deep CNNs. In *International Conference on Computer Vision and Pattern Recognition*, pp. 5955–5963, 2017.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

L. Hou and J. T. Kwok. Loss-aware weight quantization of deep networks. In *International Conference on Learning Representations*, 2018.

L. Hou, Q. Yao, and J. T. Kwok. Loss-aware binarization of deep networks. In *International Conference on Learning Representations*, 2017.

I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18:187–1.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

C. Leng, H. Li, S. Zhu, and R. Jin. Extremely low bit neural network: Squeeze the last bit out with admm. In *AAAI Conference on Artificial Intelligence*, 2018.

F. Li and B. Liu. Ternary weight networks. Preprint arXiv:1605.04711, 2016.

H. Li, S. De, Z. Xu, C. Studer, H. Samet, and Goldstein T. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems*, 2017.

M. Li, D. G. Andersen, J. W. Park, A. J. Smola, and A. Ahmed. Scaling distributed machine learning with the parameter server. In *USENIX Symposium on Operating Systems Design and Implementation*, volume 583, pp. 598, 2014a.

M. Li, D. G. Andersen, A. J. Smola, and K. Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pp. 19–27, 2014b.

X. Lin, C. Zhao, and W. Pan. Towards accurate binary convolutional neural network. In *Advances in Neural Information Processing Systems*, pp. 344–352, 2017.

Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio. Neural networks with few multiplications. In *International Conference on Learning Representations*, 2016.

A. Polino, R. Pascanu, and D. Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018.

R. Rabenseifner. Optimization of collective reduction operations. In *International Conference on Computational Science*, pp. 1–9, 2004.

M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, 2016.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and application to data-parallel distributed training of speech DNNs. In *Interspeech*, 2014.

J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. Preprint arXiv:1710.09854, 2017.

W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, 2017.

S. Wu, G. Li, F. Chen, and L. Shi. Training and inference with integers in deep neural networks. In *International Conference on Learning Representations*, 2018.

H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang. The ZipML framework for training models with end-to-end low precision: The cans, the cannots, and a little bit of deep learning. In *International Conference on Machine Learning*, 2017.

S. Zhou, Z. Ni, X. Zhou, H. Wen, Y. Wu, and Y. Zou. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. Preprint arXiv:1606.06160, 2016.

C. Zhu, S. Han, H. Mao, and W. J. Dally. Trained ternary quantization. In *International Conference on Learning Representations*, 2017.

A Lemmas

Lemma 1. $\mathbf{E}(\tilde{\mathbf{g}}_t) = \hat{\mathbf{g}}_t$*, and* $\mathbf{E}(\|\tilde{\mathbf{g}}_t - \hat{\mathbf{g}}_t\|^2) \leq \Delta_g \|\hat{\mathbf{g}}_t\|_\infty \|\hat{\mathbf{g}}_t\|_1$.

A.1 Proof for Lemma 1

Proof. For the ith element $\tilde{g}_{t,i}$ of the quantized gradient $\tilde{\mathbf{g}}_t$, denote its two adjacent quantized values as $B_{r,i}$ and $B_{r+1,i}$ with $B_{r,i} \leq |g_{t,i}|/s_t < B_{r+1,i}$, its expectation satisfies:

$$\begin{aligned}
\mathbf{E}(\tilde{g}_{t,i}) &= \mathbf{E}(s_t \cdot \text{sign}(\hat{g}_{t,i}) \cdot q_{t,i}) = s_t \cdot \text{sign}(\hat{g}_{t,i}) \cdot \mathbf{E}(q_{t,i}) \\
&= s_t \cdot \text{sign}(\hat{g}_{t,i}) \cdot (pB_{r+1,i} + (1-p)B_{r,i}) \\
&= s_t \cdot \text{sign}(\hat{g}_{t,i}) \cdot (p(B_{r+1,i} - B_{r,i}) + B_{r,i}) \\
&= s_t \cdot \text{sign}(\hat{g}_{t,i}) \cdot \frac{|\hat{g}_{t,i}|}{s_t} = \hat{g}_{t,i}.
\end{aligned}$$

Thus $\mathbf{E}(\tilde{\mathbf{g}}_t) = \hat{\mathbf{g}}_t$, and the variance of the quantized gradients satisfy

$$\begin{aligned}
\mathbf{E}(\|\tilde{\mathbf{g}}_t - \hat{\mathbf{g}}_t\|^2) &= \sum_{i=1}^{d} (s_t B_{r+1,i} - |\hat{g}_{t,i}|)(|\hat{g}_{t,i}| - s_t B_{r,i}) \\
&= s_t^2 \sum_{i=1}^{d} (B_{r+1,i} - \frac{|\hat{g}_{t,i}|}{s_t})(\frac{|\hat{g}_{t,i}|}{s_t} - B_{r,i}) \\
&\leq s_t^2 \sum_{i=1}^{d} (B_{r+1,i} - B_{r,i}) \frac{|\hat{g}_{t,i}|}{s_t} \\
&= \Delta_g \|\hat{\mathbf{g}}_t\|_\infty \|\hat{\mathbf{g}}_t\|_1.
\end{aligned}$$

□

Lemma 2. *For* $\mathbf{v}_t = \sum_{j=1}^{t} (1-\beta)\beta^{t-j} \mathbf{g}_j^2$ *with* $\forall j \in \{1, 2 \cdots, t\}$, $\|\mathbf{g}_j\|_\infty < G_\infty$*, then its ith element* $\mathbf{v}_{t,i}$ *satisfies*

$$v_{t,i} \geq (1-\beta) g_{t,i}^2, \quad \sqrt{v_{t,i}} \leq G_\infty,$$

and

$$\|\sqrt{\mathbf{v}_t}\|_2 = \sqrt{\sum_{j=1}^{t} (1-\beta)\beta^{t-j} \|\mathbf{g}_j\|^2}.$$

