OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Communication-efficient Algorithms Under Generalized Smoothness Assumptions

Sarit Khirirat SARIT.KHIRIRAT@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST)

Abdurakhmon Sadiev ABDURAKHMON.SADIEV@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST)

Artem Riabinin ARTEM.RIABININ@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST)

Eduard Gorbunov EDUARD.GORBUNOV@MBZUAI.AC.AE
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST)

Abstract

We provide the first proof of convergence for normalized error feedback algorithms across a wide range of machine learning problems. Despite their popularity and efficiency in training deep neural networks, traditional analyses of error feedback algorithms rely on the smoothness assumption that does not capture the properties of objective functions in these problems. Rather, these problems have recently been shown to satisfy generalized smoothness assumptions, and the theoretical understanding of error feedback algorithms under these assumptions remains largely unexplored. Moreover, to the best of our knowledge, all existing analyses under generalized smoothness either i) focus on centralized settings or ii) make unrealistically strong assumptions for distributed settings, such as requiring data heterogeneity, and almost surely bounded stochastic gradient noise variance. In this paper, we propose distributed error feedback algorithms that utilize normalization to achieve the $\mathcal{O}(1/\sqrt{K})$ convergence rate for nonconvex problems under generalized smoothness. Our analyses apply for distributed settings without data heterogeneity conditions, and enable stepsize tuning that is independent of problem parameters. Finally, we show that normalized EF21, due to its larger allowable stepsizes, outperforms EF21 on various tasks, including the minimization of polynomial functions, logistic regression, and ResNet-20 training.

1. Introduction

Modern machine learning models achieve impressive prediction and classification power by employing sophisticated architectures, comprising vast numbers of parameters, and requiring training on massive datasets. Distributed training has emerged as an important approach, where multiple machines collaborate to train a model efficiently within a reasonable time. Many centralized algorithms can be easily adapted for distributed training applications. For example, classical gradient descent can be modified into distributed gradient descent within a data parallelism framework, and into federated averaging algorithms [29] in a federated learning framework. However, the communication overhead of running these distributed algorithms poses a significant barrier to scaling up to large models. For example, training the VGG-16 model [37] using distributed stochastic gra-

dient descent involves communicating 138.34 million parameters, thus consuming over 500MB of storage, and posing an unmanageable burden on the communication network between machines.

One approach to mitigate the communication bottleneck is to apply compression. In this approach, the gradients are compressed using sparsifiers or quantizers to be transmitted with much smaller sizes between machines. However, while this reduces communication overhead, too coarse compression often brings substantial challenges in maintaining high training performance due to information loss, and in extreme cases, it may potentially lead to divergence. Therefore, error feedback mechanisms have been developed to improve the convergence performance of compression algorithms while ensuring high communication efficiency. Examples of error feedback mechanisms include EF14 [1, 14, 36, 38, 44], EF21 [11, 33], EF21-SGDM [12], EF21-P [16], and EControl [13]. To the best of our knowledge, these prior works developing and analyzing error feedback algorithms often assume that the objective function is smooth, i.e. its gradient is Lipschitz continuous.

However, training deep neural network often involves non-smooth problems. For instance, the gradients of the loss computed for deep neural networks, such as LSTM [46], ResNet20 [46], and transformer models [8], do not exhibit Lipschitz continuity. These empirical observations highlight the need for a new smoothness assumption. One such assumption is (L_0, L_1)-smoothness, initially introduced by Zhang et al. [46], for twice differentiable functions, and subsequently extended to differentiable functions by Chen et al. [7].

To solve (L_0, L_1)-smooth problems, clipping and normalization have been widely utilized in first-order algorithms. Gradient descent with gradient clipping was initially shown by Zhang et al. [46] to achieve lower iteration complexity (i.e., fewer iterations needed to reach a target solution accuracy) than classical gradient descent. Subsequent works have further refined the convergence theory of clipped gradient descent [21], and improved its convergence performance by employing momentum updates [45], variance reduction techniques [32], and adaptive step sizes [25, 39, 43]. Similar convergence results have been obtained for gradient descent using normalization [47] and its momentum variants [18], including generalized SignSGD [8]. However, these first-order algorithms have mostly been explored in centralized training. To the best of our knowledge, distributed algorithms under (L_0, L_1)-smoothness have been investigated in only a few studies, e.g., by Crawshaw et al. [9], Liu et al. [27], which impose restrictive assumptions, such as data heterogeneity, almost sure variance bounds, and symmetric noise distributions around their means. Moreover, these stated first-order algorithms under (L_0, L_1)-smoothness do not account for communication compression to enhance communication efficiency. Therefore, we are motivated to build *distributed communication-efficient algorithms for solving nonconvex generalized smooth problems*.

Contributions. In this paper, we develop distributed communication-efficient algorithms for nonconvex problems under generalized smoothness. We address the challenge of the generalized smoothness parameter scaling with the gradient norm by introducing gradient normalization to EF21 [33]. Our theoretical analysis of normalized EF21 is based on standard assumptions regarding functions and compressors, and is applicable in distributed settings with any degree of data heterogeneity. Under these conditions, normalized EF21 achieves an $\mathcal{O}(1/\sqrt{K})$ convergence rate in the expected gradient norm, matching the performance of EF21 under traditional smoothness. Numerical experiments on the minimization of polynomial functions, logistic regression, and ResNet20 training demonstrate the superior performance of normalized EF21 compared to the original EF21.

2. Preliminaries

In this paper, we focus on the following distributed optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where n refers to the number of clients, and $f_i(x)$ is the loss of a model parameterized by vector $x \in \mathbb{R}^d$ over its local data owned by client $i \in [n]$.

Next, we impose standard assumptions on objective functions and compression operators.

Assumption 1 *(Lower Bound of f) A function $f : \mathbb{R}^d \to \mathbb{R}$ is bounded from below, i.e., $f^{\inf} = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.*

Assumption 2 *(Lower Bound of f_i) A function $f_i : \mathbb{R}^d \to \mathbb{R}$ is bounded from below, i.e., $f_i^{\inf} = \inf_{x\in\mathbb{R}^d} f_i(x) > -\infty$.*

Assumption 3 *(Symmetric Generalized Smoothness of f_i) A function $f_i : \mathbb{R}^d \to \mathbb{R}$ is symmetric generalized smooth if there exists $L_0, L_1 > 0$ such that for $u_\theta = \theta x + (1-\theta)y$, and for $x, y \in \mathbb{R}^d$,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le \left(L_0 + L_1 \sup_{\theta\in[0,1]} \|\nabla f_i(u_\theta)\|\right)\|x - y\|. \tag{2}$$

Assumption 4 *(Contractive Compressor) An operator $\mathcal{C}^k : \mathbb{R}^d \to \mathbb{R}^d$ is an α-contractive compressor if there exists $\alpha \in (0, 1]$ such that for $k \ge 0$ and $v \in \mathbb{R}^d$,*

$$\mathrm{E}\left[\left\|\mathcal{C}^k(v) - v\right\|^2\right] \le (1-\alpha)\|v\|^2. \tag{3}$$

Assumptions 1 and 2 are standard for analyzing optimization algorithms that solve unconstrained optimization problems. Assumption 3 refers to symmetric generalized smoothness defined by Chen et al. [7], which covers asymmetric generalized smoothness [7, 21], and the original (L_0, L_1)-smoothness by [46]. Moreover, Assumption 3 with $L_1 = 0$ reduces to the traditional L_0-smoothness, under which the convergence of optimization algorithms has been extensively studied [4, 30]. Furthermore, compressors defined by Assumption 4 cover top-k sparsifiers [1, 38], low-rank approximation [34, 42], and various other compressors described in [5, 35].

Notations. We use $[n]$ to denote a set $\{1, 2, \ldots, n\}$, and $\mathrm{E}[u]$ to denote the expectation of a random variable u. Also, $\|\cdot\|$ refers to the Euclidean norm of a vector or spectral norm of a matrix, and $\langle x, y\rangle$ is the inner product between $x, y \in \mathbb{R}^d$.

