Fine-Grained Analysis of Local SGD under Intermittent Communication

Ali Zindari

October 20, 2025

Problem Definition

We are interested in solving the following finite-sum minimization problems:

$$x^\star := \arg\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) \right]. \tag{1}$$

- We have M clients which are **not** located on the same machine.
- $f_m(x)$ is the objective/loss function of client m where $x_m^\star := \arg\min_x f_m(x)$.
- $\nabla f(x^\star) = 0$ but in general $\nabla f_m(x^\star) \neq 0$.

The goal is to find parameter $x^\star$ that is a good solution for all clients.

Distributed Optimization

- Each client has their own data $\mathcal{D}_m$ located on their own device.
- We **cannot** have all the data from all clients on one machine and train a model on it.
- Gathering data from everyone in one place is infeasible and costly, violating privacy.
- If $\mathcal{D}_1 = \ldots = \mathcal{D}_M$ we say problem is Homogeneous.
- If $\mathcal{D}_1 \neq \ldots \neq \mathcal{D}_M$ we say problem is Heterogeneous.

The focus of this work is on Heterogeneous regime.

Local SGD

Local SGD is a simple yet efficient method for solving (1).

- **Local Update Rule on Each Client**:
  - At round r, each client is initialized at x_r.
  - Each client m takes K local steps: $x_{t+1}^m = x_t^m - \eta \nabla f_m(x_t^m, \xi_t^m), \quad \xi_t^m \sim \mathcal{D}_m$.
- **Global Aggregation on a Central Server**:
  - After K steps: $x_{r+1} = \frac{1}{M} \sum_{i=1}^{M} x_K^m$.
  - Server *communicates* x_{r+1} to all clients.

  We repeat all these steps for R times.

We aim to communicate as less as possible. Local steps are relatively cheap!

Assumptions

Smoothness

For each client m and for every $x, y \in \mathbb{R}^d$ we have:

$$f_m(y) \leq f_m(x) + \langle \nabla f_m(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$$

Strong Convexity

For each client m and for every $x, y \in \mathbb{R}^d$ we have:

$$f_m(x) + \langle \nabla f_m(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2 \leq f_m(y)$$

- We require $L \geq 0$.
- If $\mu = 0$, the above condition simplifies to convexity.

Assumptions

Gradient Similarity

For every $x \in \mathbb{R}^d$ we have:

$$\frac{1}{M} \sum_{m \in [M]} \|\nabla f_m(x) - \nabla f(x)\|^2 \leq \zeta^2$$

Gradient Similarity at Optimum

For every $x^\star = \arg\min_x \; f(x)$ we have:

$$\frac{1}{M} \sum_{m \in [M]} \|\nabla f_m(x^\star)\|^2 \leq \zeta_\star^2$$

- First assumption is stronger and more restrictive.

Related Works: Strongly Convex Setting with ζ

Woodworth et al. (2020)

For any $K, R, M \geq 1$ and $L, B, \sigma \geq 0$, using a decreasing stepsize $\eta_t \leq \frac{1}{2L}$, we have:

$$\mathbb{E}[f(\hat{x}_{KR})] - f(x^\star) = \mathcal{O}\left(\frac{LB^2}{LKR + \mu K^2 R^2} + \frac{\sigma^2}{\mu MKR} + \frac{L\zeta^2}{\mu^2 R^2} + \frac{L\sigma^2}{\mu^2 K R^2}\right),$$

where $\|x_0 - x^\star\| \leq B$.

- All terms go to zero with $K \to \infty$ except the third one.
- Third term is affected by heterogeneity ζ.

Related Works: Strongly Convex Setting with $\zeta_\star$

Koloskova et al. (2020)

For any $K, R, M \geq 1$ and $L, B, \sigma \geq 0$, using a decreasing stepsize $\eta_t \leq \frac{1}{2L}$, we have:

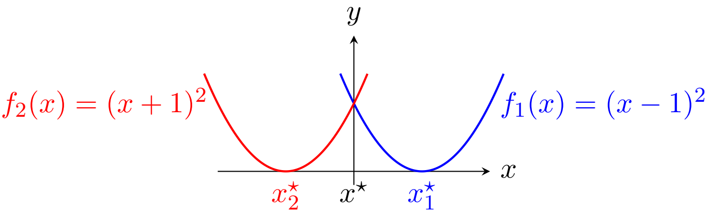
$$\mathbb{E}[f(\hat{x}_{KR})] - f(x^\star) \preceq \exp\left(-\frac{\mu R}{L}\right) LB^2 + \frac{\sigma^2}{\mu MKR} + \frac{L\zeta_\star^2}{\mu^2 R^2} + \frac{L\sigma^2}{\mu^2 KR^2},$$

where $\|x_0 - x^\star\| \leq B$.

- All terms go to zero with $K \to \infty$ except the first and third one.
- Third term is affected by heterogeneity $\zeta_\star$.

What is Missing?

- This simple scenario cannot be captured by the discussed rates.
- Extreme communication acceleration is expected.



What is Missing?

Formally:

$$f(x) = \underbrace{\frac{1}{2}(x-a)^2}_{f_1} + \underbrace{\frac{1}{2}(x+a)^2}_{f_2}$$

- **Quadratic** functions with identical **Hessians**.
- When $a \to \infty$, $\zeta_\star \to \infty$ meaning that problem is highly heterogeneous.

$$x^\star = \frac{x_1^\star + x_2^\star}{2}$$

- We just need to compute the average of clients' minima even if $\zeta_\star$ is very large.
- Only **one** communication round is needed with many local steps to converge.

What is Missing?