Proof.

$$v_{t,i} = \sum_{j=1}^{t} (1-\beta)\beta^{t-j} g_{j,i}^2 \geq (1-\beta) g_{t,i}^2.$$

$$\sqrt{v_{t,i}} = \sqrt{\sum_{j=1}^{t} (1-\beta)\beta^{t-j} g_{j,i}^2} \leq \sqrt{\sum_{j=1}^{t} (1-\beta)\beta^{t-j} G_\infty^2} \leq G_\infty.$$

$$\|\sqrt{\mathbf{v}_t}\|_2 = \sqrt{\sum_{i=1}^{d} v_{t,i}} = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{t} (1-\beta)\beta^{t-j} g_{j,i}^2} = \sqrt{\sum_{j=1}^{t} (1-\beta)\beta^{t-j} \|\mathbf{g}_j\|^2}.$$

□

Lemma 3. *[Lemma 10.3 in (Kingma & Ba, 2015)] Let* $\mathbf{g}_{1:T,i} = [g_{1,i}, g_{2,i}, \ldots, g_{T,i}]^\top \in \mathbb{R}^T$ *be the vector containing the ith element of the gradients for all iterations up to T, and* $\mathbf{g}_t$ *be bounded as in Assumption* **A3**,

$$\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} \leq 2G_\infty \|\mathbf{g}_{1:T,i}\|_2.$$

B Proofs

B.1 Proof for Theorem 1

Proof. When only weights are quantized, the update for loss-aware weight quantization is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\hat{\mathbf{v}}_t})^{-1}\hat{\mathbf{g}}_t, \tag{8}$$

Consider the ith parameter of $\mathbf{w}_t$. the update for it is $w_{t+1,i} = w_{t,i} - \eta_t \frac{\hat{g}_{t,i}}{\sqrt{\hat{v}_{t,i}}}$, which implies

$$\begin{aligned}(w_{t+1,i} - w_i^*)^2 &= (w_{t,i} - w_i^*)^2 - 2\eta_t(w_{t,i} - w_i^*)\frac{g_{t,i}}{\sqrt{\hat{v}_{t,i}}} \\ &\quad +2\eta_t(w_{t,i} - w_i^*)\frac{(g_{t,i} - \hat{g}_{t,i})}{\sqrt{\hat{v}_{t,i}}} + \eta_t^2(\frac{\hat{g}_{t,i}}{\sqrt{\hat{v}_{t,i}}})^2.\end{aligned}$$

After rearranging,

$$\begin{aligned}g_{t,i}(w_{t,i} - w_i^*) &= \frac{\sqrt{\hat{v}_{t,i}}}{2\eta_t}((w_{t,i} - w_i^*)^2 - (w_{t+1,i} - w_i^*)^2) \\ &\quad +(w_{t,i} - w_i^*)(g_{t,i} - \hat{g}_{t,i}) + \frac{\eta_t\sqrt{\hat{v}_{t,i}}}{2}(\frac{\hat{g}_{t,i}}{\sqrt{\hat{v}_{t,i}}})^2.\end{aligned} \tag{9}$$

Since f_t is convex, we have

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) \leq \mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}^*) = \sum_{i=1}^d g_{t,i}(w_{t,i} - w_i^*). \tag{10}$$

As f_t is convex (Assumption **A1**) and twice differentiable (Assumption **A2**), $\|\nabla f_t(\mathbf{u}) - \nabla f_t(\mathbf{v})\|_2 \leq L\|\mathbf{u} - \mathbf{v}\|_2$ for any $\mathbf{u}, \mathbf{v}$ is equivalent to $\nabla^2 f_t \preceq L\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Thus from the domain bound assumption $\|\mathbf{w}_m - \mathbf{w}_n\|_2 \leq D$, we have

$$\|\mathbf{w}_t - \mathbf{w}^*\|^2_{\mathbf{H}'_t} \leq L\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq LD^2, \quad \|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t} \leq L\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2. \tag{11}$$

Combining (9) and (10) sum over all the dimensions for $i \in \{1, 2 \cdots, d\}$ and over all the iterations for $t \in \{1, 2, \cdots, T\}$, we have

