Supplementary Document for SQuaFL: Sketched-Quantization based Communication Efficient Federated Learning

I. Convergence Analysis and Discussion

We provide the theoretical results of the convergence of SQuaFL for non-convex objective. First, we state the standard assumptions and definitions commonly used in the context of analyzing stochastic algorithms satisfied by our compressors $S(Q(.))$, weight matrix w and objective function f. Here, $\|.\|$ denotes the l_2-norm of a vector v.

Assumption 1. *The local objective function f_n is differentiable and L-smooth such that, $\|\nabla f_n(u) - \nabla f_n(v)\| \leq L\|u - v\|$, $\forall u, v \in \mathbb{R}^d$, for a constant $L > 0$ and is lower-bounded, i.e., $f^* = \min_{w\in\mathbb{R}^d} f(w) > -\infty$.*

Assumption 2. *Stochastic gradients $g = \nabla\tilde{f}(v)$ are unbiased and variance bounded i.e., $\mathbb{E}[g] = \bar{g}$ and $\mathbb{E}[\|g - \bar{g}\|^2] \leq \sigma^2$.*

Assumption 3. *The function f_n is λ-strongly convex if for any $u, v \in \mathbb{R}^d$ we have, $\langle\nabla f_n(u) - \nabla f_n(v), u - v\rangle \geq \lambda\|u - v\|^2$.*

We state the following lemma about the important properties Count Sketches possess, which is useful in our convergence analysis.

Lemma 1. *[?] For a Count Sketch S with r hash tables of b bins into $r \times b$ array of counters, for any input element $v_i \in v$ and query K, with probability $1-\delta$, the following respective relations of unbiasedness and bounded estimation error of Count Sketches hold,*

$$\begin{aligned} &\mathbb{E}[K(S(v))] = v, \\ &\mathbb{E}[\|K(S(v) - v\|^2)] \leq \mu^2 d\|v\|^2, \end{aligned} \tag{1}$$

where $a = \mathcal{O}(\ln(d/\delta))$ and $b = \mathcal{O}(e/\mu^2)$.

Lemma 1 states that high-magnitude entries of the vector which comprise a large portion of $\|v\|$ can be recovered from the unbiased sketched estimate $S(v)$.

Correspondingly, we introduce the following lemma to show that the approximate original values can be retrieved from the sketched vector which consists of quantized values.

Lemma 2. *For a Count Sketch estimate $S(.)$ with $r = \mathcal{O}(\ln(d/\delta))$ hash tables and $b = \mathcal{O}(e/\mu^2 + q)$ bins over quantized values $Q(.)$ of any vector v, with probability $1-\delta$, the following respective relations of unbiasedness and bounded estimation error of our compression hold,*

$$\begin{aligned} &\mathbb{E}[K(\tilde{v})] = v, \\ &\mathbb{E}[\|K(\tilde{v}) - v\|^2] \leq \Delta\|v\|^2, \end{aligned} \tag{2}$$

where $\tilde{v} = S(Q(v))$.

Proof sketch. Using the properties of the stochastic quantizer $Q(.)$ given in Section **??** and Lemma 1, we can show the properties of the query over sketched-quantized vectors. □

Definition 1. *(ϵ-Differential Privacy). A randomized mechanism $\mathcal{A}$ is ϵ-differentially private if for any two neighboring inputs D, D' that differ in at most one single element, and for any possible output s in the output space of $\mathcal{A}$, it holds that*

$$Pr(\mathcal{A}(D) = s) \leq e^{\epsilon} Pr(\mathcal{A}(D') = s)$$

Theorem 1. *(ϵ-differential privacy of Count Sketch, [?]). For a sketching algorithm using Count Sketch $S_{r\times b}$ with r arrays of b bins, for any input vector I with length κ drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and is bounded by a constant ϕ with a large probability, achieves $r.\ln\left(1 + \frac{\zeta\phi^2 b(b-1)}{\sigma^2(\kappa-2)}(1 + \ln(\kappa - b))\right)$-differential privacy with high probability, where ζ is a positive constant satisfying $\frac{\phi^2 b(b-1)}{\sigma^2(\kappa-2)}(1 + \ln(\kappa - b)) \leq \frac{1}{2} - \frac{1}{\beta}$.*

Next, we present the following intermediate lemmas to analyze the convergence of our method in Algorithm **??**.

