# Supplementary Materials for TINYSCRIPT

## A. Proofs

In this section, we provide proofs for the theorems and some omitted deduction in our paper.

### A.1. Proofs for Section 3

**Theorem 1.** *There is a unique stationary point $\hat{\boldsymbol{s}}^*$ of $V_{\hat{n}}$ in the domain of $\hat{\boldsymbol{s}}$, and $V_{\hat{n}}$ attains minimum only at $\hat{\boldsymbol{s}}^*$.*

*Proof.* We relax the domain of $\hat{\boldsymbol{s}}$ to $\hat{s}_0 \leq \hat{s}_1 \leq ... \leq \hat{s}_n = M$, denoted by $\mathcal{D}$. Since $\mathcal{D}$ is a compact set on $\mathbb{R}^{n-1}$, $V_{\hat{n}}(\hat{\boldsymbol{s}})$ must attain a minimum on $\mathcal{D}$. Also denote $p(x) = kx^{k-1}\exp(-x^k), 0 < k <= 1$. It is easy to check that $p(x)$ is convex and strictly decreasing on $[0, \infty)$.

Now, we only need to prove two claims.

**Claim 1.** *The minimum of $V_{\hat{n}}(\hat{\boldsymbol{s}})$ cannot lies on the boundary $\partial\mathcal{D}$.*

*Proof of Claim 1.* We proof this claim by contradiction.

Notice that the points on $\partial\mathcal{D}$ must satisfy $\hat{s}_i = \hat{s}_j$ for some $i \neq j$, which will in fact results in a loss of parameters. We assume w.l.o.g. that the global minimum $\hat{\boldsymbol{s}}^* = (\hat{s}_1^*, ..., \hat{s}_{n-1}^*)$ "only" satisfies $\hat{s}_1^* = \hat{s}_2^*$ (which means that $\hat{s}_3^* \neq \hat{s}_2^*$ and $\hat{s}_1^* \neq 0$).

Now denote $\hat{\boldsymbol{s}}' = (\hat{s}_1^*, \frac{\hat{s}_1^*+\hat{s}_3^*}{2}, ..., \hat{s}_{n-1}^*)$. We will show that $V_{\hat{n}}(\hat{\boldsymbol{s}}') < V_{\hat{n}}(\hat{\boldsymbol{s}}^*)$, which is a contradiction to the definition of $\hat{\boldsymbol{s}}^*$. In fact,

$$
\begin{aligned}
V_{\hat{n}}(\hat{\boldsymbol{s}}^*) - V_{\hat{n}}(\hat{\boldsymbol{s}}') = &\int_{\hat{s}_1^*}^{\hat{s}_3^*} p(x)(\hat{s}_3^* - x)(x - \hat{s}_1^*)dx \\
&- \int_{\hat{s}_1^*}^{\frac{\hat{s}_1^*+\hat{s}_3^*}{2}} p(x)(\frac{\hat{s}_1^* + \hat{s}_3^*}{2} - x)(x - \hat{s}_1^*)dx \\
&- \int_{\frac{\hat{s}_1^*+\hat{s}_3^*}{2}}^{\hat{s}_3^*} p(x)(\hat{s}_3^* - x)(x - \frac{\hat{s}_1^* + \hat{s}_3^*}{2})dx \\
= &\int_{\hat{s}_1^*}^{\frac{\hat{s}_1^*+\hat{s}_3^*}{2}} p(x)\frac{\hat{s}_3^* - \hat{s}_1^*}{2}(x - \hat{s}_1^*)dx \\
&+ \int_{\frac{\hat{s}_1^*+\hat{s}_3^*}{2}}^{\hat{s}_3^*} p(x)(\hat{s}_3^* - x)\frac{\hat{s}_3^* - \hat{s}_1^*}{2}dx \\
&> 0.
\end{aligned}
$$

Thus, Claim 1 holds. ■

**Claim 2.** *There exists an unique first-order stationary point in $\mathcal{D}$.*

*Proof of Claim 2.* By setting the derivatives to zero, we have

$$
\int_{\hat{s}_{t-1}}^{\hat{s}_t} (x - \hat{s}_{t-1})p(x)dx = \int_{\hat{s}_t}^{\hat{s}_{t+1}} (\hat{s}_{t+1} - x)p(x)dx \quad (*)
$$
$$
\forall t = 1, 2, ..., n-1.
$$

Notice that the right hand side is strictly decreasing w.r.t. $\hat{s}_{t+1}$. Thus, when $\hat{s}_{t-1}$ and $\hat{s}_t$ are fixed, $\hat{s}_{t+1}$ can be uniquely determined. By induction, $\forall t \geq 2$, $\hat{s}_t$ can be uniquely determined by $\hat{s}_1$ (since $\hat{s}_0 = 0$). Also, it is easy to see that when $\hat{s}_1 > 0$, the sequence $\{\hat{s}_t\}$ is strictly increasing w.r.t. $t$.

