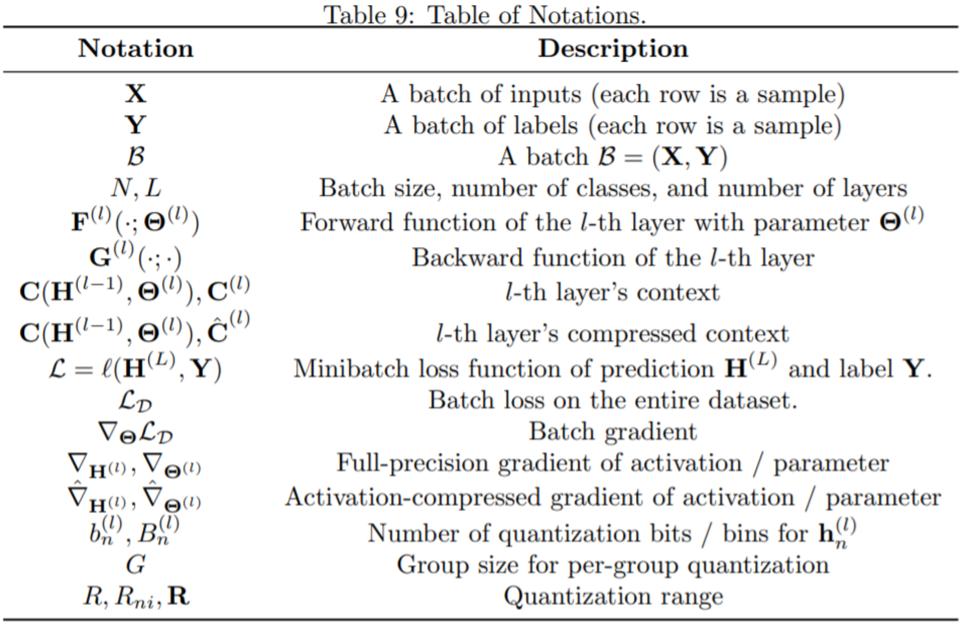
定理证明

准备工作:

注释:

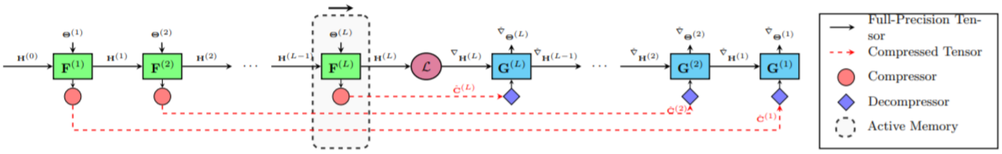
Table 9: Table of Notations.



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

定义:

1. 传输过程



根据参数的进出我们可以得到如下几个定义:

$$H^{(l)} = F^{(l)}(H^{(l-1)}; \theta^{(l)})$$

这里$H^{(l)}$的维度为$N \times D^{(l)}$，N 是 batch size，$D^{(l)}$是特征的数量

在反向传播中，我们可以把反向传播需要计算的梯度定义为:

$$\nabla_{H^{(l-1)}}, \nabla_{\theta^{(l)}} \ = \ G^{(l)}(\nabla_{H^{(l)}}, C(H^{(l-1)}, \theta^{(l)}))$$

这里 C(~)是反向传播中需要被储存的信息。l 层的$G^{(l)}$函数使用上层输出的$\nabla_{H^{(l)}}$和 C(~)计算得到 l 层的梯度。初始的，假定$\nabla_{H^{(l)}}$ 和 $\nabla_{\theta^{(l)}}$是全精度的（full precision）梯度。并且考虑一个比较简单的情形——线性层：$H^{(l)} = H^{(l-1)}\theta^{(l)}$，我们可以通过 chain rule 和梯度方程的相关计算公式得到相应的梯度:

$$\nabla_{H^{(l-1)}} = \nabla_{H^{(l)}} \theta^{(l)^T}, \nabla_{\theta^{(l)}} = H^{(l-1)^T} \nabla_{H^{(l)}}$$

这里证明不做过多赘述，写于 appendix 中

并且在这一线性情况下，$C\left(H^{(l-1)}, \theta^{(l)}\right) = (H^{(l-1)}, \theta^{(l)})$

2. ActNN 准备工作:

定义，这里我们在相关的标识符上加一个 hat 表示这是关于 ActNN 的。

在反向传播中用到储存的信息不再是$C(H^{(l-1)}, \theta^{(l)})$，而是$\hat{C}(H^{(l-1)}, \theta^{(l)})$
并且，梯度计算为$\hat{\nabla}_{H^{(l-1)}}, \hat{\nabla}_{\theta^{(l)}} = G^{(l)}(\hat{\nabla}_{H^{(l)}}, \hat{C}(H^{(l-1)}, \theta^{(l)}))$，注意到在最后一层向前传播（L层）并转为向后传播时候并没有受到存储的激活压缩信息而非全精度信息的影响，因此此时$\nabla_{H^{(L)}} = \hat{\nabla}_{H^{(L)}}$。

进一步的，如果我们把特征矩阵梯度进行细化，通过 chain rule 对每一个元素都进行计算可以得到，$\sum$对所有的 kl 偏导求和没关系，不相关没用到的都是 0：

$$\nabla_{H_{ij}^{(l-1)}} = \sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}} \nabla_{H_{kl}^{(l)}}, \nabla_{\theta_i^{(l)}} = \sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial \theta_i^{(l-1)}} \nabla_{H_{kl}^{(l)}}$$

每次迭代都会产生一个系数θ，记作$\theta_t = \{\theta^{(l)}\}_{l=1}^{L}$，作为参数的扁平向量，相应的，

$\nabla_{\theta_t} = \left\{\nabla_{\theta_t^{(l)}}\right\}_{l=1}^{L}, \hat{\nabla}_{\theta_t} = \left\{\hat{\nabla}_{\theta_t^{(l)}}\right\}_{l=1}^{L}$分别对应全精度和激活压缩的梯度。
$L_D(\theta)$是整个数据集上的小样本批次损失（batch loss），同时$\nabla_\theta, \hat{\nabla}_\theta$都是批次梯度$\hat{\nabla}_\theta L_D(\theta)$的随机估计量，因为$\nabla_\theta$的随机性进来自于 batch size，是小批量的采样，在这里我们假设这是无偏的，所以$E[\nabla_\theta] = \nabla_\theta L_D(\theta)$。而$\hat{\nabla}_\theta$的随机性还来自于 C(~)的随机量化，如果这也是无偏的，那么第一步就完成了（第二步是验证 AC 方差和全精度方差是否相近，进而判断是 ActNN 是否可以替代）。

