LocalNewton: Reducing Communication Rounds for Distributed Learning (Supplementary material)

Vipul Gupta[1] **Avishek Ghosh**[1] **Michał Dereziński**[2] **Rajiv Khanna**[2] **Kannan Ramchandran**[1] **Michael W. Mahoney**[2]

[1]Department of EECS, University of California, Berkeley
[2]Department of Statistics, University of California, Berkeley

1 AUXILIARY LEMMAS AND THEIR PROOFS

Here, we prove the auxiliary lemmas that are used in the main proofs of the paper. (For completeness, we restate the lemma statements).

Lemma 1.1. *Let $f(\cdot)$ satisfy assumptions 1-4 and $0 < \epsilon \leqslant 1/2$ and $\delta < 1$ be fixed constants. Then, if $s \geqslant \frac{4B}{\kappa\epsilon^2}\log\frac{2d}{\delta}$, the local Hessian at the k-th worker satisfies*

$$(1-\epsilon)\kappa \leqslant \nabla^2 f^k(\mathbf{w}) = \mathbf{H}^k(\mathbf{w}) \leqslant (1+\epsilon)M, \tag{1}$$

for all $\mathbf{w} \in \mathbb{R}^d$ and $k \in [K]$ with probability (w.p.) at least $1-\delta$.

Proof. At the k-th worker which samples $\mathcal{S}_k$ observations from $[n]$, the following is true by Matrix Chernoff (see Theorem 2.2 in Tropp (2011))

$$\mathbb{P}(\lambda_{\min}\left(\nabla^2 f^k(\mathbf{w})\right) \leqslant (1-\epsilon)\kappa) \leqslant \delta_1 = d\left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{s\kappa/B}, \tag{2}$$

$$\mathbb{P}(\lambda_{\max}\left(\nabla^2 f^k(\mathbf{w})\right) \geqslant (1+\epsilon)M) \leqslant \delta_2 = d\left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{sM/B}. \tag{3}$$

Now, using the inequality $\log(1-\epsilon) \leqslant \frac{-\epsilon}{\sqrt{1-\epsilon}}$ for $0 \leqslant \epsilon < 1$, we get

$$\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}} \leqslant e^{-\epsilon+\epsilon\sqrt{1-\epsilon}}.$$

Further, utilizing the fact that $\sqrt{1-\epsilon} \leqslant \frac{1}{1+\epsilon/2}$, we get

$$e^{-\epsilon+\epsilon\sqrt{1-\epsilon}} \leqslant e^{\frac{-\epsilon^2}{1+\epsilon/2}} \leqslant e^{-\epsilon^2/4}.$$

Hence, we have $\delta_1 \leqslant de^{-s\kappa\epsilon^2/4B}$. Further, using the fact that $\log(1+\epsilon) \geqslant \epsilon - \epsilon^2/2$, we get

$$\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}} \leqslant e^{-\epsilon^2/2+\epsilon^3/2} \leqslant e^{-\epsilon^2/4},$$

where the last inequality follows from the fact that $\epsilon \leqslant 1/2$. Hence, $\delta_2 \leqslant de^{-sM\epsilon^2/4B}$. Thus, by union bound and subsequently using upper bounds on δ_1 and δ_2, we get

$$\begin{aligned}\mathbb{P}\left[(1-\epsilon)\kappa\mathbf{I} \leqslant \nabla^2 f^k(\mathbf{w}) \leqslant (1+\epsilon)M\mathbf{I}\right] &\geqslant 1-(\delta_1+\delta_2)\\ &\geqslant 1-(de^{-s\kappa\epsilon^2/4B}+de^{-sM\epsilon^2/4B})\\ &\geqslant 1-(2de^{-s\kappa\epsilon^2/4B}),\end{aligned}$$

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

where the last inequality follows from the fact that $\kappa \leqslant M$. Hence, the result follows by noting that

$$(1-\epsilon)\kappa \mathbf{I} \preccurlyeq \nabla^2 f^k(\mathbf{w}) \preccurlyeq (1+\epsilon)M\mathbf{I} \text{ w. p. at least } 1-\delta,$$

and requiring that $\delta \geqslant 2de^{-s\kappa\epsilon^2/4B}$ (or $s \geqslant \frac{4B}{\kappa\epsilon^2}\log\frac{2d}{\delta}$).

□

Lemma 1.2. *Let the function $f(\cdot)$ satisfy assumptions 1-3, and step-size α_t^k that solves the line-search condition in Eq. (5). Also, let $0 < \epsilon \leqslant 1/2$ and $0 < \delta < 1$ be fixed constants. Moreover, let the sample size $s \geqslant \frac{4B}{\kappa\epsilon^2}\log\frac{2d}{\delta}$. Then, the LocalNewton update at the k-th worker satisfy*

$$f^k(\mathbf{w}_{t+1}^k) - f^k(\mathbf{w}_t^k) \leqslant -\psi||\mathbf{g}_t^k||^2 \; \forall \; k \in [K],$$

w.p. at least $1-\delta$, where $\psi = \frac{\alpha^\star\beta}{M(1+\epsilon)}$.

