Supplementary materials for "LENA: Communication-Efficient Distributed Learning with Self-Triggered Gradient Uploads"

A Useful Definitions and Lemmas

Definition 1. *We define virtual sequence* $(\widetilde{\boldsymbol{w}}^t)_{t\geq 0}$ *as*

$$\widetilde{\boldsymbol{w}}^{t+1} = \widetilde{\boldsymbol{w}}^t - \frac{1}{M}\sum_{m\in[M]} \alpha_t \boldsymbol{g}_m^t \,, \tag{A.1}$$

for all $t \geq 0$.

Lemma 1. *We have the following useful set of inequalities:*

- *For any vectors* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ *and any* $\gamma > 0$, *we have*

$$\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq (1+\gamma)\|\boldsymbol{a}\|^2 + (1+\gamma^{-1})\|\boldsymbol{b}\|^2 \,. \tag{A.2}$$

- *For any vectors* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ *and any* $\gamma > 0$, *we have*

$$2\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \left(\gamma\|\boldsymbol{a}\|^2 + \gamma^{-1}\|\boldsymbol{b}\|^2\right) \,. \tag{A.3}$$

Lemma 2. *Under Assumptions 1 and 3, stochastic gradient* $\boldsymbol{g}_m^t := \nabla f_m(\boldsymbol{w}^t) + \boldsymbol{\xi}_m^t$ *satisfies*

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{M}\sum_m \boldsymbol{g}_m^t\right\|^2\right] &= \left\|\nabla F(\boldsymbol{w}^t)\right\|^2 + \frac{1}{M^2}\sum_{m\in[M]}\mathbb{E}\left[\left\|\boldsymbol{\xi}_m^t\right\|^2\right] \\ &\overset{(8)}{\leq} \left\|\nabla F(\boldsymbol{w}^t)\right\|^2 + \frac{\sigma^2}{M}\end{aligned} \tag{A.4}$$

where the expectation is with respect to $\boldsymbol{\xi}_1^t, \ldots, \boldsymbol{\xi}_M^t$.

If in addition, the function F *is convex, then*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_m \boldsymbol{g}_m^t\right\|^2\right] \leq 2\overline{L}\left(F(\boldsymbol{w}^t) - F^\star\right) + \frac{\sigma^2}{M} \,, \tag{A.5}$$

Lemma 3. *Let* $\{\boldsymbol{w}^t, \boldsymbol{v}^t, \boldsymbol{e}^t\}_{t\geq 0}$ *follow iterates of Algorithm 1, and Assumptions 1–3 hold. Consider the definition of virtual sequence, given in* (A.1). *When* $\alpha \leq 1/4\overline{L}$ *for all* $t \geq 0$, *we have*

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^\star\right\|^2\right] \leq \left(1 - \frac{\mu\alpha}{2}\right)\mathbb{E}\left[\left\|\widetilde{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] + 3\overline{L}\alpha\mathbb{E}\left[\left\|\boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\|^2\right] - \frac{\alpha}{2}\mathbb{E}\left[F(\boldsymbol{w}^t) - F^\star\right] + \frac{\alpha^2\sigma^2}{M} \,. \tag{A.6}$$

Lemma 4. *Let* $\{\boldsymbol{w}^t, \boldsymbol{v}^t, \boldsymbol{e}^t\}_{t\geq 0}$ *follow iterates of Algorithm 1, and Assumptions 1 and 3 hold. When* $\alpha \leq 1/2\overline{L}$ *for all* $t \geq 0$, *the virtual sequence, defined in* (A.1) *follows*

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{w}}^{t+1})\right] \leq \mathbb{E}\left[F(\widetilde{\boldsymbol{w}}^t)\right] - \frac{\alpha}{4}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right] + \frac{\alpha\overline{L}^2}{2}\mathbb{E}\left[\left\|\boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\|^2\right] + \frac{\alpha^2\sigma^2\overline{L}}{2M} \,. \tag{A.7}$$

Lemma 5. *Let sequence* $(\boldsymbol{e}_m^t)_{t\geq 0}$ *follows iterates of Algorithm 1 and assume inequality* $\|\nabla f_m(\boldsymbol{w})\|^2 \leq G^2$ *hold for every* m *and* $\boldsymbol{w}$ *and some positive* G. *It holds:*

$$\mathbb{E}\left[\left\|\boldsymbol{e}_m^t\right\|^2\right] \leq \beta\alpha^2\left(G^2 + \sigma^2\right) \,. \tag{A.8}$$

Lemma 6 (Based on Appendix A.2 of (Koloskova et al., 2020)). *Let $(r_t)_{t\geq 0}$ and $(s_t)_{t\geq 0}$ be sequences of positive numbers satisfying*

$$r_{t+1} \leq (1-\alpha A)r_t - B\alpha s_t + C\alpha^2 + D\alpha^3 \,,$$

for some positive constants $A, B > 0$, $C, D \geq 0$, and for constant step-sizes $0 < \alpha \leq \frac{1}{E}$, for $E \geq 0$. Then there exists a constant stepsize $\alpha \leq \frac{1}{E}$ such that

$$\frac{B}{W_T}\sum_{t=0}^{T} w_t s_t \leq 2r_0 E \exp\left[-\frac{A(T+1)}{E}\right] + \frac{C(1+\ln\tau)}{A(T+1)} + \frac{D\ln\tau(1+\ln\tau)}{A^2(T+1)^2}$$

for $w_t := (1-\alpha A)^{-(t+1)}$, $W_T := \sum_{t=0}^{T} w_t$ and

$$\tau = \max\left\{2, \min\left\{\frac{A^2 r_0 (T+1)^2}{C}, \frac{A(T+1)}{E\exp\left[-A(T+1)/E\right]}\right\}\right\} \tag{A.9}$$