Lemma 3. *The expectation of the inner product of the stochastic gradient and batch gradient can be bounded using the smoothness and lower boundedness assumption of the local objective function as,*

$$-\mathbb{E}_{\xi,\mathrm{C}}[\langle \nabla f(w_t), g_t \rangle] \leq \frac{\eta}{2N} \sum_{n=1}^{N} [-\|\nabla f(w_t)\|^2 - \|\nabla f(w_t^n)\|^2 + L^2 \eta^2 [\sigma^2 + \|\bar{g}_t^n\|^2]]. \tag{3}$$

Proof. Expectation over our sketched-quantization compressor denoted by $\mathbb{E}_{\mathrm{C}}$ can be applied as,

$$\begin{aligned}
&- \mathbb{E}[\mathbb{E}_{\mathrm{C}}[\langle \nabla f(w_t), \hat{g}_t \rangle]] \\
&= -\mathbb{E}[\langle \nabla f(w_t), \eta \frac{1}{N} \sum_{n=1}^{N} g_t^n \rangle] \\
&= -\langle \nabla f(w_t), \frac{\eta}{N} \sum_{n=1}^{N} \mathbb{E}[g_t^n] \rangle \\
&= -\frac{\eta}{N} \sum_{n=1}^{N} \langle \nabla f(w_t), \bar{g}_t^n \rangle = -\frac{\eta}{N} \sum_{n=1}^{N} \langle \nabla f(w_t), \nabla f(w_t^n) \rangle && (4) \\
&= -\frac{1}{2} \frac{\eta}{N} \sum_{n=1}^{N} [\|\nabla f(w_t)\|^2 + \|\nabla f(w_t^n)\|^2 - \|\nabla f(w_t) - \nabla f(w_t^n)\|^2] && (5) \\
&\leq \frac{1}{2} \frac{\eta}{N} \sum_{n=1}^{N} [-\|\nabla f(w_t)\|^2 - \|\nabla f(w_t^n)\|^2 + L^2 \|w_t - w_t^n\|^2] && (6) \\
&\leq \frac{\eta}{2N} \sum_{n=1}^{N} \left[-\|\nabla f(w_t)\|^2 - \|\nabla f(w_t^n)\|^2 + L^2 \eta^2 [\sigma^2 + \|\bar{g}_t^n\|^2]\right], && (7)
\end{aligned}$$

where (4) is from Assumption 2, (5) is due to the property of $2\langle u, v \rangle = \|u\|^2 + \|v\|^2 + \|u - v\|^2$ and (6) is from the smoothness assumption 1. We use the following lemma 4 to bound the last term in (6) to hence obtain (7). □

Lemma 4. *For a given communication round t, bound for the distance between the global and local models can be given as,*

$$\mathbb{E}[\|w_t - w_t^n\|^2] \leq \eta^2 \sigma^2 + \eta^2 \|\bar{g}_t^n\|^2. \tag{8}$$

Proof. From the update rule of our algorithm we have,

$$\mathbb{E}[\|w_t - w_t^n\|^2] = \mathbb{E}[\|w_t - (w_t - \eta g_t^n)\|^2] = \mathbb{E}[\|\eta g_t^n\|^2].$$