3. Normalized EF21

We propose a distributed error feedback algorithm to minimize nonconvex generalized smooth functions. To address the issue of the generalized smoothness parameter scaling with the gradient norm, we introduce gradient normalization to EF21 [33], a well-known error feedback variant that achieves an $\mathcal{O}(1/K)$ convergence rate in the squared gradient norm for nonconvex, L-smooth problems. In

Algorithm 1 Normalized EF21

Input: Stepsize $\gamma_k > 0$, initial vectors $x^0, v_i^{-1} \in \mathbb{R}^d$ for $i = 1, 2, \ldots, n$, and α-contractive compressor $\mathcal{C}^k : \mathbb{R}^d \to \mathbb{R}^d$
for each iteration $k = 0, 1, \ldots, K$ **do**
  for each client $i = 1, 2, \ldots, n$ in parallel **do**
    Compute local gradient $\nabla f_i(x^k)$
    Transmit $\Delta_i^k = \mathcal{C}^k(\nabla f_i(x^k) - v_i^{k-1})$
    Update $v_i^k = v_i^{k-1} + \Delta_i^k$
  end for
  Central server computes $v^k = (1/n)\sum_{i=1}^n v_i^k$ via $v_i^k = v_i^{k-1} + \Delta_i^k$.
  Central server updates $x^{k+1} = x^k - \gamma_k \frac{v^k}{\|v^k\|}$.
end for
Output: Final iterates x^{K+1}

particular, normalized EF21 (Algorithm 1), unlike the standard EF21, updates the next iterate x^{k+1} using a normalized gradient descent iteration.

The next result shows that normalized EF21, like EF21 [33] under L-smoothness, achieves the $\mathcal{O}(1/\sqrt{K})$ convergence in the gradient norm under generalized smoothness.

Theorem 1 *Consider Problem (1), where Assumption 1 (lower bound of f), Assumption 2 (lower bound of f_i), Assumption 3 (Asymmetric generalized smoothness of f_i), and Assumption 4 (Contractive compressor) hold. Then, the iterates $\{x^k\}$ generated by normalized EF21 (Algorithm 1) with*

$$\gamma_k = \frac{\gamma_0}{\sqrt{K+1}}$$

for $K \geq 0$ and $\gamma_0 > 0$ satisfy

$$\min_{k=0,1,\ldots,K} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{1}{\sqrt{K+1}}\left[\frac{V^0 \exp(8c_1 L_1 \exp(L_1\gamma_0)\gamma_0^2)}{\gamma_0} + \gamma_0 b\right],$$

where $V^k := f(x^k) - f^{\inf} + \frac{2\gamma_k}{1-\sqrt{1-\alpha}}\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k) - v_i^k\right\|$, $b = 2c_0 + \frac{8L_1c_1}{n}\sum_{i=1}^n (f^{\inf} - f_i^{\inf})$, and $c_i = \left(\frac{1}{2} + 2\frac{\sqrt{1-\alpha}}{1-\sqrt{1-\alpha}}\right) L_i$ for $i = 0, 1$.

Theorem 1 establishes the $\mathcal{O}(1/\sqrt{K})$ convergence in the expectation of gradient norms for Normalized EF21 on nonconvex deterministic problems under generalized smoothness. This rate is the same as Theorem 1 of Richtárik et al. [33] for EF21 under traditional smoothness, and does not depend on data heterogeneity conditions in contrast to Crawshaw et al. [9], Liu et al. [27]. Also, our stepsize depends on any positive constant γ_0, and total iteration count K, without needing to know smoothness constants L_0, L_1 in contrast to Richtárik et al. [33]. Additionally, if we choose $\gamma_0 = 1/(8cL_1)$, then our convergence bound from Theorem 1 becomes

$$\min_{k=0,1,\ldots,K} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \quad \leq \quad \frac{32cL_1V^0 + L_0/L_1 + 2L_1\delta^{\inf}}{\sqrt{K+1}},$$

where $c = \frac{1}{2} + 2\frac{\sqrt{1-\alpha}}{1-\sqrt{1-\alpha}}$, and $\delta^{\inf} = \frac{1}{n}\sum_{i=1}^n (f^{\inf} - f_i^{\inf})$.

Comparisons between normalized EF21 and EF21 under traditional smoothness. For nonconvex, traditional smooth problems, normalized EF21 from Theorem 1 with $L_1 = 0$ achieves the same $\mathcal{O}(1/\sqrt{K})$ rate in the expectation of gradient norms as EF21 analyzed by Richtárik et al. [33], but with a larger convergence factor. We prove this by assuming $\nabla f_i(x^0) = g_i^0$ for all i. That is, Theorem 1 with $L_0 = L$, $L_1 = 0$, $\gamma_0 = \sqrt{(f(x^0) - f^{\inf})/(2b)}$, and $b = \frac{L}{2} + 2\frac{\sqrt{1-\alpha}L}{1-\sqrt{1-\alpha}}$ implies that normalized EF21 achieves

$$\begin{aligned}
\min_{k=0,1,\ldots,K} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] &\leq \frac{1}{\sqrt{K+1}}\left[\frac{f(x^0) - f^{\inf}}{\gamma_0} + 2b\gamma_0\right] \\
&\leq 2\sqrt{L\frac{(1+3\sqrt{1-\alpha})(1+\sqrt{1-\alpha})}{\alpha}}\sqrt{\frac{f(x^0) - f^{\inf}}{K+1}} \\
&\overset{\alpha\geq 0}{\leq} 4\sqrt{2}\sqrt{\frac{L}{\alpha}}\sqrt{\frac{f(x^0) - f^{\inf}}{K+1}}.
\end{aligned}$$

On the other hand, EF21 attains from Theorem 1 of [33] with $L_i = \tilde{L} = L$ (i.e., $f_i(x)$ has the same smoothness constant as $f(x)$), and $\hat{x}^K$ being chosen from the iterates $x^0, x^1, \ldots, x^K$ uniformly at random

$$\begin{aligned}
\min_{k=0,1,\ldots,K} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] &\leq \mathrm{E}\left[\left\|\nabla f(\hat{x}^K)\right\|\right] \\
&\leq \sqrt{\mathrm{E}\left[\left\|\nabla f(\hat{x}^K)\right\|^2\right]} \\
&\leq \sqrt{2L(1+\sqrt{\beta/\theta})\frac{f(x^0) - f^{\inf}}{K+1}} \\
&\overset{\sqrt{\beta/\theta}\leq 2/\alpha - 1}{\leq} 2\sqrt{\frac{L}{\alpha}}\sqrt{\frac{f(x^0) - f^{\inf}}{K+1}}.
\end{aligned}$$

In conclusion, the convergence bound of normalized EF21 is slower by a factor of $2\sqrt{2}$ than the original EF21 for nonconvex, L-smooth problems.

4. Experimental Results

Finally, we demonstrate the stronger convergence performance of normalized EF21 (`EF21-norm`) than EF21 [33] for solving nonconvex problems under asymmetric generalized smoothness. We present the results for solving logistic regression with a nonconvex regularizer over synthetic and benchmark data from LIBSVM [6]. Additional results for minimizing polynomial functions, and for training ResNet-20 over CIFAR10 can be found in Appendix F.

4.1. Logistic Regression with a Nonconvex Regularizer

We consider a logistic regression problem with a nonconvex regularizer, i.e. Problem (1) with

$$f_i(x) = \log(1 + \exp(-b_i a_i^T x)) + \frac{\lambda}{n}\sum_{j=1}^{d}\frac{x_j^2}{1+x_j^2},$$

where $a_i \in \mathbb{R}^d$ is the i^{th} feature vector of data matrix $A \in \mathbb{R}^{n\times d}$ with its class label $b_i \in \{-1, 1\}$, and $\lambda > 0$ is a regularization parameter. Here, $f(x)$ is nonconvex, and L-smooth with $L = \|A\|^2/(4n) + 2\lambda$. Also, each $f_i(x)$ is L_i-smooth with $L_i = \|a_i\|^2/4 + 2\lambda$, and asymmetric generalized smooth with $L_0 = 2\lambda + \lambda\sqrt{d}\max_i \|a_i\|$ and $L_1 = \max_i \|a_i\|$. Detailed derivations of smoothness parameters can be found in Appendix E.