Now, in order to prove Claim 2, it suffices to prove:

$$
\forall t \geq 2, \hat{s}_t \text{ is an increasing continuous function of } \hat{s}_1. \quad \text{(I)}
$$

In fact, we may ignore the constraint $\hat{s}_{n+1} = M$ for the moment and set a small enough $\hat{s}_1'$ such that $\hat{s}_{n+1}'$ (which is determined by $(*)$ and $\hat{s}_0 = 0$) satisfies $\hat{s}_{n+1}' < M$. Similarly, we can set a $\hat{s}_1''$ large enough (e.g., $\hat{s}_1'' > M$) such that $\hat{s}_{n+1}'' > M$. Note that once (I) holds, then $\hat{s}_{n+1}$ is an increasing continuous function of $\hat{s}_1$, we may apply the Intermediate Value Theorem and obtain the desired result. That is to say, there exists a unique $\hat{s}_1^* \in (\hat{s}_1', \hat{s}_1'')$ satisfying $\hat{s}_{n+1}^* = M$.

Before entering the proof of (I), we first need to do some useful transformations. Note that $(*)$ is equivalent to

$$
\begin{aligned}
\int_{\hat{s}_{t-1}}^{\hat{s}_{t+1}} xp(x)dx = \hat{s}_{t-1}&\int_{\hat{s}_{t-1}}^{\hat{s}_t} p(x)dx \\
&+ \hat{s}_{t+1}\int_{\hat{s}_t}^{\hat{s}_{t+1}} p(x)dx, \qquad (**) \\
&\forall t = 1, 2, ..., n-1
\end{aligned}
$$

Recall that $p(x) = kx^{k-1}\exp\left(-x^k\right) = \left(\exp(-x^k)\right)'$. Denote $f(x) = \exp(-x^k)$ and integrating the left hand side of $(**)$ by part, we have

$$
\begin{aligned}
&xf(x)|_{\hat{s}_{t-1}}^{\hat{s}_{t+1}} - \int_{\hat{s}_{t-1}}^{\hat{s}_{t+1}} f(x)dx = \\
&\quad \hat{s}_{t-1}f(x)|_{\hat{s}_{t-1}}^{\hat{s}_t} + \hat{s}_{t+1}f(x)|_{\hat{s}_t}^{\hat{s}_{t+1}} \qquad (***) \\
&\Leftrightarrow \frac{\int_{\hat{s}_{t-1}}^{\hat{s}_{t+1}} f(x)dx}{\hat{s}_{t+1} - \hat{s}_{t-1}} = f(\hat{s}_t) = \exp(-\hat{s}_t^k).
\end{aligned}
$$

We are now ready to prove (I). Suppose there are two sequences of $(***)$, namely $\{a_t\}, \{b_t\}$, satisfying $a_0 =$

$b_0 = 0$ and $a_1 < b_1$. By induction, it suffices to prove that $\forall t \geq 1, a_{t-1} \leq b_{t-1} \wedge a_t < b_t \Rightarrow a_{t+1} < b_{t+1}$.

Denote $F(x, y) = \frac{\int_x^y f(t)dt}{y-x}$ for $x < y$. Since $f$ is strictly decreasing, it is straightforward to see that $F(x, y)$ is strictly increasing w.r.t. $x$ and strictly decreasing w.r.t. $y$.