Lemma 7. *Let $(r_t)_{t\geq 0}$ and $(s_t)_{t\geq 0}$ be sequences of positive numbers satisfying*

$$r_{t+1} \leq r_t - B\alpha s_t + C\alpha^2 + D\alpha^3 \,,$$

for some positive constants $B > 0$, $C, D \geq 0$, and for constant step-sizes $0 < \alpha \leq \frac{1}{E}$, for $E \geq 0$. Then there exists a constant stepsize $\alpha \leq \frac{1}{E}$ such that

$$\frac{B}{T+1}\sum_{t=0}^{T} s_t \leq \frac{Er_0}{T+1} + 2D^{1/3}\left(\frac{r_0}{T+1}\right)^{2/3} + 2\left(\frac{Cr_0}{T+1}\right)^{1/2} . \tag{A.10}$$

B Proofs

B.1 Lemma 3

By definition (A.1),

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^\star\right\|^2 &\leq \left\|\widetilde{\boldsymbol{w}}^{t} - \frac{1}{M}\sum_{m\in[M]} \alpha \boldsymbol{g}_m^t - \boldsymbol{w}^\star\right\|^2 \\
&\leq \left\|\widetilde{\boldsymbol{w}}^{t} - \boldsymbol{w}^\star\right\|^2 + \alpha^2\left\|\frac{\sum_m \boldsymbol{g}_m^t}{M}\right\|^2 - \frac{2\alpha}{M}\sum_m \left\langle \boldsymbol{g}_m^t, \boldsymbol{w}^t - \boldsymbol{w}^\star\right\rangle \\
&\quad + \frac{2\alpha}{M}\sum_m \left\langle \boldsymbol{g}_m^t, \boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\rangle .
\end{aligned}$$

From (A.5) and Assumption 3,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\boldsymbol{w}}^{t+1} - \boldsymbol{w}^\star\right\|^2 \mid \widetilde{\boldsymbol{w}}^t\right] &\leq \left\|\widetilde{\boldsymbol{w}}^{t} - \boldsymbol{w}^\star\right\|^2 + 2\alpha^2\overline{L}\left(F(\boldsymbol{w}^t) - F^\star\right) + \frac{\alpha^2\sigma^2}{M} \\
&\quad - 2\alpha\left\langle \nabla F_m(\boldsymbol{w}^t), \boldsymbol{w}^t - \boldsymbol{w}^\star\right\rangle + 2\alpha\left\langle \nabla F(\boldsymbol{w}^t), \boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\rangle \\
&\overset{(a)}{\leq} \left\|\widetilde{\boldsymbol{w}}^{t} - \boldsymbol{w}^\star\right\|^2 + 2\alpha^2\overline{L}\left(F(\boldsymbol{w}^t) - F^\star\right) + \frac{\alpha^2\sigma^2}{M} \\
&\quad - \alpha\left(\mu\|\boldsymbol{w}^t - \boldsymbol{w}^\star\|^2 + 2\left(F(\boldsymbol{w}^t) - F^\star\right)\right) + 2\alpha\left\langle \nabla F(\boldsymbol{w}^t), \boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\rangle .
\end{aligned} \tag{A.11}$$

where (a) is due to the quasi-convexity of F. Moreover, inequality (A.3) yields

$$\begin{aligned}
2\alpha\left\langle \nabla F(\boldsymbol{w}^t), \boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\rangle &\leq \frac{\alpha}{2\overline{L}}\left\|\nabla F(\boldsymbol{w}^t)\right\|^2 + 2\alpha\overline{L}\left\|\boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\|^2 \\
&\leq \alpha\left(F(\boldsymbol{w}^t) - F^\star\right) + 2\alpha\overline{L}\left\|\boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\|^2 .
\end{aligned} \tag{A.12}$$

Similarly, (A.2) yields

$$-\mu\alpha\|\boldsymbol{w}^t-\boldsymbol{w}^\star\|^2 \le -\frac{\mu\alpha}{2}\left\|\widetilde{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^2+\mu\alpha\left\|\boldsymbol{w}^t-\widetilde{\boldsymbol{w}}^t\right\|^2 . \tag{A.13}$$

Substituting (A.12) and (A.13) into (A.11) and rearranging the terms give

$$\begin{aligned}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{w}}^{t+1}-\boldsymbol{w}^\star\right\|^2 \mid \widetilde{\boldsymbol{w}}^t\right] \le \left(1-\frac{\mu\alpha}{2}\right)\left\|\widetilde{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^2+\alpha\left(2\alpha\overline{L}-1\right)\left(F(\boldsymbol{w}^t)-F^\star\right)+\frac{\alpha^2\sigma^2}{M}\\ +\alpha\left(2\overline{L}+\mu\right)\|\boldsymbol{w}^t-\widetilde{\boldsymbol{w}}^t\|^2 ,\end{aligned}$$

which completes the proof noting that $\alpha \le 1/4\overline{L}$ and $\mu \le \overline{L}$.

B.2 Lemma 4

Notice that F is $\overline{L}$-smooth, so

$$\begin{aligned}F(\widetilde{\boldsymbol{w}}^{t+1}) &\le F(\widetilde{\boldsymbol{w}}^t)-\left\langle\nabla F(\widetilde{\boldsymbol{w}}^t),\frac{\alpha}{M}\sum_m \boldsymbol{g}_m^t\right\rangle+\frac{\alpha^2\overline{L}}{2}\left\|\frac{\sum_m \boldsymbol{g}_m^t}{M}\right\|^2\\ &= F(\widetilde{\boldsymbol{w}}^t)-\alpha\left\langle\nabla F(\widetilde{\boldsymbol{w}}^t),\nabla F(\boldsymbol{w}^t)+\frac{1}{M}\sum_m \boldsymbol{\xi}_m^t\right\rangle+\frac{\alpha^2\overline{L}}{2}\left\|\frac{\sum_m \boldsymbol{g}_m^t}{M}\right\|^2 .\end{aligned}$$