In these experiments, we initialized $x^0 \in \mathbb{R}^d$, where each coordinate was drawn from a standard normal distribution $\mathcal{N}(0, 1)$, set $\lambda = 0.1$, and used a top-k compressor (with $\alpha = k/d$). Here, $\lambda > \lambda_{\min}\left(A^\top A\right)/(2n)$ to ensure that $f(x)$ is nonconvex. We ran normalized EF21 and EF21 over the following datasets: (1) two from LIBSVM [6]: `Breast Cancer` ($n = 683$, $d = 10$, and scaled to $[-1, 1]$), and `a1a` ($n = 1605$, $d = 123$); and (2) a synthetically generated dataset ($n = 20$, $d = 10$), where the data matrix $A \in \mathbb{R}^{n\times d}$ had entries sampled from $\mathcal{N}(0, 1)$, and the class label b_i was set to either -1 or 1 with equal probability. For normalized EF21, we chose $\gamma^k = \gamma_0/\sqrt{K+1}$ with $\gamma_0 > 0$ from Theorem 1, by setting $\gamma_0 = 1$, $K = 100$ for the generated data and `Breast Cancer`, and $K = 400$ for `a1a`. For EF21, we selected $\gamma^k = 1/\left(L + \tilde{L}\sqrt{\frac{\beta}{\theta}}\right)$ with $\tilde{L} = \left(\sum_{i=1}^n \tilde{L}_i^2/n\right)^{1/2}$, $\theta = 1 - \sqrt{1-\alpha}$, and $\beta = (1-\alpha)/(1-\sqrt{1-\alpha})$, as given by [33, Theorem 1].

Figure 1 shows that normalized EF21 outperforms the traditional EF21 on all evaluated datasets, achieving faster convergence and higher solution accuracy. This improvement results from the fact that the theoretical stepsize for normalized EF21, as derived in Theorem 1, is larger than the stepsize for the traditional EF21 outlined by [33, Theorem 1].

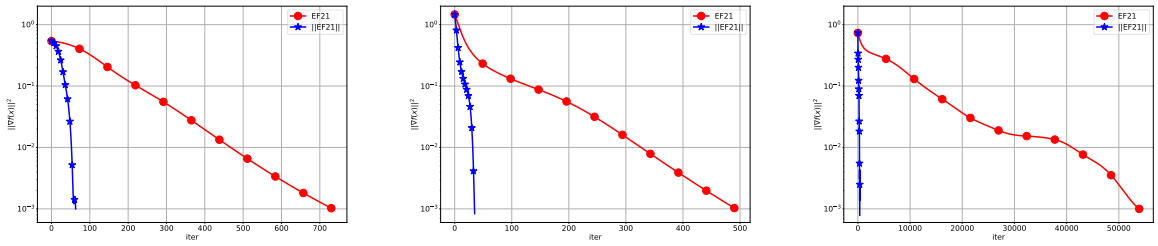

Figure 1: Logistic regression with a nonconvex regularizer using normalized EF21 (`EF21-norm`) and EF21. We reported $\left\|\nabla f(x^k)\right\|^2$ with respect to iteration count k. We used the constant stepsize $\gamma = \frac{1}{L+\tilde{L}\sqrt{\frac{\beta}{\theta}}}$ for EF21, and $\gamma = \frac{\gamma_0}{\sqrt{K+1}}$, $\gamma_0 = 1$ for normalized EF21. Here, $K = 100$ for our generated data (left), and `Breast Cancer` (middle), while $K = 400$ for `a1a` (right).

5. Acknowledgement

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

[1] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[2] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[4] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[5] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

[6] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[7] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR, 2023.

[8] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

[9] Michael Crawshaw, Yajie Bao, and Mingrui Liu. Federated learning with client subsampling, data heterogeneity, and unbounded smoothness: A new algorithm and lower bounds. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

[11] Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

[12] Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Yuan Gao, Rustem Islamov, and Sebastian Stich. EControl: fast distributed optimization with compression and error control. *arXiv preprint arXiv:2311.05645*, 2023.

[14] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. *Advances in Neural Information Processing Systems*, 33: 20889–20900, 2020.

[15] Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

[16] Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-P and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pages 11761–11807. PMLR, 2023.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[18] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR, 2024.

[19] Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34:2771–2782, 2021.

[20] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, 2019.

[21] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

[23] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

[24] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. *arXiv preprint arXiv:2002.11364*, 2020.

[27] Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. *Advances in Neural Information Processing Systems*, 35:26204–26217, 2022.

[28] Vien V Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR, 2021.

[29] 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*, pages 1273–1282. PMLR, 2017.

[30] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[31] Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. *Advances in Neural Information Processing Systems*, 34:30401–30413, 2021.

[32] Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. *arXiv preprint arXiv:2303.00883*, 2023.

[33] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34: 4384–4396, 2021.

[34] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. FedNL: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021.

[35] Mher Safaryan, Egor Shulgin, and Peter Richtárik. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *Information and Inference: A Journal of the IMA*, 11(2):557–580, 2022.

[36] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth annual conference of the international speech communication association*, 2014.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[38] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

[39] Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Polyak meets parameter-free clipped gradient descent. *arXiv preprint arXiv:2405.15010*, 2024.

[40] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, 2019.

[41] Hanlin Tang, Yao Li, Ji Liu, and Ming Yan. Errorcompensatedx: error compensation for variance reduced algorithms. *Advances in Neural Information Processing Systems*, 34:18102–18113, 2021.

[42] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2960–2969, 2024.

[44] Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In *International conference on machine learning*, pages 5325–5333. PMLR, 2018.

[45] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33: 15511–15521, 2020.

[46] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[47] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

Contents

Appendix A. Related Works

In this section, we review prior literature on error feedback, non-smoothness conditions, and clipping and normalization operators.

Error feedback. Error feedback mechanisms have been integrated into various optimization algorithms using compression, leading to significant improvements in solution accuracy. The first of these mechanisms, EF14, was introduced by Seide et al. [36], and later analyzed for first-order algorithms in both centralized [20, 38] and distributed settings [1, 2, 14, 26, 31, 40, 41, 44]. Richtárik et al. [33] proposed EF21, which offers strong convergence guarantees for distributed gradient algorithms with any contractive compressors, without requiring bounded gradient norm or bounded data heterogeneity assumptions. EF21 can also be adapted for distributed stochastic optimization through sufficiently large mini-batches [11] or momentum updates [12]. More recently, Gao et al. [13] have developed EControl that achieves superior complexity results for distributed stochastic optimization compared to prior approaches [12]. To the best of our knowledge, existing research on error feedback has focused solely on optimization problems assuming traditional L-smoothness. In this paper, we introduce a normalized variant of the EF21 methods [33] for solving nonconvex, asymmetric generalized smooth optimization problems. We demonstrate that the normalized EF21 method permits larger step sizes, thus leading to faster convergence rates than the original EF21.

Non-smoothness assumptions. Empirical findings suggest that the traditional smooth assumption used for analyzing optimization algorithms does not capture the properties of objective functions in many machine learning problems, such as distributionally robust optimization and deep neural network training. This motivates researchers to consider different assumptions to reaplce this traditional smoothness condition. First introduced by Zhang et al. [46], the (L_0, L_1)-smoothness condition on a twice differentiable function $f(x)$ is defined by $\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$ for $x \in \mathbb{R}^d$. This (L_0, L_1)-smoothness has been extended to differentiable functions. For instance, asymmetric generalized smoothness [19], the smoothness with a differentiable function $\ell(x)$ [24], and symmetric generalized smoothness [7] covers the (L_0, L_1)-smoothness when the Hessian exists, and includes many important machine learning problems, such as phase retrieval problems [7], and distributionally robust optimization [23]. Other classes of non-smoothness assumptions, which are not related to the (L_0, L_1)-smoothness but capture other optimization problems, include Hölder's Lipschitz continuity of the gradient [10], the relative smoothness [3], and the polynomial growth of the gradient norm [28]. In this paper, we assume the asymmetric generalized smoothness to establish the convergence of normalized EF21.