Now, we have

$$F(b_{t-1}, b_{t+1}) = f(b_t) < f(a_t) = F(a_{t-1}, a_{t+1}) \leq F(b_{t-1}, a_{t+1}).$$

Thus, $a_{t+1} < b_{t+1}$ and (I) is proved. Therefore, Claim 2 holds. ■

Combining Claim 1 and Claim 2 and recall the existence of global minimum on $\mathcal{D}$, it is straightforward to see that the unique first-order stationary point is also the unique global minimum. Further, by Claim 1 we know the unique global minimum lies on the interior $\mathcal{D}^\circ$. This finishes the proof of Theorem 1. □

**Claim 3.** [Proof of Equation 2] *The partial derivative for the minimum quantization variance problem has the form of*

$$\begin{aligned}\frac{\partial V_{\hat{n}}}{\partial \hat{s}_t} =& -\Gamma(1+1/k, \hat{s}_{t+1}^k) + \Gamma(1+1/k, \hat{s}_{t-1}^k) \\ &+ \hat{s}_{t+1}\exp(-\hat{s}_{t+1}^k) - \hat{s}_{t-1}\exp(-\hat{s}_{t-1}^k) \\ &- (\hat{s}_{t+1} - \hat{s}_{t-1})\exp(-\hat{s}_t^k),\end{aligned}$$

*for* $t = 1, 2, ..., n-1$

*Proof.* Let $p(x) = kx^{k-1}\exp(-x^k)$, the partial derivative is given as

$$\begin{aligned}\frac{\partial V_{\hat{n}}}{\partial \hat{s}_t} =& \int_{\hat{s}_{t-1}}^{\hat{s}_{t+1}} xp(x)dx \\ &- \left(\hat{s}_{t+1}\int_{\hat{s}_t}^{\hat{s}_{t+1}} p(x)dx + \hat{s}_{t-1}\int_{\hat{s}_{t-1}}^{\hat{s}_t} p(x)dx\right).\end{aligned} \tag{$*$}$$

By substituting $u = x^k$, we have

$$\begin{aligned}\int_{\hat{s}_{t-1}}^{\hat{s}_{t+1}} xp(x)dx &= \int_{\hat{s}_{t-1}^k}^{\hat{s}_{t+1}^k} u^{1/k}\exp(-u)du \\ &= -\Gamma(1+1/k, \hat{s}_{t+1}^k) + \Gamma(1+1/k, \hat{s}_{t-1}^k),\end{aligned}$$

where $\Gamma(\cdot, \cdot)$ is the incomplete gamma function.

Since $\int p(x)dx = -\exp(-x^k) + \textit{Constant}$, we have

$$\begin{aligned}&\hat{s}_{t+1}\int_{\hat{s}_t}^{\hat{s}_{t+1}} p(x)dx + \hat{s}_{t-1}\int_{\hat{s}_{t-1}}^{\hat{s}_t} p(x)dx \\ &= -\hat{s}_{t+1}\exp(-\hat{s}_{t+1}^k) + \hat{s}_{t-1}\exp(-\hat{s}_{t-1}^k) \\ &+ (\hat{s}_{t+1} - \hat{s}_{t-1})\exp(-\hat{s}_t^k)\end{aligned}$$

Consequently, by substituting the above equations into Equation ($*$), we achieve the final result. □

### A.2. Proofs for Section 4.1

We first migrate the definition of asymptotic equivalence moment from (Vladimirova et al., 2018):

**Definition 1.** *Two sequences $a_r$ and $b_r$ are called asymptotic equivalent and denoted as $a_r \asymp b_r$ if there exist constants $A > 0$ and $B > 0$ such that*

$$A \leq a_r / b_r \leq B, \textit{ for } \forall r \in \mathbb{N}.$$

According to the equivalent sub-Weibull distribution properties defined in (Vladimirova et al., 2018), if the $r$-th moment of r.v. $X$ satisfies

$$||X||_r = (\mathbb{E}[|X|^r]))^{1/r} \asymp r^\theta,$$

then $X \sim \text{subW}(\theta)$.

Before we start to prove the theorems in Section 4.1, we need to introduce the following lemmas.

**Lemma 1.** *Assume $w_1, w_2, ..., w_n$ are independent variables drawn from a normal distribution with zero mean, and $x_1, x_2, ..., x_n$ are variables drawn from an identical distribution. Let $y = \sum_{i=1}^n w_i x_i$, then $y$ follows a distribution symmetric about 0.*

*Proof.* Since $w_1, w_2, ..., w_n$ are independent with each other, we have

$$\begin{aligned}&\mathbb{P}(w_1 = w_{1,0}, ..., w_n = w_{n,0}) = \prod_{i=1}^n \mathbb{P}(w_i = w_{i,0}) \\ &= \prod_{i=1}^n \mathbb{P}(w_i = -w_{i,0}) = \mathbb{P}(w_1 = -w_{1,0}, ..., -w_n = w_{n,0})\end{aligned}$$