Proof. From Lemma 1.1, we know that $f^k(\cdot)$ is $M(1-\epsilon)$ smooth with probability $1-\delta$. M-smoothness of a function $g(\cdot)$ implies

$$g(\mathbf{y}) - g(\mathbf{x}) \leqslant (\mathbf{y}-\mathbf{x})^T\nabla g(\mathbf{x}) + \frac{M}{2}||\mathbf{y}-\mathbf{x}||^2 \; \forall \; \mathbf{x},\mathbf{y} \in \mathbb{R}^d. \tag{4}$$

Hence,

$$f^k(\mathbf{w}_t^k - \alpha\mathbf{p}_t^k) - f^k(\mathbf{w}_t^k) \leqslant (-\alpha\mathbf{p}_t^k)^T\mathbf{g}^k(\mathbf{w}_t^k) + \frac{M(1-\epsilon)}{2}\alpha^2||\mathbf{p}_t^k||^2. \tag{5}$$

The above inequality is satisfied for all $\alpha \in \mathbb{R}$. We know that α_t^k, the local step-size at worker k, satisfies the line-search constraint in Eq. (5). Thus, for $\alpha_t^k \in (0,1]$ to exist that satisfies the line-search condition, it is enough to find $\alpha > 0$ that satisfies

$$-\alpha(\mathbf{p}_t^k)^T\mathbf{H}_t^k\mathbf{p}_t^k + \frac{M(1-\epsilon)}{2}\alpha^2\|\mathbf{p}_t^k\|^2 \leqslant -\alpha\beta(\mathbf{p}_t^k)^T\mathbf{H}_t^k\mathbf{p}_t^k, \tag{6}$$

where we have used the fact that $\mathbf{g}_t^k = \mathbf{H}_t^k\mathbf{p}_t^k$. Thus, α must satisfy

$$\frac{M(1-\epsilon)}{2}\alpha\|\mathbf{p}_t^k\|^2 \leqslant (1-\beta)(\mathbf{p}_t^k)^T\mathbf{H}_t^k\mathbf{p}_t^k. \tag{7}$$

Using lemma 1.1, we know that for sufficiently large sample-size at the k-th worker, we get

$$(1-\epsilon)\nabla^2 f(\mathbf{w}) \preceq \nabla^2 f^k(\mathbf{w}) \preceq (1+\epsilon)\nabla^2 f(\mathbf{w}) \tag{8}$$

with probability $1-\delta$. Also, by κ-strong convexity of $f(\cdot)$, we know that $\nabla^2 f(\mathbf{w}) \succeq \kappa\mathbf{I}$. Thus, the local line-search constraint is always satisfied for

$$\alpha \leqslant \frac{2(1-\beta)\kappa(1-\epsilon)}{M(1+\epsilon)}.$$

Hence, if we choose $\alpha^\star \leqslant \frac{2(1-\beta)\kappa(1-\epsilon)}{M(1+\epsilon)}$, or $\alpha^\star \leqslant \frac{\kappa(1-\beta)}{M}$ for $\epsilon < 1/2$, we are guaranteed to have the line-search condition from Eq. (5) satisfied with $\alpha_t^k = \alpha^\star$. This is satisfied by the line search equation in Eq. (5). Hence, from the line-search guarantee, we get

$$\begin{aligned} f^k(\mathbf{w}_{t+1}^k) - f^k(\mathbf{w}_t^k) &\leqslant -\alpha^\star\beta(\mathbf{p}_t^k)^T\mathbf{g}_t^k & (9)\\ &= \alpha^\star\beta(\mathbf{g}_t^k)^T(\mathbf{H}_t^k)^{-1}\mathbf{g}_t^k, & (10)\\ &\leqslant -\alpha^\star\beta\frac{1}{M(1+\epsilon)}\|\mathbf{g}_t^k\|^2, & (11)\end{aligned}$$

w.p. $1-\delta$. Here, the last inequality uses the fact that $f^k(\cdot)$ is $M(1+\epsilon)$–smooth, that is, $\mathbf{H}_t^k \preceq M(1+\epsilon)\mathbf{I}$. This proves the desired result. □

2 PROOF OF THEOREM 4.3

The proofs for theorems in this paper use the auxiliary lemmas in Appendix 1.

Proof. The proof of the theorem is based on the following two high probability lower bounds:

Case 1:

$$f(\bar{\mathbf{w}}_t) - f(\bar{\mathbf{w}}_{t+1}) \geqslant C\|\mathbf{g}(\bar{\mathbf{w}}_t)\|^2, \tag{12}$$

where $C = \frac{\alpha^\star \beta(1-\epsilon)}{2M(1+\epsilon)}$ is a constant, and

Case 2

$$f(\bar{\mathbf{w}}_t) - f(\bar{\mathbf{w}}_{t+1}) \geqslant C_1\|\mathbf{g}(\bar{\mathbf{w}}_t)\|^2 - \frac{\eta\Gamma}{\kappa(1-\epsilon)}, \tag{13}$$

where C_1 is a constant (> 0) and $\eta = (1 + \sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\Gamma$.

We will prove the above result shortly, but let us complete the proof of the theorem assuming that Eq. (12) and Eq. (13) are true.

Case 1 (using Eq. (12)) Invoking the κ strong convexity of the the function f we have

$$f(\bar{\mathbf{w}}_t) - f(\mathbf{w}^*) \leqslant \frac{1}{2\kappa}\|\mathbf{g}(\bar{\mathbf{w}}_t)\|^2, \tag{14}$$

where $\bar{\mathbf{w}}^*$ is the unique global minimizer of the function f. Combining the last lower bound with equation (12) we obtain

$$f(\bar{\mathbf{w}}_{t+1}) - f(\bar{\mathbf{w}}_t) \leqslant (1 - 2\kappa C)(f(\bar{\mathbf{w}}_t) - f(\mathbf{w}^*)), \tag{15}$$

with probability $1 - \delta$. Also note that

$$1 > 1 - 2\kappa C = 1 - \frac{\kappa\alpha^\star\beta(1-\epsilon)}{M(1+\epsilon)} > 0,$$

where the last inequality uses the definition of $\alpha^\star$ from Eq. (6). The completes the proof of Theorem 3.2.

Case 2 (using Eq. (13)) Using the same steps as before, and using the condition of Eq. (13), we obtain Theorem 3.2.

It remains to prove the claim (12) and (13).

