A. Proof of the Theorems

A.1. Table of Notations

Table 9. Table of Notations.

Notation	Description
$\mathbf{X}$	A batch of inputs (each row is a sample)
$\mathbf{Y}$	A batch of labels (each row is a sample)
$\mathcal{B}$	A batch $\mathcal{B} = (\mathbf{X}, \mathbf{Y})$
N, L	Batch size, number of classes, and number of layers
$\mathbf{F}^{(l)}(\cdot; \mathbf{\Theta}^{(l)})$	Forward function of the l-th layer with parameter $\mathbf{\Theta}^{(l)}$
$\mathbf{G}^{(l)}(\cdot; \cdot)$	Backward function of the l-th layer
$\mathbf{C}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)}), \mathbf{C}^{(l)}$	l-th layer's context
$\mathbf{C}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)}), \hat{\mathbf{C}}^{(l)}$	l-th layer's compressed context
$\mathcal{L} = \ell(\mathbf{H}^{(L)}, \mathbf{Y})$	Minibatch loss function of prediction $\mathbf{H}^{(L)}$ and label $\mathbf{Y}$.
$\mathcal{L}_{\mathcal{D}}$	Batch loss on the entire dataset.
$\nabla_{\mathbf{\Theta}} \mathcal{L}_{\mathcal{D}}$	Batch gradient
$\nabla_{\mathbf{H}^{(l)}}, \nabla_{\mathbf{\Theta}^{(l)}}$	Full-precision gradient of activation / parameter
$\hat{\nabla}_{\mathbf{H}^{(l)}}, \hat{\nabla}_{\mathbf{\Theta}^{(l)}}$	Activation-compressed gradient of activation / parameter
$b_n^{(l)}, B_n^{(l)}$	Number of quantization bits / bins for $\mathbf{h}_n^{(l)}$
G	Group size for per-group quantization
$R, R_{ni}, \mathbf{R}$	Quantization range

A.2. Theorem 1

The FP gradient is defined by the recursion

$$\nabla_{\mathbf{H}^{(l-1)}}, \nabla_{\mathbf{\Theta}^{(l)}} = \mathbf{G}^{(l)}\left(\nabla_{\mathbf{H}^{(l)}}, \mathbf{C}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)})\right),$$

and the ActNN gradient is defined by

$$\hat{\nabla}_{\mathbf{H}^{(l-1)}}, \hat{\nabla}_{\mathbf{\Theta}^{(l)}} = \mathbf{G}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}}, \hat{\mathbf{C}}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)})\right),$$

where $\hat{\nabla}_{\mathbf{H}^{(L)}} = \nabla_{\mathbf{H}^{(L)}}$. The batch loss is $\mathcal{L}_{\mathcal{D}}(\mathbf{\Theta})$ and the batch gradient is $\nabla_{\mathbf{\Theta}} \mathcal{L}_{\mathcal{D}}(\mathbf{\Theta})$. Assume that the FP gradient is unbiased, i.e., $\mathbb{E}\left[\nabla_{\mathbf{\Theta}^{(l)}}\right] = \nabla_{\mathbf{\Theta}^{(l)}} \mathcal{L}_{\mathcal{D}}(\mathbf{\Theta})$ for all l.

We first prove the following lemma

Lemma 1. *If* $\mathbb{E}\left[\hat{\nabla}_{\mathbf{H}^{(l)}}\right] = \mathbb{E}\left[\nabla_{\mathbf{H}^{(l)}}\right]$, *then there exists* $\hat{\mathbf{C}}^{(l)}$, *s.t.,* $\mathbb{E}\left[\hat{\nabla}_{\mathbf{H}^{(l-1)}}\right] = \mathbb{E}\left[\nabla_{\mathbf{H}^{(l-1)}}\right]$ *and* $\mathbb{E}\left[\hat{\nabla}_{\mathbf{\Theta}^{(l)}}\right] = \mathbb{E}\left[\nabla_{\mathbf{\Theta}^{(l)}}\right]$.

Proof. By the chain rule of differentiation, we have

$$\nabla_{H_{ij}^{(l-1)}} = \sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}} \nabla_{H_{kl}^{(l)}}, \quad \nabla_{\Theta_i^{(l)}} = \sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial \Theta_i^{(l)}} \nabla_{H_{kl}^{(l)}}. \tag{11}$$

Therefore, we can write

$$\mathbf{G}^{(l)}\left(\nabla_{\mathbf{H}^{(l)}}, \mathbf{C}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)})\right) = \{\sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}} \nabla_{H_{kl}^{(l)}}\}_{ij}, \{\sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial \Theta_i^{(l)}} \nabla_{H_{kl}^{(l)}}\}_i,$$

where $\mathbf{C}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)}) = \{\partial H_{kl}^{(l)} / \partial H_{ij}^{(l-1)}, \partial H_{kl}^{(l)} / \partial \Theta_i^{(l)}\}_{ijkl}$. Let $\hat{\mathbf{C}}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)}) = Q(\mathbf{C}(\mathbf{H}^{(l-1)}, \mathbf{\Theta}^{(l)}))$, where $Q(\cdot)$ is an unbiased quantizer, i.e., for all $\mathbf{x}$, $\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}$. Then, we have