$$\begin{aligned}R(T) &= \sum\nolimits_{t=1}^T f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) \\ &\leq \sum_{i=1}^d \frac{\sqrt{\hat{v}_{1,i}}}{2\eta_1}(w_{1,i} - w_i^*)^2 + \sum_{i=1}^d\sum_{t=2}^T(\frac{\sqrt{\hat{v}_{t,i}}}{2\eta_t} - \frac{\sqrt{\hat{v}_{t-1,i}}}{2\eta_{t-1}})(w_{t,i} - w_i^*)^2 \\ &\quad +\sum_{t=1}^T\langle\mathbf{w}_t - \mathbf{w}^*, \mathbf{g}_t - \hat{\mathbf{g}}_t\rangle + \sum_{i=1}^d\sum_{t=1}^T\frac{\eta_t}{2}(\frac{\hat{g}_{t,i}^2}{\sqrt{(1-\beta)\hat{g}_{t,i}^2}}) \\ &\leq \frac{D_\infty^2}{2\eta}\sum_{i=1}^d\sqrt{T\hat{v}_{T,i}} + \sum_{t=1}^T\langle\mathbf{H}_t^{'\frac{1}{2}}(\mathbf{w}_t - \mathbf{w}^*), \mathbf{H}_t^{'-\frac{1}{2}}(\mathbf{g}_t - \hat{\mathbf{g}}_t)\rangle + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^d\|\hat{\mathbf{g}}_{1:T,i}\|_2 \\ &\leq \frac{D_\infty^2}{2\eta}\sum_{i=1}^d\sqrt{T\hat{v}_{T,i}} + \sum_{t=1}^T\sqrt{\|\mathbf{w}_t - \mathbf{w}^*\|^2_{\mathbf{H}'_t}}\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t}} + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^d\|\hat{\mathbf{g}}_{1:T,i}\|_2 \\ &\leq \frac{D_\infty^2}{2\eta}\sum_{i=1}^d\sqrt{T\hat{v}_{T,i}} + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^d\|\hat{\mathbf{g}}_{1:T,i}\|_2 + \sqrt{L}D\sum_{t=1}^T\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t}} \\ &\leq \frac{D_\infty^2}{2\eta}\sqrt{dT}\sqrt{\sum_{t=1}^T(1-\beta)\beta^{T-t}\|\hat{\mathbf{g}}_t\|^2} + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sqrt{d}\sqrt{\sum_{t=1}^T\|\hat{\mathbf{g}}_t\|^2} \\ &\quad +\sqrt{L}D\sum_{t=1}^T\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t}}.\end{aligned} \tag{12}$$

The first inequality comes from Lemma 2. The second inequality comes from the definition of $\mathbf{H}'_t$. The third inequality comes from the definition of $\mathbf{H}'_t$ and Cauchy's inequality. The fourth inequality comes from (11). The last inequality comes from Lemma 2.

For m-bit ($m > 1$) loss-aware weight quantization in (8), as $\hat{\mathbf{w}}_t = \arg\min_{\mathbf{w}} \|\mathbf{w}-\mathbf{w}_t\|^2_{\text{Diag}(\sqrt{\hat{\mathbf{v}}_{t-1}})} = \sum_{i=1}^{d} \sqrt{\hat{v}_{t-1,i}}(w - w_{t,i})^2$. If $w_{t,i}$ is within the representable range, i.e., $\alpha_t M_{k-1} \leq w_{t,i} \leq \alpha_t M_k$, as $\sqrt{\hat{v}_{t-1,i}} > 0$, the optimal $\hat{w}_{t,i}$ satisfies $\hat{w}_{t,i} \in \{\alpha_t M_{r,i}, \alpha_t M_{r+1,i}\}$, where r is the index that satisfies $\alpha_t M_{r,i} \leq w_{t,i} \leq \alpha_t M_{r+1,i}$. Since $\alpha = \max\{\alpha_1, \cdots, \alpha_T\}$, we have

$$\begin{aligned}(\hat{w}_{t,i} - w_{t,i})^2 &= \min\{(\alpha_t M_{r+1,i} - |w_{t,i}|)^2, (\alpha_t M_{r,i} - |w_{t,i}|)^2\} \\ &\leq (\frac{\alpha_t M_{r+1,i} - \alpha_t M_{r,i}}{2})^2 \leq \frac{\alpha^2 \Delta_w^2}{4}. \end{aligned} \tag{13}$$

In the other case, if $w_{t,i}$ is exterior of the representable range, the optimal $\hat{w}_{t,i}$ is just the nearest representable value of $w_{t,i}$, thus

$$(\hat{w}_{t,i} - w_{t,i})^2 = (|w_{t,i}| - \alpha_t M_k)^2 \leq w_{t,i}^2. \tag{14}$$

From (13) and (14), and sum over all the dimensions we have

$$\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2 \leq \frac{d\alpha^2\Delta_w^2}{4} + \|\mathbf{w}_t\|^2 \leq \frac{d\alpha^2\Delta_w^2}{4} + D^2. \tag{15}$$

From (11) and (15),

$$\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|^2_{\mathbf{H}'_t} \leq L(D^2 + \frac{d\alpha^2\Delta_w^2}{4}). \tag{16}$$

From (16) and Lemma 2, we have from (12)

$$R(T) \leq \frac{D_\infty^2 d G_\infty \sqrt{T}}{2\eta} + \frac{\eta d G_\infty^2 \sqrt{T}}{\sqrt{1-\beta}} + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}T.$$

Thus the average regret

$$R(T)/T \leq \left(\frac{D_\infty^2 G_\infty}{2\eta} + \frac{\eta G_\infty^2}{\sqrt{1-\beta}}\right)\frac{d}{\sqrt{T}} + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}. \tag{17}$$

□

B.2 Proof for Proposition 1

Proof. From Lemma 1,

$$\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2) = \mathbf{E}(\|\tilde{\mathbf{g}}_t - \hat{\mathbf{g}}_t\|^2) + \|\hat{\mathbf{g}}_t\|^2 \leq \Delta_g \|\hat{\mathbf{g}}_t\|_\infty \|\hat{\mathbf{g}}_t\|_1 + \|\hat{\mathbf{g}}_t\|^2.$$

Denote $\{x_1, x_2, \ldots, x_d\}$ as the absolute values of the elements in $\hat{\mathbf{g}}_t$ sorted in ascending order. From Cauchy inequality, we have

$$\begin{aligned}\|\hat{\mathbf{g}}_t\|_\infty \|\hat{\mathbf{g}}\|_1 &= x_d \sum_{i=1}^{d} x_i = \sum_{i=1}^{d} x_d x_i \leq \sum_{i=1}^{d-1}(\frac{1+\sqrt{2d-1}}{2}x_i^2 + \frac{1}{1+\sqrt{2d-1}}x_d^2) + x_d^2 \\ &= \frac{1+\sqrt{2d-1}}{2}\sum_{i=1}^{d} x_i^2 = \frac{1+\sqrt{2d-1}}{2}\|\hat{\mathbf{g}}_t\|^2. \end{aligned}$$

