Eco-Fedsplit: FEDERATED LEARNING WITH ERROR-COMPENSATED COMPRESSION

Sarit Khirirat[1]*, Sindri Magnússon*[2]*, and Mikael Johansson*[1]

[1]KTH Royal Institute of Technology, [2]Stockholm University

ABSTRACT

Federated learning is an emerging framework for collaborative machine-learning on devices which do not want to share local data. State-of-the art methods in federated learning reduce the communication frequency, but are not guaranteed to converge to the optimal model parameters. These methods also experience a communication bottleneck, especially when the devices are power-constrained and communicate over a shared medium. This paper presents `ECO-FedSplit`, an algorithm that increases the communication efficiency of federated learning without sacrificing solution accuracy. The key is to compress inter-device communication and to compensate for information losses in a theoretically justified manner. We prove strong convergence properties of `ECO-FedSplit` on strongly convex optimization problems and show that the algorithm yields a highly accurate solution with dramatically reduced communication. Extensive numerical experiments validate our theoretical result on real data sets.

***Index Terms*—** Optimization methods, operator splitting schemes, quantization, federated learning

1. INTRODUCTION

Federated learning (FL) has become a popular distributed machine learning framework as it allows multiple nodes to optimize problem parameters without sharing local data [1, 2, 3]. However, communication easily becomes a major bottleneck in FL because the nodes need to iteratively share models which often have millions of parameters [4, 5, 6]. This easily renders FL algorithms infeasible in high-dimensional settings, especially if the nodes are power-constrained or communicate over bandwidth-restricted networks [2, 7, 8].

A popular approach to limit the communication overhead in FL is to increase the local computation on the devices. This is typically done by having the devices perform additional deterministic, stochastic, or proximal gradient updates locally before communicating its local parameters [9, 10, 11]. Experimental results have shown that this approach does accelerate convergence, but it also causes the algorithm to converge to a sub-optimal solution. In fact, it is easily shown that increasing the number of local gradient/proximal updates can make the algorithm converge to an increasingly sub-optimal solution [12]. To mitigate this problem, [12] developed `FedSplit`, an FL algorithm based on operator splitting. `FedSplit` adapts Peaceman-Rachford splitting [13] for solving distributed problems, and it enjoys fast linear convergence toward the *exact* optimal solution.

Communication efficiency can also be improved by reducing the number of bits exchanged per transmission round. This can, e.g., be done by compressing the information using, for example, *sparsification* [14, 15] or *quantization* [8, 16, 17]. Even though these compression operators have improved the communication efficiency of distributed optimization algorithms, they generally suffer in terms of solution accuracy. However, recent work has illustrated how the solution accuracy of compressed algorithms can be improved by error-compensation, exploiting a feedback mechanism based on previous compression errors. Error-compensation is shown to enable algorithms with even coarse compression to retain the convergence rate of the full-precision algorithm, and also obtain highly accurate solutions [6, 18, 19].

In this paper we propose `Eco-FedSplit`, a fast communication efficient federated algorithm that does not sacrifice solution accuracy. By using both operator splitting schemes and error-compensated compression, our algorithm guarantees global linear convergence towards a high-accuracy solution. Our key results prove that error-compensated compression has far superior performance to naive compression, and fully eliminates the accumulation of compression errors. In particular, we demonstrate that error compensation enables our algorithm to obtain an arbitrary solution accuracy as we decrease its tuning parameter λ. Finally, our numerical experiments confirm strong performance of `Eco-FedSplit` on logistic regression problems over benchmark data sets.

Notation and definitions: We let $\mathbb{N}$,$\mathbb{N}_0$, $\mathbb{Z}$, and $\mathbb{R}$ be the set of natural numbers, the set of natural numbers including zero, the set of integers, and the set of real numbers, respectively. For $x \in \mathbb{R}^d$, $\|x\|$ and $\|x\|_1$ are the ℓ_2 norm and the ℓ_1 norm, respectively, and $\lceil x \rceil_+ = \max\{0, x\}$. For a symmetric matrix $A \in R^{d\times d}$, we denote by $\lambda_1(A), \ldots, \lambda_d(A)$ its eigenvalues in an increasing order (including multiplicities), and its spectral norm is defined by $\|A\| = \max_i |\lambda_i(A)|$. For the