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\nabla}_{\mathbf{H}^{(l-1)}}, \hat{\nabla}_{\boldsymbol{\Theta}^{(l)}}\right] = \mathbb{E}\left[\mathbf{G}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}}, \hat{\mathbf{C}}(\mathbf{H}^{(l-1)}, \boldsymbol{\Theta}^{(l)})\right)\right] = \mathbb{E}\left[\{\sum_{kl} Q(\frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}})\hat{\nabla}_{H_{kl}^{(l)}}\}_{ij}, \{\sum_{kl} Q(\frac{\partial H_{kl}^{(l)}}{\partial \Theta_i^{(l)}})\hat{\nabla}_{H_{kl}^{(l)}}\}_i\right] \\
=&\{\sum_{kl} \mathbb{E}Q(\frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}})\mathbb{E}\hat{\nabla}_{H_{kl}^{(l)}}\}_{ij}, \{\sum_{kl} \mathbb{E}Q(\frac{\partial H_{kl}^{(l)}}{\partial \Theta_i^{(l)}})\mathbb{E}\hat{\nabla}_{H_{kl}^{(l)}}\}_i = \{\sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}}\nabla_{H_{kl}^{(l)}}\}_{ij}, \{\sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial \Theta_i^{(l)}}\nabla_{H_{kl}^{(l)}}\}_i \\
=&\mathbf{G}^{(l)}\left(\nabla_{\mathbf{H}^{(l)}}, \mathbf{C}(\mathbf{H}^{(l-1)}, \boldsymbol{\Theta}^{(l)})\right) = \nabla_{\mathbf{H}^{(l-1)}}, \nabla_{\boldsymbol{\Theta}^{(l)}}.
\end{aligned}
$$

□

Now we can prove Theorem 1.

Proof. By definition, $\hat{\nabla}_{\mathbf{H}^{(L)}} = \nabla_{\mathbf{H}^{(L)}}$, so $\mathbb{E}\left[\hat{\nabla}_{\mathbf{H}^{(L)}}\right] = \mathbb{E}\left[\nabla_{\mathbf{H}^{(L)}}\right]$. Using Lemma 1 and mathematical induction, we get $\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{\Theta}^{(l)}}\right] = \mathbb{E}\left[\nabla_{\boldsymbol{\Theta}^{(l)}}\right]$, for all $l \in \{1, \dots, L\}$, so $\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{\Theta}}\right] = \mathbb{E}\left[\nabla_{\boldsymbol{\Theta}}\right]$.

On the other hand, $\mathbb{E}\left[\nabla_{\boldsymbol{\Theta}}\right] = \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\Theta})$, by the assumption. Put it all together, we have $\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{\Theta}}\right] = \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\Theta})$. □

A.3. Theorem 2

Let $\boldsymbol{\Theta}_t = \{\boldsymbol{\Theta}^{(l)}\}_{l=1}^{L}$ be a flattened vector of the parameters at the t-th iteration, and $\hat{\nabla}_{\boldsymbol{\Theta}_t} = \{\hat{\nabla}_{\boldsymbol{\Theta}_t^{(l)}}\}_{l=1}^{L}$ be the corresponding AC gradient. For any vector $\mathbf{x}$, let $\mathrm{Var}\left[\mathbf{x}\right] := \mathbb{E}\left\|\mathbf{x}\right\|^2 - \left\|\mathbb{E}\left[\mathbf{x}\right]\right\|^2$. Let $\mathcal{L}(\boldsymbol{\Theta}_t)$ be the batch loss at the t-th iteration, where $\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\Theta}_t) = \mathbb{E}\left[\hat{\nabla}_{\boldsymbol{\Theta}_t}\right]$. Assume the SGD iteration $\boldsymbol{\Theta}_{t+1} \leftarrow \boldsymbol{\Theta}_t - \alpha\hat{\nabla}_{\boldsymbol{\Theta}_t}$. Further, let $\mathbb{E}\left[\cdot \mid t\right]$ and $\mathrm{Var}\left[\cdot \mid t\right]$ be the expectation and variance taken only over the minibatch and random quantizations at the t-th iteration.

Proof. According to Bottou et al. (2018), A1 implies that for any $\boldsymbol{\Theta}_t$ and $\boldsymbol{\Theta}_{t+1}$,

$$\mathcal{L}(\boldsymbol{\Theta}_{t+1}) - \mathcal{L}(\boldsymbol{\Theta}_t) \le \nabla\mathcal{L}(\boldsymbol{\Theta}_t)^\top(\boldsymbol{\Theta}_{t+1} - \boldsymbol{\Theta}_t) + \frac{1}{2}\beta\left\|\boldsymbol{\Theta}_{t+1} - \boldsymbol{\Theta}_t\right\|^2.$$

Plugging the SGD iteration, we have

$$\mathcal{L}(\boldsymbol{\Theta}_{t+1}) - \mathcal{L}(\boldsymbol{\Theta}_t) \le -\alpha\mathcal{L}(\boldsymbol{\Theta}_t)^\top\hat{\nabla}_{\boldsymbol{\Theta}_t} + \frac{1}{2}\alpha^2\beta\left\|\hat{\nabla}_{\boldsymbol{\Theta}_t}\right\|^2,$$

taking expectation w.r.t. iteration $t+1$ on both sides, and utilizing A3, the definition of variance, the step size inequality, and the unbiased AC gradient,