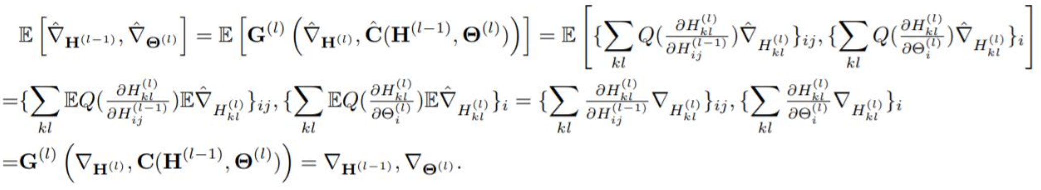
3. T1 定理描述（无偏梯度）：$\hat{C}$存在随机量化策略使得$E\left[\hat{\nabla}_\theta\right] = \nabla_\theta L_D(\theta)$
证明：
先有引理：Lemma1：如果$E\left[\hat{\nabla}_{H^{(l)}}\right] = E\left[\nabla_{H^{(l)}}\right]$，那么存在$\hat{C}^{(l)}$，使得$E\left[\hat{\nabla}_{H^{(l-1)}}\right] = E\left[\nabla_{H^{(l-1)}}\right]$并且$E\left[\hat{\nabla}_{\theta^{(l)}}\right] = E\left[\nabla_{\theta^{(l)}}\right]$
证：先前提到由链式法则我们有

$$\nabla_{H_{ij}^{(l-1)}} = \sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}} \nabla_{H_{kl}^{(l)}}, \nabla_{\theta_i^{(l)}} = \sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial \theta_i^{(l-1)}} \nabla_{H_{kl}^{(l)}}$$

因此，我们有$G^{(l)}\left(\nabla_{H^{(l)}}, C\left(H^{(l-1)}, \theta^{(l)}\right)\right) = \left\{\sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}} \nabla_{H_{kl}^{(l)}}\right\}_{ij}, \left\{\sum_{kl} \frac{\partial H_{kl}^{(l)}}{\partial \theta_i^{(l-1)}} \nabla_{H_{kl}^{(l)}}\right\}_{i}$，

$C\left(H^{(l-1)}, \theta^{(l)}\right) = \left\{\frac{\partial H_{kl}^{(l)}}{\partial H_{ij}^{(l-1)}}, \frac{\partial H_{kl}^{(l)}}{\partial \theta_i^{(l)}}\right\}_{ijkl}$

令$\hat{C}\left(H^{(l-1)}, \theta^{(l)}\right) = Q(C\left(H^{(l-1)}, \theta^{(l)}\right))$，$Q(\cdot)$是一个无偏算子，s.t.对于任意 x 都有$E[Q(x)] = x$
可以得到



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

然后我们回到定理 1 中，因为先前说到$\nabla_{H^{(L)}} = \hat{\nabla}_{H^{(L)}}$，所以$E[\nabla_{H^{(L)}}] = E[\hat{\nabla}_{H^{(L)}}]$，然后我们根据引理 1 和数学归纳法，可以得到对任意的$l \in \{1, \ldots, L\}$，均有$E\left[\hat{\nabla}_{\theta^{(l)}}\right] = E\left[\nabla_{\theta^{(l)}}\right]$，所以

$E\left[\,\hat{\nabla}_\theta\right] = E[\nabla_\theta]$

又因为根据假设我们有$E[\,\nabla_\theta] = \nabla_\theta L_D(\theta)$，所以$E\left[\,\hat{\nabla}_\theta\right] = \nabla_\theta L_D(\theta)$

在 T2 开始前, 我们假定 SGD 迭代形如$\theta_{t+1} \leftarrow \theta_t - \alpha\hat{\nabla}_{\theta_t}\,(t \geq 1)$, 并有如下三个条件的存在:

A1：损失方程$L_D(\theta)$连续可微，且其梯度是β利普西茨连续的，即存在β使得

$$||\,\nabla L_D\left(\theta_{t_1}\right) - \nabla L_D\left(\theta_{t_2}\right)|| \leq \beta||\,\theta_{t_1} - \theta_{t_2}||$$

A2：$L_D(\theta)$有下界，定为L_{inf}

A3 有界性：存在$\sigma^2 > 0$，使得对任意θ，$Var\left[\hat{\nabla}_\theta\right] \leq \sigma^2$，向量的方差为$\mathrm{Var}\,[\mathrm{x}] \coloneqq \mathrm{E}||\mathrm{x}||^2 - ||\mathrm{E}[\mathrm{x}]||^2$

4. T2 定理描述（收敛性）：如果满足 A1-A3，并且$0 < \alpha \leq \frac{1}{\beta}$，如果 t 从{1,⋯,T}中取得（T 为最大迭代次数）且取任意值的概率相同，那么有：

$$E||\,\nabla L_D(\theta_t)\,||^2 \leq \frac{2\left(L(\theta_1) - L_{inf}\right)}{\alpha T} + \alpha\beta\sigma^2$$

可以直观的从中看出，随着最大迭代次数 T 的增加，+号前的部分会趋于 0，因此这个梯度会收敛于一个值，由方差控制。

证明：

在论文 *Optimization methods for large-scale machine learning* 中，有一个结论

$$L(\theta_{t+1}) - L(\theta_t) \leq \nabla L(\theta_t)^T(\theta_{t+1} - \theta_t) + \frac{1}{2}\beta||\,\theta_{t+1} - \theta_t\,||^2$$

我们将 SGD 迭代插进去后可以得到，

$$L(\theta_{t+1}) - L(\theta_t) \leq -\alpha\nabla L(\theta_t)^T\hat{\nabla}_{\theta_t} + \frac{1}{2}\alpha^2\beta||\,\hat{\nabla}_{\theta_t}\,||^2$$

使用 A3 并在第 t+1 次迭代中取期望可以得到（所有 t 次迭代相关的参数与 t+1 相互独立）

$$E[L(\theta_{t+1})|t+1] - L(\theta_t) \leq -\alpha||\,\nabla L(\theta_t)\,||^2 + \frac{1}{2}\alpha^2\beta(Var\left[\hat{\nabla}_{\theta_t}\middle|t+1\right] + ||\,E\left[\hat{\nabla}_{\theta_t}\middle|t+1\right]||^2)$$

$$= -\mathrm{\alpha}(1 - \frac{1}{2}\mathrm{\alpha\beta})||\,\nabla L(\theta_t)\,||^2 + \frac{1}{2}\alpha^2\beta\sigma^2$$

$$\leq -\frac{1}{2}\alpha||\,\nabla L(\theta_t)\,||^2 + \frac{1}{2}\alpha^2\beta\sigma^2$$