Taking expectation with respect to $\{\boldsymbol{\xi}_m^t\}_{m\in[M]}$ and using (A.4) yield

$$\begin{aligned}\mathbb{E}\left[F(\widetilde{\boldsymbol{w}}^{t+1}) \mid \boldsymbol{w}^t\right] &\le F(\widetilde{\boldsymbol{w}}^t)-\alpha\left\langle\nabla F(\widetilde{\boldsymbol{w}}^t),\nabla F(\boldsymbol{w}^t)\right\rangle+\frac{\alpha^2\overline{L}}{2}\left\|\nabla F(\boldsymbol{w}^t)\right\|^2+\frac{\sigma^2\alpha^2\overline{L}}{2M}\\ &\le F(\widetilde{\boldsymbol{w}}^t)+\alpha\left(\frac{\alpha\overline{L}}{2}-1\right)\left\|\nabla F(\boldsymbol{w}^t)\right\|^2+\alpha\left\langle\nabla F(\boldsymbol{w}^t)-\nabla F(\widetilde{\boldsymbol{w}}^t),\nabla F(\boldsymbol{w}^t)\right\rangle\\ &\quad+\frac{\sigma^2\alpha^2\overline{L}}{2M} .\end{aligned}$$

From (A.3),

$$\begin{aligned}\alpha\left\langle\nabla F(\boldsymbol{w}^t)-\nabla F(\widetilde{\boldsymbol{w}}^t),\nabla F(\boldsymbol{w}^t)\right\rangle &\le \frac{\alpha}{2}\left(\left\|\nabla F(\boldsymbol{w}^t)-\nabla F(\widetilde{\boldsymbol{w}}^t)\right\|^2+\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right)\\ &\overset{\text{(a)}}{\le}\frac{\alpha\overline{L}^2}{2}\left\|\boldsymbol{w}^t-\widetilde{\boldsymbol{w}}^t\right\|^2+\frac{\alpha}{2}\left\|\nabla F(\boldsymbol{w}^t)\right\|^2 ,\end{aligned}$$

where (a) is due to the smoothness of F. Therefore,

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{w}}^{t+1}) \mid \boldsymbol{w}^t\right] \le F(\widetilde{\boldsymbol{w}}^t)+\frac{\alpha}{2}\left(\alpha\overline{L}-1\right)\left\|\nabla F(\boldsymbol{w}^t)\right\|^2+\frac{\alpha\overline{L}^2}{2}\left\|\boldsymbol{w}^t-\widetilde{\boldsymbol{w}}^t\right\|^2+\frac{\sigma^2\alpha^2\overline{L}}{2M} .$$

which completes the proof noting that $\alpha \le 1/2\overline{L}$.

B.3 Lemma 5

From Algorithm 1,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{e}_m^{t+1}\right\|^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{e}_m^t+\alpha\boldsymbol{g}_m^t-\boldsymbol{v}_m^t\right\|^2\right] \overset{(4)}{\le} \beta\mathbb{E}\left[\left\|\alpha\boldsymbol{g}_m^{t-1}\right\|^2\right]\\ &= \beta\alpha^2\mathbb{E}\left[\left\|\nabla f_m(\boldsymbol{w}_{t-1})+\boldsymbol{\xi}_m^{t-1}\right\|^2\right]\\ &\overset{(a)}{\le}\beta\alpha^2\left(G^2+\sigma^2\right) ,\end{aligned} \tag{A.14}$$

where (a) is due to $\|\nabla f_m(\boldsymbol{w})\| \le G^2$ for all $\boldsymbol{w}$ and $m \in [M]$.

Similarly, if we use the average over the last silent window, given by (7), for the RHS of the update rule, from (A.4), we end up to the same bound $\mathbb{E}\left[\left\|\boldsymbol{e}_m^{t+1}\right\|^2\right] \le \beta\alpha^2\left(G^2+\sigma^2\right)$ by noticing that $\sigma^2/(t_m^i-t_m^{i-1}) \le \sigma^2$.

B.4 Lemma 6

After rearranging and multiplying by w_t we obtain

$$Bw_t s_t \leq \frac{(1-\alpha A)w_t r_t}{\alpha} - \frac{w_t r_{t+1}}{\alpha} + \alpha C + \alpha^2 D \,.$$

Observing that that $w_t(1-\alpha A) = w_{t-1}$ we obtain a telescoping sum,

$$\frac{B}{W_T}\sum_{t=0}^{T} w_t s_t \leq \frac{1}{\alpha W_T}\left((1-\alpha A)w_0 r_0 - w_T r_{T+1}\right) + \alpha C + \alpha^2 D \leq \frac{r_0}{\alpha W_T} - \frac{w_T r_{T+1}}{\alpha W_T} + \alpha C + \alpha^2 D \,.$$

Using that $W_T = w_T \sum_{t=0}^{T}(1-\alpha A)^t \leq \frac{w_T}{\alpha A}$ and $W_T \geq w_T = (1-\alpha A)^{-(T+1)}$ we can simplify

$$\begin{aligned}\frac{B}{W_T}\sum_{t=0}^{T} w_t s_t + A r_{T+1} &\leq \frac{(1-\alpha A)^{T+1} r_0}{\alpha} + \alpha C + \alpha^2 D \\ &\leq \frac{r_0}{\alpha}\exp\left[-\alpha A(T+1)\right] + \alpha C + \alpha^2 D =: \Psi_T \end{aligned} \tag{A.15}$$

Now the lemma follows by tuning α in the same way as in (Stich, 2019b, Lemma 2) (slightly more carefully):