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\boldsymbol{\Theta}_{t+1}) \mid t+1\right] - \mathcal{L}(\boldsymbol{\Theta}_t) &\le -\alpha\left\|\mathcal{L}(\boldsymbol{\Theta}_t)\right\|^2 + \frac{1}{2}\alpha^2\beta(\mathrm{Var}\left[\hat{\nabla}_{\boldsymbol{\Theta}_t} \;\middle|\; t+1\right] + \left\|\mathbb{E}\left[\hat{\nabla}_{\boldsymbol{\Theta}_t} \;\middle|\; t+1\right]\right\|^2) \\
&= -\alpha(1 - \frac{1}{2}\alpha\beta)\left\|\mathcal{L}(\boldsymbol{\Theta}_t)\right\|^2 + \frac{1}{2}\alpha^2\beta\sigma^2 \\
&\le -\frac{1}{2}\alpha\left\|\mathcal{L}(\boldsymbol{\Theta}_t)\right\|^2 + \frac{1}{2}\alpha^2\beta\sigma^2.
\end{aligned}
$$

Taking expectation on both sides, we have

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\Theta}_{t+1})\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\Theta}_t)\right] \le -\frac{1}{2}\alpha\mathbb{E}\left\|\mathcal{L}(\boldsymbol{\Theta}_t)\right\|^2 + \frac{1}{2}\alpha^2\beta\sigma^2. \tag{12}$$

Summing Eq. (12) up across iterations $\{1, \dots, T\}$, and utilize A2, we have

$$L_{\inf} - \mathcal{L}(\boldsymbol{\Theta}_1) \le \mathbb{E}\left[\mathcal{L}(\boldsymbol{\Theta}_{T+1})\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\Theta}_1)\right] \le -\frac{1}{2}\alpha\sum_{t=1}^{T}\mathbb{E}\left\|\mathcal{L}(\boldsymbol{\Theta}_t)\right\|^2 + \frac{1}{2}T\alpha^2\beta\sigma^2.$$

Rearrange the terms, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\mathcal{L}(\mathbf{\Theta}_t)\right\|^2 \le \frac{2(\mathcal{L}(\Theta_1)-\mathcal{L}_{inf})}{\alpha T}+\alpha\beta\sigma^2.$$

Viewing t as a random variable, we have Eq. (6).

□

A.4. Theorem 3

Let $\mathbb{E}\left[X \mid Y\right]$ and $\mathrm{Var}\left[X \mid Y\right]$ be the conditional expectation / variance of X given Y. We use the following proposition.

Proposition 1. *(Law of Total Variance)*

$$\mathrm{Var}\left[X\right]=\mathbb{E}\left[\mathrm{Var}\left[X \mid Y\right]\right]+\mathrm{Var}\left[\mathbb{E}\left[X \mid Y\right]\right].$$

Define

$$\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}},\hat{\mathbf{C}}^{(m)}\right)=\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\mathbf{G}_{\mathbf{H}}^{(l+1)}\left(\cdots\mathbf{G}_{\mathbf{H}}^{(m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}},\hat{\mathbf{C}}^{(m)}\right)\cdots,\mathbf{C}^{(l+1)}\right),\mathbf{C}^{(l)}\right),$$
$$\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}\right)=\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\mathbf{G}_{\mathbf{H}}^{(l+1)}\left(\cdots\mathbf{G}_{\mathbf{H}}^{(m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}},\mathbf{C}^{(m)}\right)\cdots,\mathbf{C}^{(l+1)}\right),\mathbf{C}^{(l)}\right),$$

Proof. (of Theorem 3) First, we have

$$\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim L)}\left(\hat{\nabla}_{\mathbf{H}^{(L)}}\right)\right]=\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim L)}\left(\nabla_{\mathbf{H}^{(L)}}\right)\right]=\mathrm{Var}\left[\nabla_{\mathbf{H}^{(l)}}\right]. \tag{13}$$

For all $m<L$, by definition of $\hat{\nabla}_{\mathbf{H}^{(m)}}$

$$\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}\right)\right]=\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\mathbf{G}_{\mathbf{H}}^{(m+1)}\left(\hat{\nabla}_{\mathbf{H}^{(m+1)}},\hat{\mathbf{C}}^{(m+1)}\right)\right)\right],$$

by law of total variance

$$\begin{aligned}&\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\mathbf{G}_{\mathbf{H}}^{(m+1)}\left(\hat{\nabla}_{\mathbf{H}^{(m+1)}},\hat{\mathbf{C}}^{(m+1)}\right)\right)\right]\\=&\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\mathbf{G}_{\mathbf{H}}^{(m+1)}\left(\hat{\nabla}_{\mathbf{H}^{(m+1)}},\hat{\mathbf{C}}^{(m+1)}\right)\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(m+1)}}\right]\right]+\mathrm{Var}\left[\mathbb{E}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\mathbf{G}_{\mathbf{H}}^{(m+1)}\left(\hat{\nabla}_{\mathbf{H}^{(m+1)}},\hat{\mathbf{C}}^{(m+1)}\right)\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(m+1)}}\right]\right]\end{aligned}$$

by definition of $\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}},\hat{\mathbf{C}}^{(m)}\right)$, definition of $\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}\right)$ and Theorem 1

$$=\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m+1)}\left(\hat{\nabla}_{\mathbf{H}^{(m+1)}},\hat{\mathbf{C}}^{(m+1)}\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(m+1)}}\right]\right]+\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m+1)}\left(\hat{\nabla}_{\mathbf{H}^{(m+1)}}\right)\right]. \tag{14}$$