当我们对上面的等式再取一次期望时，条件期望的期望就变成了无条件期望。

得到：$E[L(\theta_{t+1})] - E[L(\theta_t)] \leq -\frac{1}{2}\alpha E||\,\nabla L(\theta_t)\,||^2 + \frac{1}{2}\alpha^2\beta\sigma^2$

当我们把上述不等式从 t=1 累加到 t=T 时候（经过 T 次迭代）就会得到：

$$L_{inf} - L(\theta_1) \leq E[L(\theta_T)] - E[L(\theta_1)] \leq -\frac{1}{2}\alpha\sum_{t=1}^{T}E||\,\nabla L(\theta_t)\,||^2 + \frac{1}{2}T\alpha^2\beta\sigma^2$$

对于等式的最左端和最右端经过移项并同除以αT处理即可得到 T2

$$E||\,\nabla L_D(\theta_t)\,||^2 \leq \frac{2\left(L(\theta_1) - L_{inf}\right)}{\alpha T} + \alpha\beta\sigma^2$$

T2 证毕。

5. T3 准备：让$G_H(\cdot)$和$\mathrm{G}_\theta(\cdot)$作为$G(\cdot)$的一部分，分别对应∇_H和∇_θ

定义$G_\theta^{(l\sim m)}\left(\hat{\nabla}_{H^{(m)}}, \hat{C}^{(m)}\right) = G_\theta^{(l)}(G_H^{(l+1)}\left(\cdots G_H^{(m)}\left(\hat{\nabla}_{H^{(m)}}, \hat{C}^{(m)}\right)\cdots, C^{(l+1)}\right), C^{(l)})$, 这里计算的是从

$\hat{\nabla}_{H^{(m)}}$处计算得到的$\hat{\nabla}_{\theta^{(l)}}$，这里仅仅在 m 层采用激活压缩精度，其他层均采用全精度。

Proposition 1：

$$\mathrm{Var}\,[\mathrm{X}] \;=\; \mathrm{E}\,[\mathrm{Var}\,[\mathrm{X} \mid \mathrm{Y}\,]] \;+\; \mathrm{Var}\,[\mathrm{E}\,[\mathrm{X} \mid \mathrm{Y}\,]]\,,$$

$$G_\theta^{(l\sim m)}\left(\hat{\nabla}_{H^{(m)}}, \hat{C}^{(m)}\right) = G_\theta^{(l)}\left(G_H^{(l+1)}\left(\cdots G_H^{(m)}\left(\hat{\nabla}_{H^{(m)}}, \hat{C}^{(m)}\right)\cdots, C^{(l+1)}\right), C^{(l)}\right), \qquad (*)$$

$$G_\theta^{(l\sim m)}\left(\hat{\nabla}_{H^{(m)}}\right) = G_\theta^{(l)}\left(G_H^{(l+1)}\left(\cdots G_H^{(m)}\left(\hat{\nabla}_{H^{(m)}}, C^{(m)}\right)\cdots, C^{(l+1)}\right), C^{(l)}\right), \qquad (**)$$

(*) m 层 C(~)压缩了，而(**)没有

T3 定理描述（梯度方差）：$Var\left[\hat{\nabla}_{\theta^{(l)}}\right] = Var\left[\nabla_{\theta^{(l)}}\right] + \sum_{m=l}^{L} E\left[Var\left[G_\theta^{(l\sim m)}\left(\hat{\nabla}_{H^{(m)}}, \hat{C}^{(m)}\right)\middle|\hat{\nabla}_{H^{(m)}}\right]\right]$

这里我们可以看到，采用激活压缩的梯度方差比全精度梯度方差多了一部分，如果多出来的那部分相较于全精度梯度方差非常小，那么就可以说明降低数值精度是可行的。并且多出来的部分方差是明确的，这就可以通过继续设计量化压缩策略来最小化这个值。

证明：

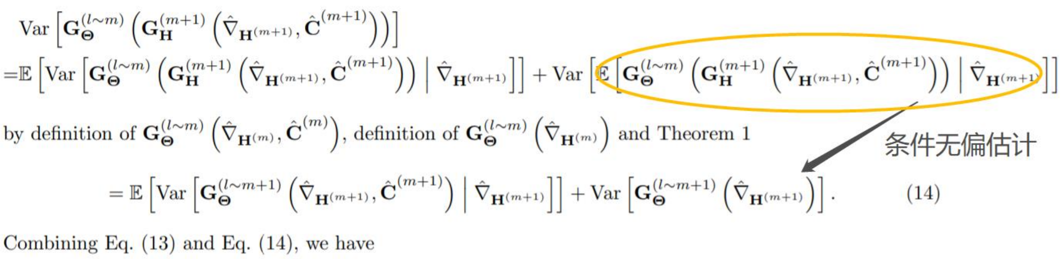
首先，因为$\nabla_{H^{(L)}} = \hat{\nabla}_{H^{(L)}}$，按照定义有$Var\left[G_\theta^{(l\sim L)}\left(\hat{\nabla}_{H^{(L)}}\right)\right] = Var\left[G_\theta^{(l\sim L)}\left(\nabla_{H^{(L)}}\right)\right] = Var\left[\nabla_{\theta^{(l)}}\right]$ (13)

其次，对于所有的 m<L，由$\hat{\nabla}_{H^{(m)}}$定义$\hat{\nabla}_{H^{(m)}} = G_H^{(m+1)}\left(\hat{\nabla}_{H^{(m+1)}}, \hat{C}^{(m+1)}\right)$

得到，

$Var\left[G_\theta^{(l\sim m)}\left(\hat{\nabla}_{H^{(m)}}\right)\right] = Var\left[G_\theta^{(l\sim m)}\left(G_H^{(m+1)}\left(\hat{\nabla}_{H^{(m+1)}}, \hat{C}^{(m+1)}\right)\right)\right] = Var\left[G_\theta^{(l\sim m+1)}\left(\hat{\nabla}_{H^{(m+1)}}, \hat{C}^{(m+1)}\right)\right]$,