Gradient clipping and normalization. Clipping and normalization are commonly employed in gradient-based methods for solving generalized smooth problems. Clipped (stochastic) gradient descent has been studied in both nonconvex and convex settings under (L_0, L_1)-smoothness conditions by Koloskova et al. [21], Zhang et al. [46]. Extensions to clipped gradient algorithms have been proposed, including momentum updates [45], variance reduction methods [32], and adaptive step sizes [25, 39, 43]. Comparable complexities have been achieved for normalized gradient descent [47] and its momentum-based variants [18], such as generalized SignSGD [8]. Convergence properties of gradient-based algorithms have also been explored under more generalized forms of non-uniform smoothness, extending beyond (L_0, L_1)-smoothness to cover a wider range of optimization problems. For example, variants of (stochastic) gradient descent have been analyzed under α-symmetric generalized smoothness by Chen et al. [7] and under ℓ-smoothness involving certain differentiable functions $\ell(\cdot)$ by Li et al. [24, 25]. However, the majority of these analyses focus on centralized settings. To the best of our knowledge, only a limited number of works, such as those by Crawshaw et al. [9], Liu et al. [27], have examined federated averaging algorithms for nonconvex problems under (L_0, L_1)-smoothness. These studies, however, often rely on restrictive assumptions, including data heterogeneity, almost sure variance bounds, and symmetric noise distributions centered around their means. In this paper, we develop distributed error feedback algorithms, which eliminate the need for the restrictive assumptions mentioned above, and rely on standard assumptions on objective functions and compressors.

Appendix B. Novel Proof Techniques for Normalized EF21

Our analysis demonstrates that normalized EF21 achieves a convergence rate under generalized smoothness equivalent to EF21 under traditional smoothness. However, our proof techniques differ from the previous work.

Lyapunov Function Innovation. We rely on a different Lyapunov function. For EF21, we use the Lyapunov function $V^k := f(x^k) - f^{\inf} + \frac{A}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - v_i^k\|$, unlike Richtárik et al. [33], which uses $V^k := f(x^k) - f^{\inf} + \frac{B}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - v_i^k\|^2$.

Convergence Rate Derivation. This novel Lyapunov function necessitates the development of new techniques to derive the convergence rate that matches those presented in the prior work. We employ Lemma 2 to handle generalized smoothness. To obtain the convergence rate for EF21, we use Lemma 4, which aligns with the rate achieved by Richtárik et al. [33].

Appendix C. Lemmas

In this section, we introduce useful lemmas for our analysis. Lemmas 2 and 3 introduce inequalities by generalized smoothness, while Lemmas 4 and 5 present the descent inequality and convergence rate, respectively, when the normalized gradient descent update is applied.

Lemma 2 *Let each $f_i(x)$ be generalized smooth with parameters $L_0, L_1 > 0$, and lower bounded by $f_i^{\inf}$, and let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$. Then, for any $x, y \in \mathbb{R}^d$*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq (L_0 + L_1 \|\nabla f_i(y)\|) \exp(L_1 \|x-y\|) \|x-y\|, \tag{4}$$

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y-x\rangle + \frac{L_0 + L_1\|\nabla f_i(x)\|}{2} \exp(L_1\|x-y\|)\|y-x\|^2, \tag{5}$$

$$\frac{\|\nabla f_i(x)\|^2}{4(L_0 + L_1\|\nabla f_i(x)\|)} \leq f_i(x) - f_i^{\inf}, \text{ and} \tag{6}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L_0 + \frac{L_1}{n}\sum_{i=1}^n \|\nabla f_i(x)\|}{2} \exp(L_1\|x-y\|)\|y-x\|^2 \tag{7}$$

Proof The first and second statements are derived in Chen et al. [7, Proposition 3.2]. Next, the third inequality follows from Gorbunov et al. [15, Lemma 2.2]. Finally, averaging (5) for $i = 1, \ldots, n$ and taking into account that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, we get (7). ■

Lemma 3 *Let $f_i(x)$ be generalized smooth with parameters $L_0, L_1 > 0$, and lower bounded by $f_i^{\inf}$, and let $f(x)$ be lower bounded by $f^{\inf}$. Then, for any $x \in \mathbb{R}^d$*

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)\| \leq 8L_1(f(x) - f^{\inf}) + \frac{8L_1}{n}\sum_{i=1}^n (f^{\inf} - f_i^{\inf}) + L_0/L_1. \tag{8}$$

Proof By the (L_0, L_1)-smoothness of $f_i(x)$,

$$4(f_i(x) - f_i^{\inf}) \overset{(6)}{\geq} \frac{\|\nabla f_i(x)\|^2}{L_0 + L_1\|\nabla f_i(x)\|} \geq \begin{cases} \frac{\|\nabla f_i(x)\|^2}{2L_0} & \text{if } \|\nabla f_i(x)\| \leq \frac{L_0}{L_1} \\ \frac{\|\nabla f_i(x)\|}{2L_1} & \text{otherwise.} \end{cases}$$

This condition implies

$$\begin{aligned} \|\nabla f_i(x)\| &\leq \max(8L_1(f_i(x) - f_i^{\inf}), L_0/L_1) \\ &\leq 8L_1(f_i(x) - f_i^{\inf}) + L_0/L_1 \\ &\leq 8L_1(f_i(x) - f^{\inf}) + 8L_1(f^{\inf} - f_i^{\inf}) + L_0/L_1. \end{aligned}$$

Finally, by the fact that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$,

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)\| \leq 8L_1(f(x) - f^{\inf}) + \frac{8L_1}{n}\sum_{i=1}^n (f^{\inf} - f_i^{\inf}) + L_0/L_1.$$

■

Lemma 4 *Let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, where each $f_i(x)$ is generalized smooth with parameters $L_0, L_1 > 0$. Let $x^{k+1} = x^k - \frac{\gamma_k}{\|v^k\|}v^k$ for $\gamma_k > 0$. Then,*

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \gamma_k \left\|\nabla f(x^k)\right\| + 2\gamma_k \left\|\nabla f(x^k) - v^k\right\| \\ &\quad + \frac{\gamma_k^2}{2}\exp\left(\gamma_k L_1\right)\left(L_0 + \frac{L_1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|\right). \end{aligned}$$

Proof Let each $f_i(x)$ be generalized smooth with $L_0, L_1 > 0$, and $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$. By (7) of Lemma 2, and by the fact that $x^{k+1} = x^k - \frac{\gamma_k}{\|v^k\|}v^k$ for $\gamma_k > 0$,

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \frac{\gamma_k}{\|v^k\|}\langle \nabla f(x^k), v^k\rangle + \frac{\gamma_k^2}{2}\exp(\gamma_k L_1)\left(L_0 + \frac{L_1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|\right) \\ &= f(x^k) - \frac{\gamma_k}{\|v^k\|}\langle \nabla f(x^k) - v^k, v^k\rangle - \gamma_k\left\|v^k\right\| \\ &\quad + \frac{\gamma_k^2}{2}\exp(\gamma_k L_1)\left(L_0 + \frac{L_1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|\right) \\ &\leq f(x^k) + \gamma_k\left\|\nabla f(x^k) - v^k\right\| - \gamma_k\left\|v^k\right\| \\ &\quad + \frac{\gamma_k^2}{2}\exp(\gamma_k L_1)\left(L_0 + \frac{L_1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|\right), \end{aligned}$$

where we reach the last inequality by Cauchy-Schwarz inequality. Next, since

$$-\left\|v^k\right\| \overset{\text{triangle ineq.}}{\leq} -\left\|\nabla f(x^k)\right\| + \left\|\nabla f(x^k) - v^k\right\|,$$

we get

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \gamma_k \left\|\nabla f(x^k)\right\| + 2\gamma_k \left\|\nabla f(x^k) - v^k\right\| \\ &\quad + \frac{\gamma_k^2}{2}\exp(\gamma_k L_1)\left(L_0 + \frac{L_1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k)\right\|\right). \end{aligned}$$