Combining Eq. (13) and Eq. (14), we have

$$\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}\right)\right]=\mathrm{Var}\left[\nabla_{\mathbf{H}^{(l)}}\right]+\sum_{j=m+1}^{L}\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim j)}\left(\hat{\nabla}_{\mathbf{H}^{(j)}},\hat{\mathbf{C}}^{(j)}\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(j)}}\right]\right]. \tag{15}$$

Similarly, by definition and the law of total variance,

$$\mathrm{Var}\left[\hat{\nabla}_{\mathbf{\Theta}^{(l)}}\right]=\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}},\hat{\mathbf{C}}^{(l)}\right)\right]=\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}},\hat{\mathbf{C}}^{(l)}\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(l)}}\right]\right]+\mathrm{Var}\left[\mathbb{E}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}},\hat{\mathbf{C}}^{(l)}\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(l)}}\right]\right].$$

. by definition of $\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}},\hat{\mathbf{C}}^{(m)}\right)$, definition of $\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}\right)$ and Theorem 1

$$=\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}},\hat{\mathbf{C}}^{(l)}\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(l)}}\right]\right]+\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}},\mathbf{C}^{(l)}\right)\right] \tag{16}$$
$$=\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}},\hat{\mathbf{C}}^{(l)}\right)\;\middle|\;\hat{\nabla}_{\mathbf{H}^{(l)}}\right]\right]+\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}}\right)\right]. \tag{17}$$

Plugging Eq. (15) into Eq. (16), we have

$$\begin{aligned}\mathrm{Var}\left[\hat{\nabla}_{\boldsymbol{\Theta}^{(l)}}\right] = & \mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\boldsymbol{\Theta}}^{(l\sim l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}}, \hat{\mathbf{C}}^{(l)}\right) \;\middle|\; \hat{\nabla}_{\mathbf{H}^{(l)}}\right]\right] + \mathrm{Var}\left[\nabla_{\mathbf{H}^{(l)}}\right] + \sum_{j=l+1}^{L} \mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\boldsymbol{\Theta}}^{(l\sim j)}\left(\hat{\nabla}_{\mathbf{H}^{(j)}}, \hat{\mathbf{C}}^{(j)}\right) \;\middle|\; \hat{\nabla}_{\mathbf{H}^{(j)}}\right]\right] \\ = & \mathrm{Var}\left[\nabla_{\mathbf{H}^{(l)}}\right] + \sum_{m=l}^{L} \mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\boldsymbol{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}, \hat{\mathbf{C}}^{(m)}\right) \;\middle|\; \hat{\nabla}_{\mathbf{H}^{(m)}}\right]\right].\end{aligned}$$

□

B. Bias and Variance for Individual Layers

B.1. Convolutional Layers

Consider an arbitrary dimensional convolutional layer

$$\mathbf{y}_{nia} = \sum_{\Delta_i} \mathbf{W}_{\Delta_i,a}\mathbf{x}_{n,si+d\Delta_i,a}. \tag{18}$$

where $\mathbf{X} = \{\mathbf{x}_{nia}\}$ is the input, $\mathbf{Y} = \{\mathbf{y}_{nia}\}$ is the output, s is a stride, and d is a dilation, $a \in [0, A)$ is the group index for depthwise-separable convolution. $\mathbf{x}_{ni}$ is the feature vector of the n-th sample at the i-th location, where i can be a tuple of arbitrary dimensions, e.g., 2 or 3. The kernel has $K = |\Delta_i|$ locations, and for each location, the kernel $\mathbf{W}_{\Delta_i,a}$ is a $(D_{out}/A) \times (D_{in}/A)$ matrix.

The gradients are

$$\nabla_{\mathbf{W}_{\Delta_i},a} = \sum_{ni} \nabla_{\mathbf{y}_{nia}} \mathbf{x}_{n,si+d\Delta_i,a}^\top, \quad \nabla_{\mathbf{x}_{nia}} = \sum_{\Delta_i} \nabla_{\mathbf{y}_{n,(i-d\Delta_i)/s,a}} \mathbf{W}_{\Delta_i,a}^\top. \tag{19}$$

Define the approximate context as $\hat{\mathbf{C}} = (Q(\mathbf{X}), \mathbf{W})$, where $Q(\cdot)$ is an unbiased quantizer. Then,

$$\mathbb{E}\left[\hat{\nabla}_{\mathbf{W}_{\Delta_i,a}}\right] = \sum_{ni} \mathbb{E}\left[\hat{\nabla}_{\mathbf{y}_{nia}}\right] \mathbb{E}\left[Q(\mathbf{x}_{n,si+d\Delta_i,a}^\top)\right] = \sum_{ni} \nabla_{\mathbf{y}_{nia}} \mathbf{x}_{n,si+d\Delta_i,a}^\top. = \mathbb{E}\left[\nabla_{\mathbf{W}_{\Delta_i,a}}\right].$$

Therefore, the gradient is unbiased.