因为 Proposition 1，因为是条件无偏估计，所以期望可以直接去掉($E\left[\hat{C}\right] = C$, 按照定义就省略了)，所以



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

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

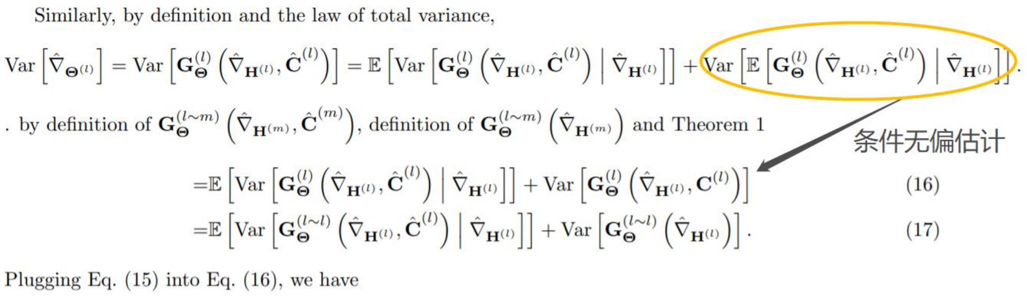
条件无偏估计

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

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

$$Var\left[G_\theta^{(l\sim m)}\left(\hat{\nabla}_{H^{(m)}}\right)\right] = Var\left[\nabla_{\theta^{(l)}}\right] + \sum_{j=m+1}^{L} E[Var[G_\theta^{(l\sim j)}(\hat{\nabla}_{H^{(j)}}, \hat{C}^{(j)})|\hat{\nabla}_{H^{(j)}}\,]] \qquad (15)$$

注：对于等式(14)迭代至 L 求和即可得到等式(15)。



Similarly, by definition and the law of total variance,

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

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

条件无偏估计

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

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

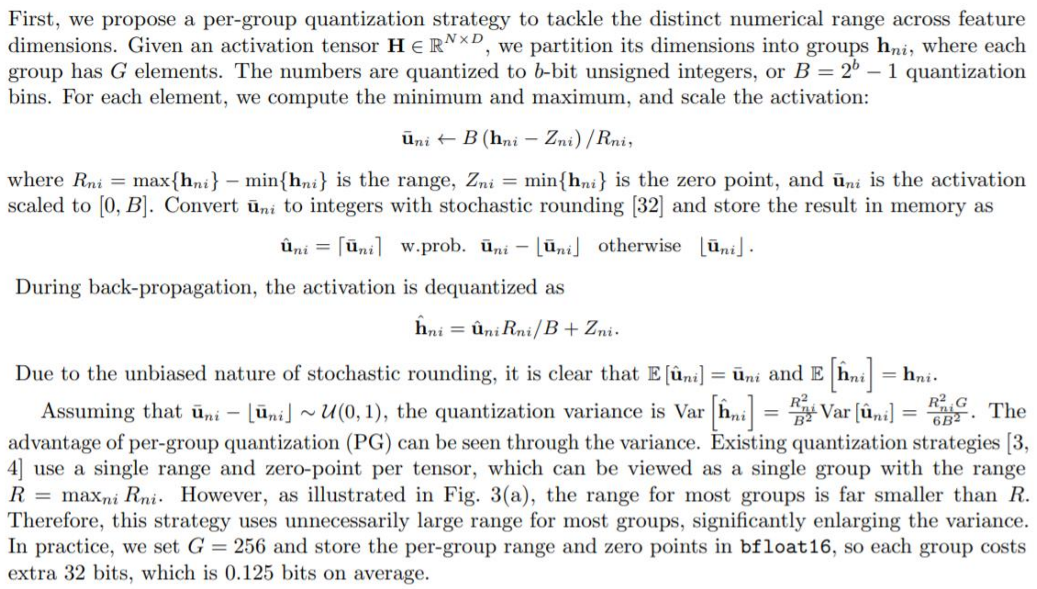
$$Var\left[\hat{\nabla}_{\theta^{(l)}}\right] = E\left[Var\left[G_{\theta}^{(l\sim l)}\left(\hat{\nabla}_{H^{(l)}}, \hat{C}^{(l)}\right)\middle|\hat{\nabla}_{H^{(l)}}\right]\right] + Var\left[\nabla_{\theta^{(l)}}\right] + \sum_{j=l+1}^{L} E\left[Var\left[G_{\theta}^{(l\sim j)}\left(\hat{\nabla}_{H^{(j)}}, \hat{C}^{(j)}\right)\middle|\hat{\nabla}_{H^{(j)}}\right]\right]$$

$$= Var\left[\nabla_{\theta^{(l)}}\right] + \sum_{m=l}^{L} E\left[Var\left[G_{\theta}^{(l\sim m)}\left(\widehat{\nabla}_{H^{(m)}}, \widehat{C}^{(m)}\right)\middle|\widehat{\nabla}_{H^{(m)}}\right]\right]$$

T3 证毕。

6. 压缩策略

分组量化：



First, we propose a per-group quantization strategy to tackle the distinct numerical range across feature dimensions. Given an activation tensor $\mathbf{H} \in \mathbb{R}^{N\times D}$, we partition its dimensions into groups $\mathbf{h}_{ni}$, where each group has G elements. The numbers are quantized to b-bit unsigned integers, or $B = 2^b - 1$ quantization bins. For each element, we compute the minimum and maximum, and scale the activation:

$$\bar{\mathbf{u}}_{ni} \leftarrow B\left(\mathbf{h}_{ni} - Z_{ni}\right)/R_{ni},$$

where $R_{ni} = \max\{\mathbf{h}_{ni}\} - \min\{\mathbf{h}_{ni}\}$ is the range, $Z_{ni} = \min\{\mathbf{h}_{ni}\}$ is the zero point, and $\bar{\mathbf{u}}_{ni}$ is the activation scaled to $[0, B]$. Convert $\bar{\mathbf{u}}_{ni}$ to integers with stochastic rounding [32] and store the result in memory as

$$\hat{\mathbf{u}}_{ni} = \lceil\bar{\mathbf{u}}_{ni}\rceil \quad \text{w.prob.} \quad \bar{\mathbf{u}}_{ni} - \lfloor\bar{\mathbf{u}}_{ni}\rfloor \quad \text{otherwise} \quad \lfloor\bar{\mathbf{u}}_{ni}\rfloor.$$

During back-propagation, the activation is dequantized as

$$\hat{\mathbf{h}}_{ni} = \hat{\mathbf{u}}_{ni}R_{ni}/B + Z_{ni}.$$

Due to the unbiased nature of stochastic rounding, it is clear that $\mathbb{E}\left[\hat{\mathbf{u}}_{ni}\right] = \bar{\mathbf{u}}_{ni}$ and $\mathbb{E}\left[\hat{\mathbf{h}}_{ni}\right] = \mathbf{h}_{ni}$.