Proof of the claim (12): Recall that for $L = 1$, we have

$$\mathbf{w}_{t+1}^k = \bar{\mathbf{w}}_t - \alpha_t^k \mathbf{p}_t^k, \quad \text{and} \quad \bar{\mathbf{w}}_{t+1} := \frac{1}{K}\sum_{k=1}^K \mathbf{w}_{t+1}^k = \bar{\mathbf{w}}_t - \frac{1}{K}\sum_{k=1}^K \alpha_t^k \mathbf{p}_t^k,$$

where the $\mathbf{p}_t^k = (\mathbf{H}_t^k)^{-1}\mathbf{g}_t^k, \mathbf{H}_t^k = (\mathbf{H}^k)^{-1}(\bar{\mathbf{w}}_t)$ and $\mathbf{g}_t^k = \mathbf{g}^k(\bar{\mathbf{w}}_t)$. Invoking the M-smoothness of the function $f(\cdot)$ we have

$$\begin{aligned}
f(\bar{\mathbf{w}}_t) - f(\bar{\mathbf{w}}_{t+1}) &\geqslant \frac{-M}{2K^2}\|\bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t+1}\|^2 + \langle \mathbf{g}(\bar{\mathbf{w}}_t), \bar{\mathbf{w}}_t - \bar{\mathbf{w}}_{t+1}\rangle \\
&\geqslant \frac{-M}{2K^2}\left\|\sum_{k=1}^K (\alpha_t^k)\mathbf{p}_t^k\right\|^2 + \langle \mathbf{g}(\bar{\mathbf{w}}_t), \frac{1}{K}\sum_{k=1}^K \alpha_t^k \mathbf{p}_t^k\rangle \\
&\overset{(i)}{\geqslant} \frac{-M}{2K}\sum_{k=1}^K (\alpha_t^k)^2\|\mathbf{p}_t^k\|^2 + \langle \mathbf{g}(\bar{\mathbf{w}}_t), \frac{1}{K}\sum_{k=1}^K \alpha_t^k \mathbf{p}_t^k\rangle \\
&= \frac{1}{K}\sum_{k=1}^K \left(\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) - \frac{M}{2}(\alpha_t^k)^2\|\mathbf{p}_t^k\|^2\right)
\end{aligned} \tag{16}$$

where the inequality (i) uses the following fact

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \mathbf{a}^k \right\|^2 \leqslant \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{a}^k\|^2, \tag{17}$$

for all vectors $\mathbf{a}^1, \mathbf{a}^2, \cdots, \mathbf{a}^K \in \mathbb{R}^d$.

We now complete the proof by using the following bound on the first term in Eq. (16). In particular, In the first case, we show that, for all $k \in [K]$ provided

$$s \gtrsim \left(\frac{\Gamma^2}{\epsilon_1^2 G^2} \log(d/\delta) \right),$$

and $\|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \geqslant G$, where $\epsilon_1 > 0$ (small number), we have

$$\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) \geqslant \left(\psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)(\alpha_t^k)^2}{2} \|\mathbf{p}_t^k\|^2 \tag{18}$$

with probability at least $1 - 4\delta$.

Let us substitute Eq. (18) in equation (16), we get

$$\begin{aligned} f(\bar{\mathbf{w}}_t) - f(\bar{\mathbf{w}}_{t+1}) &\geqslant \frac{1}{K} \sum_{k=1}^{K} \left[\left(\psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|\mathbf{g}_t^k\|^2 - \frac{(M - \kappa(1-\epsilon))(\alpha_t^k)^2}{2} \|\mathbf{p}_t^k\|^2 \right] \\ &\geqslant \frac{1}{K} \sum_{k=1}^{K} \left[\left(\psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|\mathbf{g}_t^k\|^2 - \frac{(M - \kappa(1-\epsilon))(\alpha_t^k)^2}{2\kappa^2(1-\epsilon)^2} \|\mathbf{g}_t^k\|^2 \right] \end{aligned} \tag{19}$$

where the last inequality follows from the fact that the function f^k is $\kappa(1-\epsilon)$ strongly convex with probability $1 - \delta$, and thus

$$\|\mathbf{p}_t^k\|^2 := \|(\mathbf{H}_t^k)^{-1} \mathbf{g}_t^k\|_2^2 \leqslant \|(\mathbf{H}_t^k)^{-1}\|_2^2 \|\mathbf{g}_t^k\|^2 \leqslant \frac{1}{\kappa^2(1-\epsilon)^2} \|\mathbf{g}_t^k\|^2. \tag{20}$$

with probability $1 - \delta$. Now, using the upper bound on α_t^k, we have

$$\begin{aligned} f(\bar{\mathbf{w}}_t) - f(\bar{\mathbf{w}}_{t+1}) &\geqslant \frac{1}{K} \sum_{k=1}^{K} \left[\left(\psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|\mathbf{g}_t^k\|^2 - \frac{(M - \kappa(1-\epsilon)^2)}{2} \frac{\alpha^{\star 2}}{\kappa^2(1-\epsilon)^2} \|\mathbf{g}_t^k\|^2 \right] \\ &= \left(\psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} - \frac{(M - \kappa(1-\epsilon)^2)}{2} \frac{\alpha^{\star 2}}{\kappa^2(1-\epsilon)^2} \right) \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{g}_t^k\|^2 \\ &\geqslant C \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{g}_t^k\|^2, \end{aligned} \tag{21}$$

with probability exceeding $1 - 6\delta$, where $C = \frac{(1-\epsilon)\psi}{2} - \frac{\epsilon_1}{\kappa(1-\epsilon)}$, and the last bound follows by substituting the value of α^* from equation (6) and using the fact that $0 < \epsilon < 1/2$. Moreover, using Eq. (17), we get

$$\|\mathbf{g}(\cdot)\|^2 \leqslant \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{g}^k(\cdot)\|^2,$$

which prove Eq. (12).

It now remains to prove bound (18).