The equality holds iff $x_1 = x_2 = \cdots = \frac{\sqrt{2}}{1+\sqrt{2d-1}}x_d$. Thus we have

$$\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2) \leq (\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1)\|\sqrt{\hat{\mathbf{g}}_t}\|^2.$$

□

B.3 Proof for Theorem 2

Proof. When both weights and gradients are quantized, the update is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\tilde{\mathbf{v}}_t})^{-1}\tilde{\mathbf{g}}_t. \tag{18}$$

Similar to the proof for Theorem 1, and using that $\mathbf{E}(\tilde{\mathbf{g}}_t) = \hat{\mathbf{g}}_t$, we can get

$$\begin{aligned}
\mathbf{E}(R(T)) \leq\; & \sum_{i=1}^{d}\mathbf{E}(\frac{\sqrt{\tilde{v}_{1,i}}}{2\eta_1}(w_{1,i}-w_i^*)^2) + \sum_{i=1}^{d}\sum_{t=2}^{T}\mathbf{E}\left((\frac{\sqrt{\tilde{v}_{t,i}}}{2\eta_t} - \frac{\sqrt{\tilde{v}_{t-1,i}}}{2\eta_{t-1}})(w_{t,i}-w_i^*)^2\right) \\
& + \sum_{t=1}^{T}\mathbf{E}(\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{g}_t - \tilde{\mathbf{g}}_t\rangle) + \sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\eta_t}{2}\mathbf{E}(\left(\frac{\tilde{g}_{t,i}^2}{\sqrt{(1-\beta)\tilde{g}_{t,i}^2}})\right) \\
\leq\; & \frac{D_\infty^2}{2\eta}\sum_{i=1}^{d}\mathbf{E}(\sqrt{T\tilde{v}_{T,i}}) + \sum_{t=1}^{T}\mathbf{E}\langle \mathbf{H}_t^{'\frac{1}{2}}(\mathbf{w}_t - \mathbf{w}^*), \mathbf{H}_t^{'-\frac{1}{2}}(\mathbf{g}_t - \hat{\mathbf{g}}_t)\rangle \\
& + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^{d}\mathbf{E}(\|\tilde{\mathbf{g}}_{1:T,i}\|_2) \\
\leq\; & \frac{D_\infty^2}{2\eta}\sum_{i=1}^{d}\mathbf{E}(\sqrt{T\tilde{v}_{T,i}}) + \sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \mathbf{w}^*\|_{\mathbf{H}_t'}^2}\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2}) \\
& + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^{d}\mathbf{E}(\|\tilde{\mathbf{g}}_{1:T,i}\|_2) \\
\leq\; & \frac{D_\infty^2}{2\eta}\sqrt{dT}\mathbf{E}(\|\sqrt{\tilde{\mathbf{v}}_T}\|_2) + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^{d}\mathbf{E}(\|\tilde{\mathbf{g}}_{1:T,i}\|_2) \\
& + \sqrt{L}D\sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2}) \\
\leq\; & \frac{D_\infty^2}{2\eta}\sqrt{dT}\sqrt{\sum_{t=1}^{T}(1-\beta)\beta^{T-t}\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2)} + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sqrt{d}\sqrt{\sum_{t=1}^{T}\mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2)} \\
& + \sqrt{L}D\sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2}).
\end{aligned}$$

As (16) in the proof for Theorem 1 still hold, using Proposition 1 and Assumption **A3**, we have

$$\begin{aligned}
\mathbf{E}(R(T)) \leq\; & \left(\frac{D_\infty^2}{2\eta}\sqrt{dT}\sqrt{\sum_{t=1}^{T}(1-\beta)\beta^{T-t}\|\hat{\mathbf{g}}_t\|^2} + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sqrt{d}\sqrt{\sum_{t=1}^{T}\|\hat{\mathbf{g}}_t\|^2}\right) \\
& \times\sqrt{\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1} + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}} \\
\leq\; & \left(\frac{D_\infty^2 G_\infty}{2\eta} + \frac{\eta G_\infty^2}{\sqrt{1-\beta}}\right)d\sqrt{T}\sqrt{\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1} + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}.
\end{aligned}$$

$$\mathbf{E}(R(T)/T) \leq \left(\frac{D_\infty^2 G_\infty}{2\eta} + \frac{\eta G_\infty^2}{\sqrt{1-\beta}}\right)\sqrt{\frac{1+\sqrt{2d-1}}{2}\Delta_g + 1}\frac{d}{\sqrt{T}} + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}}.$$

□

B.4 Proof for Proposition 2

Proof. From Lemma 1,

$$\mathbf{E}(Q_g(\text{Clip}(\hat{\mathbf{g}}_t))) = \text{Clip}(\hat{\mathbf{g}}_t),$$

and

$$\mathbf{E}(\|Q_g(\text{Clip}(\hat{\mathbf{g}}_t))\|^2) \leq \mathbf{E}(\Delta_g \|\text{Clip}(\hat{\mathbf{g}}_t)\|_\infty \|\text{Clip}(\hat{\mathbf{g}}_t)\|_1 + \|\text{Clip}(\hat{\mathbf{g}}_t)\|^2).$$