Assuming that $\bar{\mathbf{u}}_{ni} - \lfloor\bar{\mathbf{u}}_{ni}\rfloor \sim \mathcal{U}(0,1)$, the quantization variance is $\mathrm{Var}\left[\hat{\mathbf{h}}_{ni}\right] = \frac{R_{ni}^2}{B^2}\mathrm{Var}\left[\hat{\mathbf{u}}_{ni}\right] = \frac{R_{ni}^2 G}{6B^2}$. The advantage of per-group quantization (PG) can be seen through the variance. Existing quantization strategies [3, 4] use a single range and zero-point per tensor, which can be viewed as a single group with the range $R = \max_{ni} R_{ni}$. However, as illustrated in Fig. 3(a), the range for most groups is far smaller than R. Therefore, this strategy uses unnecessarily large range for most groups, significantly enlarging the variance. In practice, we set $G = 256$ and store the per-group range and zero points in `bfloat16`, so each group costs extra 32 bits, which is 0.125 bits on average.

注：这里$Var[\hat{u}_{ni}] = \frac{1}{6}$需要证明下：

证：假设$\hat{u}_{ni} \in [0,1]$，这里$\hat{u}_{ni}$具体的取值无关紧要，因为算的是方差

因为有 w.prob.后面那个概率服从(0,1)的均匀分布，令这个值为 p

翻译一下就是$\hat{u}_{ni}$以 p 的概率取 1，以 $1-\mathrm{p}$ 的概率取 0，并且 p 服从 U(0,1)均匀分布。

按照方差的定义：$Var[\hat{u}_{ni}] = E[\hat{u}_{ni}^2] - (E[\hat{u}_{ni}])^2$

$\hat{u}_{ni}^2$是一个新的变量，同样以 p 的概率取 1，以 1-p 的概率取 0，p 服从(0,1)均匀分布（离散化的情况结论很明显）。

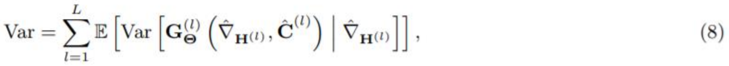
$$Var[\hat{u}_{ni}] = E[E[\hat{u}_{ni}^2] - (E[\hat{u}_{ni}])^2] = E[p - p^2] = E[p] - E[p^2] \qquad (*)$$

$$E[p^2] = \int_0^1 p^2 * 1\, dp = \frac{p^3}{3}|_0^1 = \frac{1}{3}$$

$\therefore (*)$可化为$\frac{1}{2} - \frac{1}{3} = \frac{1}{6}$

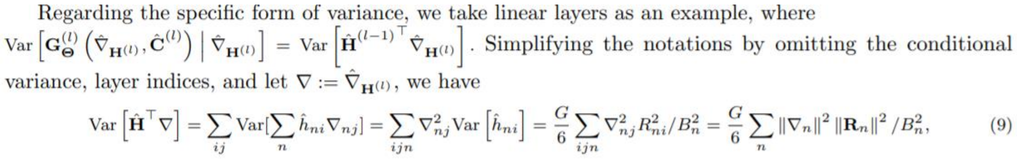
综上$Var[\hat{u}_{ni}] = \frac{1}{6}$，证毕。

目标：最小化



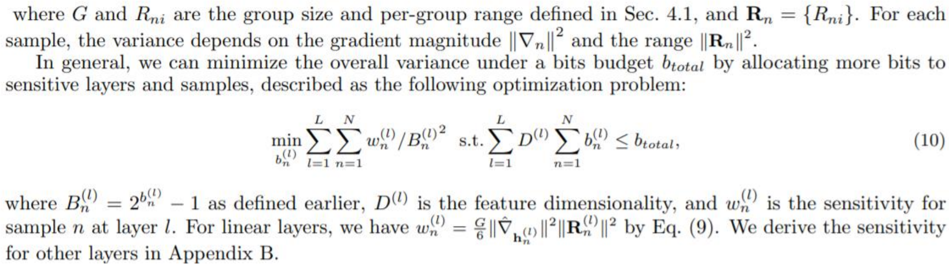
$$\mathrm{Var} = \sum_{l=1}^{L} \mathbb{E}\left[\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}}, \hat{\mathbf{C}}^{(l)}\right) \,\middle|\, \hat{\nabla}_{\mathbf{H}^{(l)}}\right]\right], \tag{8}$$

首先考虑线性层:



Regarding the specific form of variance, we take linear layers as an example, where $\mathrm{Var}\left[\mathbf{G}_{\mathbf{\Theta}}^{(l)}\left(\hat{\nabla}_{\mathbf{H}^{(l)}}, \hat{\mathbf{C}}^{(l)}\right) \,\middle|\, \hat{\nabla}_{\mathbf{H}^{(l)}}\right] = \mathrm{Var}\left[\hat{\mathbf{H}}^{(l-1)^\top} \hat{\nabla}_{\mathbf{H}^{(l)}}\right]$. Simplifying the notations by omitting the conditional variance, layer indices, and let $\nabla := \hat{\nabla}_{\mathbf{H}^{(l)}}$, we have

$$\mathrm{Var}\left[\hat{\mathbf{H}}^\top \nabla\right] = \sum_{ij} \mathrm{Var}[\sum_{n} \hat{h}_{ni} \nabla_{nj}] = \sum_{ijn} \nabla_{nj}^2 \mathrm{Var}\left[\hat{h}_{ni}\right] = \frac{G}{6} \sum_{ijn} \nabla_{nj}^2 R_{ni}^2 / B_n^2 = \frac{G}{6} \sum_{n} \|\nabla_n\|^2 \|\mathbf{R}_n\|^2 / B_n^2, \tag{9}$$



where G and R_{ni} are the group size and per-group range defined in Sec. 4.1, and $\mathbf{R}_n = \{R_{ni}\}$. For each sample, the variance depends on the gradient magnitude $\|\nabla_n\|^2$ and the range $\|\mathbf{R}_n\|^2$.