Using the expression of variance, $\mathrm{var}[u] = \mathbb{E}[u^2] + [\mathbb{E}[u]^2]$,

$$\begin{aligned}
\mathbb{E}[\|\eta g_t\|^2] &= \mathbb{E}[\|\eta \mathrm{var}[g_t^n]\|^2] + [\|\eta \bar{g}_t^n\|^2] \\
&= \eta^2 \mathbb{E}[\|g_t^n - \bar{g}_t^n\|^2] + \eta^2 \|\bar{g}_t^n\|^2 \\
&= \eta^2 \sigma^2 + \eta^2 \|\bar{g}_t^n\|^2, && (9)
\end{aligned}$$

where (9) also uses variance boundedness of the gradient from Assumption 2. □

Lemma 5. *The mean of quantized and sketched stochastic gradients can be bounded using unbiased compression properties of our compressor and bounded variance assumptions of the gradient as,*

$$\mathbb{E}[\mathbb{E}_{\mathrm{C}}[\|\hat{g}_t\|^2]] \leq [\Delta + 1] \frac{\sigma^2}{N} + \left[\frac{\Delta}{N} + 1\right] \frac{1}{N} \sum_{n=1}^{N} \|\bar{g}_t^n\|^2. \tag{10}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{\mathrm{C}}[\|\hat{g}_t\|^2]\right] &= \mathbb{E}\left[\mathbb{E}_{\mathrm{C}}\left[\left\|\frac{1}{N} \sum_{n=1}^{N} S(\mathcal{Q}(g_t^n))\right\|^2\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\mathrm{C}}\left[\left\|\frac{1}{N} \sum_{n=1}^{N} \hat{g}_t^n\right\|^2\right]\right]
\end{aligned}$$

$$= \mathbb{E}\left[\mathbb{E}_{\mathrm{C}}\left[\left\|\frac{1}{N}\sum_{n=1}^{N}\hat{g}_t^n - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathrm{C}}[\hat{g}_t^n]\right\|^2\right] + \left\|\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathrm{C}}[\hat{g}_t^n]\right\|^2\right] \tag{11}$$

$$= \mathbb{E}\left[\mathbb{E}_{\mathrm{C}}\left[\frac{1}{N^2}\sum_{n=1}^{N}\|\hat{g}_t^n - g_t^n\|^2\right] + \left\|\frac{1}{N}\sum_{n=1}^{N}g_t^n\right\|^2\right] \tag{12}$$

$$\leq \mathbb{E}\left[\frac{1}{N^2}\sum_{n=1}^{N}\Delta\|g_t^n\|^2\right] + \left\|\frac{1}{N}\sum_{n=1}^{N}g_t^n\right\|^2 \tag{13}$$

$$= \sum_{n=1}^{N}\frac{\Delta}{N^2}\left[\mathrm{var}[g_t^n] + \|\bar{g}_t^n\|^2\right] + \left[\mathrm{var}\left(\frac{1}{N}\sum_{n=1}^{N}g_t^n\right) + \left\|\frac{1}{N}\sum_{n=1}^{N}\bar{g}_t^n\right\|^2\right] \tag{14}$$

$$\leq \sum_{n=1}^{N}\frac{\Delta}{N^2}[\sigma^2 + \|\bar{g}_t^n\|^2] + \left[\frac{1}{N^2}\sum_{n=1}^{N}\sigma^2 + \frac{1}{N}\sum_{n=1}^{N}\|\bar{g}_t^n\|^2\right] \tag{15}$$

$$\leq [\Delta+1]\frac{\sigma^2}{N} + \left[\frac{\Delta}{N}+1\right]\frac{1}{N}\sum_{n=1}^{N}\|\bar{g}_t^n\|^2.$$

where (11) and (14) follow from the variance expression, results from Lemma 2 are used in (12) and (13). (15) is obtained using the variance bound of the gradient from Assumption 2. Hence the proof is completed. □