As $[\hat{\mathbf{g}}_t]_i \sim \mathcal{N}(0, \sigma^2)$, its pdf is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, thus

$$\begin{aligned}\mathbf{E}(\|\text{Clip}(\hat{\mathbf{g}}_t)\|_1) \leq \mathbf{E}(\|\hat{\mathbf{g}}_t\|_1) &= \sum_{i=1}^d \mathbf{E}(|[\hat{\mathbf{g}}_t]_i|) = d\int_{-\infty}^{+\infty} |x| f(x)\mathbf{d}x = 2d\int_0^{+\infty} x f(x)\mathbf{d}x \\ &= d\int_0^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \mathbf{d}x^2 = d\frac{-2\sigma^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{u}{2\sigma^2}}\Big|_0^{+\infty} = (2/\pi)^{\frac{1}{2}} d\sigma.\end{aligned}$$

As $\mathbf{E}(\|\text{Clip}(\hat{\mathbf{g}}_t)\|^2) \leq \mathbf{E}(\|\hat{\mathbf{g}}_t\|^2)$, and that $\mathbf{E}(\|\hat{\mathbf{g}}_t\|^2) = d\sigma^2$, we have

$$\begin{aligned}\mathbf{E}(\|Q_g(\text{Clip}(\hat{\mathbf{g}}_t))\|^2) &\leq \mathbf{E}(\Delta_g \|\text{Clip}(\hat{\mathbf{g}}_t)\|_\infty \|\text{Clip}(\hat{\mathbf{g}}_t)\|_1 + \|\text{Clip}(\hat{\mathbf{g}}_t)\|^2) \\ &\leq \Delta_g c\sigma (2/\pi)^{\frac{1}{2}} d\sigma + \mathbf{E}(\|\text{Clip}(\hat{\mathbf{g}}_t)\|^2) \\ &\leq ((2/\pi)^{\frac{1}{2}} c\Delta_g + 1)\mathbf{E}(\|\hat{\mathbf{g}}_t\|^2).\end{aligned}$$

□

B.5 Proof for Proposition 3

Proof.

$$\begin{aligned}\mathbf{E}(\|\text{Clip}(\hat{\mathbf{g}}_t) - \hat{\mathbf{g}}_t\|^2) &= d(\int_{c\sigma}^{+\infty} (x - c\sigma)^2 f(x)\mathbf{d}x + \int_{-\infty}^{-c\sigma} (-x - c\sigma)^2 f(x)\mathbf{d}x) \\ &\leq 2d\int_{c\sigma}^{+\infty} (x - c\sigma)^2 f(x)\mathbf{d}x \\ &= \frac{2d}{\sqrt{2\pi\sigma^2}} \sigma^3 (-ce^{-\frac{c^2}{2}} + \sqrt{\frac{\pi}{2}}(1 + c^2)(1 - \text{erf}(\frac{c}{\sqrt{2}}))) \\ &= (2/\pi)^{\frac{1}{2}} d\sigma^2 F(c).\end{aligned}$$

□

B.6 Proof for Theorem 3

Proof. When both weights and gradients are quantized, and gradient clipping is applied before gradient quantization, the update then becomes

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Diag}(\sqrt{\check{\mathbf{v}}_t})^{-1} \check{\mathbf{g}}_t. \tag{19}$$

Similar to the proof for Theorem 1, and using that $\mathbf{E}(Q_g(\text{Clip}(\hat{\mathbf{g}}_t))) = \text{Clip}(\hat{\mathbf{g}}_t)$, we can get

$$\begin{aligned}\mathbf{E}(R(T)) \leq\ & \sum_{i=1}^d \mathbf{E}(\frac{\sqrt{\check{v}_{1,i}}}{2\eta_1}(w_{1,i} - w_i^*)^2) + \sum_{i=1}^d \sum_{t=2}^T \mathbf{E}\left((\frac{\sqrt{\check{v}_{t,i}}}{2\eta_t} - \frac{\sqrt{\check{v}_{t-1,i}}}{2\eta_{t-1}})(w_{t,i} - w_i^*)^2\right) \\ & + \sum_{t=1}^T \mathbf{E}(\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{g}_t - \check{\mathbf{g}}_t \rangle) + \sum_{i=1}^d \sum_{t=1}^T \frac{\eta_t}{2} \mathbf{E}\left(\frac{\check{g}_{t,i}^2}{\sqrt{(1-\beta)\check{g}_{t,i}^2}}\right) \\ \leq\ & \frac{D_\infty^2}{2\eta} \sum_{i=1}^d \mathbf{E}(\sqrt{T\check{v}_{T,i}}) + \sum_{t=1}^T \mathbf{E}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{g}_t - \hat{\mathbf{g}}_t \rangle + \sum_{t=1}^T \mathbf{E}\langle \mathbf{w}_t - \mathbf{w}^*, \hat{\mathbf{g}}_t - \text{Clip}(\hat{\mathbf{g}}_t) \rangle \\ & + \frac{\eta G_\infty}{\sqrt{1-\beta}} \sum_{i=1}^d \mathbf{E}(\|\check{\mathbf{g}}_{1:T,i}\|_2)\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{D_\infty^2}{2\eta}\sum_{i=1}^{d}\mathbf{E}(\sqrt{T\check{v}_{T,i}}) + \sum_{t=1}^{T}\mathbf{E}\langle \mathbf{H}_t^{'\frac{1}{2}}(\mathbf{w}_t - \mathbf{w}^*), \mathbf{H}_t^{'-\frac{1}{2}}(\mathbf{g}_t - \hat{\mathbf{g}}_t)\rangle \\
&\quad + \sum_{t=1}^{T}\mathbf{E}\langle \mathbf{w}_t - \mathbf{w}^*, \hat{\mathbf{g}}_t - \text{Clip}(\hat{\mathbf{g}}_t)\rangle + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^{d}\mathbf{E}(\|\check{\mathbf{g}}_{1:T,i}\|_2) \\
&\leq \frac{D_\infty^2\sqrt{dT}}{2\eta}\mathbf{E}(\|\sqrt{\check{\mathbf{v}}_T}\|_2) + \frac{\eta G_\infty}{\sqrt{1-\beta}}\sum_{i=1}^{d}\mathbf{E}(\|\check{\mathbf{g}}_{1:T,i}\|_2) \\
&\quad + \sqrt{L}D\sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2}) + \sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \mathbf{w}^*\|^2}\sqrt{\|(\hat{\mathbf{g}}_t - \text{Clip}(\hat{\mathbf{g}}_t))\|^2}) \\
&\leq \frac{D_\infty^2\sqrt{dT}}{2\eta}\sqrt{\sum_{t=1}^{T}(1-\beta)\beta^{T-t}\mathbf{E}(\|\check{\mathbf{g}}_t\|^2)} + \frac{\eta G_\infty\sqrt{d}}{\sqrt{1-\beta}}\sqrt{\sum_{t=1}^{T}\mathbf{E}(\|\check{\mathbf{g}}_t\|^2)} \\
&\quad + \sqrt{L}D\sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2}) + D\sum_{t=1}^{T}\mathbf{E}(\sqrt{\|\hat{\mathbf{g}}_t - \text{Clip}(\hat{\mathbf{g}}_t)\|^2}).
\end{aligned}
$$