In general, we can minimize the overall variance under a bits budget b_{total} by allocating more bits to sensitive layers and samples, described as the following optimization problem:

$$\min_{b_n^{(l)}} \sum_{l=1}^{L} \sum_{n=1}^{N} w_n^{(l)} / B_n^{(l)^2} \quad \text{s.t.} \sum_{l=1}^{L} D^{(l)} \sum_{n=1}^{N} b_n^{(l)} \leq b_{total}, \tag{10}$$

where $B_n^{(l)} = 2^{b_n^{(l)}} - 1$ as defined earlier, $D^{(l)}$ is the feature dimensionality, and $w_n^{(l)}$ is the sensitivity for sample n at layer l. For linear layers, we have $w_n^{(l)} = \frac{G}{6} \|\hat{\nabla}_{\mathbf{h}_n^{(l)}}\|^2 \|\mathbf{R}_n^{(l)}\|^2$ by Eq. (9). We derive the sensitivity for other layers in Appendix B.

SOME LAYERS 的偏差和方差（特殊性）:

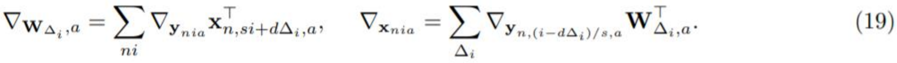
就是重复定理 1-3，然后看期望相等，方差较小，灵敏度

卷积层:

定义:

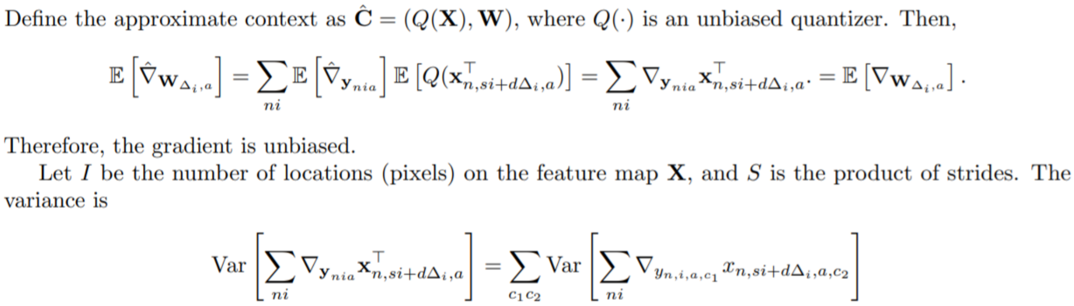
$$\mathbf{y}_{nia} = \sum_{\Delta_i} \mathbf{W}_{\Delta_i, a} \mathbf{x}_{n, si + d\Delta_i, a}. \tag{18}$$

梯度表示:



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

计算期望和方差:



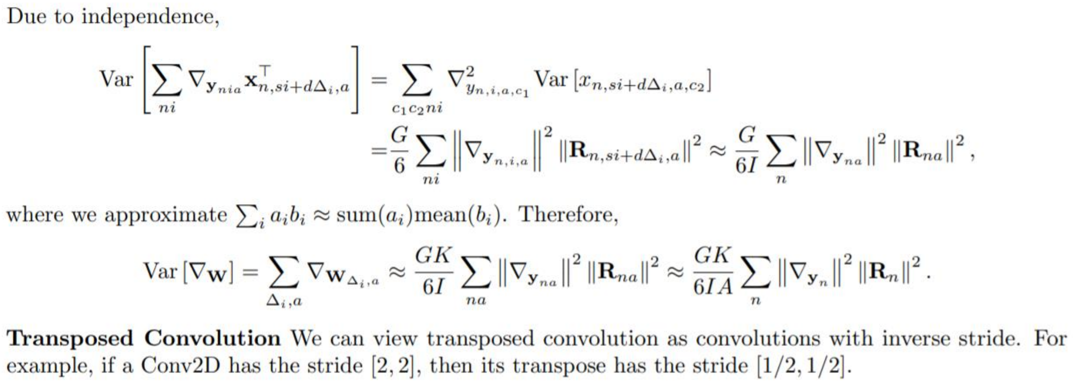
Define the approximate context as $\hat{\mathbf{C}} = (Q(\mathbf{X}), \mathbf{W})$, where $Q(\cdot)$ is an unbiased quantizer. Then,

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

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

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



Due to independence,

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

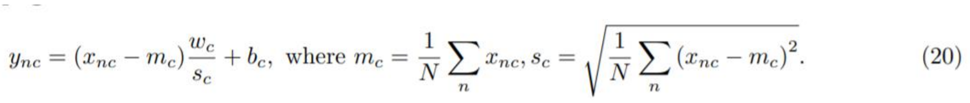
where we approximate $\sum_i a_ib_i\approx\operatorname{sum}(a_i)\operatorname{mean}(b_i)$. Therefore,

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

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 $[1/2,1/2]$.