Local SGD for convex quadratics with identical hessians

For any $K, R, M \geq 1$ and $\sigma \geq 0$ with a stepsize $\eta \leq \frac{1}{\lambda_{\max}(A)}$, we have:

$$\mathbb{E}\left[\|\bar{x}_{KR} - x^\star\|^2\right] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}KR\right) + \frac{\sigma^2}{\lambda_{\min}^2(A)MKR}\right).$$

- Both terms go to zero only by choosing $K \to \infty$.
- Local SGD achieves extreme communication acceleration.

None of the discussed rates recovers this scenario!

What is Missing? New Assumptions

Recall the special properties needed to achieve extreme acceleration:

- Quadratic clients.
- Identical Hessians.

Questions:

- How to identify if a function is quadratic?
- How to identify how similar are the Hessians of clients?

What is Missing? New Assumptions

Lipschitz Hessian

Function f is Q-Lipschitz if for every $x, y \in \mathbb{R}^d$:

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \leq Q \cdot \|x - y\| .$$

Hessian Similarity

For every $x \in \mathbb{R}^d$, we assume:

$$\sup_{m,n\in[M]} \left\|\nabla^2 f_m(x) - \nabla^2 f_n(x)\right\| \leq \tau.$$

- If a function is quadratic then $Q = 0$.
- If all clients have the same hessian then $\tau = 0$.

New Rates: Convex and Strongly Convex

New rates for convex and strongly convex objectives [Patel et al. (2024)]

For convex smooth clients using a constant stepsize $\eta \le \frac{1}{2L}$, we have:

$$\mathcal{O}\left(\frac{LB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(\tau\sigma B^3)^{\frac{1}{2}}}{K^{\frac{1}{4}}R^{\frac{1}{2}}} + \frac{(Q\sigma^2 B^5)^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{(\tau\zeta B^3)^{\frac{1}{2}}}{R^{\frac{1}{2}}} + \frac{(Q\zeta^2 B^5)^{\frac{1}{3}}}{R^{\frac{2}{3}}}\right),$$

and for strongly convex clients:

$$\tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu KR}{L}\right)LB^2 + \frac{\sigma^2}{\mu MKR} + \frac{Q^2\sigma^4}{\mu^5 K^2 R^4} + \frac{Q^2\zeta^4}{\mu^5 R^4} + \frac{\tau^2\sigma^2}{\mu^3 KR^2} + \frac{\tau^2\zeta^2}{\mu^3 R^2}\right).$$

where the rates hold for $\mathbb{E}\left[f(\bar{x}_{KR}) - f(x^\star)\right]$.

New Rates: Convex and Strongly Convex

- In the new rates, we have $Q\zeta$ and $\tau\zeta$ instead of $L\zeta$.
- We require a few communication rounds in the regime where τ, Q are very small.
- Note that τ, Q can be arbitrary small and even zero.
- When $\tau = Q = 0$, we achieve extreme communication efficiency:

$$\mathcal{O}\left(\frac{LB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}}\right) \qquad \textbf{and} \qquad \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu KR}{L}\right) LB^2 + \frac{\sigma^2}{\mu MKR}\right).$$

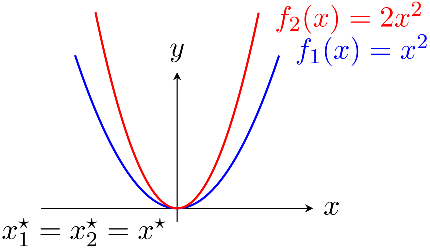
However, we still need to rely on ζ assumption.

What is Missing?

Let us assume clients are quadratics that share their minima but have different Hessians.

- This implies $\zeta_\star = Q = 0$ but $\tau, \zeta \neq 0$.
- It is clear that $x_1^\star = \ldots = x_M^\star = x^\star$.
- We can just minimize one function and converge.
- No need for any communication.

Our previous rate does not capture this scenario.



New Rates: Replacing ζ by $\zeta_\star$ and $\phi_\star$

New rate for strongly convex objectives with $\zeta_\star$ and $\phi_\star$ [Patel et al. (2025)]

For strongly convex smooth clients using a constant stepsize $\eta \leq \frac{1}{2L}$, we have:

$$\tilde{\mathcal{O}}\left(e^{-\frac{\mu KR}{2L}}B^2 + \frac{\sigma^2}{\mu^2 MKR} + \frac{\tau^2 L^2 \phi_\star^2}{\mu^4 R^2} + \frac{L^4 \zeta_\star^2}{\mu^4 R^2} + \frac{L^2 \tau^2 \sigma^2}{\mu^6 KR^3} + \frac{L^2 \sigma^2}{\mu^4 KR^2}\right).$$

where the rate holds for $\mathbb{E}\left[\|x_{KR} - x^\star\|^2\right]$ and $\sup_{m\in[M]} \|x_m^\star - x^\star\| \leq \phi_\star$.

- For quadratics with identical minima $\zeta_\star = \phi_\star = 0$.
- If $\zeta_\star = \phi_\star = 0$, we achieve extreme communication efficiency.

References

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. (2020). A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR.

Patel, K. K., Glasgow, M., Zindari, A., Wang, L., Stich, S. U., Cheng, Z., Joshi, N., and Srebro, N. (2024). The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. In Agrawal, S. and Roth, A., editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4115–4157. PMLR.

Patel, K. K., Zindari, A., Stich, S., and Wang, L. (2025). Revisiting consensus error: A fine-grained analysis of local sgd under second-order data heterogeneity. *arxiv*.

Woodworth, B. E., Patel, K. K., and Srebro, N. (2020). Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292.