Let I be the number of locations (pixels) on the feature map $\mathbf{X}$, and S is the product of strides. The variance is

$$\mathrm{Var}\left[\sum_{ni} \nabla_{\mathbf{y}_{nia}} \mathbf{x}_{n,si+d\Delta_i,a}^\top\right] = \sum_{c_1c_2} \mathrm{Var}\left[\sum_{ni} \nabla_{y_{n,i,a,c_1}} x_{n,si+d\Delta_i,a,c_2}\right]$$

Due to independence,

$$\begin{aligned}\mathrm{Var}\left[\sum_{ni} \nabla_{\mathbf{y}_{nia}} \mathbf{x}_{n,si+d\Delta_i,a}^\top\right] = & \sum_{c_1c_2ni} \nabla_{y_{n,i,a,c_1}}^2 \mathrm{Var}\left[x_{n,si+d\Delta_i,a,c_2}\right] \\ = & \frac{G}{6} \sum_{ni} \left\|\nabla_{\mathbf{y}_{n,i,a}}\right\|^2 \left\|\mathbf{R}_{n,si+d\Delta_i,a}\right\|^2 \approx \frac{G}{6I} \sum_{n} \left\|\nabla_{\mathbf{y}_{na}}\right\|^2 \left\|\mathbf{R}_{na}\right\|^2,\end{aligned}$$

where we approximate $\sum_i a_i b_i \approx \mathrm{sum}(a_i)\mathrm{mean}(b_i)$. Therefore,

$$\mathrm{Var}\left[\nabla_{\mathbf{W}}\right] = \sum_{\Delta_i,a} \nabla_{\mathbf{W}_{\Delta_i,a}} \approx \frac{GK}{6I} \sum_{na} \left\|\nabla_{\mathbf{y}_{na}}\right\|^2 \left\|\mathbf{R}_{na}\right\|^2 \approx \frac{GK}{6IA} \sum_{n} \left\|\nabla_{\mathbf{y}_n}\right\|^2 \left\|\mathbf{R}_n\right\|^2.$$

Transposed Convolution We can view transposed convolution as convolutions with inverse stride. For example, if a Conv2D has the stride $[2, 2]$, then its transpose has the stride $[1/2, 1/2]$.

B.2. Normalization Layers

Suppose the input is $\mathbf{X} \in \mathbb{R}^{N\times C}$, where there are C features to be normalized. The layer has the weight $\mathbf{w} \in \mathbb{R}^{C}$ and the bias $\mathbf{b} \in \mathbb{R}^{C}$. The output is $\mathbf{Y} \in \mathbb{R}^{N\times C}$. For example, for `BatchNorm2d`, C is the number of channels, and N is the product of the batch size and the number of pixels per image. This formulation applies to both batch normalization and layer normalization, for arbitary-shape tensors.

Forward Propagation

$$y_{nc} = (x_{nc} - m_c)\frac{w_c}{s_c} + b_c, \text{ where } m_c = \frac{1}{N}\sum_n x_{nc}, s_c = \sqrt{\frac{1}{N}\sum_n (x_{nc} - m_c)^2}. \tag{20}$$

Back Propagation

$$\nabla_{x_{nc}} = \frac{w_c}{s_c}\left(\nabla_{y_{nc}} - \frac{1}{N}\sum_{c'}\nabla_{y_{nc'}} - \frac{1}{Ns_c^2}(x_{nc} - m_c)\sum_{n'}(x_{n'c} - m_c)\nabla_{y_{n'c}}\right)$$

The context is $(\mathbf{X}, \mathbf{m}, \mathbf{s}, \mathbf{w})$. We only quantize $\mathbf{X}$ here since all the other vectors are negligible in size.

Unbiased Quantization We can keep two independently quantized copies of x_{nc}: $\hat{x}_{nc}$ and $\dot{x}_{nc}$, such that $\mathbb{E}\left[\hat{x}_{nc}\right] = x_{nc}$, $\mathbb{E}\left[\dot{x}_{nc}\right] = x_{nc}$. In this way,

$$\mathbb{E}\left[\hat{\nabla}_{x_{nc}}\right] = \mathbb{E}\left[\frac{w_c}{s_c}\left(\hat{\nabla}_{y_{nc}} - \frac{1}{N}\sum_{c'}\hat{\nabla}_{y_{nc'}} - \frac{1}{Ns_c^2}(\hat{x}_{nc} - m_c)\sum_{n'}(\dot{x}_{n'c} - m_c)\hat{\nabla}_{y_{n'c}}\right)\right]$$

Due to independence,

$$\begin{aligned}\mathbb{E}\left[\hat{\nabla}_{x_{nc}}\right] &= \frac{w_c}{s_c}\left(\mathbb{E}\left[\hat{\nabla}_{y_{nc}}\right] - \frac{1}{N}\sum_{c'}\mathbb{E}\left[\hat{\nabla}_{y_{nc'}}\right] - \frac{1}{Ns_c^2}\mathbb{E}\left[\hat{x}_{nc} - m_c\right]\sum_{n'}\mathbb{E}\left[\dot{x}_{n'c} - m_c\right]\mathbb{E}\left[\hat{\nabla}_{y_{n'c}}\right]\right)\\ &= \frac{w_c}{s_c}\left(\nabla_{y_{nc}} - \frac{1}{N}\sum_{c'}\nabla_{y_{nc'}} - \frac{1}{Ns_c^2}(x_{nc} - m_c)\sum_{n'}(x_{n'c} - m_c)\nabla_{y_{n'c}}\right) = \nabla_{x_{nc}}.\end{aligned}$$

Therefore, the gradient of the input is unbiased.