This work was supported partially by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation, and also by the Swedish Research Council (Vetenskapsrådet) under grant 2020-03607. Emails: `sindri.magnusson@dsv.su.se`, `sarit@kth.se`, `mikaelj@kth.se`.

fixed-point operator $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ and a positive integer p, we denote $\mathcal{T}^p(x) = \mathcal{T} \circ \mathcal{T} \circ \ldots \circ \mathcal{T} x$ (p times).The proximal operator and the reflected proximal operator are defined by

$$\mathbf{prox}_{\gamma F}(z) = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ F(x) + \frac{1}{2\gamma}\|x - z\|^2 \right\},$$

and $\mathbf{refl}_{\gamma F}(z) = 2\mathbf{prox}_{\gamma F}(z) - z$, respectively. A continuously differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, is μ-strongly convex if there exists a positive constant μ such that for all $x, y \in \mathbb{R}^d$ $F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$, and L-smooth if its gradient is L-Lipscthiz continuous i.e. for all $x, y \in \mathbb{R}^d$ $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$. Finally, we denote $x^\star$ and $F^\star$ are an optimal solution and optimal value to the problem of minimizing $F(x)$.

2. λ-FedSplit

Federated learning is a distributed framework for solving finite-sum optimization problems on the form

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where each $f_i(\cdot)$ is an objective function privately known by worker i. Typically, a master updates the solution, based on the local variables aggregated from all the workers. In particular, for a given initial value x^0, the master performs the following update

$$x^{k+1} = (1 - \lambda)x^k + \lambda \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}^p_{\gamma f_i}(x^k), \tag{2}$$

where k is the iteration index, $\mathcal{T}_{\gamma f_i}(\cdot)$ is a local operator performed by worker i, and $\gamma > 0$ is the learning-rate. The parameter $\lambda \in [0, 1]$ gives the master node flexibility to put some weights on its own previous iterations. This degree of freedom is not typical in FL algorithms. However, we show that for error-compensated compression, λ allows us to balance a trade-off between the solution accuracy and convergence rate. In particular, decreasing λ will improve the solution accuracy at the cost of more iterations.

At each iteration k, every worker i performs p local iterations by applying the local operator $\mathcal{T}_{\gamma f_i}(\cdot)$ p-times on the parameter x^k. The benchmark federated algorithms `FedAvg` [9] and `FedProx` [11] are covered by Equation (2) with $\mathcal{T}_{\gamma f_i}(x) = x - \gamma \nabla f_i(x)$ and with $\mathcal{T}_{\gamma f_i}(x) = \mathbf{prox}_{\gamma f_i}(x)$, respectively, and $\lambda = 1$. However, `FedAvg` and `FedProx` do not generally converge to the global optimum, even in the deterministic case that we study. In particular, the stationary/fixed points of these algorithms can be distant from the optimal solution even on simple problems, as shown in [12].

Recently, `FedSplit` [12] was developed to find the exact optimal solution of the problem. This algorithm relies on classical splitting schemes, and has the following iteration. Each worker node i updates its local vector z_i^k based on its private function $f_i(\cdot)$ via:

$$z_i^{k+1} = \mathbf{refl}_{\gamma f_i}(2\bar{z}^k - z_i^k). \tag{3}$$

Then, the master node aggregates the local vectors from all worker nodes, and performs the following update

$$x^{k+1} = (1 - \lambda)x^k + \lambda \bar{z}^{k+1}, \tag{4}$$

where $\bar{z}^k = \sum_{i=1}^n z_i^k/n$ and $\lambda \in [0, 1]$ is the tuning parameter. Note that when $n = 1$, `FedSplit` reduces to Douglas-Rachford splitting [20] for $\lambda = 1/2$ and to Peaceman-Rachford splitting [13] for $\lambda = 1$.

The linear convergence of `FedSplit` for $\lambda = 1$ is proved in [12]. For the purposes of this paper, it will be useful to have the freedom to consider any $\lambda \in (0, 1]$. We begin by establishing a similar linear convergence result in this case.