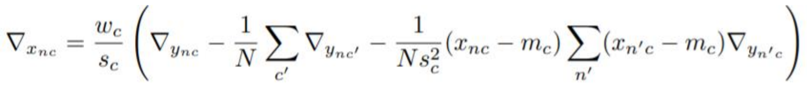
归一化层，标准化层

向前传播公式（BatchNorm2d 函数）：



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

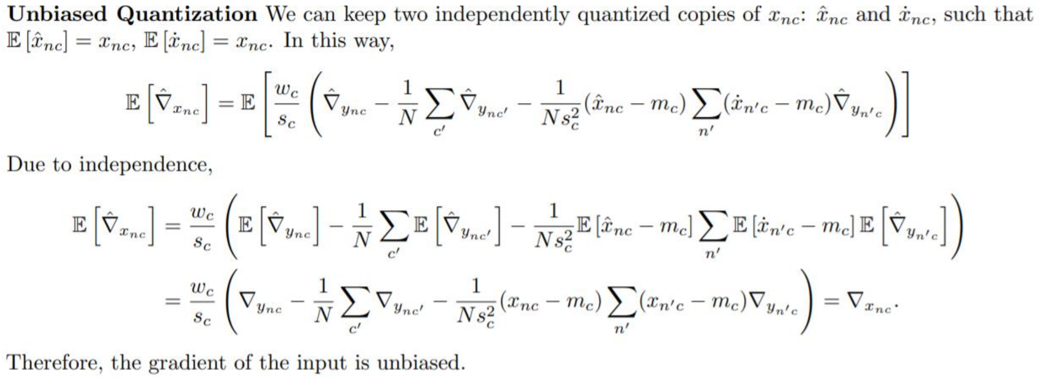
反向传播公式：



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

对应之前讲到的 C(~)这里是(X,m,s,w)，这里可以只讨论 X，别的量级都是非常小的。

无偏性：



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,

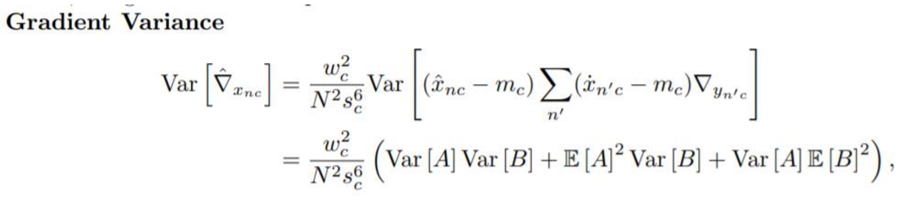
$$\mathbb{E}\left[\hat{\nabla}_{x_{nc}}\right]=\mathbb{E}\left[\frac{w_c}{s_c}\left(\hat{\nabla}_{y_{nc}}-\frac{1}{N}\sum_{c'}\hat{\nabla}_{y_{nc'}}-\frac{1}{Ns_c^2}(\hat{x}_{nc}-m_c)\sum_{n'}(\dot{x}_{n'c}-m_c)\hat{\nabla}_{y_{n'c}}\right)\right]$$

Due to independence,

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

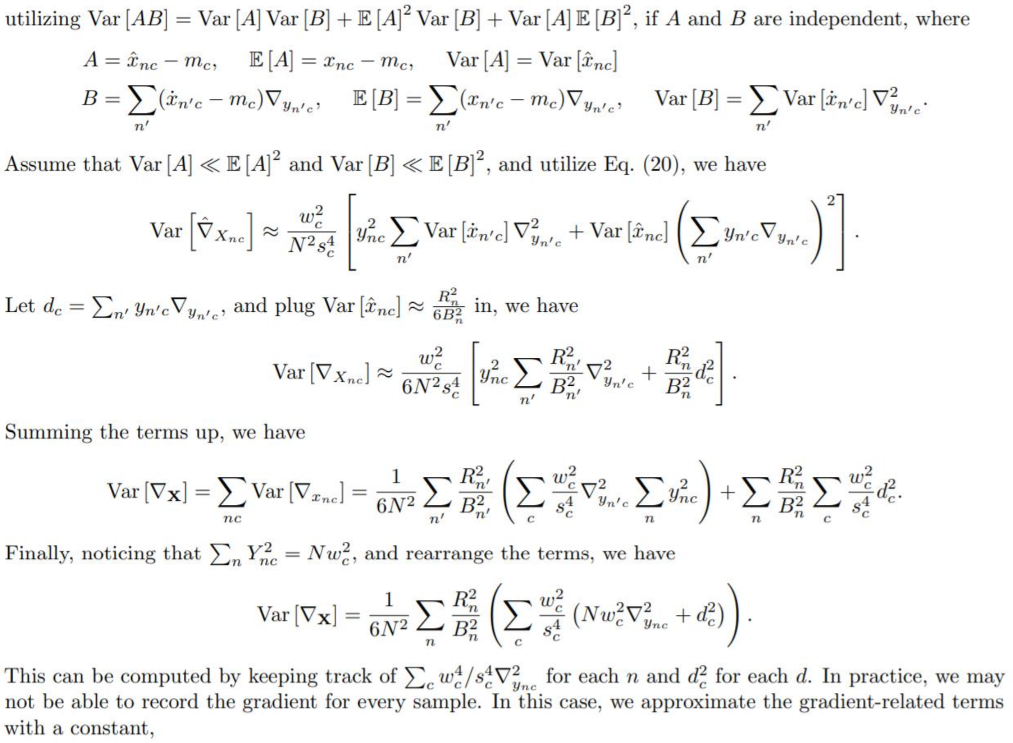
Therefore, the gradient of the input is unbiased.

梯度方差：



Gradient Variance

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



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

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

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

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

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

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

Summing the terms up, we have

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

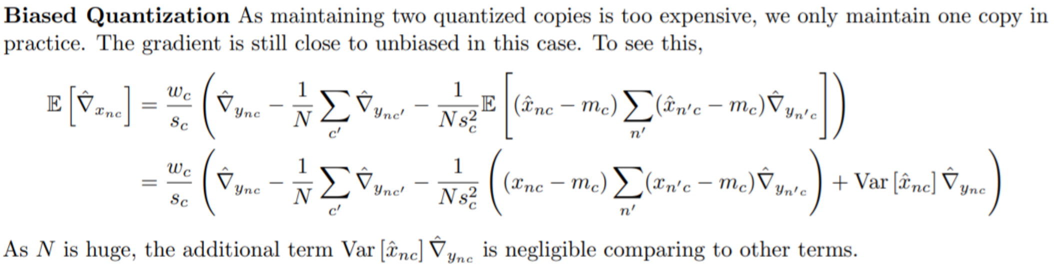
Finally, noticing that $\sum_n Y_{nc}^2 = N w_c^2$, and rearrange the terms, we have

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

This can be computed by keeping track of $\sum_c w_c^4/s_c^4 \nabla^2_{y_{nc}}$ 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,

$$Var[\nabla_X] \propto \sum_n \frac{R_n^2}{B_n^2}$$

偏差（但是可以忽略）:



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,

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

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

当 N 很大时候方差是可以忽视的。

激活层

池化层

Appendix：

1. 已知$H^{(l)} = H^{(l-1)}\theta^{(l)}$，那么$\nabla_{H^{(l-1)}} = \nabla_{H^{(l)}}\theta^{(l)^T}$,$\nabla_{\theta^{(l)}} = H^{(l-1)^T}\nabla_{H^{(l)}}$

证明：这里$H^{(l)}$，$H^{(l-1)}$，$\theta^{(l)}$其实为三个矩阵，为了方便表示，简写为 C, A, B。
那么有 C=AB
假设存在一个标量函数 f，使得$y = f(AB)$，其中 A 维度为$n \times m$，B 维度为$m \times k$

那么梯度的问题转化为要求$\frac{\partial y}{\partial A}$ 和 $\frac{\partial y}{\partial B}$

∵C=AB，那么有$y=f(C)$，同时 y 对于 C 中每个元素$C_{i,j}$的偏导数为$\frac{\partial y}{\partial C_{i,j}}$