Proof of bound (18)**:** From the uniform subsampling property (similar to Lemma 1.1, see Appendix 4.2), we get

$$|(\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) - (\mathbf{p}_t^k)^T \mathbf{g}^k(\bar{\mathbf{w}}_t)| \leqslant \epsilon_1 \|(\mathbf{p}_t^k)\| \|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \text{ w.p. } 1 - \delta. \tag{22}$$

Thus,

$$(\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) \geqslant (\mathbf{p}_t^k)^T \mathbf{g}^k(\bar{\mathbf{w}}_t) - \epsilon_1 \|(\mathbf{p}_t^k)\| \|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \tag{23}$$

w.p. $1-\delta$. Now, since the function f^k is $\kappa(1-\epsilon)$ strongly-convexity with probability $1-\delta$, we have the following bound w.p. at least $1-\delta$:

$$\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}_t^k \geqslant (f^k(\bar{\mathbf{w}}_t) - f^k(\mathbf{w}_{t+1}^k)) + \frac{\kappa(1-\epsilon)}{2} (\alpha_t^k)^2 \|\mathbf{p}_t^k\|^2 \tag{24}$$

Combing the equations (23)-(24) and using Lemma 1.2 we have

$$\begin{aligned}
\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) &\geqslant (f^k(\bar{\mathbf{w}}_t) - f^k(\mathbf{w}_{t+1}^k)) + \frac{\kappa(1-\epsilon)}{2} (\alpha_t^k)^2 \|\mathbf{p}_t^k\|^2 - \epsilon_1 \|(\mathbf{p}_t^k)\| \|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \\
&\overset{(i)}{\geqslant} \psi \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)}{2} (\alpha_t^k)^2 \|\mathbf{p}_t^k\|^2 - \epsilon_1 \|(\mathbf{p}_t^k)\| \|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \\
&\overset{(ii)}{\geqslant} \psi \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)(\alpha_t^k)^2}{2} \|\mathbf{p}_t^k\|^2 - \frac{\epsilon_1}{\kappa(1-\epsilon)} \|\mathbf{g}^k(\bar{\mathbf{w}}_t)\|^2 \\
&= \left(\psi - \frac{\epsilon_1}{\kappa(1-\epsilon)} \right) \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)(\alpha_t^k)^2}{2} \|\mathbf{p}_t^k\|^2
\end{aligned}$$

with probability exceeding $1-4\delta$, where the inequality (i) follows from Lemma 1.2 and inequality (ii) follows from (20).

Note that the bound in (18) hold for all $k \in [K]$ with probability $1-\delta_1$ (thus, the sample size increases by a factor of K in the $\log(\cdot)$ term). This concludes the Case 1 of our proof. We now move to Case 2.

Proof of the claim (13)**:** We now continue with the same analysis and show the following

$$f(\bar{\mathbf{w}}_t) - f(\bar{\mathbf{w}}_{t+1}) \geqslant C_1 \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{g}_t^k\|^2 - \frac{\eta\Gamma}{\kappa(1-\epsilon)}, \tag{25}$$

with probability at least $1-4\delta$.

In this case, we show that the requirement of a lower bound on $\|g^k(\bar{\mathbf{w}}_t)\|$ and s can be relaxed at the expense of getting hit by an error floor. In particular, we show that

$$\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) \geqslant \psi \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)(\alpha_t^k)^2}{2} \|\mathbf{p}_t^k\|^2 - \frac{\eta\Gamma}{\kappa(1-\epsilon)} \tag{26}$$

with probability at least $1-4\delta$, where $\eta = (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\Gamma$. Substituting this yields the bound of Eq. (25).

Proof of bound Eq. (26) : From the uniform subsampling property (see Appendix 4.1), we get

$$|(\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) - (\mathbf{p}_t^k)^T \mathbf{g}^k(\bar{\mathbf{w}}_t)| \leqslant \eta \|(\mathbf{p}_t^k)\| \text{ w.p. } 1-\delta. \tag{27}$$

where $\eta = (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\Gamma$. Thus,

$$(\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) \geqslant (\mathbf{p}_t^k)^T \mathbf{g}^k(\bar{\mathbf{w}}_t) - \eta \|(\mathbf{p}_t^k)\| \tag{28}$$

w.p. $1-\delta$. Now, since the function f^k is $\kappa(1-\epsilon)$ strongly-convexity with probability $1-\delta$, we have the following bound w.p. at least $1-\delta$:

$$\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}_t^k \geqslant (f^k(\bar{\mathbf{w}}_t) - f^k(\mathbf{w}_{t+1}^k)) + \frac{\kappa(1-\epsilon)}{2} (\alpha_t^k)^2 \|\mathbf{p}_t^k\|^2 \tag{29}$$

Combing the equations (28)-(29) and using Lemma 1.2 we have

$$\begin{aligned}
\alpha_t^k (\mathbf{p}_t^k)^T \mathbf{g}(\bar{\mathbf{w}}_t) &\geqslant (f^k(\bar{\mathbf{w}}_t) - f^k(\mathbf{w}_{t+1}^k)) + \frac{\kappa(1-\epsilon)}{2} (\alpha_t^k)^2 \|\mathbf{p}_t^k\|^2 - \eta \|(\mathbf{p}_t^k)\| \\
&\overset{(i)}{\geqslant} \psi \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)}{2} (\alpha_t^k)^2 \|\mathbf{p}_t^k\|^2 - \eta \|(\mathbf{p}_t^k)\| \\
&\overset{(ii)}{\geqslant} \psi \|\mathbf{g}_t^k\|^2 + \frac{\kappa(1-\epsilon)(\alpha_t^k)^2}{2} \|\mathbf{p}_t^k\|^2 - \frac{\eta}{\kappa(1-\epsilon)}\Gamma
\end{aligned}$$

with probability exceeding $1-4\delta$, where the inequality (i) follows from Lemma 1.2 and inequality (ii) follows from (20) and the fact that $\|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \leqslant \Gamma$.