Theorem 1. *Consider `FedSplit` in (3) and (4) to solve Problem (1), where each $f_i(\cdot)$ is μ-strongly convex and L-smooth. If $\gamma = 1/\sqrt{\mu L}$ and $x^0 = z_1^0 = \ldots = z_n^0$, then*

$$\mathbf{w}^{k+1} \leq A^k \mathbf{w}^0, \quad \textit{where} \quad A := \begin{bmatrix} 1 - \lambda & \lambda\rho/\sqrt{n} \\ 0 & \rho \end{bmatrix}, \tag{5}$$

$\lambda \in (0, 1]$, $\rho = 1 - 2/(\sqrt{L/\mu} + 1)$, and

$$\mathbf{w}^k = \begin{bmatrix} \|x^k - x^\star\| \\ \|z^k - z^\star\| \end{bmatrix} \tag{6}$$

for $z = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^{d \cdot n}$.

Theorem 1 shows that `FedSplit` with $\lambda \in (0, 1]$ converges linearly towards the exact optimum, unlike FL methods like `FedAvg` and `FedProx`. When $\lambda = 1$, our result recovers the `FedSplit` convergence presented in Theorem 1 with $b = 0$ in [12]. However, now the convergence rate of `FedSplit` is also affected by the tuning parameter λ. In particular, decreasing λ slows down the convergence rate of `FedSplit`. However, we show that when we compress the local variables with error-compensation, then decreasing λ allows us to achieve better solution accuracy.

3. DIRECT COMPRESSION - LIMITATIONS

To limit the communication in `FedSplit` we need to compress the communicated information. Since in `FedSplit` z_i is the variable that is communicated by each worker node, a natural compressed version of `FedSplit` is to have the master perform the following update at each iteration k:

$$x^{k+1} = (1 - \lambda)x^k + \lambda \bar{z}^{k+1}, \tag{7}$$

where $\bar{z}^{k+1}$ is updated according to

$$\begin{aligned} \bar{z}^{k+1} &= \sum_{i=1}^{n} Q(z_i^{k+1})/n, \quad \text{and} \\ z_i^{k+1} &= \textbf{refl}_{\gamma f_i}(2\bar{z}^k - z_i^k), \end{aligned} \tag{8}$$

where $Q(\cdot)$ is a compression operator. To make the conclusions of our analysis broad we consider a general class compressors.

Definition 1 (Bounded Error Compression)**.** *The operator $Q : \mathbb{R}^d \to \mathbb{R}^d$ is an* ϵ-compressor *if there exists a positive constant ϵ such that $\|Q(v) - v\| \leq \epsilon$ for all $v \in \mathbb{R}^d$.*

For compression operators to satisfy Definition 1 they need only to be bounded error magnitude. Here $\epsilon > 0$ quantifies the precision of the compression operator. In particular, a compression with a large ϵ is coarse/imprecise and a compression with a small ϵ is precise (e.g. $Q(v) = v$ in the extreme case when $\epsilon = 0$). Most popular compressors in machine learning such as the sign compression [17], the Top-K sparsification [15] and the sparsification together with quantization [8] are in fact the ϵ-compressor if the full-precision vector has a bounded norm.

Our next result characterizes the linear convergence of `FedSplit` with the ϵ-compressor in Equation (7).

Theorem 2. *Consider compressed `FedSplit` in (7) and (8) to solve Problem (1), where each $f_i(\cdot)$ is μ-strongly convex and L-smooth. Also, let $\textbf{refl}_{I_E}(z) = 2\bar{z} \cdot \mathbf{1} - z$ be non-expansive. If $\gamma = 1/\sqrt{\mu L}$ and $x^0 = z_1^0 = \ldots = z_n^0$, then*

$$\mathbf{w}^{k+1} \leq A\mathbf{w}^k + \lambda\epsilon \cdot [1, 0]^T, \tag{9}$$

where $\lambda \in (0, 1]$, $\rho = 1 - 2/(\sqrt{L/\mu} + 1)$, and where A and $\mathbf{w}^k$ are defined in (5) and (6). In addition, we have