- If $\frac{1}{E} \geq \frac{\ln \tau}{A(T+1)}$ then we choose $\alpha = \frac{\ln \tau}{A(T+1)}$ and get (after some simple calculations) that

$$\Psi_T \leq \frac{1}{\ln \tau}\max\left\{\frac{C}{A(T+1)}, r_0 E \exp\left[-\frac{A(T+1)}{E}\right]\right\} + \frac{C\ln\tau}{A(T+1)} + \frac{D\ln^2\tau}{A^2(T+1)^2}$$

- Otherwise $\frac{1}{E} \leq \frac{\ln \tau}{A(T+1)}$ and we pick $\alpha = \frac{1}{E}$ and get that

$$\begin{aligned}\Psi_T &\leq r_0 E \exp\left[-\frac{A(T+1)}{E}\right] + \frac{C}{E} + \frac{D}{E^2} \\ &\leq r_0 E \exp\left[-\frac{A(T+1)}{E}\right] + \frac{C\ln\tau}{A(T+1)} + \frac{D\ln\tau}{A^2(T+1)^2}\end{aligned}$$

B.5 Lemma 7

Rearranging and dividing by $\alpha > 0$ gives

$$Bs_t \leq \frac{r_t}{\alpha} - \frac{r_{t+1}}{\alpha} + C\alpha + D\alpha^2$$

and summing from $t = 0$ to T yields

$$\frac{B}{T+1}\sum_{t=0}^{T} \leq \frac{r_0}{\alpha(T+1)} + C\alpha + D\alpha^2 \,.$$

Now the claim follows from (Koloskova et al., 2020, Lemma 15).

B.6 Upload Intervals

In this we will estimate how often uploads are triggered in Algorithm 1 for a specific worker m. If we denote by t_m^i the iteration indices when node m sends an update to the server, $t_m^i + 1 \leq t_m^{i+1}$, then we are thus interested to estimate $t_m^{i+1} - t_m^i$.

Instead of considering the momentum based estimator as proposed in (6), let us first consider the choice $\boldsymbol{q}_m^{t_m^i} := \nabla f_m(\boldsymbol{w}^{t_m^i})$ for simplicity.

In this section, let $a_m^i := \max\left\{\left\|\boldsymbol{w}^{t_m^i} - \boldsymbol{w}^\star\right\|, \ldots, \left\|\boldsymbol{w}^{t_m^{i+1}} - \boldsymbol{w}^\star\right\|\right\}$.

It holds $\boldsymbol{e}_m^{t_m^i+1} = \mathbf{0}$ for all i, and $\boldsymbol{v}_m^t = \alpha \nabla f_m(\boldsymbol{w}^{t_m^i})$ for $t_m^i + 1 \leq k \leq t_m^{i+1}$, by definition of the algorithm. We will now estimate how fast the left hand side in our criterion grows.

Observation 1. It holds

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{e}_m^{t_m^{i+1}} + \alpha \boldsymbol{g}_m^{t_m^{i+1}} - \boldsymbol{v}_m^{t_m^{i+1}}\right\|^2\right] &= \mathbb{E}\left[\left\|\sum_{t=t_m^i+1}^{t_m^{i+1}} \alpha\left(\nabla f_m(\boldsymbol{w}^t) + \boldsymbol{\xi}_m^t - \nabla f_m(\boldsymbol{w}_{t_m^i})\right)\right\|^2\right] \\
&= \alpha^2 \mathbb{E}\left[\left\|\sum_{t=t_m^i+1}^{t_m^{i+1}} \left(\nabla f_m(\boldsymbol{w}^t) - \nabla f_m(\boldsymbol{w}^{t_m^i})\right)\right\|^2\right] \\
&\quad + \alpha^2 \mathbb{E}\left[\left\|\sum_{t=t_m^i+1}^{t_m^{i+1}} \boldsymbol{\xi}_m^t\right\|^2\right] \\
&\leq \alpha^2 (t_m^{i+1} - t_m^i) \sum_{t=t_m^i+1}^{t_m^{i+1}} \mathbb{E}\left[\left\|\nabla f_m(\boldsymbol{w}^t) - \nabla f_m(\boldsymbol{w}^{t_m^i})\right\|^2\right] \\
&\quad + \alpha^2 (t_m^{i+1} - t_m^i)\sigma^2 \\
&\leq 4\alpha^2 (t_m^{i+1} - t_m^i)^2 L_m^2 (a_m^i)^2 + \alpha^2 (t_m^{i+1} - t_m^i)\sigma^2 \qquad \text{(A.16)}
\end{aligned}$$

using

$$\begin{aligned}
\left\|\nabla f_m(\boldsymbol{w}^t) - \nabla f_m(\boldsymbol{w}^{t_m^i})\right\|^2 &\leq 2\left\|\nabla f_m(\boldsymbol{w}^t) - \nabla f_m(\boldsymbol{w}^\star)\right\|^2 + 2\left\|\nabla f_m(\boldsymbol{w}^{t_m^i}) - \nabla f_m(\boldsymbol{w}^\star)\right\|^2 \\
&\leq 4L_m^2 (a_m^i)^2
\end{aligned}$$

Observation 2. On other hand,

$$\begin{aligned}
\mathbb{E}\left[\left\|\alpha \boldsymbol{g}_m^{t_m^{i+1}}\right\|^2\right] &= \left(\mathbb{E}\left[\left\|\alpha \nabla f_m(\boldsymbol{w}^{t_m^{i+1}})\right\|^2\right] + \alpha^2\sigma^2\right) \\
&\geq \frac{\alpha^2}{2}\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 + \alpha^2\sigma^2 - \alpha^2\left\|\nabla f_m(\boldsymbol{w}^{t_m^{i+1}}) - \nabla f_m(\boldsymbol{w}^\star)\right\|^2 \\
&\geq \frac{\alpha^2}{2}\left(\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 + 2\sigma^2 - 2L_m^2(a_m^i)^2\right) \\
&\geq \frac{\alpha^2}{2}\left(\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 - 2(t_m^{i+1} - t_m^i)^2 L_m^2 (a_m^i)^2\right) \qquad \text{(A.17)}
\end{aligned}$$