Therefore, $y$ follows a distribution symmetric about 0. □

**Lemma 2.** *Assume there is a distribution $x \sim X$, and variable $y = px$, where $p \sim Bernouli(\frac{1}{2})$ is independent from $X$. Then we have*
*(a) if $x_i, x_j \sim X$, and $\text{Cov}[(x_i)^s, (x_j)^t] \geq 0$ holds for any $s, t \in \mathbb{N}$, then $\text{Cov}[(y_i)^s, (y_j)^t] \geq 0$;*
*(b) if $||X||_r \asymp r^\theta$, then $||Y||_r \asymp r^\theta$.*

*Proof.* (a) Since $p$ is independent of $x_i, x_j$ and $\mathbb{E}[p^i] = \frac{1}{2^{i+1}}$ for any integer $i$, we have

$$\begin{aligned}&\text{Cov}[(y_i)^s, (y_j)^t] = \mathbb{E}[(y_i)^s(y_j)^t] - \mathbb{E}[(y_i)^s]\mathbb{E}[(y_j)^t] \\ &= \mathbb{E}[p^{s+t}(x_i)^s(x_j)^t] - \mathbb{E}[p^s]\mathbb{E}[(x_i)^s]\mathbb{E}[p^t]\mathbb{E}[(x_j)^t] \\ &= \mathbb{E}[p^{s+t}]\mathbb{E}[(x_i)^s(x_j)^t] - (\mathbb{E}[p^s]\mathbb{E}[p^t])\mathbb{E}[(x_i)^s]\mathbb{E}[(x_j)^t] \\ &= \frac{1}{2^{s+t+1}}(\mathbb{E}[(x_i)^s(x_j)^t] - \mathbb{E}[(x_i)^s]\mathbb{E}[(x_j)^t]) \\ &= \frac{1}{2^{s+t+1}}\text{Cov}[(x_i)^s, (x_j)^t] \\ &\geq 0.\end{aligned}$$

(b) According to the definition of asymptotic equivalence moment, there exist constants $A, B$ such that $A \leq ||X||_r/r^\theta \leq B$. Since $p$ and $X$ are independent, we have

$$||Y||_r = (\mathbb{E}|Y|^r)^{(1/r)} = (\mathbb{E}|pX|^r)^{(1/r)}$$
$$= (\mathbb{E}|p|^r\mathbb{E}|X|^r)^{(1/r)} = \frac{1}{2}(\mathbb{E}|X|^r)^{(1/r)} = \frac{1}{2}||X||_r.$$

Therefore, $A/2 \leq ||Y||_r/r^\theta \leq B/2$. □

**Lemma 3.** *Assume $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n, \tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n$ are independent variables drawn from the same normal distribution with zero mean and $x_1, x_2, \ldots, x_n$ are variables an identical distribution satisfying $\text{Cov}[x_i^s, x_j^t] \geq 0$ for any $s, t \in \mathbb{N}$. Let $\hat{y} = \sum_{i=1}^n \hat{w}_i x_i$ and $\tilde{y} = \sum_{i=1}^n \tilde{w}_i x_i$. Then we have $\text{Cov}[\hat{y}^s, \tilde{y}^t] \geq 0$.*

*Proof.* This lemma is actually one part of the Theorem 3.2 of (Vladimirova et al., 2019). Therefore, we refer to their proof for simplicity. □

Now we present the proofs for theorems in Section 4.1.

**Theorem 3.** *Assume the MLP model in Theorem 2 satisfies Assumption 2, then for any $1 \leq l < L, 1 \leq i \leq C_l$, we have $m_i^{(l)}, n_i^{(l)} \sim \text{subW}((L-l)/2)$.*

*Proof.* Throughout our proof, we use $m_a, m_b$ ($n_a, n_b$) to indicate arbitrary element of $\boldsymbol{m}$ ($\boldsymbol{n}$).

With Assumption 2, $\boldsymbol{m}^{(l)}$ is independent with $\boldsymbol{u}^{(l)}$. According to Lemma 1, $\boldsymbol{u}^{(l)}$ has a symmetric distribution, hence we introduce a r.v. $p \sim Bernouli(\frac{1}{2})$ which is independent with $\boldsymbol{m}^{(l)}$, and assume $n_i^{(l)} = pm_i^{(l)}$.