$$\limsup_{k\to\infty} \|x^k - x^\star\| \leq \epsilon.$$

Theorem 2 establishes a convergence bound on the compressed `FedSplit`. In particular, the convergence rate in Equation (9) has two parts. The first part is the linear rate to the exact optimal solution, similar to the non-compressed `FedSplit` in Theorem 1. The second part is a compression error depending on $\epsilon > 0$ that is not improved as the algorithm progresses. Note that even if we decrease λ the upper bound on $\lim_{k\to\infty} ||x^k - x^\star||$ is not improved. This is because decreasing λ increases the eigenvalues of the matrix A. In particular, the fixed point of the linear dynamical system in Equation (9) is

$$\mathbf{w}^{\texttt{Fix}} = (I - A)^{-1} \begin{bmatrix} \lambda\epsilon & 0 \end{bmatrix}^T = \begin{bmatrix} \epsilon & 0 \end{bmatrix}^T.$$

This means that no matter how we choose γ or λ, we cannot improve the solution accuracy.

The upper bound in Theorem 2 is tight as we show next.

Proposition 1. *There exists a FL problem on the form of (1) such that $\lim_{k\to\infty} ||x^k - x^\star|| = \epsilon$.*

We can prove the proposition with the following example.

Example 1 (Lower Bound)**.** *Consider the FL optimization problem in (1) with $i = 1$, $d = 1$, and $f_1(x) = (\mu/2)x^2$. Let x^k, $\bar{z}^k$, and z_i^k be the iterates of compressed `FedSplit` given in (7) and (8) where $Q(\cdot)$ is the ϵ-compression*

$$Q(z) = \begin{cases} z - \epsilon\frac{z}{|z|} & \text{if } z \neq 0 \\ \epsilon & \text{otherwise.} \end{cases}$$

Suppose also that $z^k = x^k$, $\lambda \in (0, 1]$, $\gamma > 0$ and $x^0 > \theta\epsilon$ where $\theta = 1 + 1/(\mu\gamma)$. Then for all $k \in \mathbb{N}$, we have that $\textbf{refl}_{\gamma f}(2x^k - z^k) = (1/(1 + \mu\gamma))x^k$, and

$$\begin{aligned} |x^{k+1} - x^\star| &= \rho|x^k| + \lambda\epsilon = \rho^{k+1}|x^0| + \lambda\epsilon\sum_{i=0}^{k}\rho^i \\ &= \rho^{k+1}|x^0| + \lambda\epsilon\frac{1-\rho^{k+1}}{1-\rho} = \rho^{k+1}\left(|x^0| - \theta\epsilon\right) + \theta\epsilon \geq \epsilon, \end{aligned}$$

where $\rho = 1 - \lambda + \lambda/(1 + \mu\gamma)$ and $x^\star = 0$. The last inequality follows from the fact that $\theta \geq 1$.

These results show that the solution accuracy of `FedSplit` with direct compression can never be better than the compression precision ϵ. In the next section we illustrate how we can achieve arbitrarily good solution accuracy no matter how large the compression precision ϵ is.

4. Eco-FedSplit

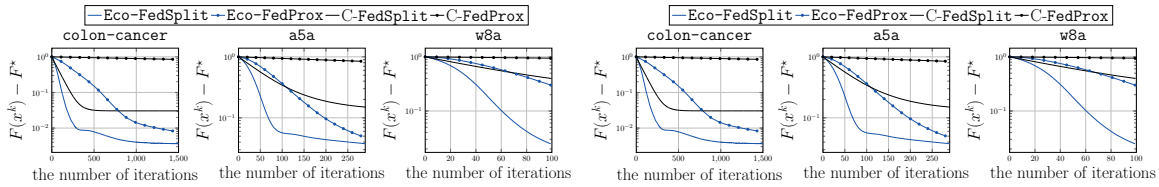
We now illustrate how error-compensation (`Eco`) enables `FedSplit` to achieve significant solution accuracy improvements. We call this algorithm `Eco-FedSplit`, and describe it as follows. The master receives the compensated vectors from all worker nodes, and performs the iteration

$$x^{k+1} = (1 - \lambda)x^k + \lambda\bar{z}^{k+1}, \tag{10}$$

while each worker updates the compression error e_i^k by the following iteration: given $e_i^0 = 0$ for $i = 1, 2, \ldots, n$,

$$\begin{aligned} \bar{z}^{k+1} &= \sum_{i=1}^{n} Q(z_i^{k+1} + (1 - \lambda)e_i^k)/n, \\ z_i^{k+1} &= \textbf{refl}_{\gamma f_i}(2\bar{z}^k - z_i^k), \quad \text{and} \\ e_i^{k+1} &= (z_i^{k+1} + (1 - \lambda)e_i^k) - Q(z_i^{k+1} + (1 - \lambda)e_i^k). \end{aligned} \tag{11}$$