Gradient Variance

$$\begin{aligned}\mathrm{Var}\left[\hat{\nabla}_{x_{nc}}\right] &= \frac{w_c^2}{N^2 s_c^6}\mathrm{Var}\left[(\hat{x}_{nc} - m_c)\sum_{n'}(\dot{x}_{n'c} - m_c)\nabla_{y_{n'c}}\right]\\ &= \frac{w_c^2}{N^2 s_c^6}\left(\mathrm{Var}\left[A\right]\mathrm{Var}\left[B\right] + \mathbb{E}\left[A\right]^2\mathrm{Var}\left[B\right] + \mathrm{Var}\left[A\right]\mathbb{E}\left[B\right]^2\right),\end{aligned}$$

utilizing $\mathrm{Var}\left[AB\right] = \mathrm{Var}\left[A\right]\mathrm{Var}\left[B\right] + \mathbb{E}\left[A\right]^2\mathrm{Var}\left[B\right] + \mathrm{Var}\left[A\right]\mathbb{E}\left[B\right]^2$, if A and B are independent, where

$$\begin{aligned}&A = \hat{x}_{nc} - m_c, \quad \mathbb{E}\left[A\right] = x_{nc} - m_c, \quad \mathrm{Var}\left[A\right] = \mathrm{Var}\left[\hat{x}_{nc}\right]\\ &B = \sum_{n'}(\dot{x}_{n'c} - m_c)\nabla_{y_{n'c}}, \quad \mathbb{E}\left[B\right] = \sum_{n'}(x_{n'c} - m_c)\nabla_{y_{n'c}}, \quad \mathrm{Var}\left[B\right] = \sum_{n'}\mathrm{Var}\left[\dot{x}_{n'c}\right]\nabla_{y_{n'c}}^2.\end{aligned}$$

Assume that $\mathrm{Var}\left[A\right] \ll \mathbb{E}\left[A\right]^2$ and $\mathrm{Var}\left[B\right] \ll \mathbb{E}\left[B\right]^2$, and utilize Eq. (20), we have

$$\mathrm{Var}\left[\hat{\nabla}_{X_{nc}}\right] \approx \frac{w_c^2}{N^2 s_c^4}\left[y_{nc}^2\sum_{n'}\mathrm{Var}\left[\dot{x}_{n'c}\right]\nabla_{y_{n'c}}^2 + \mathrm{Var}\left[\hat{x}_{nc}\right]\left(\sum_{n'}y_{n'c}\nabla_{y_{n'c}}\right)^2\right].$$

Let $d_c = \sum_{n'} y_{n'c}\nabla_{y_{n'c}}$, and plug $\mathrm{Var}\left[\hat{x}_{nc}\right] \approx \frac{R_n^2}{6B_n^2}$ in, we have

$$\mathrm{Var}\left[\nabla_{X_{nc}}\right] \approx \frac{w_c^2}{6N^2 s_c^4}\left[y_{nc}^2 \sum_{n'} \frac{R_{n'}^2}{B_{n'}^2}\nabla_{y_{n'c}}^2 + \frac{R_n^2}{B_n^2} d_c^2\right].$$

Summing the terms up, we have

$$\mathrm{Var}\left[\nabla_{\mathbf{X}}\right] = \sum_{nc} \mathrm{Var}\left[\nabla_{x_{nc}}\right] = \frac{1}{6N^2}\sum_{n'} \frac{R_{n'}^2}{B_{n'}^2}\left(\sum_c \frac{w_c^2}{s_c^4}\nabla_{y_{n'c}}^2 \sum_n y_{nc}^2\right) + \sum_n \frac{R_n^2}{B_n^2}\sum_c \frac{w_c^2}{s_c^4} d_c^2.$$

Finally, noticing that $\sum_n Y_{nc}^2 = N w_c^2$, and rearrange the terms, we have

$$\mathrm{Var}\left[\nabla_{\mathbf{X}}\right] = \frac{1}{6N^2}\sum_n \frac{R_n^2}{B_n^2}\left(\sum_c \frac{w_c^2}{s_c^4}\left(N w_c^2 \nabla_{y_{nc}}^2 + d_c^2\right)\right).$$

This can be computed by keeping track of $\sum_c w_c^4 / s_c^4 \nabla_{y_{nc}}^2$ for each n and d_c^2 for each d. In practice, we may not be able to record the gradient for every sample. In this case, we approximate the gradient-related terms with a constant,

$$\mathrm{Var}\left[\nabla_{\mathbf{X}}\right] \propto \sum_n \frac{R_n^2}{B_n^2}.$$

Biased Quantization As maintaining two quantized copies is too expensive, we only maintain one copy in practice. The gradient is still close to unbiased in this case. To see this,

$$\begin{aligned}\mathbb{E}\left[\hat{\nabla}_{x_{nc}}\right] &= \frac{w_c}{s_c}\left(\hat{\nabla}_{y_{nc}} - \frac{1}{N}\sum_{c'}\hat{\nabla}_{y_{nc'}} - \frac{1}{Ns_c^2}\mathbb{E}\left[(\hat{x}_{nc} - m_c)\sum_{n'}(\hat{x}_{n'c} - m_c)\hat{\nabla}_{y_{n'c}}\right]\right)\\ &= \frac{w_c}{s_c}\left(\hat{\nabla}_{y_{nc}} - \frac{1}{N}\sum_{c'}\hat{\nabla}_{y_{nc'}} - \frac{1}{Ns_c^2}\left((x_{nc} - m_c)\sum_{n'}(x_{n'c} - m_c)\hat{\nabla}_{y_{n'c}}\right) + \mathrm{Var}\left[\hat{x}_{nc}\right]\hat{\nabla}_{y_{nc}}\right)\end{aligned}$$

As N is huge, the additional term $\mathrm{Var}\left[\hat{x}_{nc}\right]\hat{\nabla}_{y_{nc}}$ is negligible comparing to other terms.