□

3 PROOF OF THEOREM 4.4

Proof. Recall from perturbed iterate analysis

$$\bar{\mathbf{w}}_{t+1} = \bar{\mathbf{w}}_{t_0} - \sum_{\tau=t_0}^{t} \bar{\mathbf{p}}_\tau, \tag{30}$$

where $\bar{\mathbf{p}}_\tau = \frac{1}{K}\sum_{k=1}^{K} \alpha_\tau^k \mathbf{p}_\tau^k$ is the average descent direction and $\mathbf{p}_\tau^k = (\mathbf{H}_\tau^k)^{-1}\mathbf{g}_\tau^k$ is the local descent direction at the k-th worker at time τ.

Similar to the proof of theorem 3.2, we next invoke the M-smoothness property of $f(\cdot)$ to get

$$\begin{aligned} f(\bar{\mathbf{w}}_{t_0}) - f(\bar{\mathbf{w}}_{t+1}) &\geqslant \frac{-M}{2}\|\sum_{\tau=t_0}^{t}\bar{\mathbf{p}}_\tau\|^2 + \langle \mathbf{g}(\bar{\mathbf{w}}_{t_0}), \sum_{\tau=t_0}^{t}\bar{\mathbf{p}}_\tau\rangle \\ &= \frac{-M}{2}\|\frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\alpha_\tau^k\mathbf{p}_\tau^k\|^2 + \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\langle \mathbf{g}(\bar{\mathbf{w}}_{t_0}), \alpha_\tau^k\mathbf{p}_\tau^k\rangle \\ &\geqslant \frac{-M}{2K}\sum_{k=1}^{K}\|\sum_{\tau=t_0}^{t}\alpha_\tau^k\mathbf{p}_\tau^k\|^2 + \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\langle \mathbf{g}(\bar{\mathbf{w}}_{t_0}), \alpha_\tau^k\mathbf{p}_\tau^k\rangle, \end{aligned} \tag{31}$$

where the last inequality uses the fact

$$\left(\sum_{k=1}^{K}\|\mathbf{a}_k\|\right)^2 \leqslant K\sum_{k=1}^{K}\|\mathbf{a}_k\|^2,\ \forall\ \mathbf{a}_k \in \mathbb{R}^d, k \in [K]. \tag{32}$$

Similarly, by $\kappa(1-\epsilon)$ strong-convexity of $f^k(\cdot)$, we get

$$f^k(\mathbf{w}_{t_0}^k) - f^k(\mathbf{w}_{t+1}^k) \leqslant \frac{-\kappa(1-\epsilon)}{2}\|\sum_{\tau=t_0}^{t}\alpha_\tau^k\mathbf{p}_t^k\|^2 + \langle \mathbf{g}_t^k, \sum_{\tau=t_0}^{t}\alpha_\tau^k\mathbf{p}_\tau^k\rangle, \tag{33}$$

with probability $1-\delta$. The above inequality, when averaged across k, becomes

$$\frac{1}{K}\sum_{k=1}^{K}\left(f^k(\mathbf{w}_{t_0}^k) - f^k(\mathbf{w}_{t+1}^k)\right) \leqslant \frac{-\kappa(1-\epsilon)}{2K}\sum_{k=1}^{K}\|\sum_{\tau=t_0}^{t}\alpha_\tau^k\mathbf{p}_t^k\|^2 + \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\langle \mathbf{g}_t^k, \alpha_\tau^k\mathbf{p}_\tau^k\rangle \tag{34}$$

Moreover, similar to Eq. (27), we get

$$|\mathbf{r}^T\mathbf{g}(\bar{\mathbf{w}}_t) - \mathbf{r}^T\mathbf{g}^k(\bar{\mathbf{w}}_t)| \leqslant \eta\|\mathbf{r}\| \text{ w.p. } 1-\delta. \tag{35}$$

where $\eta = (1+\sqrt{2\log(\frac{m}{\delta})})\sqrt{\frac{1}{s}}\Gamma$. Keeping $\mathbf{r} = \alpha_\tau^k\mathbf{p}_\tau^k$ and $\mathbf{w} = \bar{\mathbf{w}}_{t_0}$, we get

$$(\alpha_\tau^k\mathbf{p}_\tau^k)^T\mathbf{g}(\bar{\mathbf{w}}_{t_0}) \geqslant (\alpha_\tau^k\mathbf{p}_\tau^k)^T\mathbf{g}^k(\bar{\mathbf{w}}_{t_0}) - \eta\alpha_\tau^k\|\mathbf{p}_\tau^k\|, \tag{36}$$

w. p. $1-\delta$, where $\eta = (1+\sqrt{2\log(\frac{m}{\delta})})\sqrt{\frac{1}{s}}\Gamma$.

Now, after combining inequalities (31) and (34) using (36) to eliminate the terms $\frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\langle \mathbf{g}(\bar{\mathbf{w}}_{t_0}), \alpha_\tau^k\mathbf{p}_\tau^k\rangle$ and $\frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\langle \mathbf{g}^k(\bar{\mathbf{w}}_{t_0}), \alpha_\tau^k\mathbf{p}_\tau^k\rangle$, we get

$$\begin{aligned} f(\bar{\mathbf{w}}_{t_0}) - f(\bar{\mathbf{w}}_{t+1}) \geqslant \frac{1}{K}\sum_{k=1}^{K}(f^k(\bar{\mathbf{w}}_{t_0}^k) - f^k(\bar{\mathbf{w}}_{t+1}^k) - \frac{(M-\kappa(1-\epsilon))}{2K}\sum_{k=1}^{K}(\|\sum_{\tau=t_0}^{t}\alpha_\tau^k\mathbf{p}_t^k\|^2) \\ - \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t}\eta\alpha_\tau^k\|\mathbf{p}_\tau^k\|. \end{aligned} \tag{37}$$

Also, from Lemma 1.2, we have

$$f^k(\bar{\mathbf{w}}_{t_0}^k) - f^k(\bar{\mathbf{w}}_{t+1}^k) \geqslant \psi \sum_{\tau=t_0}^{t} \|\mathbf{g}_\tau^k\|^2. \tag{38}$$