To understand why `Eco-FedSplit` achieves higher solution accuracy than `FedSplit` with direct compression, we compare the closed form of both algorithms. On the one



(a) Sign quantizer

(b) Top-K quantizer with $K = 0.3 \cdot d$

Fig. 1: Convergence of `FedSplit` and `FedProx` with compression (C) and error compensation (Eco).

hand, compressed `FedSplit` in Equation (7) has the following equivalent closed-form

$$x^k = (1-\lambda)^k x^0 + \lambda \sum_{l=0}^{k} (1-\lambda)^{k-l}(\tilde{z}^l + \tilde{e}^l),$$

where $\tilde{z}^k = \sum_{i=1}^n z_i^k/n$, $\tilde{e}^k = \sum_{i=1}^n e_i^k/n$ and $e_i^k = Q(z_i^k) - z_i^k$ for $i = 1, 2, \ldots, n$. On the other hand, `Eco-FedSplit` (10) can be written equivalently as

$$x^k = (1-\lambda)^k x^0 + \lambda \sum_{l=0}^{k} (1-\lambda)^{k-l}\tilde{z}^l + \lambda \tilde{e}^k.$$

Notice that `Eco-FedSplit` fully avoids the accumulations of previous compression errors. In fact, our algorithm can recover the solution with high accuracy by properly adjusting the tuning parameter λ. We illustrate this below:

Theorem 3. *Consider `Eco-FedSplit` in (10) and (11) to solve Problem (1), where each $f_i(\cdot)$ is μ-strongly convex and L-smooth. Also, let $\mathbf{refl}_{I_E}(z) = 2\bar{z} \cdot \mathbf{1} - z$ be non-expansive. If $\gamma = 1/\sqrt{\mu L}$ and $x^0 = z_1^0 = \ldots = z_n^0$, then*

$$\mathbf{w}^{k+1} \leq A\mathbf{w}^k + \lambda^2 \epsilon \cdot [1, 0]^T \tag{12}$$

where $\lambda \in (0, 1]$, $\rho = 1 - 2/(\sqrt{L/\mu} + 1)$, and where A and $\mathbf{w}^k$ are defined in (5) and (6). In addition,

$$\limsup_{k\to\infty} \|x^k - x^\star\| \leq \lambda \cdot \epsilon.$$

Theorem 3 shows that `Eco-FedSplit` converges to the neighborhood of the optimal solution with the same linear rate as the full-precision and compressed `FedSplit`. Similarly as Theorem 2 for compressed `FedSplit`, the convergence bound has the residual term due to the compression ϵ. Compared to direct compression, this upper bound on $\lim_{k\to\infty} ||x^k - x^\star||$ for `Eco-FedSplit` can be made arbitrarily small by reducing λ. For instance, `Eco-FedSplit` obtains the approximate solution with higher accuracy than compressed `FedSplit` when $\lambda < 1$, and with the same accuracy as full-precision `FedSplit` when λ is close to zero. `Eco-FedSplit` with the small λ guarantees significant solution improvements at the cost of the slower rate. This highlights the trade-off between the solution accuracy and the convergence speed for `Eco-FedSplit`, similarly for error-compensated distributed gradient algorithms in [19].

5. EXPERIMENTAL RESULTS

We now illustrate `Eco-FedSplit` on the regularized logistic regression problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \sum_{i=1}^{m} \log(1 + \exp(-b_i \cdot a_i^T x)) + \frac{\sigma}{2}\|x\|^2, \tag{13}$$

where σ is a regularization parameter, and (a_i, b_i) are data points where $a_i \in \mathbb{R}^d$ is the feature vector and $b_i \in \{-1, 1\}$ is the associated class label. We implemented `FedSplit` and `FedProx` with direct compression (C) and error compensation (Eco). We set $\sigma = 10^{-1}$, $\lambda = 10^{-2}$, and $\gamma = 0.06, 0.1$ and 5 for `a5a`, `w8a` and `colon-cancer`, respectively.