We first prove that the following inequality

$$\text{Cov}[(m_a^{(l)})^s, (m_b^{(l)})^t] \geq 0, \text{Cov}[(n_a^{(l)})^s, (n_b^{(l)})^t] \geq 0$$

hold for any $s, t \in \mathbb{N}$ and $1 \leq l \leq L-1$. The proof is performed by induction w.r.t. network layers.

*Base Step:*
Consider the backward units of the $(L-1)$-th layer. Since model weights $\boldsymbol{W}_a^{(L)}$ and $\boldsymbol{W}_b^{(L)}$ are independently drawn from an normal distribution, and $\boldsymbol{n}^{(L)}$ is independent with $\boldsymbol{W}^{(L)}$. Therefore, $m_a^{(L-1)} = \boldsymbol{W}_a^{(L-1)^T}\boldsymbol{n}^{(L)}$ and $m_b^{(L-1)} = \boldsymbol{W}_b^{(L)^T}\boldsymbol{n}^{(L)}$ are independent, which means $(m_a^{(L-1)})^s$ and $(m_b^{(L-1)})^t$ are also independent. Hence $\text{Cov}[(m_a^{(L-1)})^s, (m_b^{(L-1)})^t] = 0$. According to Lemma 2(a), $\text{Cov}[(n_a^{(L-1)})^s, (n_b^{(L-1)})^t] = 0$.

*Induction step:*
According to Lemma 3 and Lemma 2(a), we know the non-negative conditions also hold for the $l$-th layer.

Now we prove Theorem 3. We only need to prove $||m_a^{(l)}||_r \asymp r^{(L-l)/2}$ and $||n_a^{(l)}||_r \asymp r^{(L-l)/2}$ for $1 \leq l \leq L-1$ and $r \in \mathbb{N}$ by induction w.r.t. to network layers.

*Base step:*
Consider the backward units of the $(L-1)$-th layer.

Since $m_a^{(L-1)} = \boldsymbol{W}_a^{(L)^T}\boldsymbol{n}^{(L)}$, where the weights $\boldsymbol{W}_a^{(L)}$ follow normal distribution and $\boldsymbol{n}^{(L)}$ is a vector independent with $\boldsymbol{W}_a^{(L)}$, therefore

$$m_a^{(L-1)} = \boldsymbol{W}_a^{(L)^T}\boldsymbol{n}^{(L)} \sim \mathcal{N}(0, {\sigma^{(l)}}^2||\boldsymbol{n}^{(L)}||^2).$$

According to Lemma A.1 (lemma of Gaussiam moments) from (Vladimirova et al., 2019), we have $||m_a^{(L-1)}||_r \asymp \sqrt{r}$. Then in accord to Lemma 2(b), we have $||n_a^{(L-1)}||_r \asymp \sqrt{r}$.

*Induction step:*
Suppose the backward units of the $(l+1)-th$ layer satisfy $||m_a^{(l+1)}||_r \asymp r^{(L-l-1)/2}$, $||n_a^{(l+1)}||_r \asymp r^{(L-l-1)/2}$. Since $\boldsymbol{n}^{(l+1)}$ satisfies the non-negative covariance condition and $m_a^{(l)} = \boldsymbol{W}_a^{(l+1)^T}\boldsymbol{n}^{(l+1)}$, according to Lemma A.2 (lemma of multiplication moments) from (Vladimirova et al., 2019), we have $||m_a^{(l)}||_r \asymp r^{(L-l)/2}$. By Lemma 2(b), $||n_a^{(l)}||_r \asymp r^{(L-l)/2}$. This completes the proof. □

**Theorem 4.** *Considering the MLP model in Theorem 2, 3, for any $1 < l < L, 1 \leq i \leq C_{l-1}, 1 \leq j \leq C_l$, we have $\boldsymbol{\nabla W}_{ij}^{(l)} \sim \text{subW}((L-1)/2)$.*

*Proof.* Assume $X \sim \text{subW}(\theta_1)$, $Y \sim \text{subW}(\theta_2)$ are independent variables. Let $Z = XY$, then we have

$$||Z||_r = \mathbb{E}(|Z|^r)^{1/r} = \mathbb{E}(|XY|^r)^{1/r}$$
$$= \mathbb{E}(|X|^r)^{1/r}\mathbb{E}(|Y|^r)^{1/r}$$
$$= ||X||_r||Y||_r.$$

By the definition of asymptotic equivalence moment, there exist constants $A_1, B_1, A_2, B_2$ such that

$$A_1 r^{\theta_1} \leq ||X||_r \leq B_1 r^{\theta_1}, A_2 r^{\theta_2} \leq ||X||_r \leq B_2 r^{\theta_2}.$$

Therefore, we have $A_1A_2 r^{\theta_1+\theta_2} \leq ||Z||_r \leq B_1B_2 r^{\theta_1+\theta_2}$, which means $Z \sim \text{subW}(\theta_1+\theta_2)$.