Using above, we get

$$f(\bar{\mathbf{w}}_{t_0} - f(\bar{\mathbf{w}}_{t+1}) \geqslant \frac{1}{K}\psi \sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \|\mathbf{g}_\tau^k\|^2 - \frac{(M-\kappa(1-\epsilon))}{2K} \sum_{k=1}^{K}(\| \sum_{\tau=t_0}^{t} \alpha_\tau^k \mathbf{p}_t^k\|^2) - \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \eta\alpha_\tau^k \|\mathbf{p}_\tau^k\|. \tag{39}$$

Using triangle inequality above, we get

$$f(\bar{\mathbf{w}}_{t_0}) - f(\bar{\mathbf{w}}_{t+1}) \geqslant \frac{1}{K}\psi \sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \|\mathbf{g}_\tau^k\|^2 - \frac{(M-\kappa(1-\epsilon))}{2K} \sum_{k=1}^{K}\sum_{\tau=t_0}^{t} (\alpha_\tau^k)^2 \|\mathbf{p}_\tau^k\|^2 - \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \eta\alpha_\tau^k \|\mathbf{p}_\tau^k\|. \tag{40}$$

Also, since $\alpha_t^k \leqslant 1$ and $\|\mathbf{p}_\tau^k\| \leqslant \frac{1}{\kappa(1-\epsilon)}\|\mathbf{g}_\tau^k\|$, we get

$$\begin{aligned} f(\bar{\mathbf{w}}_{t_0} - f(\bar{\mathbf{w}}_{t+1}) &\geqslant \frac{1}{K}\psi \sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \|\mathbf{g}_\tau^k\|^2 - \frac{(M-\kappa(1-\epsilon))}{2K\kappa^2(1-\epsilon)^2} \sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \|\mathbf{g}_\tau^k\|^2 \\ &- \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \frac{\eta}{\kappa(1-\epsilon)}\|\mathbf{g}_\tau^k\| \\ &= \frac{C}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t} \|\mathbf{g}_\tau^k\|^2 - \frac{\eta L\Gamma}{\kappa(1-\epsilon)} \end{aligned} \tag{41}$$

where $C = \psi - \frac{(M-\kappa(1-\epsilon))}{2K\kappa^2(1-\epsilon)^2}$, which proves the claim.

□

4 CONCENTRATION INEQUALITIES: WITH AND WITHOUT ERROR FLOOR

Consider a vector $v \in \mathbb{R}^d$. We have defined the following: $\mathbf{g}(\bar{\mathbf{w}}_t) = \frac{1}{n}\sum_i \mathbf{g}_i(\bar{\mathbf{w}}_t)$ and $\mathbf{g}^k(\bar{\mathbf{w}}_t) = \frac{1}{s}\sum_{i\in\mathcal{S}} \mathbf{g}_i(\bar{\mathbf{w}}_t)$, where $\mathbf{g}_i$ denotes the local gradient in worker machine i, and $\mathcal{S}$ is the random set consisting data points for machine k. Let us do the calculation in two settings:

4.1 WITH ERROR FLOOR

Here we have the error floor. Note that having an error floor is not restrictive, if we go for the adaptive variation of the algorithm, where we run GIANT for the final iterations. Since GIANT has no error floor, the final accuracy won't be affected by the error floor obtained in the first few steps of the algorithm (check if this is true).

Lemma 4.1 (McDiarmid's Inequality)**.** *Let $X = X_1, \ldots, X_m$ be m independent random variables taking values from some set A, and assume that $f : A^m \to \mathbb{R}$ satisfies the following condition (bounded differences):*

$$\sup_{x_1,\ldots,x_m,\hat{x}_i} |f(x_i, \ldots, x_i, \ldots, x_m) - f(x_i, \ldots, \hat{x}_i, \ldots, x_m)| \leqslant c_i,$$

for all $i \in \{1, \ldots, m\}$. Then for any $\epsilon > 0$ we have

$$P\left[f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)] \geqslant \epsilon\right] \leqslant \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

The property described in the following is useful for uniform row sampling matrix.

Let $\mathbf{S} \in \mathbb{R}^{n\times s}$ be any uniform sampling matrix, then for any matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n] \in \mathbb{R}^{d\times n}$ with probability $1-\delta$ for any $\delta > 0$ we have,

$$\|\frac{1}{n}\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\| \leqslant (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\max_i \|\mathbf{b}_i\|, \tag{42}$$

where $\mathbf{1}$ is all ones vector.

Let us first see the justification of the above statement.The vector $\mathbf{B}\mathbf{1}$ is the sum of column of the matrix $\mathbf{B}$ and $\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1}$ is the sum of uniformly sampled and scaled column of the matrix $\mathbf{B}$ where the scaling factor is $\frac{1}{\sqrt{sp}}$ with $p = \frac{1}{n}$. If $(i_1, \ldots, i_s)$ is the set of sampled indices then $\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} = \sum_{k\in(i_1,\ldots,i_s)} \frac{1}{sp}\mathbf{b}_k$.

Define the function $f(i_1, \ldots, i_s) = \|\frac{1}{n}\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\|$. Now consider a sampled set $(i_1, \ldots, i_{j'}, \ldots, i_s)$ with only one item (column) replaced then the bounded difference is

$$\begin{aligned}\Delta &= |f(i_1, \ldots, i_j, \ldots, i_s) - f(i_1, \ldots, i_{j'}, \ldots, i_s)| \\ &= |\frac{1}{n}\|\frac{1}{sp}\mathbf{b}_{i'_j} - \frac{1}{sp}\mathbf{b}_{i_j}\|| \leqslant \frac{2}{s}\max_i \|\mathbf{b}_i\|.\end{aligned}$$

Now we have the expectation

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{n}\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\|^2] \leqslant \frac{n}{sn^2}\sum_{i=1}^n \|\mathbf{b}_i\|^2 = \frac{1}{s}\max_i \|\mathbf{b}_i\|^2 \\ \Rightarrow &\mathbb{E}[\|\frac{1}{n}\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\|] \leqslant \sqrt{\frac{1}{s}}\max_i \|\mathbf{b}_i\|.\end{aligned}$$