From Figure 1(a) and 1(b), `FedSplit` has faster convergence than `FedProx` with both direct and error-compensated compression. Also, error compensation gains significant solution accuracy improvements than direct compression for `FedProx` and `FedSplit`. In particular, `Eco-FedSplit` outperforms other federated methods, in terms of both the speed and accuracy. When training over `colon-cancer`, `Eco-FedSplit` obtains a higher accurate solution than `Eco-FedProx` by an order of magnitude for the sign quantization and two orders of magnitude for the top-K sparsification with $K = 0.3 \cdot d$. To reach accuracy at 10^{-1}, `Eco-FedSplit` requires twice less iteration counts than `Eco-FedProx` to solve the problems on `a5a` and `w8a`.

6. CONCLUSION

We proposed `Eco-FedSplit`, a fast federated learning algorithm with error-compensated compression that improves communication efficiency without sacrificing solution accuracy. Our theoretical results suggest that `Eco-FedSplit` improves both communication efficiency and solution accuracy of traditional federated learning algorithms. In fact, unlike direct compression, error-compensated compression helps `FedSplit` converge toward the approximate solution with arbitrarily high accuracy as we reduce the value of the tuning parameter λ. The superior performance of `Eco-FedSplit` compared to other compressed algorithms in federated learning was illustrated empirically on logistic regression problems on several benchmark data-sets.

7. REFERENCES

[1] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[2] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[3] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[4] Joseph Azar, Abdallah Makhoul, Mahmoud Barhamgi, and Raphaël Couturier, "An energy efficient IoT data compression approach for edge machine learning," *Future Generation Computer Systems*, vol. 96, pp. 168–175, 2019.

[5] Mohammad Abdur Razzaque, Chris Bleakley, and Simon Dobson, "Compression in wireless sensor networks: A survey and comparative evaluation," *ACM Transactions on Sensor Networks (TOSN)*, vol. 10, no. 1, pp. 1–44, 2013.

[6] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli, "The convergence of sparsified gradient methods," in *Advances in Neural Information Processing Systems*, 2018, pp. 5973–5983.

[7] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8253–8265.

[8] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.

[9] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang, "On the convergence of FedAvg on Non-IID data," in *International Conference on Learning Representations*, 2020.

[10] Sebastian U. Stich, "Local SGD converges fast and communicates little," in *International Conference on Learning Representations*, 2019.

[11] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[12] Reese Pathak and Martin J Wainwright, "Fedsplit: an algorithmic framework for fast federated optimization," in *Advances in Neural Information Processing Systems*, 2020, vol. 33.

[13] Donald W Peaceman and Henry H Rachford, Jr, "The numerical solution of parabolic and elliptic differential equations," *Journal of the Society for industrial and Applied Mathematics*, vol. 3, no. 1, pp. 28–41, 1955.

[14] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[15] Sarit Khirirat, Mikael Johansson, and Dan Alistarh, "Gradient compression for communication-limited convex optimization," in *2018 IEEE Conference on Decision and Control (CDC)*, Dec 2018, pp. 166–171.

[16] Sindri Magnússon, Chinwendu Enyioha, Na Li, Carlo Fischione, and Vahid Tarokh, "Convergence of limited communications gradient methods," *IEEE Transactions on Automatic Control*, 2017.

[17] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar, "signSGD: Compressed optimisation for non-convex problems," pp. 560–569, 2018.

[18] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi, "Sparsified SGD with memory," in *Advances in Neural Information Processing Systems*, 2018, pp. 4447–4458.

[19] Sarit Khirirat, Sindri Magnússon, and Mikael Johansson, "Compressed gradient methods with hessian-aided error compensation," *IEEE Transactions on Signal Processing*, vol. 69, pp. 998–1011, 2020.

[20] Jonathan Eckstein, *Splitting methods for monotone operators with applications to parallel optimization*, Ph.D. thesis, Massachusetts Institute of Technology, 1989.

[21] Shi Pu and Angelia Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, pp. 1–49, 2020.