If we change the RHS of the upload criteria to $\frac{1}{t_m^{i+1} - t_m^i}\sum_{t=t_m^i+1}^{t_m^{i+1}} \alpha \boldsymbol{g}_m^t$, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{t_m^{i+1} - t_m^i}\sum_{t=t_m^i+1}^{t_m^{i+1}} \alpha \boldsymbol{g}_m^t\right\|^2\right] &= \left(\mathbb{E}\left[\left\|\frac{\alpha}{t_m^{i+1} - t_m^i}\sum_{t=t_m^i+1}^{t_m^{i+1}} \nabla f_m(\boldsymbol{w}^{t_m^{i+1}})\right\|^2\right] + \frac{\alpha^2\sigma^2}{t_m^{i+1} - t_m^i}\right) \\
&\geq \frac{\alpha^2}{2}\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 \\
&\quad - \frac{\alpha^2}{t_m^{i+1} - t_m^i}\sum_{t=t_m^i+1}^{t_m^{i+1}} \left\|\nabla f_m(\boldsymbol{w}^t) - \nabla f_m(\boldsymbol{w}^\star)\right\|^2 \\
&\geq \frac{\alpha^2}{2}\left(\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 - 2L_m^2(a_m^i)^2\right) \\
&\geq \frac{\alpha^2}{2}\left(\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 - 2(t_m^{i+1} - t_m^i)^2 L_m^2 (a_m^i)^2\right),
\end{aligned}$$

which is the same as of (A.17).

Summary. Combining Observation 1 and 2, we conclude that as long as

$$\beta \left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2 \geq 10 \left(t_m^{i+1} - t_m^i\right)^2 L_m^2 (a_m^i)^2 + 2(t_m^{i+1} - t_m^i)\sigma^2 \,, \tag{A.18}$$

then

$$\mathbb{E}\left[\left\|\boldsymbol{e}_m^{t_m^{i+1}} + \alpha \boldsymbol{g}_m^{t_m^{i+1}} - \boldsymbol{v}_m^{t_m^{i+1}}\right\|^2\right] \leq \beta\, \mathbb{E}\left[\left\|\alpha \boldsymbol{g}_m^{t_m^{i+1}}\right\|^2\right] \,,$$

and also

$$\mathbb{E}\left[\left\|\boldsymbol{e}_m^{t_m^{i+1}} + \alpha \boldsymbol{g}_m^{t_m^{i+1}} - \boldsymbol{v}_m^{t_m^{i+1}}\right\|^2\right] \leq \beta\, \mathbb{E}\left[\left\|\frac{1}{t_m^{i+1} - t_m^i} \sum_{t=t_m^i+1}^{t_m^{i+1}} \alpha \boldsymbol{g}_m^t\right\|^2\right] \,,$$

Consequently, *in expectation* no upload is triggered. In other words, by rearranging (A.18), we get the estimate

$$\frac{t_m^{i+1} - t_m^i}{\beta} = \mathcal{O}\left(\frac{\left\|\nabla f_m(\boldsymbol{w}^\star)\right\|}{L_m a_m^i} + \frac{\beta \left\|\nabla f_m(\boldsymbol{w}^\star)\right\|^2}{\sigma^2}\right) \tag{A.19}$$

.

So far, we have argued only on the upload frequency of the idealistic choice $\boldsymbol{q}_m^{t_m^i} := \nabla f_m(\boldsymbol{w}^{t_m^i})$, i.e. the gradient evaluated on a full batch. It is easy to check, that our estimates will not significantly change when using a biased version, $\boldsymbol{q}_m^{t_m^i} = \nabla f_m(\boldsymbol{w}^{t_m^i}) + \boldsymbol{\xi}$ instead, as long as the bias is sufficiently small, that is $\|\boldsymbol{\xi}\|^2 \leq \frac{1}{t_m^{i+1} - t_m^i}\sigma^2$. This can for instance achieved by averaging $(t_m^{i+1} - t_m^i)$ stochastic gradients (which can be cheaper than evaluating on the full batch). Furthermore, by smoothness, we observe that it is not essential to estimate the gradient at the point $\boldsymbol{w}^{t_m^i}$ precisely, instead gradients evaluated at points close by will behave similary. Therefore, we suggest average-based estimators, where either gradients are (locally on node m) averaged over a window of increasing size, or, using weighted average of past gradients. Our proposed momentum based estimator for $\boldsymbol{q}_m^t$ as in equation (6) performs an exponentially weighted averaging over past iterates, thereby averaging the noise over a horizon of roughly steps $\frac{1}{\gamma}$.

B.7 Proof of Proposition 1

Note that

$$\left\|\boldsymbol{w}^t - \widetilde{\boldsymbol{w}}^t\right\|^2 = \left\|\frac{1}{M}\sum_{m\in[M]} \boldsymbol{e}_m^t\right\|^2 \leq \frac{1}{M}\sum_{m\in[M]} \|\boldsymbol{e}_m^t\|^2 \tag{A.20}$$

holds by the convexity of squared norm operator.

Claim 1: Define $s_t := \mathbb{E}\left[F(\boldsymbol{w}^t) - F^\star\right]$ and $r_t := \mathbb{E}\left[\|\widetilde{\boldsymbol{w}}^t - \boldsymbol{w}^\star\|^2\right]$. From Lemma 3, (A.20), and Lemma 5, we obtain the inequality

$$r_{t+1} \leq \left(1 - \frac{\alpha\mu}{2}\right) r_t - \alpha\frac{s_t}{2} + \alpha^2\frac{\sigma^2}{M} + \alpha^3 3\beta\overline{L}\left(G^2 + \sigma^2\right) \tag{A.21}$$

The proof follows from Lemma 6.