Using McDiarmid inequality (Lemma 4.1) we have

$$P\left[\|\frac{1}{n}\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\| \geqslant \sqrt{\frac{1}{s}}\max_i \|\mathbf{b}_i\| + t\right] \leqslant \exp\left(-\frac{2t^2}{s\Delta^2}\right).$$

Equating the probability with δ we have

$$\begin{aligned}&\exp(-\frac{2t^2}{s\Delta^2}) = \delta \\ \Rightarrow &t = \Delta\sqrt{\frac{s}{2}\log(\frac{1}{\delta})} = \max_i \|\mathbf{b}_i\|\sqrt{\frac{2}{s}\log(\frac{1}{\delta})}.\end{aligned}$$

Finally we have with probability $1-\delta$

$$\|\frac{1}{n}\mathbf{B}\mathbf{S}\mathbf{S}^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\| \leqslant (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\max_i \|\mathbf{b}_i\|,$$

and hence equation (42) is justified.

We now apply the above in distributed gradient estimation. For the k-th worker machine, we have

$$\|\frac{1}{n}\mathbf{B}\mathbf{S}_k\mathbf{S}_k^\top\mathbf{1} - \frac{1}{n}\mathbf{B}\mathbf{1}\| \leqslant (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\max_i \|\mathbf{b}_i\|,$$

with probability $1-\delta$, which implies

$$\|\mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\| \leqslant (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}\Gamma,$$

with probability at least $1-\delta$ provided $\|\mathbf{g}_i(\bar{\mathbf{w}}_t)\| \leqslant \Gamma$ for all $i \in [m]$

Writing, $\eta = (1+\sqrt{2\log(\frac{1}{\delta})})\sqrt{\frac{1}{s}}L$, we succinctly write

$$|\langle v, \mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\rangle| \leqslant \|v\|\|\mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\| \leqslant \eta\|v\|$$

with probability at least $1-\delta$, where $\eta = \mathcal{O}(1/\sqrt{s})$ is small.

4.2 WITHOUT ERROR FLOOR

In this section, we analyze the same quantity using vector Bernstein inequality. Intuitively, we show that unless $\mathbf{g}(\bar{\mathbf{w}}_t)$ is too small, we can overcome the error floor shown in the previous calculation. In particular, we assume that

$$\|\mathbf{g}^k(\bar{\mathbf{w}}_t)\| \geqslant G.$$

The idea here is to use the vector Bernstein inequality. Using the notation of Appendix 4.1, $\mathbf{g}^k(\bar{\mathbf{w}}_t) = \frac{1}{n}\mathbf{BSS}^\top\mathbf{1}$, where $\mathbf{S}$ is appropriately defined sampling matrix. Also $\mathbf{g}(\bar{\mathbf{w}}_t) = \frac{1}{n}\mathbf{B1}$. For the k-th machine,

$$\mathbf{g}^k(\bar{\mathbf{w}}_t) = \frac{1}{s}\sum_{i\in\mathcal{S}}\mathbf{g}_i(\bar{\mathbf{w}}_t),$$

and so,

$$\mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t) = \frac{1}{s}\sum_{i\in\mathcal{S}}(\mathbf{g}_i(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)),$$

with $|\mathcal{S}| = s$. We also have $\|\mathbf{g}_i(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\| \leqslant \Gamma + \Gamma = 2\Gamma$, and $\mathbb{E}\|\mathbf{g}_i(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\|^2 \leqslant 4\Gamma^2$. Using vector Bernstein inequality with $t = \epsilon_1\|\mathbf{g}^k\|$, we obtain

$$\mathbb{P}\left(\|\mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\| \geqslant \epsilon_1\|\mathbf{g}^k(\bar{\mathbf{w}}_t)\|\right) \leqslant d\exp(-s\frac{\epsilon_1^2\|\mathbf{g}^k\|^2}{32\Gamma^2} + 1/4) \leqslant d\exp(-s\frac{\epsilon_1^2 G^2}{32L^2} + 1/4).$$

So, as long as

$$G^2 = \Omega\left(\frac{\Gamma^2}{\epsilon_1^2 s}\log(d/\delta)\right),$$

or,

$$s \gtrsim \left(\frac{\Gamma^2}{\epsilon_1^2 G^2}\log(d/\delta)\right),$$

we have,

$$|\langle v, \mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\rangle| \leqslant \|v\|\|\mathbf{g}^k(\bar{\mathbf{w}}_t) - \mathbf{g}(\bar{\mathbf{w}}_t)\| \leqslant \epsilon_1\|v\|\|\mathbf{g}^k\|$$

with probability at least $1-\delta$.

5 EXPERIMENTS: ADDITIONAL DETAILS AND PLOTS

In this section, we include additional details regarding the experiments that couldn't be added in the main paper due to space constraints.

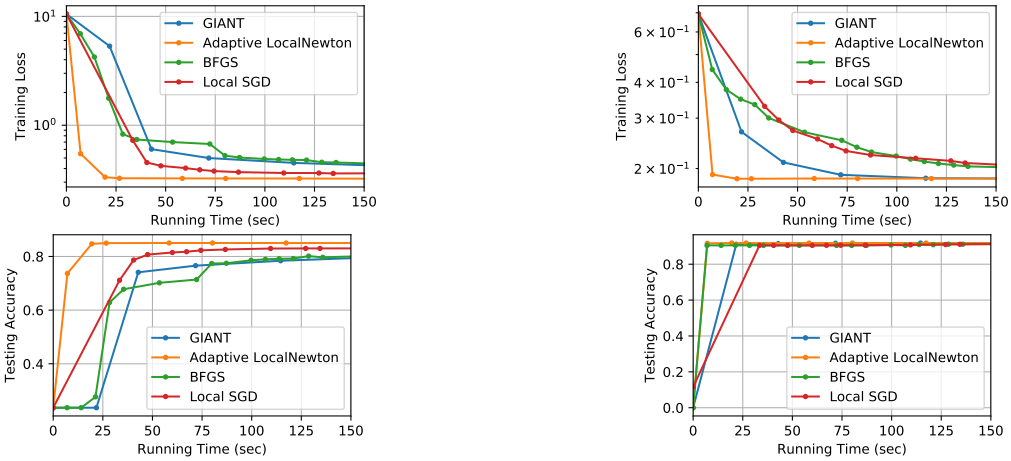
Hyperparameters for Local SGD and BFGS: In Table 1, we provide the step-sizes for local SGD and BFGS that were obtained through hyperparameter tuning, where $s = n/K$, n is the number of training examples in the dataset and $K = 100$.