B.3. Activation Layers

Activation layers take the form

$$y_i = f(x_i), \quad \nabla_{x_i} = f'(x_i)\nabla_{y_i},$$

where $f(\cdot)$ is an activation function. We can simply store a quantized version of $\{f'(x_i)\}$ as the context for unbiased gradient.

ReLU layers are particularly simple, which have $f'(x_i) = \mathbb{I}(x_i > 0)$. Therefore, ReLU layers only take a single bit per dimension to store, without any approximation.

B.4. Pooling Layers

Pooling layers are computed without any approximation. We don't need to quantize their context because their context only takes little memory. We discuss how many bits are required for their context below.

Average pooling layers An average pooling layer can be seen as a special case of convolution layer with a constant kernel. Because the kernel is a constant, we do not need to compute the gradient of kernel. To compute the gradient of input, according to Eq. 19, we only need the gradient of output. Therefore, we can compute average pooling pooling layers exactly without saving anything in the context.

Max Pooling Following the notation in Eq. 18. A max pooling layer takes the form

$$y_{nij} = \max_{\Delta_i} x_{n,si+d\Delta_i,j}$$

The gradient of input is

$$\nabla_{x_{nij}} = \sum_{\Delta_i} \nabla_{y_{n,(i-d\Delta_i)/s,j}} \mathbb{I}(\operatorname*{argmax}_k x_{n,i-d\Delta_i+dk,j} = \Delta_i)$$

For each output location y_{nij}, we need to store an integer value $k_{nij} = \mathrm{argmax}_{\Delta_i} x_{n,si+d\Delta_i,j}$. Note that $0 \le k_{nij} < K$, where K is the kernel size of pooling. To store k_{nij}, we need $\lceil log(K) \rceil$ bits per output location (pixel). In common neural networks, the kernel size of a max pooling layer is less than $8 \times 8 = 256$. We use 8 bits per output location in our implementation.

C. Experimental Setup and Additional Experiments

C.1. Experimental Setup

We use standard open-source model architecture and training recipes for all the tasks.

Quantization Strategy We use the ResNet50-v1.5 repo[1] for ImageNet experiments. The batch size is $32 \times 8 = 256$, and the initial learning rate is 0.256. We train 90 epochs with 4 warmup epochs. The repo further has cosine learning rate schedule and label smoothing.

Computational Overhead We use ResNet-50, ResNet-152, WideResNet-101 and DenseNet-201 from the **torchvision** package. We convert them to use ActNN's layers by our model convertor. We run five training batches on ImageNet and report the median of the training throughput (images per second).

Semantic Segmentation and Object Detection We use the open-source frameworks MMSegmentation (Contributors, 2020) and MMDetection (Chen et al., 2019) for these tasks, and follow the original training recipes. For semantic segmentation, the crop size is 512×1024. The methods are FCN (Shelhamer et al., 2017) and FPN (Kirillov et al., 2019). Backbones are chosen from HRNetV2W48 (HRNet) (Wang et al., 2020), ResNet-50 dilation 8 (Dilation8), and original ResNet-50 (FPN). For object detection, the input size is 800 pixels for the short edge. The model is RetinaNet (Lin et al., 2017) with FPN as the backbone.

C.2. Variance Profile

We visualize all the terms in Thm. 3 in Fig. C.2 for a ResNet-50 trained on ImageNet at the 50-th epoch. The quantization strategy is 2-bit per-group quantization (ActNN L2 in Tab. 2). In the figure, each row is a stochasticity and each column is a parameter. For example, the entry at the m-th row and the l-th column is the term $\mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l\sim m)}\left(\hat{\nabla}_{\mathbf{H}^{(m)}}, \hat{\mathbf{C}}^{(m)}\right) \,\middle|\, \hat{\nabla}_{\mathbf{H}^{(m)}}\right]\right]$, i.e., the impact of the m-th layer's quantized activation to the gradient variance of the l-th layer's parameter. The last row is the minibatch sampling variance $\mathrm{Var}\left[\nabla_{\mathbf{\Theta}^{(l)}}\right]$. From the figure we can observe that

1. Minibatch sampling variance is much higher than the quantization variance. Therefore, it is possible to train with compressed activations, without impacting the final accuracy;
2. The quantization variance for each layer is dominantly impacted by compressing the activation at the same layer. Therefore, our strategy in Sec. 4.2, which omits all the distant terms, approximates the exact variance well.

C.3. CIFAR-10 Results

In Fig. 9, we additional present results on CIFAR-10. The conclusion remains the same with the CIFAR-100 and ImageNet experiments in the main text. BLPA converges only at 4 bits, while ActNN converges at 2 bits.