Theorem 2. *(Non-convex). Considering the iterates w_t generated from Algorithm* **??** *for the total number of communication rounds T, suppose that the conditions in Assumptions $1-2$ hold, w_* being the global optimal solution with function value $f(w_*)$ and we set the step size as $\eta = \frac{1}{L}\sqrt{\frac{N}{T(\Delta+1)}}$ and given bins $b = \mathcal{O}(\frac{e}{\mu^2+q})$, then the following condition holds,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2 \leq \frac{2(f(w_0)-f(w_*))}{\eta^2 T} + L^2\eta^2\sigma^2 + \frac{L\eta^2\sigma^2(\Delta+1)}{N}. \tag{16}$$

Proof. The sequence of iterates computed from the update step of Algorithm **??** is given by,

$$w_{t+1} = w_t - \eta\hat{g}_t \tag{17}$$

Since we quantize and sketch the local updates, values returned from query used to update the global model can be written as,

$$w_{t+1} = w_t - \eta\left(\frac{\eta}{N}\sum_{n=1}^{N}\left[S\left(\mathcal{Q}\left(\frac{w_t - w_t^n}{\eta}\right)\right)\right]\right)$$

$$w_{t+1} = w_t - \eta\left(\frac{\eta}{N}\sum_{n=1}^{N}S\left(\mathcal{Q}\left(g_t^n\right)\right)\right) \tag{18}$$

Next, taking expectation over the randomness of quantization and count sketch as $\mathbb{E}_{\mathrm{C}}$ and using the unbiased properties,

$$\mathbb{E}_{\mathrm{C}}[\hat{g}_t] = \mathbb{E}\left[\frac{\eta}{N}\sum_{n=1}^{N}\left[S\left(\mathcal{Q}\left(\frac{w_t - w_t^n}{\eta}\right)\right)\right]\right]$$

$$= \frac{1}{N}\sum_{n\in\mathcal{N}}\left[-\eta\mathbb{E}_{\mathrm{C}}\left[S\left(\mathcal{Q}\left(g_t^n\right)\right)\right]\right] \triangleq g_t. \tag{19}$$

Using the results from (17), (19) and L-smoothness assumption of the gradient we have,

$$f(w_{t+1}) - f(w_t) \leq -\eta\langle\nabla f(w), g_t\rangle + \frac{\eta^2 L}{2}\|g_t\|^2. \tag{20}$$

Taking expectation over sampling on both sides, we obtain:

$$\mathbb{E}[\mathbb{E}_{\mathrm{C}}\left[f(w_{t+1}) - f(w_t)\right]] \leq -\eta\mathbb{E}[\mathbb{E}_{\mathrm{C}}\left[\langle\nabla f(w), \hat{g}_t\rangle\right]] + \frac{\eta^2 L}{2}\mathbb{E}[\mathbb{E}_{\mathrm{C}}\left[\|\hat{g}_t\|^2\right]$$

$$= -\eta\mathbb{E}[\langle\nabla f(w), g_t\rangle] + \frac{\eta^2 L}{2}\mathbb{E}[\mathbb{E}_{\mathrm{C}}\left[\|\hat{g}_t\|^2\right]. \tag{21}$$

Using the results from Lemmas 3, 4 and 5 to bound the terms in (21) we obtain,

$$\mathbb{E}[\mathbb{E}_{\mathrm{C}}\left[f(w_{t+1})-f(w_t)\right]] \leq \frac{L\eta^4\sigma^2}{2N}[\Delta+1+LN] - \frac{\eta^2}{2N}\sum_{n=1}^{N}\|\bar{g}_t^n\|^2\left[-L\eta^2\left(\frac{\Delta}{N}+1\right)-L^2\eta^2+1\right] - \frac{\eta^2}{2}\|\nabla f(w_t)\|^2. \quad (22)$$

Suppose the value of $[-L\eta^2\left(\frac{\Delta}{N}+1\right)-L^2\eta^2+1] \leq 1$, we can rewrite (22) as,

$$\mathbb{E}[\mathbb{E}_{\mathrm{C}}\left[f(w_{t+1})-f(w_t)\right]] \leq -\frac{\eta^2}{2}\|\nabla f(w_t)\|^2 + \frac{L\eta^2\sigma^2}{2N}[\Delta+1+L\eta N].$$