Claim 2: By setting $\mu = 0$ in Equation (A.21) we obtain the inequality

$$r_{t+1} \leq r_t - \alpha\frac{s_t}{2} + \alpha^2\frac{\sigma^2}{M} + \alpha^3 3\beta\overline{L}\left(G^2 + \sigma^2\right)$$

The proof follows from Lemma 7.

Claim 3: Define $r_t := \mathbb{E}\left[F(\widetilde{\boldsymbol{w}}^t) - F^\star\right]$ and $s_t := \mathbb{E}\left[\|\nabla F(\boldsymbol{w}^t)\|^2\right]$. From Lemma 4, (A.20), and Lemma 5, we obtain the inequality

$$r_{t+1} \leq r_t - \alpha\frac{s_t}{4} + \alpha^2\frac{\sigma^2\overline{L}}{2M} + \alpha^3\frac{\overline{L}^2(1-\beta)}{\beta}\left(\frac{2G^2}{\beta} + \sigma^2\right) \,.$$

The proof follows from Lemma 7.

C Additional Experiments

C.1 Impact of β

Figure A.1 shows the impact of β. Adopting a high value for β would lead to a unnecessarily high upload rate, especially toward the end of iterations. Extension to adaptive β would is an interesting future direction.

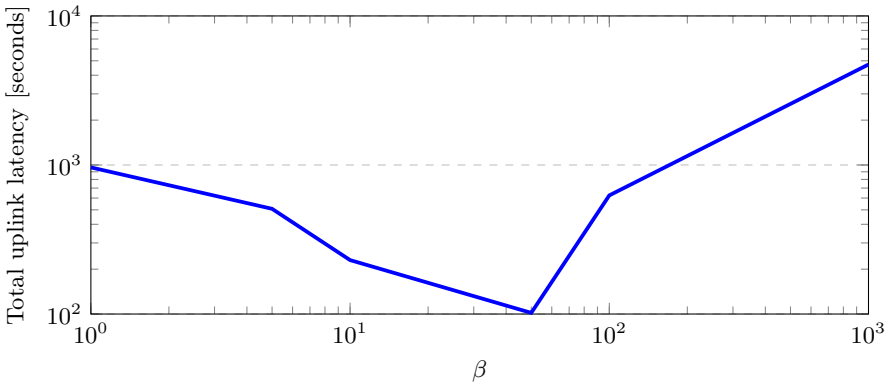


Figure A.1: Impact of β for a network of $M = 500$ nodes.

C.2 Reducing Noisy Uploads

LENA algorithm may trigger the upload criteria due to some unfortunate noisy gradients. To make it more robust, we have changed the line 8 of Algorithm 1 to activate the upload trigger once that condition is multiple times (over a window of some size). Figure A.2 shows the impact of this window size on the upload decision criteria. From the figure, a window size of 2 is enough to effectively eliminate unnecessary uploads.

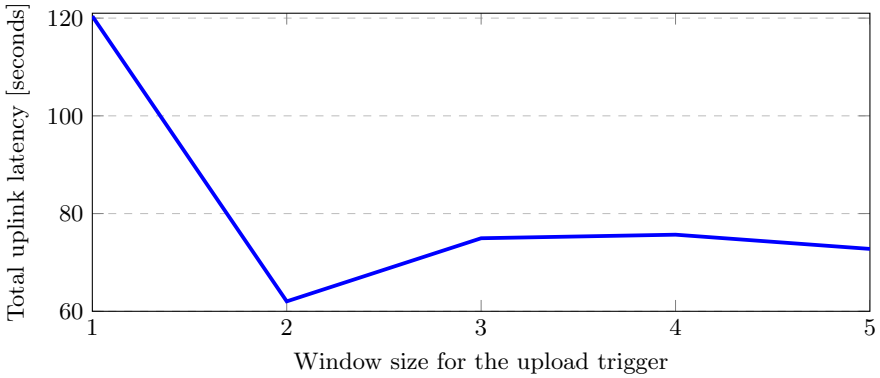
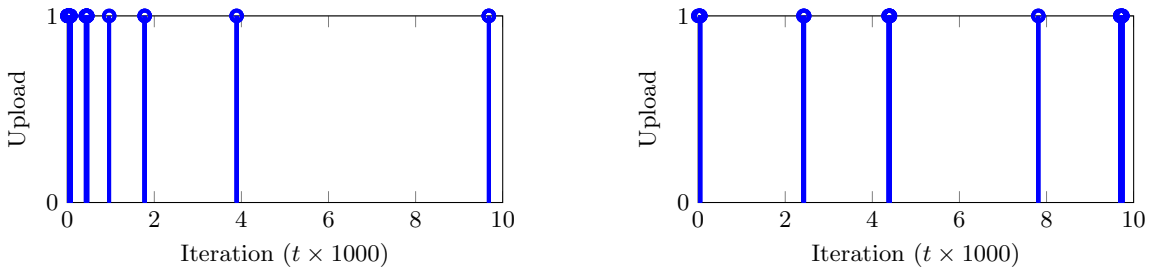


Figure A.2: Impact of upload trigger window size on the latency.

C.3 Impact of local smoothness on upload frequency

Figure A.3 shows the upload events of two nodes with different local smoothness parameters, L_m. We have computed the maximum eigenvalue of the hessian matrix for both nodes, with respect to their private datasets, leading to $L_{20} = 3.9$ and $L_{180} = 4.7$. A higher L_m has led to a higher upload rate. Notice that when momentum is applied to $\boldsymbol{q}_m$, it may average the noise and change the upload rate. Moreover, for the case of IID datasets where $\|\nabla f_m\| \to 0$ for all m, we have observed a higher upload frequency for all nodes, irrespective of their local smoothness, which is in agreement with (9).