[1] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/Classification/ConvNets/resnet50v1.5

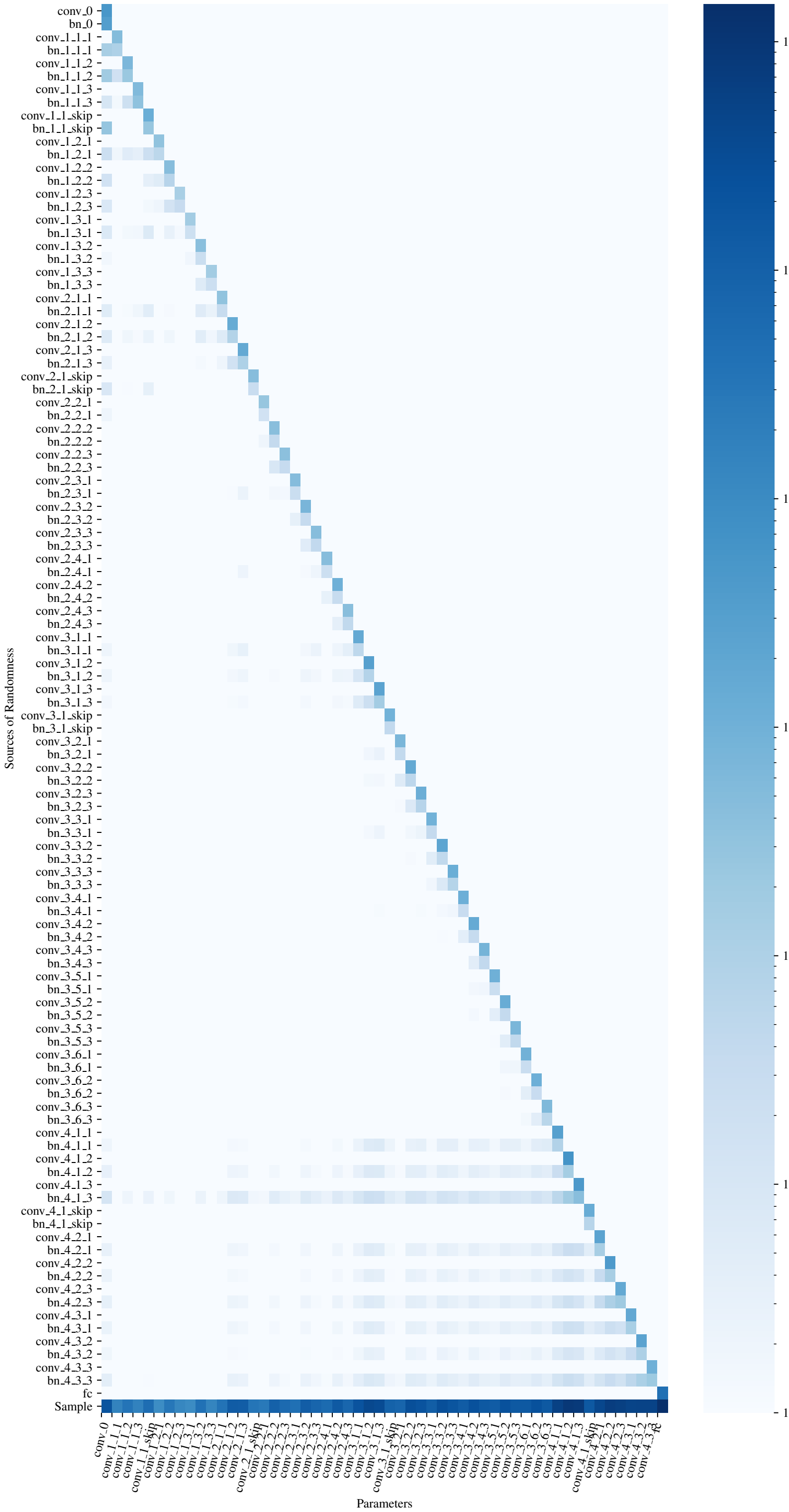
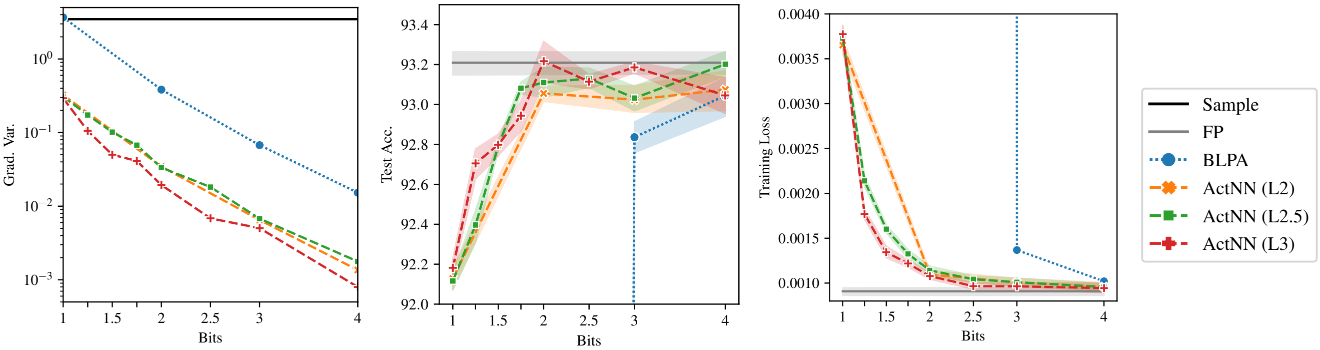


Figure 8. A decomposition of variance according to Thm. 3. Each row is a source of randomness, either quantization or minibatch sampling (last row). Each column is a parameter gradient, which we would like measure variance of. Each entry is the impact of one source of randomness to one layer's parameter gradient.



(a) Gradient variance on CIFAR-10

(b) Testing accuracy on CIFAR-10

(c) Testing loss on CIFAR-10

Figure 9. Ablation study on the quantization strategy on CIFAR-10. BLPA diverges with 1 and 2 bits. The gradient variance is calculated at the 10th epoch.