■

Lemma 5 *Let $\{V^k\}_{k\geq 0}, \{W^k\}_{k\geq 0}$ be non-negative sequences satisfying*

$$V^{k+1} \leq (1 + b_1\exp(L_1\gamma)\gamma^2)V^k - b_2\gamma W^k + b_3\exp(L_1\gamma)\gamma^2,$$

for $\gamma, b_1, b_2, b_3 > 0$. Then,

$$\min_{k=0,1,\ldots,K} W^k \leq \frac{V^0\exp(b_1\exp(L_1\gamma)\gamma^2(K+1))}{b_2\gamma(K+1)} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma.$$

Proof Define $\beta_k = \frac{\beta_{k-1}}{1+b_1\exp(L_1\gamma)\gamma^2}$ for $k = 0, 1, \ldots$ and $\beta_{-1} = 1$. Then, we can show that $\beta_k = \frac{1}{(1+b_1\exp(L_1\gamma)\gamma^2)^{k+1}}$ for $k = 0, 1, \ldots$, and that

$$\begin{aligned}\beta_k V^{k+1} &\leq (1 + b_1\exp(L_1\gamma)\gamma^2)\beta_k V^k - b_2\gamma\beta_k W^k + b_3\exp(L_1\gamma)\gamma^2\beta_k \\ &= \beta_{k-1}V^k - b_2\gamma\beta_k W^k + b_3\exp(L_1\gamma)\gamma^2\beta_k.\end{aligned}$$

Therefore,

$$\begin{aligned}\min_{k=0,1,\ldots,K} W^k &\leq \frac{1}{\sum_{k=0}^K \beta_k}\sum_{k=0}^K \beta_k W^k \\ &\leq \frac{\sum_{k=0}^K(\beta_{k-1}V^k - \beta_k V^{k+1})}{b_2\gamma\sum_{k=0}^K\beta_k} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma \\ &= \frac{\beta_{-1}V^0 - \beta_K V^{k+1}}{b_2\gamma\sum_{k=0}^K\beta_k} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma.\end{aligned}$$

By the fact that $\beta_{-1} = 1$, $\beta_K > 0$, and $V^{k+1} \geq 0$,

$$\min_{k=0,1,\ldots,K} W^k \leq \frac{V^0}{b_2\gamma\sum_{k=0}^K\beta_k} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma.$$

Next, since

$$\sum_{k=0}^K \beta_k \geq (K+1)\min_{k=0,1,\ldots,K}\beta_k = \frac{K+1}{(1+b_1\exp(L_1\gamma)\gamma^2)^{K+1}},$$

we have

$$\begin{aligned}\min_{k=0,1,\ldots,K} W^k &\leq \frac{V^0(1+b_1\exp(L_1\gamma)\gamma^2)^{K+1}}{b_2\gamma(K+1)} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma \\ &\overset{1+x\leq\exp(x)}{\leq} \frac{V^0\exp(b_1\exp(L_1\gamma)\gamma^2(K+1))}{b_2\gamma(K+1)} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma.\end{aligned}$$

■

Appendix D. Convergence Proof for Normalized EF21 (Theorem 1)

In this section, we derive the convergence rate results of normalized EF21. We start with the following lemma technical lemma.

Lemma 6 *Let Assumptions 3 and 4 hold. Then, the iterates $\{x^k\}$ generated by normalized EF21 (Algorithm 1) satisfy*

$$\begin{aligned}\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\|\right] &\leq \sqrt{1-\alpha}\,\mathrm{E}\left[\left\|\nabla f_i(x^k) - g_i^k\right\|\right] \\ &\quad + \sqrt{1-\alpha}\exp(L_1\gamma_k)\gamma_k(L_0 + L_1\mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right]). \qquad (9)\end{aligned}$$

Proof From the definition of the Euclidean norm, and by taking the expectation conditioned on x^{k+1}, g_i^k, and by the update of g_i^k from Algorithm 1

$$\begin{aligned}
&\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\| \middle| x^{k+1}, g_i^k\right] \\
&= \mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^k - \mathcal{C}^k(\nabla f_i(x^{k+1}) - g_i^k)\right\| \middle| x^{k+1}, g_i^k\right] \\
&\leq \sqrt{\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^k - \mathcal{C}(\nabla f_i(x^{k+1}) - g_i^k)\right\|^2 \middle| x^{k+1}, g_i^k\right]},
\end{aligned}$$

where we use the concavity of the square root function, and Jensen's inequality for the concave function, i.e., $\mathrm{E}\left[f(x)\right] \leq f(\mathrm{E}\left[x\right])$ if $f(x)$ is concave. By the α-contractive property of compressors in (3), by the fact that $\left\|\nabla f_i(x^{k+1}) - g_i^k\right\|$ is a constant conditioned on x^{k+1}, g_i^k, and then by the triangle inequality, we have

$$\begin{aligned}
\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\| \middle| x^{k+1}, g_i^k\right] &\leq \sqrt{(1-\alpha)\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^k\right\|^2 \middle| x^{k+1}, g_i^k\right]} \\
&= \sqrt{1-\alpha}\left\|\nabla f_i(x^{k+1}) - g_i^k\right\| \\
&\leq \sqrt{1-\alpha}\left\|\nabla f_i(x^k) - g_i^k\right\| + \sqrt{1-\alpha}\left\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\right\|.
\end{aligned}$$

By the generalized smoothness of $f_i(x)$ in (2), and by the fact that $x^{k+1} = x^k - \gamma_k \frac{g^k}{\|g^k\|}$,

$$\begin{aligned}
\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\| \middle| x^{k+1}, g_i^k\right] &\leq \sqrt{1-\alpha}\left\|\nabla f_i(x^k) - g_i^k\right\| \\
&\quad + \sqrt{1-\alpha}(L_0 + L_1\left\|\nabla f_i(x^k)\right\|)\exp(L_1\gamma_k)\gamma_k.
\end{aligned}$$

Let $\gamma_k > 0$ be constants conditioned on x^{k+1}, g_i^k. Then, by the tower property, i.e.,

$$\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\|\right] = \mathrm{E}\left[\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\| \middle| x^{k+1}, g_i^k\right]\right],$$

we have

$$\begin{aligned}
\mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\|\right] &\leq \sqrt{1-\alpha}\mathrm{E}\left[\left\|\nabla f_i(x^k) - g_i^k\right\|\right] \\
&\quad + \sqrt{1-\alpha}\exp(L_1\gamma_k)\gamma_k(L_0 + L_1\mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right]).
\end{aligned}$$

This concludes the proof. ■

Next, we present the following descent lemma for normalized EF21.

Lemma 7 *Let Assumptions 1-4 hold. Then, the iterates $\{x^k\}$ generated by normalized EF21 (Algorithm 1) satisfy*

$$\mathrm{E}\left[V^{k+1}\right] \leq \mathrm{E}\left[V^k\right] + c_1\gamma_k^2\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right] - \gamma_k\mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + c_0\gamma_k^2,$$

where $V^k := f(x^k) - f^{\inf} + \frac{2\gamma_k}{1-\sqrt{1-\alpha}}\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k) - g_i^k\right\|$, and $c_i = \frac{L_i}{2} + 2\frac{\sqrt{1-\alpha}L_i}{1-\sqrt{1-\alpha}}$ for $i = 0, 1$.