(16) in the proof for Theorem 1 still hold. Similar to the proof for Theorem 2, using the domain bound assumption ($\|\mathbf{w}_m - \mathbf{w}_n\|_2 \leq D$ and $\|\mathbf{w}_m - \mathbf{w}_n\|_\infty \leq D_\infty$ for any $\mathbf{w}_m, \mathbf{w}_n \in \mathcal{S}$), Proposition 3 and that $\sum_{t=1}^{T}\frac{1}{t} \leq 2\sqrt{T}$, we have

$$
\begin{aligned}
\mathbf{E}(R(T)/T) \quad &\leq \quad \left(\frac{D_\infty^2 G_\infty}{2\eta} + \frac{\eta G_\infty^2}{\sqrt{1-\beta}}\right)\sqrt{(2/\pi)^{\frac{1}{2}}c\Delta_g + 1}\frac{d}{\sqrt{T}} \\
&\quad + LD\sqrt{D^2 + \frac{d\alpha^2\Delta_w^2}{4}} + \sqrt{d}D\sigma(2/\pi)^{\frac{1}{4}}\sqrt{F(c)}.
\end{aligned}
$$

□

C Experimental Setup

The detailed setup for the CIFAR-10 and ImageNet data sets are shown in Table 4.

Table 4: Experimental setup. Here, B is the total batch size summed over N workers, and T is the number of iterations.

dataset	optimizer	network	weight decay	N	initial learning rate	B	T
CIFAR-10	Adam	Cifarnet	-	2	0.0002	128	300k
				4	0.0004	256	150k
				8	0.0004	512	75k
				16	0.0008	1024	37.5k
ImageNet	Adam	AlexNet	0.0005	2	0.0001	256	370k
				4	0.0002	512	185k
				8	0.0004	1024	92.5k

D Comparison with Stochastic Weight Quantization

D.1 Linear Models

For stochastic weight quantization, the gradient is only unbiased for linear models i.e., $f_t(\mathbf{w}_t) = (\mathbf{x}_t^\top \mathbf{w}_t - y_t)^2)$. Specifically, denote the unbiased stochastic weight quantization function as Q_w (i.e. $\mathbf{E}(Q_w(\mathbf{w})) = \mathbf{w}$ for any $\mathbf{w} \in \mathbb{R}^d$), then the expectation of the gradient w.r.t. the quantized weights is also unbiased

$$
\mathbf{E}(\hat{\mathbf{g}}_t) = \mathbf{E}(2\mathbf{x}_t(\mathbf{x}_t^\top Q_w(\mathbf{w}_t) - y_t)) = 2\mathbf{x}_t(\mathbf{x}_t^\top \mathbf{E}(Q_w(\mathbf{w}_t)) - y_t) = \mathbf{g}_t,
$$

where $\mathbf{g}_t = 2\mathbf{x}_t(\mathbf{x}_t^\top \mathbf{w}_t - y_t)$ is the full-precision gradient w.r.t. the full-precision weight. Moreover, if we further use an unbiased stochastic gradient quantization function (i.e., $\mathbf{E}(Q_g(\mathbf{g})) = \mathbf{g}$ for any $\mathbf{g} \in \mathbb{R}^d$), then the quantized gradient w.r.t. the quantized weight is also unbiased

$$\mathbf{E}(\tilde{\mathbf{g}}_t) = \mathbf{E}(Q_g(2\mathbf{x}_t(\mathbf{x}_t^\top Q_w(\mathbf{w}_t) - y_t))) = \mathbf{E}(2\mathbf{x}_t(\mathbf{x}_t^\top Q_w(\mathbf{w}_t) - y_t)) = \mathbf{g}_t.$$

In this case, convergence analysis of gradient quantization Wen et al. (2017); Alistarh et al. (2017) and ZipML Zhang et al. (2017) that based on unbiased approximation of SGD gradients, can be easily extended here. Moreover, the error term will also disappear.

D.2 Non-linear models

For non-linear models, however, using an unbiased stochastic weight quantization does not necessarily guarantee the gradient w.r.t. the quantized weight or the quantized gradient w.r.t. the quantized weight to be unbiased. Take the linear stochastic weight quantization as an example, where the quantization function also satisfies $\mathbf{E}(Q_w(\mathbf{w})) = \mathbf{w}$, the convergence and error when using full-precision, quantized, quantized clipped gradients are shown in Theorems 4, 5 and 6, respectively.