同时，根据多元函数的链式法则有$\frac{\partial y}{\partial A_{p,q}}=\sum_{i,j}\frac{\partial y}{\partial C_{i,j}}*\frac{\partial C_{i,j}}{\partial A_{p,q}}$

又∵$C_{i,j}=\sum_h A_{i,h}B_{h,j}$

∴$\frac{\partial C_{i,j}}{\partial A_{p,q}}=\begin{cases}B_{q,j} & i=p\\0 & i\neq p\end{cases}\quad(*)$

将(*)带入$\frac{\partial y}{\partial A_{p,q}}$中有，$\frac{\partial y}{\partial A_{p,q}}=\sum_{i,j}\frac{\partial y}{\partial C_{i,j}}*\frac{\partial C_{i,j}}{\partial A_{p,q}}=\sum_j\frac{\partial y}{\partial C_{p,j}}B_{q,j}=\sum_j\frac{\partial y}{\partial C_{p,j}}B^T_{j,q}$

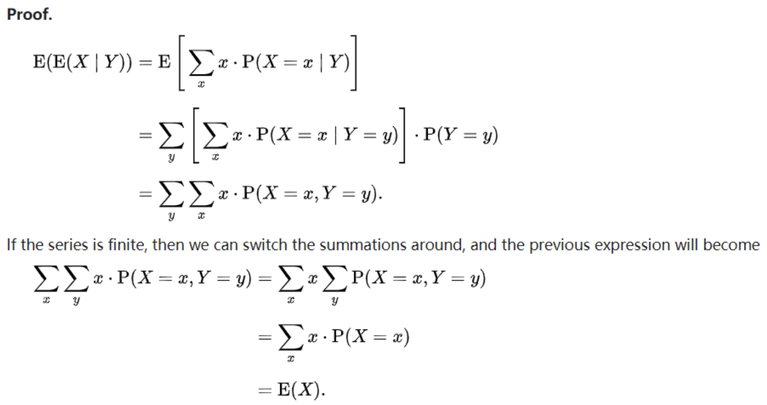
∴$\frac{\partial y}{\partial A_{p,q}}=\frac{\partial y}{\partial C}B^T$即有$\nabla A=\nabla C*B^T$

同理可得$\nabla B=A^T\nabla C$

将$H^{(l)}$，$H^{(l-1)}$，$\theta^{(l)}$回代即可得到$\nabla_{H^{(l-1)}}=\nabla_{H^{(l)}}\theta^{(l)^T},\nabla_{\theta^{(l)}}=H^{(l-1)^T}\nabla_{H^{(l)}}$

2. 条件期望的期望等于无条件的期望（总期望定理）

因为我们这里是离散化的情形，所以对离散化情况做出证明（连续函数同理，用积分）



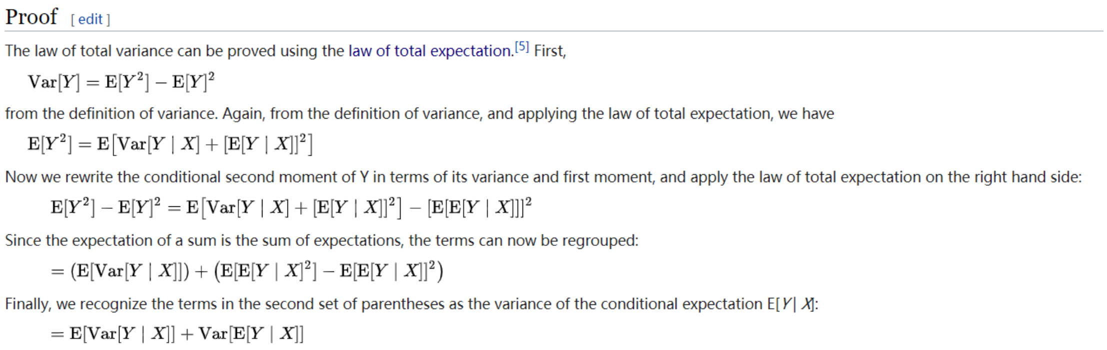
Proof.

$$\begin{aligned}\mathrm{E}(\mathrm{E}(X\mid Y))&=\mathrm{E}\left[\sum_x x\cdot\mathrm{P}(X=x\mid Y)\right]\\&=\sum_y\left[\sum_x x\cdot\mathrm{P}(X=x\mid Y=y)\right]\cdot\mathrm{P}(Y=y)\\&=\sum_y\sum_x x\cdot\mathrm{P}(X=x,Y=y).\end{aligned}$$

If the series is finite, then we can switch the summations around, and the previous expression will become

$$\begin{aligned}\sum_x\sum_y x\cdot\mathrm{P}(X=x,Y=y)&=\sum_x x\sum_y\mathrm{P}(X=x,Y=y)\\&=\sum_x x\cdot\mathrm{P}(X=x)\\&=\mathrm{E}(X).\end{aligned}$$

3. 总方差定理Var [X] ＝ E [Var [X | Y]] ＋ Var [E [X | Y]]



Proof [edit]

The law of total variance can be proved using the law of total expectation.[5] First,

$$\mathrm{Var}[Y]=\mathrm{E}[Y^2]-\mathrm{E}[Y]^2$$

from the definition of variance. Again, from the definition of variance, and applying the law of total expectation, we have

$$\mathrm{E}[Y^2]=\mathrm{E}\left[\mathrm{Var}[Y\mid X]+[\mathrm{E}[Y\mid X]]^2\right]$$

Now we rewrite the conditional second moment of Y in terms of its variance and first moment, and apply the law of total expectation on the right hand side:

$$\mathrm{E}[Y^2]-\mathrm{E}[Y]^2=\mathrm{E}\left[\mathrm{Var}[Y\mid X]+[\mathrm{E}[Y\mid X]]^2\right]-[\mathrm{E}[\mathrm{E}[Y\mid X]]]^2$$

Since the expectation of a sum is the sum of expectations, the terms can now be regrouped:

$$=(\mathrm{E}[\mathrm{Var}[Y\mid X]])+\left(\mathrm{E}[\mathrm{E}[Y\mid X]^2]-\mathrm{E}[\mathrm{E}[Y\mid X]]^2\right)$$

Finally, we recognize the terms in the second set of parentheses as the variance of the conditional expectation E[*Y*| *X*]:

$$=\mathrm{E}[\mathrm{Var}[Y\mid X]]+\mathrm{Var}[\mathrm{E}[Y\mid X]]$$