EF21 with Bells & Whistles: Practical Algorithmic Extensions of Modern Error Feedback

Ilyas Fatkhulin [1,2] Igor Sokolov [1] Eduard Gorbunov [3,4] Zhize Li [1] Peter Richtárik[1]

[1]KAUST [2]TU Munich [3]MIPT [4]Yandex

The problem

Nonconvex *distributed* optimization problem:

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

- n – number of clients
- $f_i(x)$ – smooth local loss function, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \le L_i\|x-y\|$ for all $x, y \in \mathbb{R}^d$, $f^{\inf} := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$

Goal: find $\hat{x}$ such that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \le \varepsilon^2$

Compressed learning

Biased compressor: a (possibly randomized) map $\mathcal{C}:\mathbb{R}^d \to \mathbb{R}^d$ is called a *biased compressor*, if there exists a constant $0 < \alpha \le 1$:

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \le (1-\alpha)\|x\|^2, \qquad \forall x \in \mathbb{R}^d.$$

Top-k (greedy) sparsification operator is defined via

$$\mathcal{C}(x) := \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)},$$

where $\left|x_{(1)}\right| \le \left|x_{(2)}\right| \le \dots \le \left|x_{(d)}\right|$. Then $\alpha = \frac{k}{d}$.

Development of error feedback mechanism

◇ Motivation for error feedback – the method of type

$$x^{t+1} = x^t - \gamma\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}\left(\nabla f_i(x^t)\right)$$

- **may diverge** [1] for a biased compressor $\mathcal{C}$ and $n > 1$.

◇ Original error feedback (EF)

$$x^{t+1} = x^t - \gamma w^t, \quad w^t = \frac{1}{n}\sum_{i=1}^{n} w_i^t,$$

$$e_i^{t+1} = e_i^t + \gamma\nabla f_i(x^t) - w_i^t,$$

$$w_i^{t+1} = \mathcal{C}\left(e_i^{t+1} + \gamma\nabla f_i(x^{t+1})\right):$$

- bounded gradients assumption $\|\nabla f_i(x)\| \le G$
- not optimal complexity $\mathcal{O}\left(1/\varepsilon^3\right)$

◇ Modern error feedback [2]:

Algorithm 1: EF21

for $t = 0, 1, \dots, T-1$ **do**
  Master computes
  $$x^{t+1} = x^t - \gamma g^t$$
  and broadcasts x^{t+1} to all nodes
  for *all nodes* $i = 1, \dots, n$ *in parallel* **do**
    Compress $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
    Update local state $g_i^{t+1} = g_i^t + c_i^t$
  end
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^{n} g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^{n} c_i^t$
end

- easy to implement and analyze
- optimal complexity $\mathcal{O}\left(1/\varepsilon^2\right)$
- better in practice

Main contribution

We propose six practical extensions of EF21 method, obtaining state-of-the-art theoretical results for error feedback mechanism.

1. EF21 with stochastic gradients

$$f_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right].$$

EF21-SGD method: $x^{t+1} = x^t - \gamma g^t, \quad g^t = \frac{1}{n}\sum_{i=1}^{n} g_i^t,$

$$g_i^{t+1} = g_i^t + \mathcal{C}\left(\frac{1}{|I_i^t|}\sum_{j\in I_i^t}\nabla f_{ij}(x^{t+1}) - g_i^t\right).$$

Assumption 1. [General assumption for stochastic gradients.] There exist parameters $A_i, C_i \ge 0, B_i \ge 1$ such that

$$\mathbb{E}\left[\left\|\nabla f_{ij}(x^t)\right\|^2 \mid x^t\right] \le 2A_i\left(f_i(x^t) - f_i^{\inf}\right) + B_i\left\|\nabla f_i(x^t)\right\|^2 + C_i,$$

where $j \in I_i^t$, $f_i^{\inf} = \inf_{x\in\mathbb{R}^d} f_i(x) > -\infty$.

- **UBV** assumption is a special case with $A_i = 0, B_i = 1, C_i = \sigma_i^2$
- holds for **arbitrary samplings**, e.g., independent sampling (IS)

2. EF21 with Variance Reduction

$$f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x). \tag{1}$$

EF21-PAGE method: $x^{t+1} = x^t - \gamma g^t, \quad g^t = \frac{1}{n}\sum_{i=1}^{n} g_i^t,$

$$v_i^{t+1} = \begin{cases}\nabla f_i(x^{t+1}), & \mathrm{Be}(p_i) = 1,\\ v_i^t + \frac{1}{|I_i^t|}\sum_{j\in I_i^t}\left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right), & \mathrm{Be}(p_i) = 0,\end{cases}$$
$$g_i^{t+1} = g_i^t + \mathcal{C}\left(v_i^{t+1} - g_i^t\right),$$

Assumption 2. [Average $\mathcal{L}$-smoothness] Let every f_i have the form (1). Assume that for all $t \ge 0$, all nodes $i = 1, \dots, n$, and batch I_i^t (of size τ_i), the minibatch stochastic gradients difference $\widetilde{\Delta}_i^t := \frac{1}{\tau_i}\sum_{j\in I_i^t}(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t))$ satisfies $\mathbb{E}\left[\widetilde{\Delta}_i^t \mid x^t, x^{t+1}\right] = \Delta_i^t$ and

$$\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t - \Delta_i^t\right\|^2 \mid x^t, x^{t+1}\right] \le \frac{\mathcal{L}_i^2}{\tau_i}\|x^{t+1} - x^t\|^2$$

with some $\mathcal{L}_i \ge 0$, where $\Delta_i^t := \nabla f_i(x^{t+1}) - \nabla f_i(x^t)$.