Back to the proof of Theorem 4. By Theorem 2 and 3, for $1 \leq i \leq C_{l-1}, 1 \leq j \leq C_l$, we have

$$v_j^{(l-1)} \sim \text{subW}((l-1)/2), n_i^{(l)} \sim \text{subW}((L-l)/2).$$

With Assumption 2, $v_j^{(l-1)}, n_i^{(l)}$ are independent from each other. Thus we have

$$\boldsymbol{\nabla W}_{ij}^{(l)} = v_j^{(l-1)}n_i^{(l)} \sim \text{subW}((L-1)/2).$$

□

### A.3. Proofs for Section 4.2

**Theorem 5.** *Suppose we run SGD with quantization on an object satisfying Assumption 3 with constant step size $\alpha < 2/L_2$. Assume quantization satisfies $\mathbb{E}[||\tilde{\boldsymbol{x}} - \boldsymbol{x}||] \leq Q\mathbb{E}[||\boldsymbol{x}||]$ for any $x \in \mathbb{R}^D$. After $T$ runs, select $\overline{\boldsymbol{w}}_T$ randomly from $\{\boldsymbol{w}_0, \boldsymbol{w}_1, ..., \boldsymbol{w}_{T-1}\}$. Then we have*

$$\mathbb{E}[||\nabla f(\overline{\boldsymbol{w}}_T)||^2] \leq \frac{2(f(\boldsymbol{w}_0) - f^*)}{(2\alpha - \alpha^2 L_2)T} + \frac{\alpha L_2(\sigma^2 + Q^2\sigma_0^2)}{2 - \alpha L_2}.$$

*Proof.* By Taylor's Expansion Formula with Lagrangian Remainder,

$$\begin{aligned}
f(\boldsymbol{w}_{t+1}) &= f(\boldsymbol{w}_t - \alpha\tilde{\boldsymbol{g}}_{i_t}) \\
&= f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t} + \alpha(\boldsymbol{g}_{i_t} - \tilde{\boldsymbol{g}}_{i_t})) \\
&= f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t}) + \alpha(\boldsymbol{g}_{i_t} - \tilde{\boldsymbol{g}}_{i_t})^T\nabla f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t}) \\
&\quad + \frac{1}{2}\alpha^2(\boldsymbol{g}_{i_t} - \tilde{\boldsymbol{g}}_{i_t})^T\nabla^2 f(\xi_t)(\boldsymbol{g}_{i_t} - \tilde{\boldsymbol{g}}_{i_t}) \\
&\leq f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t}) + \alpha(\boldsymbol{g}_{i_t} - \tilde{\boldsymbol{g}}_{i_t})^T\nabla f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t}) \\
&\quad + \frac{1}{2}\alpha^2 L_2||\boldsymbol{g}_{i_t} - \tilde{\boldsymbol{g}}_{i_t}||^2,
\end{aligned}$$

where the last inequality is due to the property of Lipschitz continuity. Notice that $\mathbb{E}[\tilde{\boldsymbol{g}}_{i_t}] = \tilde{\boldsymbol{g}}_{i_t}$, taking expectation on both sides, we have

$$\mathbb{E}[f(\boldsymbol{w}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t})] + \frac{1}{2}\alpha^2 L_2 Q^2\sigma_0^2. \quad (*)$$