Denote Q_w as the stochastic weight quantization function. Q_w is defined in a similar way as the gradient quantization function in Section 2.3: $Q_w(\mathbf{w}_t) = s_t \cdot \text{sign}(\mathbf{w}_t) \odot \mathbf{q}_t$, where $\mathbf{w}_t$ is the weight before quantization, $s_t = \|\mathbf{w}_t\|_\infty$, $\mathbf{q}_t \in (\mathcal{S}_w)^d$, and $\mathcal{S}_w = \{-M_k, \dots, -M_1, M_0, M_1, \dots, M_k\}$, with $k = 2^{m-1} - 1$, $0 = M_0 < M_1 < \cdots < M_k$ are uniformly spaced, and the weight quantization resolution $\Delta_w = M_{r+1} - M_r$. The ith element $q_{t,i}$ in $\mathbf{q}_t$ is equal to M_{r+1} with probability $(|w_{t,i}|/s_t - M_r)/(M_{r+1} - M_r)$, and M_r otherwise. Here, r is an index satisfying $M_r \leq |w_{t,i}|/s_t < M_{r+1}$.

When only weights are quantized, the update with full-precision gradient is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f_t(Q_w(\mathbf{w}_t)). \tag{20}$$

We still denote the full-precision gradient $\nabla_{\mathbf{w}} f_t(Q_w(\mathbf{w}_t))$ as $\hat{\mathbf{g}}_t$. As f_t is twice differentiable (Assumption **A2**), using the mean value theorem, there exists $p \in (0, 1)$ such that $\|\mathbf{g}_t - \hat{\mathbf{g}}_t\|_2 = \|\nabla_{\mathbf{w}} f_t(\mathbf{w}_t) - \nabla_{\mathbf{w}} f_t(\hat{\mathbf{w}}_t)\|_2 = \|\nabla^2_{\mathbf{w}} f_t(\hat{\mathbf{w}}_t + p(\mathbf{w}_t - \hat{\mathbf{w}}_t))(\mathbf{w}_t - \hat{\mathbf{w}}_t)\|_2$. Let $\mathbf{H}'_t = \nabla^2_{\mathbf{w}} f_t(\hat{\mathbf{w}}_t + p(\mathbf{w}_t - \hat{\mathbf{w}}_t))$ be the Hessian at $\hat{\mathbf{w}}_t + p(\mathbf{w}_t - \hat{\mathbf{w}}_t)$.

Theorem 4. *For stochastic weight quantization (scheme (20)) with full-precision weights, and $\eta_t = \eta/\sqrt{t}$,*

$$R(T) \leq \frac{D_\infty^2}{2\eta} d\sqrt{T} + \sum_{t=1}^{T} \frac{\eta}{2\sqrt{t}} \|\hat{\mathbf{g}}_t\|^2 + \sqrt{L} D \sum_{t=1}^{T} \sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}'_t}}.$$

$$\frac{R(T)}{T} \leq O(d/\sqrt{T}) + LD\sqrt{\frac{dD_\infty^2 \Delta_w^2}{4}}.$$

When both weights and gradients are quantized, the update is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t Q_g(\nabla_{\mathbf{w}} f_t(Q_w(\mathbf{w}_t))). \tag{21}$$

Theorem 5. *For stochastic weight quantization with quantized gradients (scheme (21)) and $\eta_t = \eta/\sqrt{t}$,*

$$\mathbf{E}(R(T)) \leq \frac{D_\infty^2}{2\eta} d\sqrt{T} + \sum_{t=1}^{T} \frac{\eta}{2\sqrt{t}} \mathbf{E}(\|\tilde{\mathbf{g}}_t\|^2) + \sqrt{L} D \sum_{t=1}^{T} \sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t}}.$$

$$\mathbf{E}(\frac{R(T)}{T}) \leq O((\frac{1 + \sqrt{2d - 1}}{2}\Delta_g + 1)d/\sqrt{T}) + LD\sqrt{\frac{dD_\infty^2 \Delta_w^2}{4}}.$$

Theorem 6. *Assume that $\hat{\mathbf{g}}_t$ follows $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. For stochastic weight quantization with quantized clipped gradients and $\eta_t = \eta/\sqrt{t}$,*

$$\begin{aligned}
\mathbf{E}(R(T)) \quad &\leq \quad \frac{D_\infty^2}{2\eta} d\sqrt{T} + \sum_{t=1}^{T} \frac{\eta}{2\sqrt{t}} \mathbf{E}(\|\check{\mathbf{g}}_t\|^2) \\
&\quad + \sqrt{L} D \textstyle\sum_{t=1}^{T} \mathbf{E}(\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2_{\mathbf{H}'_t}}) + D \textstyle\sum_{t=1}^{T} \mathbf{E}(\sqrt{\|\mathit{Clip}(\hat{\mathbf{g}}_t) - \hat{\mathbf{g}}_t\|^2}), \\
\mathbf{E}\left(\frac{R(T)}{T}\right) \quad &\leq \quad O\left(((2/\pi)^{\frac{1}{2}} c\Delta_g + 1)\frac{d}{\sqrt{T}}\right) + LD\sqrt{\frac{dD_\infty^2 \Delta_w^2}{4}} + \sqrt{d} D\sigma(2/\pi)^{\frac{1}{4}} \sqrt{F(c)}.
\end{aligned}$$

As can be seen from Theorems 4-6, there is also an error term $LD\sqrt{\frac{dD_\infty^2\Delta_w^2}{4}}$ related to the weight quantization resolution and dimension. Moreover, this error term can be potentially larger than the one induced by the loss-aware weight quantization in Theorem 1, as D_∞ can be much larger than α and dD_∞^2 be much larger than D^2.