Summing up over communication rounds $t=0,\cdots,T-1$ and rearranging the terms yields the following, hence proving the theorem,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2 \leq \frac{2(f(w_0)-f(w_*))}{\eta^2 T} + L^2\eta^2\sigma^2 + \frac{L\eta^2\sigma^2(\Delta+1)}{N}. \quad (23)$$

□

Remark 1. *Theorem 2 shows that the sequence of iterates attain convergence and the algorithm can achieve an ϵ-suboptimal solution $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2 \leq \epsilon$, using the results from Corollary 1 of setting $T=\mathcal{O}(\frac{\Delta+1}{\epsilon})$.*

Corollary 1. *A lower bound can be obtained on the communication rounds by rewriting the step size condition as,*

$$\begin{aligned} L^2\eta^2 + L\eta^2\left(\frac{\Delta}{N}+1\right) &\leq 1 \\ &= \frac{\sqrt{\left(\frac{\Delta}{N}+1\right)^2+4}-\left(\frac{\Delta}{N}+1\right)}{2L}. \end{aligned} \quad (24)$$

Using the step size value as indicated in Theorem 2 as $\eta=\frac{1}{L\eta}\sqrt{\frac{N}{T(\Delta+1)}}$ in (24) we obtain,

$$\begin{aligned} T \geq 4N/\Bigg[(\Delta+1)\eta^2\Bigg[\sqrt{\left(\frac{\Delta}{N}+1\right)^2\eta^2+4} \\ -\left(\frac{\Delta}{N}+1\right)\eta^2\Bigg]^2\Bigg] \quad (25) \\ \geq \mathcal{O}\left(\frac{N}{(\mu^2 d+q)+1}\right). \quad (26) \end{aligned}$$

Hence, for a target accuracy ϵ, we can have $T=\mathcal{O}\left(\frac{\Delta+1}{\epsilon}\right)$ number of communication rounds.

Remark 2. *(Total communication cost and differential privacy [?]). As a result of Corollary 1, the total communication cost per client can be written as,*

$$\mathcal{O}(T\nu) = \mathcal{O}\left(\frac{nb}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right), \quad (27)$$

where ν is the number of bits per communication round per device. We can infer the following from this result. First, it implies an improvement in communication complexity of federated learning compared to state-of-the-art result of $\mathcal{O}(\frac{d}{\epsilon})$ by methods such as [?]. In addition, this implies differential privacy in accordance with Theorem 1 satisfying Definition 1.

Theorem 3. *(Strongly Convex). From the iterates generated from Algorithm ?? for the total number of communication rounds T, suppose that the conditions in Assumptions $1-3$ hold and we set the step size as $\eta=\frac{1}{2L\eta(\frac{\Delta}{N}+1)}$ and given bins $b=\mathcal{O}(\frac{e}{\mu^2+q})$, then the following condition holds,*

$$\begin{aligned} \mathbb{E}[f(w_T)-f(w_*)] \leq (1-\eta^2\lambda)^T(f(w_0)-f(w_*)) + \frac{1}{\lambda} \\ \left[\frac{1}{2}L^2\eta^2\sigma^2 + (\Delta+1)\frac{\mu^2 L\sigma^2}{2N}\right]. \end{aligned} \quad (28)$$

Remark 3. *(Compression noise) The following can be noted from the theoretical results of Theorem 2 and 3 about the compression noise Δ,*

- *Variance of the compression scheme is scaled down by a factor of* $1/N$*, which infers that learning collaboratively helps to lower the effect of induced compression noise.*
- *Further, in case if* $b \to d$*, we can control the compression noise by using large size of the hash table which makes the compression noise smaller.*

Hence, for non-convex and strongly convex settings, we theoretically show the convergence properties of SQuaFL along with its privacy preserving attributes.

APPENDIX