(a) Node number 20. (b) Node number 180.

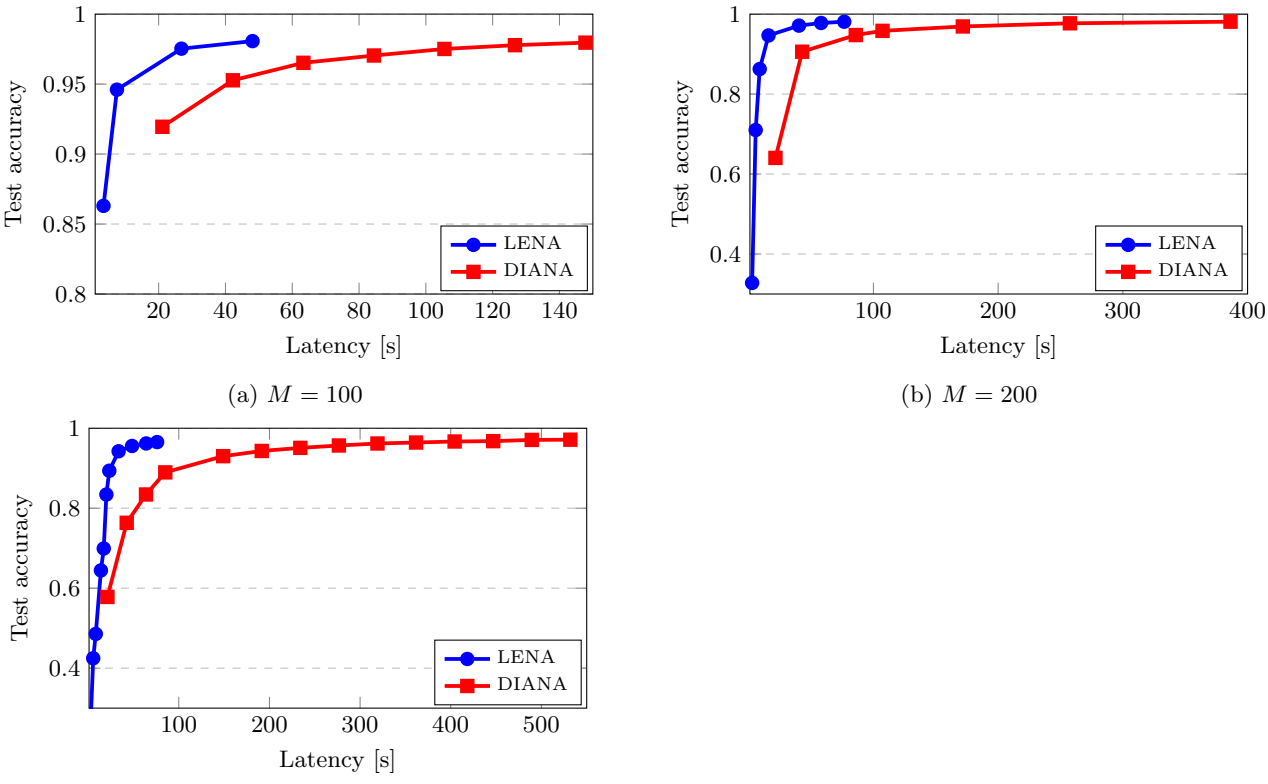
Figure A.3: Upload events for node 20 and 180, $M = 200$ and minibatch size of 15 with $\boldsymbol{q}_m^t = \boldsymbol{v}_m^{t-1}$ upon being silent.

C.4 Quantization

To further save the bandwidth, one can modify (4) as

$$\boldsymbol{v}_m^t = \begin{cases} \text{Quantize}\left(\boldsymbol{e}_m^t + \alpha \boldsymbol{g}_m^t\right) & \text{if } \|\boldsymbol{e}_m^t + \alpha \boldsymbol{g}_m^t - \boldsymbol{v}_m^{t-1}\|^2 \geq \beta \|\boldsymbol{g}_m^{t-1}\|^2 \\ \text{Quantize}\left(\boldsymbol{q}_m^t\right) & \text{otherwise}\,, \end{cases} \tag{A.22}$$

where Quantize is a quantization operator. Figure A.4 shows the performance of LENA after this modification. Optimizing over the quantizer is left as a future work.



(a) $M = 100$ (b) $M = 200$

(c) $M = 500$

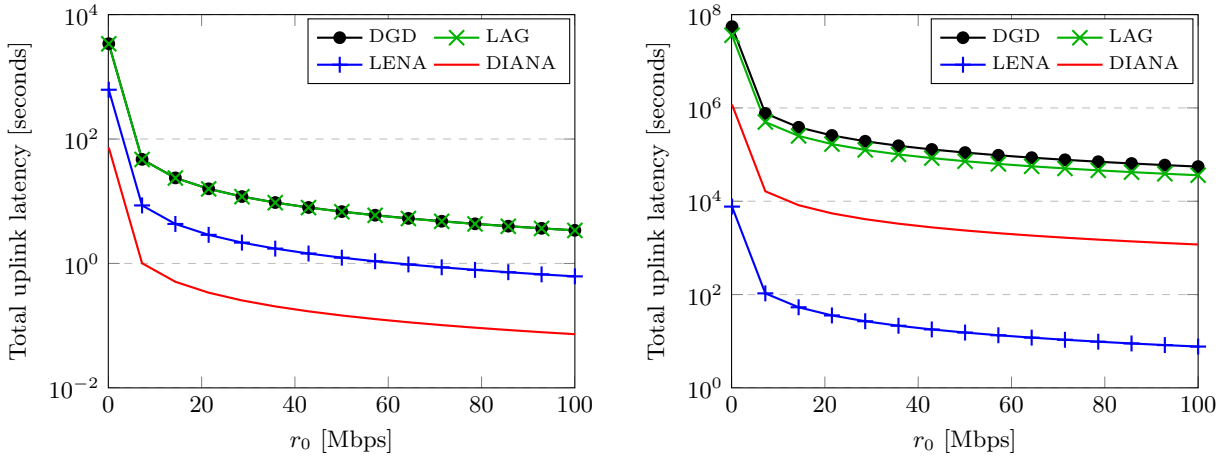
Figure A.4: Comparison of LENA with new upload criteria of (A.22) and DIANA on a deep learning model for MNIST after hyper-parameter optimization on both models and applying QSGD with 4 quantization levels to LENA uploads.