Again, using Taylor's Expansion Formula with Lagrangian Remainder,

$$\begin{aligned}
f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t}) &= f(\boldsymbol{w}_t) - \alpha\boldsymbol{g}_{i_t}^T\nabla f(\boldsymbol{w}_t) \\
&\quad + \frac{1}{2}\alpha^2\boldsymbol{g}_{i_t}^T\nabla^2 f(\eta_t)\boldsymbol{g}_{i_t} \\
&\leq f(\boldsymbol{w}_t) - \alpha\boldsymbol{g}_{i_t}^T\nabla f(\boldsymbol{w}_t) + \frac{1}{2}\alpha^2 L_2||\boldsymbol{g}_{i_t}||^2 \\
&= f(\boldsymbol{w}_t) - \alpha\boldsymbol{g}_{i_t}^T\nabla f(\boldsymbol{w}_t) \\
&\quad + \frac{1}{2}\alpha^2 L_2||\nabla f(\boldsymbol{w}_t) + (\boldsymbol{g}_{i_t} - \nabla f(\boldsymbol{w}_t))||^2 \\
&\leq f(\boldsymbol{w}_t) - \alpha\boldsymbol{g}_{i_t}^T\nabla f(\boldsymbol{w}_t) \\
&\quad + \frac{1}{2}\alpha^2 L_2(||\nabla f(\boldsymbol{w}_t)||^2 + ||\boldsymbol{g}_{i_t} - \nabla f(\boldsymbol{w}_t)||^2).
\end{aligned}$$

Notice that $\mathbb{E}[\boldsymbol{g}_{i_t}] = \nabla f(\boldsymbol{w}_t)$, taking the expectation on both sides, we have

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}_t - \alpha\boldsymbol{g}_{i_t})] &\leq \mathbb{E}[f(\boldsymbol{w}_t)] - (\alpha - \frac{1}{2}\alpha^2 L_2)\mathbb{E}[||\nabla f(\boldsymbol{w}_t)||^2] \\
&\quad + \frac{1}{2}\alpha^2 L_2\mathbb{E}[||\boldsymbol{g}_{i_t} - \nabla f(\boldsymbol{w}_t)||^2] \\
&\leq \mathbb{E}[f(\boldsymbol{w}_t)] - (\alpha - \frac{1}{2}\alpha^2 L_2)\mathbb{E}[||\nabla f(\boldsymbol{w}_t)||^2] \\
&\quad + \frac{1}{2}\alpha^2 L_2\sigma^2.
\end{aligned}$$

$(**)$

Plugging $(**)$ into $(*)$ and summing over $t$ from 0 to $T-1$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}(\alpha - \frac{1}{2}\alpha^2 L_2)\mathbb{E}[||\nabla f(\boldsymbol{w}_t)||^2] \leq \\
&\qquad f(\boldsymbol{w}_0) - \mathbb{E}[f(\boldsymbol{w}_T)] + \frac{1}{2}T\alpha^2 L_2(\sigma^2 + Q^2\sigma_0^2).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[||\nabla f(\overline{\boldsymbol{w}})||^2] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\boldsymbol{w}_t)||^2] \\
&\leq \frac{2(f(\boldsymbol{w}_0) - \mathbb{E}[f(\boldsymbol{w}_T)])}{(2\alpha - \alpha^2 L_2)} + \frac{\alpha L_2(\sigma^2 + Q\sigma_0^2)}{2 - \alpha L_2} \\
&\leq \frac{2(f(\boldsymbol{w}_0) - f*)}{(2\alpha - \alpha^2 L_2)} + \frac{\alpha L_2(\sigma^2 + Q^2\sigma_0^2)}{2 - \alpha L_2}.
\end{aligned}$$

□

## B. More Experiments

**Sensitivity** To address the concerns on the sensitivity to hyper-parameters, We further examine TINYSCRIPT on varying learning rates from 0.005 to 0.1 (we also change batch size accordingly) and the do not observe notable accuracy drop (within 1% of FP32).

**Runtime** Quantization inevitably incurs extra overhead. For instance, training ResNet50 on ImageNet without quantization (to be fair, we also apply the checkpointing technique) takes 124 hours, whilst TINYSCRIPT ($n = 8$) and Vanilla ($n = 16$) take 167 and 165, respecitively. This verifies our analysis that our method has the same complexity as uniform quantization. Therefore, we believe it is valuable to shirnk more memory with similar cost, especially when memory is limited.

**Number of GPUs** To assess the influence of number of GPUs, we train WRN34 on Cifar10 with 2 and 8 GPUs. TINYSCRIPT ($n = 4$) achieves 5.85 and 5.62 error, and the results of Vanilla ($n = 8$) are 5.80 and 5.71, respectively. It shows that the accuracy of training with quantization does not vary significantly w.r.t. number of GPUs.