Proof

For brevity, let $A_k = \frac{2\gamma_k}{1-\sqrt{1-\alpha}}$. Then, we have $V^k := f(x^k) - f^{\inf} + A_k \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x^k) - v_i^k\right\|$, and from Lemma 4, we derive

$$\begin{aligned}
\mathrm{E}\left[V^{k+1}\right] \quad \leq \quad & \mathrm{E}\left[f(x^k) - f^{\inf}\right] - \gamma_k \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \\
& + \exp(L_1\gamma_k)\gamma_k^2 \frac{L_1}{2n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right] + \exp(L_1\gamma_k)\gamma_k^2\frac{L_0}{2} \\
& + 2\gamma_k \mathrm{E}\left[\left\|\nabla f(x^k) - g^k\right\|\right] + A_{k+1}\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\|\right].
\end{aligned}$$

Identities $\nabla f(x^k) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^k)$ and $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$ and the triangle inequality imply

$$\begin{aligned}
\mathrm{E}\left[V^{k+1}\right] \quad \leq \quad & \mathrm{E}\left[f(x^k) - f^{\inf}\right] - \gamma_k \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \\
& + \exp(L_1\gamma_k)\gamma_k^2 \frac{L_1}{2n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right] + \exp(L_1\gamma_k)\gamma_k^2\frac{L_0}{2} \\
& + 2\gamma_k \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k) - g_i^k\right\|\right] + A_{k+1}\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^{k+1}) - g_i^{k+1}\right\|\right].
\end{aligned}$$

Next, we apply (9):

$$\begin{aligned}
\mathrm{E}\left[V^{k+1}\right] \quad \leq \quad & \mathrm{E}\left[f(x^k) - f^{\inf}\right] - \gamma_k \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + \left(\frac{\gamma_k^2}{2} + A_{k+1}\sqrt{1-\alpha}\gamma_k\right)\exp(L_1\gamma_k)L_0 \\
& + \left(\frac{\gamma_k^2}{2} + A_{k+1}\sqrt{1-\alpha}\gamma_k\right)\exp(L_1\gamma_k)L_1\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right] \\
& + \left(2\gamma_k + A_{k+1}\sqrt{1-\alpha}\right)\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k) - g_i^k\right\|\right].
\end{aligned}$$

If $A_k = \frac{2\gamma_k}{1-\sqrt{1-\alpha}}$, and γ_k satisfies $\gamma_{k+1} \leq \gamma_k$, then

$$2\gamma_k + A_{k+1}\sqrt{1-\alpha} \leq 2\gamma_k + A_k\sqrt{1-\alpha} = A_k.$$

Therefore,

$$\begin{aligned}
\mathrm{E}\left[V^{k+1}\right] \quad \leq \quad & \mathrm{E}\left[V^k\right] + c_1\exp(L_1\gamma_k)\gamma_k^2\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[\left\|\nabla f_i(x^k)\right\|\right] \\
& - \gamma_k \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + c_0\exp(L_1\gamma_k)\gamma_k^2,
\end{aligned}$$

where $c_i = \frac{L_i}{2} + 2\frac{\sqrt{1-\alpha}L_i}{1-\sqrt{1-\alpha}}$ for $i = 0, 1$. ■

D.1. Proof of Theorem 1

Now, we are ready to prove Theorem 1. From Lemma 7 and 3, and by the fact that $c_1 L_0/L_1 = c_0$, we have

$$\begin{aligned}\mathrm{E}\left[V^{k+1}\right] \quad \leq \quad & \mathrm{E}\left[V^{k}\right] + 8c_1 L_1 \exp(L_1\gamma_k)\gamma_k^2 \mathrm{E}\left[f(x^k) - f^{\inf}\right] \\ & -\gamma_k \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + B\exp(L_1\gamma_k)\gamma_k^2,\end{aligned}$$

where $B = 2c_0 + \frac{8c_1L_1}{n}\sum_{i=1}^n (f^{\inf} - f_i^{\inf})$. Using the fact that $f(x^k) - f^{\inf} \leq V^k$, we derive

$$\mathrm{E}\left[V^{k+1}\right] \leq (1 + 8c_1L_1\exp(L_1\gamma_k)\gamma_k^2)\mathrm{E}\left[V^k\right] - \gamma_k \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + B\exp(L_1\gamma_k)\gamma_k^2.$$

Applying Lemma 5 with $V^k = \mathrm{E}\left[V^k\right]$, $W^k = \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right]$, $b_1 = 8c_1L_1$, $b_2 = 1$, and $b_3 = B$, we get

$$\min_{k=0,1,\ldots,K} W^k \leq \frac{V^0 \exp(b_1 \exp(L_1\gamma)\gamma^2(K+1))}{b_2\gamma(K+1)} + \frac{b_3}{b_2}\exp(L_1\gamma)\gamma.$$

Finally, if $\gamma = \frac{\gamma_0}{\sqrt{K+1}}$ with $\gamma_0 > 0$, then $\exp(L_1\gamma_k) \leq \exp(L_1\gamma_0)$, and thus

$$\min_{k=0,1,\ldots,K} W^k \leq \frac{V^0 \exp(b_1 \exp(L_1\gamma_0)\gamma_0^2)}{b_2\gamma_0\sqrt{K+1}} + \frac{b_3}{b_2}\frac{\gamma_0\exp(L_1\gamma_0)}{\sqrt{K+1}}.$$

Appendix E. Omitted Proof for Generalized Smoothness of Logistic Regression

In this section, we prove the generalized smoothness parameters L_0, L_1 for logistic regression problems with a nonconvex regularizer, which are the following problems

$$\min_{x\in\mathbb{R}^d}\left\{ f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) := \frac{1}{n}\sum_{i=1}^n \underbrace{\log(1+\exp(-b_i a_i^T x))}_{=:\tilde{f}_i(x)} + \underbrace{\lambda\sum_{j=1}^d \frac{x_j^2}{1+x_j^2}}_{=:h(x)}\right\},$$

where $a_i \in \mathbb{R}^d$ is the i^{th} feature vector of matrix A with its class label $b_i \in \{-1, 1\}$, $\lambda > 0$.

First, we can prove that $f(x)$ is L-smooth with $L = \frac{1}{4n}\|A\|^2 + 2\lambda$, and that each $f_i(x)$ is $\tilde{L}_i$-smooth with $\tilde{L}_i = \frac{1}{4}\|a_i\|^2 + 2\lambda$.

Next, we show that each $f_i(x)$ is asymmetric generalized smooth with $L_0 = 2\lambda + \lambda\sqrt{d}\max_i \|a_i\|$ and $L_1 = \max_i \|a_i\|$, when the Hessian exists. By the fact that

$$\nabla \tilde{f}_i(x) = -\frac{\exp(-b_i a_i^T x)}{1+\exp(-b_i a_i^T x)} b_i a_i, \quad \text{and} \quad \nabla^2 \tilde{f}_i(x) = \frac{\exp(-b_i a_i^T x)}{(1+\exp(-b_i a_i^T x))^2} b_i^2 a_i a_i^T,$$

we have

$$\begin{aligned}\left\|\nabla^2 \tilde{f}_i(x)\right\| &\overset{b_i\in\{-1,1\}}{=} \frac{\exp(-b_i a_i^T x)}{(1+\exp(-b_i a_i^T x))^2}\lambda_{\max}(a_i a_i^T)\\ &= \frac{\exp(-b_i a_i^T x)}{(1+\exp(-b_i a_i^T x))^2}\left\|a_i\right\|^2\\ &= \frac{\|a_i\|}{1+\exp(-b_i a_i^T x)}\left\|\nabla \tilde{f}_i(x)\right\|\\ &\leq \|a_i\|\left\|\nabla \tilde{f}_i(x)\right\|. \end{aligned} \tag{10}$$

After adding the nonconvex regularizer $h(x)$, we can show the following inequalities:

$$\begin{aligned}\left\|\nabla^2 f_i(x)\right\| &\leq \left\|\nabla^2 \tilde{f}_i(x)\right\| + \left\|\nabla^2 h(x)\right\|\\ &\leq \left\|\nabla^2 \tilde{f}_i(x)\right\| + 2\lambda, \end{aligned} \tag{11}$$

and

$$\begin{aligned}\|\nabla f_i(x)\| \geq \left\|\nabla \tilde{f}_i(x)\right\| - \|\nabla h(x)\| &= \left\|\nabla \tilde{f}_i(x)\right\| - \sqrt{\left(\frac{2\lambda x_1}{(1+x_1^2)^2}\right)^2 + \ldots + \left(\frac{2\lambda x_d}{(1+x_d^2)^2}\right)^2}\\ &\geq \left\|\nabla \tilde{f}_i(x)\right\| - \sqrt{\lambda^2 + \ldots + \lambda^2}\\ &= \left\|\nabla \tilde{f}_i(x)\right\| - \lambda\sqrt{d}. \end{aligned} \tag{12}$$

By combining inequalities (10), (11), and (12), we obtain

$$\left\|\nabla^2 f_i(x)\right\| \leq \underbrace{2\lambda + \lambda\sqrt{d}\left\|a_i\right\|}_{=L_0} + \underbrace{\|a_i\|}_{=L_1}\|\nabla f_i(x)\| = L_0 + L_1\|\nabla f_i(x)\|.$$

Appendix F. Additional Experiments

In this section, we present additional experimental results for minimizing nonconvex polynomial functions, and training ResNet20 models on the CIFAR-10 dataset.