Dataset	Samples per worker (s)	Local SGD	BFGS
w8a	480	$10/s$	100
Covtype	5000	$10/s$	1
EPSILON	4000	$500/s$	10
a9a	320	$10/s$	1
ijcnn1	490	$100/s$	10

Table 1: Step-sizes obtained using tuning for Local SGD and BFGS for several datasets

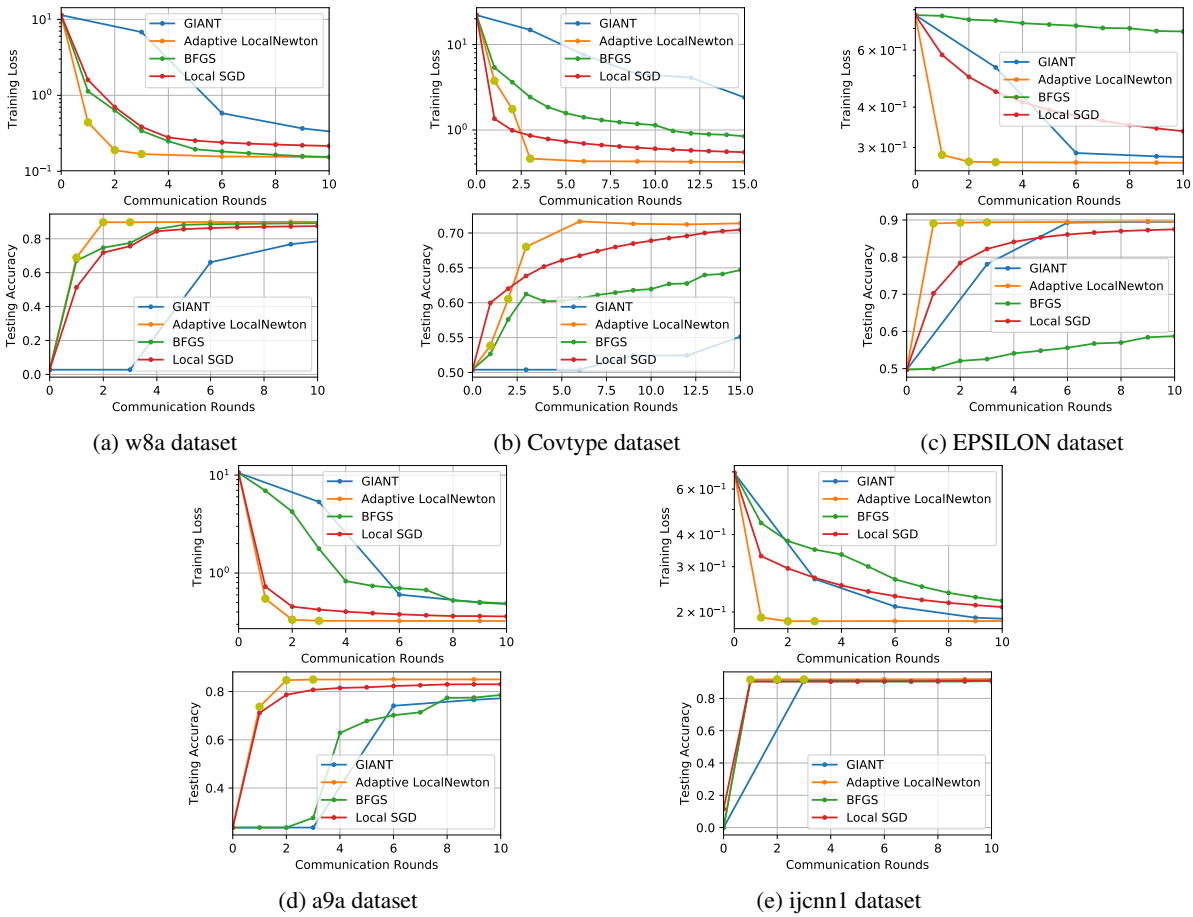
Running Times for a9a and ijcnn1 Datasets: In Figure 1, we plot the results on AWS Lambda for a9a and ijcnn1 datasets. Again, adaptive LocalNewton considerable outperforms Local SGD, GIANT and BFGS in terms of end-to-end runtimes.

Convergence w.r.t. Communication Rounds: In our main paper, we skipped the plots for convergence behavior w.r.t. communication rounds due to space constraints. In Figure 2, we show the convergence of adaptive LocalNewton, GIANT, Local SGD and BFGS with communication rounds for all the five datasets considered in this paper. Again, LocalNewton significantly outperforms existing schemes by reducing the communication rounds by at least 60% to reach the same training loss.



(a) a9a dataset (b) ijcnn1 dataset

Figure 1: Experiments on the a9a and ijcnn1 datasets on AWS Lambda



(a) w8a dataset (b) Covtype dataset (c) EPSILON dataset

(d) a9a dataset (e) ijcnn1 dataset

Figure 2: Comparing LocalNewton with competing schemes w.r.t. communication rounds. Yellow dots on adaptive LocalNewton denote transition from larger to smaller values of L (or to GIANT if $L = 1$).