We provide a sketch of proof for Theorem 4 in the next section, the proofs for Theorems 5 and 6 can be extended from it in a similar way as the loss-aware weight quantization.

D.2.1 Proof for Theorem 4

Proof. Consider the ith parameter of $\mathbf{w}_t$. the update for it is $w_{t+1,i} = w_{t,i} - \eta_t \hat{g}_{t,i}$ which implies

$$(w_{t+1,i} - w_i^*)^2 = (w_{t,i} - w_i^*)^2 - 2\eta_t(w_{t,i} - w_i^*)g_{t,i} + 2\eta_t(w_{t,i} - w_i^*)(g_{t,i} - \hat{g}_{t,i}) + \eta_t^2\hat{g}_{t,i}^2.$$

After rearranging,

$$g_{t,i}(w_{t,i} - w_i^*) = \frac{1}{2\eta_t}((w_{t,i} - w_i^*)^2 - (w_{t+1,i} - w_i^*)^2) + (w_{t,i} - w_i^*)(g_{t,i} - \hat{g}_{t,i}) + \frac{\eta_t}{2}\hat{g}_{t,i}^2. \quad (22)$$

Since f_t is convex, we have

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) \leq \mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}^*) = \sum_{i=1}^d g_{t,i}(w_{t,i} - w_i^*).$$

As f_t is convex (Assumption **A1**) and twice differentiable (Assumption **A2**), $\|\nabla f_t(\mathbf{u}) - \nabla f_t(\mathbf{v})\|_2 \leq L\|\mathbf{u} - \mathbf{v}\|_2$ for any $\mathbf{u}, \mathbf{v}$ is equivalent to $\nabla^2 f_t \preceq L\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Thus from the domain bound assumption $\|\mathbf{w}_m - \mathbf{w}_n\|_2 \leq D$, we have

$$\|\mathbf{w}_t - \mathbf{w}^*\|_{\mathbf{H}_t'}^2 \leq L\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq LD^2, \quad \|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2 \leq L\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2.$$

Combining (22) and (D.2.1), and sum over all the dimensions for $i \in \{1, 2\cdots, d\}$ and over all the iterations for $t \in \{1, 2, \cdots, T\}$, we have

$$\begin{aligned}
R(T) &= \sum_{t=1}^T f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) \\
&\leq \sum_{i=1}^d \frac{1}{2\eta_1}(w_{1,i} - w_i^*)^2 + \sum_{i=1}^d\sum_{t=2}^T(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}})(w_{t,i} - w_i^*)^2 \\
&\quad + \sum_{t=1}^T\langle\mathbf{w}_t - \mathbf{w}^*, \mathbf{g}_t - \hat{\mathbf{g}}_t\rangle + \sum_{i=1}^d\sum_{t=1}^T\frac{\eta_t}{2}\hat{g}_{t,i}^2 \\
&\leq \frac{dD_\infty^2}{2\eta}\sqrt{T} + \sum_{t=1}^T\frac{\eta_t}{2}\|\hat{\mathbf{g}}_t\|^2 + \sum_{t=1}^T\langle\mathbf{H}_t^{'\frac{1}{2}}(\mathbf{w}_t - \mathbf{w}^*), \mathbf{H}_t^{'-\frac{1}{2}}(\mathbf{g}_t - \hat{\mathbf{g}}_t)\rangle \\
&\leq \frac{dD_\infty^2}{2\eta}\sqrt{T} + \sum_{t=1}^T\frac{\eta_t}{2}\|\hat{\mathbf{g}}_t\|^2 + \sum_{t=1}^T\sqrt{\|\mathbf{w}_t - \mathbf{w}^*\|_{\mathbf{H}_t'}^2}\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2} \quad (23) \\
&\leq \frac{D_\infty^2}{2\eta}d\sqrt{T} + \eta d\sqrt{T}G_\infty^2 + \sqrt{L}D\sum_{t=1}^T\sqrt{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_{\mathbf{H}_t'}^2}. \quad (24)
\end{aligned}$$

The second inequality comes from the definition of $\mathbf{H}_t'$. The third inequality comes from the definition of $\mathbf{H}_t'$ and Cauchy's inequality. The last inequality comes from $\sum_{t=1}^T\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$.

For m-bit ($m > 1$) stochastic linear weight quantization, as $s_t M_{k-1} \leq w_{t,i} \leq s_t M_k$, and $s_t \leq D_\infty$ we have

$$\begin{aligned}
(\hat{w}_{t,i} - w_{t,i})^2 &= \min\{(s_t M_{r+1,i} - |w_{t,i}|)^2, (s_t M_{r,i} - |w_{t,i}|)^2\} \\
&\leq (\frac{s_t M_{r+1,i} - s_t M_{r,i}}{2})^2 \leq \frac{D_\infty^2\Delta_w^2}{4}.
\end{aligned}$$

Thus the regret satisfies

$$R(T) \leq \frac{D_\infty^2}{2\eta} d\sqrt{T} + \eta d\sqrt{T} G_\infty^2 + LD\sqrt{\frac{dD_\infty^2 \Delta_w^2}{4}} T.$$

Thus the average regret

$$R(T)/T \leq (\frac{D_\infty^2}{2\eta} + \eta G_\infty^2)\frac{d}{\sqrt{T}} + LD\sqrt{\frac{dD_\infty^2 \Delta_w^2}{4}}. \tag{25}$$

□