F.1. Minimization of Nonconvex Polynomial Functions

We ran normalized EF21 (`EF21-norm`), and traditional EF21 in a centralized setting ($n = 1$) for solving the following problem:

$$\min_{x\in\mathbb{R}^d}\left\{f(x) := \underbrace{\sum_{i=1}^d a_i x_i^4}_{=:g(x)} + \underbrace{\lambda\sum_{i=1}^d \frac{x_i^2}{1+x_i^2}}_{=:h(x)}\right\},$$

where $a_i > 0$, $i = 1, \ldots, d$, $\lambda > 0$.

Let us show that $f(x)$ is non-convex (for the specific choice of a_i) and (L_0, L_1)-smooth. First, we prove that $f(x)$ is non-convex. Indeed,

$$\begin{aligned}\nabla^2 f(x) &= \nabla^2 g(x) + \nabla^2 h(x) \\ &= 12\,\mathrm{diag}\left\{a_1 x_1^2, \ldots, a_d x_d^2\right\} + 2\lambda\,\mathrm{diag}\left\{\frac{1-3x_1^2}{\left(1+x_1^2\right)^3}, \ldots, \frac{1-3x_d^2}{\left(1+x_d^2\right)^3}\right\},\end{aligned}$$

is not positive definite matrix if we choose $a_i = \frac{\lambda}{24}$, $x_i = \pm 1$ for $i = 1, \ldots, d$.

Second, we find $L_0, L_1 > 0$ such that $\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|, \quad \forall x \in \mathbb{R}^d$. Let us fix some $L_1 > 0$ and choose $L_0 = \frac{9\lambda d^2}{2L_1^2} + 2\lambda$. Since $\nabla^2 h(x) \preccurlyeq 2\lambda I$,

$$\begin{aligned}\left\|\nabla^2 f(x)\right\| = \left\|\nabla^2 g(x) + \nabla^2 h(x)\right\| &\leq \left\|\nabla^2 g(x)\right\| + \left\|\nabla^2 h(x)\right\| \\ &\leq 12\sqrt{a_1^2 x_1^4 + \ldots + a_d^2 x_d^4} + 2\lambda \\ &\leq 12\left(a_1 x_1^2 + \ldots + a_d x_d^2\right) + 2\lambda.\end{aligned}$$

Also, notice that

$$\begin{aligned}\|\nabla f(x)\| = \|\nabla g(x) + \nabla h(x)\| &= \sqrt{\left(4a_1 x_1^2 + \frac{2\lambda}{(1+x_1^2)^2}\right)^2 x_1^2 + \ldots + \left(4a_d x_d^2 + \frac{2\lambda}{(1+x_d^2)^2}\right)^2 x_d^2} \\ &\geq 4\sqrt{a_1^2 x_1^6 + \ldots + a_d^2 x_d^6} \\ &\geq \frac{4}{\sqrt{d}}\left(a_1 |x_1|^3 + \ldots + a_d |x_d|^3\right).\end{aligned}$$

Our goal is to show that

$$12\left(a_1 x_1^2 + \ldots + a_d x_d^2\right) \leq \tilde{L}_0 + \frac{4L_1}{\sqrt{d}}\left(a_1 |x_1|^3 + \ldots + a_d |x_d|^3\right), \quad \tilde{L}_0 = L_0 - 2\lambda.$$

To show this, we consider two cases: if $|x_i| \leq \frac{3\sqrt{d}}{L_1}$, and otherwise.

1. If $|x_i| \leq \frac{3\sqrt{d}}{L_1}$ for all $i = 1, \ldots, d$, then $12 a_i x_i^2 \leq \frac{108 a_i d}{L_1^2}$. Thus, $12\left(a_1 x_1^2 + \ldots + a_d x_d^2\right) \leq \frac{108\lambda d^2}{24 L_1^2} = \tilde{L}_0$.
2. If $|x_j| > \frac{3\sqrt{d}}{L_1}$ for some $j = 1, \ldots, d$, then $12 a_j x_j^2 < \frac{4L_1}{\sqrt{d}} a_j |x_j|^3$, and the sum of the remaining terms (such that $|x_i| \leq \frac{3\sqrt{d}}{L_1}$) in $12\left(a_1 x_1^2 + \ldots + a_d x_d^2\right)$ can be upper bounded by $\tilde{L}_0$.

In conclusion, $f(x)$ is (L_0, L_1)-smooth, where L_1 is any positive constant and $L_0 = \frac{9\lambda d^2}{2L_1^2} + 2\lambda$.

Additionally, we can show that under certain additional constraints, $f(x)$ is L-smooth with $L = \frac{\lambda\sqrt{d}D^2}{2} + 2\lambda$. If $|x_i| \leq D$ for all $i = 1, \ldots, d$, then

$$\left\|\nabla^2 f(x)\right\| \leq 12\sqrt{a_1^2 x_1^4 + \ldots + a_d^2 x_d^4} + 2\lambda \leq \frac{\lambda\sqrt{d}D^2}{2} + 2\lambda = L,$$

In the experiments, we estimate D based on the initial point $x^0 \in \mathbb{R}^d$.

In the following experiments, we utilized a top-k compressor with $k = 1$ and $\alpha = k/d$, setting $d = 4$, $L_1 = \{1, 4, 8\}$, and $L_0 = 4$ (adjusting λ to maintain a constant L_0). The initial values x^0 were drawn from a normal distribution, $x_i^0 \sim \mathcal{N}(20, 1)$ for $i = 1, \ldots, d$, with D estimated as 20. For EF21, we set $\gamma_k = \frac{1}{L+L\sqrt{\frac{\beta}{\theta}}}$, using $\theta = 1 - \sqrt{1-\alpha}$ and $\beta = \frac{1-\alpha}{1-\sqrt{1-\alpha}}$ as outlined in Theorem 1 of [33]. For normalized EF21, we selected $\gamma_k = \frac{1}{2c_1}$, where $c_1 = \frac{L_1}{2} + 2\frac{\sqrt{1-\alpha}L_1}{1-\sqrt{1-\alpha}}$, and also used $\gamma_k = \frac{\gamma_0}{\sqrt{K+1}}$ with $\gamma_0 > 0$, as specified in Theorem 1.