C.5 Shared Wireless Channel

Define the so-called offered load $\ell^t > 0$ as the average number of packets (gradients) that the nodes inject to the wireless channel at iteration t (Bertsekas et al., 2004). The successful transmission rate model at iteration t follows $r^t = r_0 \ell^t \exp\{-\ell^t / r_1\}$ for some positive constants r_0 and r_1 describing the wireless channel and communication protocol (Bertsekas et al., 2004). This rate model implies that having too many active transmitters may make the channel "congested" and increasing the number of transmitted packets (offered loads) would only add to the congestion and packet drops, leading to a lower success rate. For DIANA, $\ell = M$ for every iteration, as all the nodes always upload their gradients. By defining ν_m^t as the number of bits uploaded by node m at iteration t, the communication latency is then $\sum_{m\in[M]} \sum_{t\in[T]} \nu_m^t / r^t$ for the first T iterations. Notice that we have ignored the download latency as the capacity of the broadcast channel would not be affected by the number of nodes.

To simulate various network models, we consider $r_1 = 10$ and two scenarios: small network ($M = 10$) and large network ($M = 100$). Figure A.5 shows the results. In all scenarios, the latency reduces as the network capacity (r_0) increases. With $M = 10$, the network is under-utilized, and activating all transmitters sending fewer bits over the channel is feasible at almost no extra penalty in terms of rate or latency. Consequently, DIANA achieves the best performance. However, as the channel contention level increases, by adopting a larger M, being silent in LENA shows a clear benefit over a mere compression. Notice that the gain of LENA can be further improved by complementing it with compression.

Downlink Efficiency: The LAG algorithm has the option of multicasting, namely sending the parameter to a subset of nodes in the downlink. The LENA algorithm, on the other hand, needs to braodcast it to all nodes, which locally decide to send back their computations or remain silent. In certain communication systems, the broadcasting may consume more network resources than multicasting. However, in some other cases, like wireless communications, broadcasting is for free, and therefore sending the parameter to a subset of users in downlink has almost the same cost as sending it to all users.



(a) $M = 10$. (b) $M = 100$.

Figure A.5: Communication latency over a shared wireless network, with minibatch size of 2 at every node.

C.6 CIFAR10 Results

Figure A.6 compares the performance of LENA and DIANA on VGG13 model, trained on CIFAR10 dataset. Again, LENA substantially outperforms DIANA. This performance gain can be further boosted buy adding quantization step to LENA, as formalized in (A.22).

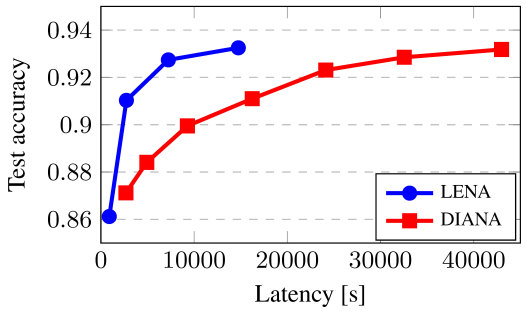


Figure A.6: Performance of LENA and DIANA on VGG13 model, trained on CIFAR10 dataset.

C.7 Comparison to FedAvg

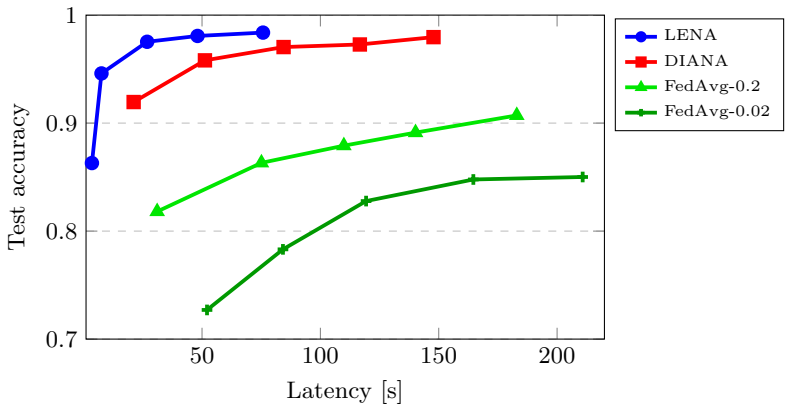


Figure A.7: Comparison to FedAvg with $M = 100$. 'FedAvg-x' shows x percent of nodes (randomly selected) are active in every iteration.

Figure A.7 shows the comparison of DIANA and LENA to the vanilla FedAvg algorithm (McMahan et al., 2017). Simulation setting and hyper-parameter optimization of FedAvg are the same as of Appendix C.4. FedAvg-0.02 has the same overall upload rate as of LENA, but achieves a much lower performance. The reason is the non-IID data. In fact, FedAvg activates a subset of nodes but it does not take into account importance of their info in the scheduling/activation strategy. It can be the case the we miss important updates in many iterations. However, LENA intelligently chooses the silent nodes based on their upload significance. The performance of FedAvg can be improved by employing a higher upload rate (higher chance of collecting important updates). Yet, it cannot still beat LENA or DIANA due to their ability to handle severe quantization.