The impact of γ_0 and K on the convergence of normalized EF21. First, we investigate the impact of γ_0 and K on the convergence of normalized EF21. We evaluated γ_0 from the set $\{0.1, 1, 10\}$, and plotted the histogram representing the number of iterations required to achieve the target accuracy of $\|\nabla f(x)\|^2 < \epsilon$ with $\epsilon = 10^{-4}$, using the stepsize rule $\gamma = \frac{\gamma_0}{\sqrt{K+1}}$. For each γ_0, we determined K as the minimum number of iterations required to achieve the desired accuracy, found through a grid search with step sizes of 500 for $\gamma_0 = 1, 10$ and 5000 for $\gamma_0 = 0.1$. From Figure 2,

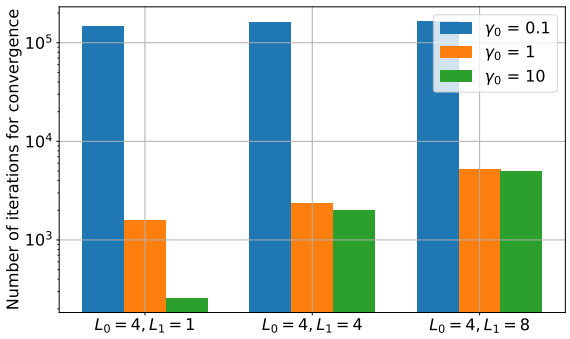

Figure 2: Number of iterations required to achieve the desired accuracy, $\|\nabla f(x)\|^2 < \epsilon$, $\epsilon = 10^{-4}$, using normalized EF21 (`EF21-norm`) with $\gamma = \frac{\gamma_0}{\sqrt{K+1}}$ for different values of L_0 and L_1.

for small values of γ_0, such as 0.1, significantly more iterations are required to reach convergence compared to γ_0 values of 1 and 10, which show similar performance (with the exception of the $L_0 = 4$, $L_1 = 1$ case, where $\gamma_0 = 10$ converges faster). Based on this observation, we use $\gamma_0 = 1$ in all subsequent experiments and adjust only K to achieve convergence, identifying the minimum number of iterations needed to reach the target accuracy through a grid search with a step size of 500.

Comparisons between EF21 and normalized EF21. Next, we evaluate the performance of EF21 and normalized EF21 for a fixed $L_0 = 4$ and varying L_1 values of {1, 4, 8}. From Figure 3, normalized EF21, regardless of the chosen stepsize γ, achieves the desired accuracy $\|\nabla f(x)\|^2 < \epsilon$ with $\epsilon = 10^{-4}$ faster than the original EF21. Initially, however, EF21 converges more quickly, likely because normalized EF21 employs normalized gradients, which can be slower at the start due to the large gradients when the initial point is far from the stationary point. Moreover, as L_1 increases, both methods show slower convergence. Finally, Figure 4 illustrates that the original EF21 algorithm may diverge if the function $f(x)$ is (L_0, L_1)-smooth, while normalized EF21 still converges successfully.

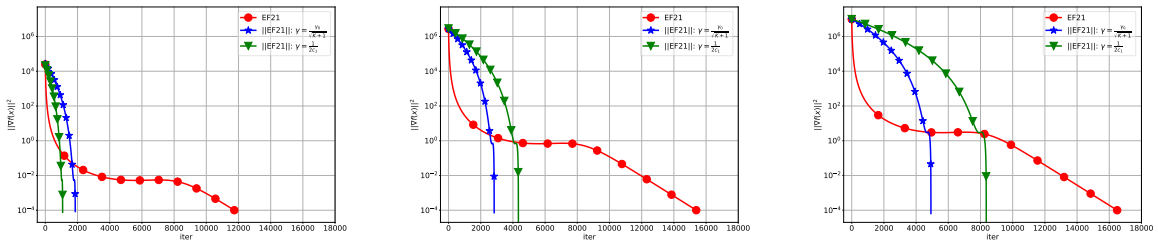

Figure 3: The minimization of polynomial functions using EF21 with $\gamma = \frac{1}{L+L\sqrt{\frac{\beta}{\theta}}}$ and normalized EF21 (`EF21-norm`) with $\gamma = \frac{\gamma_0}{\sqrt{K+1}}$, $\gamma_0 = 1$ (blue line) and $\gamma = \frac{1}{2c_1}$ (green line). Here, we ran the algorithms for (1) $L_0 = 4$, $L_1 = 1$, and $K = 2000$ (left), (2) $L_0 = 4$, $L_1 = 4$, and $K = 5000$ (middle), and (3) $L_0 = 4$, $L_1 = 8$, and $K = 16000$ (right).

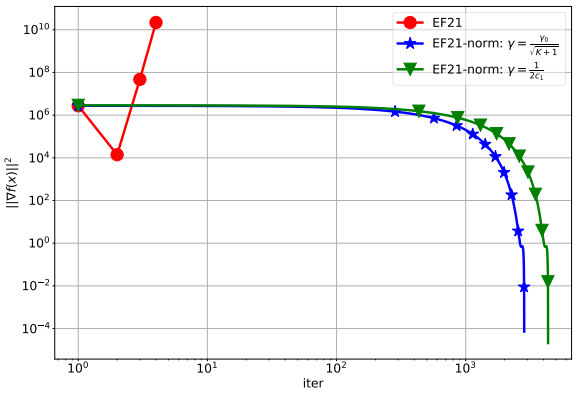

Figure 4: The minimization of polynomial functions using EF21 with $\gamma = \frac{1}{L_0+L_0\sqrt{\frac{\beta}{\theta}}}$ and normalized EF21 (`EF21-norm`) with $\gamma = \frac{\gamma_0}{\sqrt{K+1}}$, $\gamma_0 = 1$ (blue line) and $\gamma = \frac{1}{2c_1}$ (green line), $L_0 = 4$, $L_1 = 4$, $K = 5000$.

F.2. Neural Network Training Over CIFAR-10

We conduct training on the ResNet20 [17] model on the CIFAR-10 [22] dataset, which was demonstrated empirically by [46] to satisfy the (L_0, L_1)-smoothness condition.

To compare the performance between EF21 and normalized EF21, we trained ResNet20 using a top-k compressor over $50,000$ training images, with evaluation on $10,000$ test images. The dataset was evenly distributed among 5 clients, each using a mini-batch size of 128. Both algorithms were run for 100 epochs with a constant stepsize $\gamma = 5$. Here, one epoch refers to a full pass through the entire dataset processed by all clients.

From Figure 5, under the same constant stepsize and the top-k compressor with $k = 0.01d$, normalized EF21 outperforms EF21, in both training and test accuracy relative to the number of bits communicated from each client to the server. Specifically, normalized EF21 achieved accuracy gains of up to 10% over EF21. Similar patterns were observed when we vary the top-k compression parameter, such as $k = 0.1d$ as shown in Figure 6.

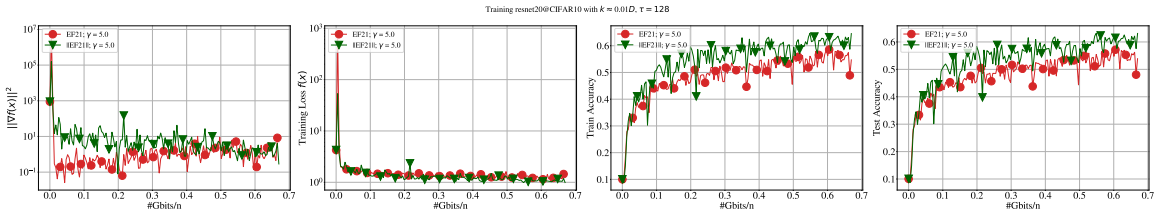

Figure 5: ResNet20 training on CIFAR-10 by using EF21 and normalized EF21 (`EF21-norm`) under the same stepsize $\gamma = 5$ and $k = 0.01d$ for a top-k compressor.

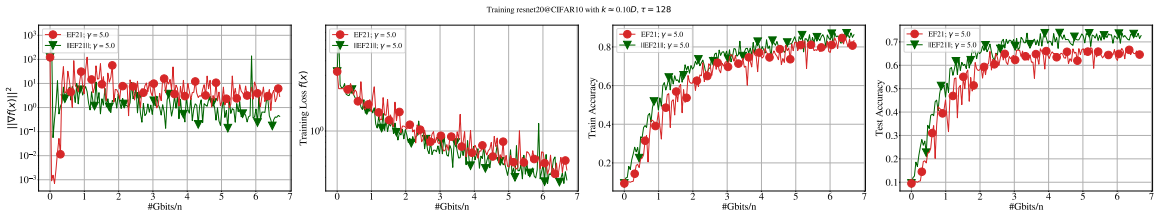

Figure 6: ResNet20 training on CIFAR-10 by using EF21 and normalized EF21 (`EF21-norm`) under the same stepsize $\gamma = 5$ and $k = 0.1d$ for a top-k compressor.