- if I_i^t is a full batch, then $\mathcal{L}_i = 0$
- for uniform sampling and f_{ij} is L_{ij}-smooth, $\mathcal{L}_i \le \max_{1\le j\le m} L_{ij}$

3. EF21 with Bidirectional Compression

EF21-BC method: $x^{t+1} = x^t - \gamma g^t,$

$$g^{t+1} = g^t + \mathcal{C}_M\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i^{t+1} - g^t\right),$$
$$\tilde{g}_i^{t+1} = \tilde{g}_i^t + \mathcal{C}_w\left(\nabla f_i(x^{t+1}) - \tilde{g}_i^t\right).$$

- some applications require compression in both directions

Convergence theory

Setup	Method	$\overline{\text{\#grads}}$	Comment
Full grads	EF21 [2]	$\frac{1}{\alpha\varepsilon^2}$	
Stochastic gradients	Choco-SGD [3]	$\frac{1}{\varepsilon^2} + \frac{G}{\alpha\varepsilon^3} + \frac{\sigma^2}{n\varepsilon^4}$	$\|\nabla f_i(x)\| \le G$
	EF21-SGD [2]	$\frac{1}{\alpha\varepsilon^2} + \frac{\sigma^2}{\alpha^3\varepsilon^4}$	UBV
	EF21-SGD	$\frac{1}{\alpha\varepsilon^2} + \frac{1+\Delta^{\inf}}{\alpha^3\varepsilon^4}$	IS
	EF21-PAGE	$m + \frac{\sqrt{m}+1/\alpha}{\varepsilon^2}$	$f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$
BC	DoubleSqueeze [4]	$\frac{1}{\varepsilon^2} + \frac{\Delta}{\varepsilon^3} + \frac{\sigma^2}{n\varepsilon^4}$	$\mathbb{E}\left[\|\mathcal{C}(x) - x\|\right] \le \Delta$
	EF21-BC	$\frac{1}{\alpha_w\alpha_M\varepsilon^2}$	
PP	EF21-PP	$\frac{1}{p\alpha\varepsilon^2}$	
Mom.	M-CSER [5]	$\frac{1}{\varepsilon^2} + \frac{G}{(1-\eta)\alpha\varepsilon^3}$	$\|\nabla f_i(x)\| \le G$
	EF21-HB	$\frac{1}{\varepsilon^2}\left(\frac{1}{1-\eta} + \frac{1}{\alpha}\right)$	
Prox	EF21-Prox	$\frac{1}{\alpha\varepsilon^2}$	

EF21-SGD, EF21-PAGE, EF21-BC, EF21-PP, EF21-Prox were also analyzed under PŁ condition, i.e., $f(x) - f(x^\star) \le \frac{1}{2\mu}\|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^d$, where $x^\star = \arg\min_{x\in\mathbb{R}^d} f(x)$.

4. EF21 with Partial Participation of Devices

EF21-PP method: $x^{t+1} = x^t - \gamma g^t, \quad g^t = \frac{1}{n}\sum_{i=1}^{n} g_i^t,$ sample a subset of devices $S_t \subset \{1, \dots, n\}$: $\mathrm{Prob}\,(i \in S_t) = p_i > 0$

$$g_i^{t+1}\begin{cases} g_i^t + \mathcal{C}\left(\nabla f_i\left(x^{t+1}\right) - g_i^t\right) & \text{if } i \in S_t,\\ g_i^t & \text{if } i \notin S_t.\end{cases}$$

- full participation is impractical in federated learning
- arbitrary proper samplings
- first analysis with error feedback

5. EF21 with Heavy Ball Momentum

EF21-HB method: $x^{t+1} = x^t - \gamma v^t,$

$$g_i^{t+1} = g_i^t + \mathcal{C}\left(\nabla f_i(x^{t+1}) - g_i^t\right),$$
$$v^{t+1} = \eta v^t + \gamma g^{t+1}, \quad g^{t+1} = \frac{1}{n}\sum_{i=1}^{n} g_i^{t+1}.$$

- works well in practice, especially in training NNs for CV

6. EF21 with Proximal Step

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

where each $f_i(\cdot)$ is L_i-smooth, $r(\cdot)$ is convex, and $\Phi^{\inf} = \inf_{x\in\mathbb{R}^d}\Phi(x) > -\infty$.

EF21-Prox method:

$$x^{t+1} = \mathrm{prox}_{\gamma r}\left(x^t - \gamma g^t\right), \quad g^t = \frac{1}{n}\sum_{i=1}^{n} g_i^t,$$
$$g_i^{t+1} = g_i^t + \mathcal{C}\left(\nabla f_i(x^{t+1}) - g_i^t\right).$$

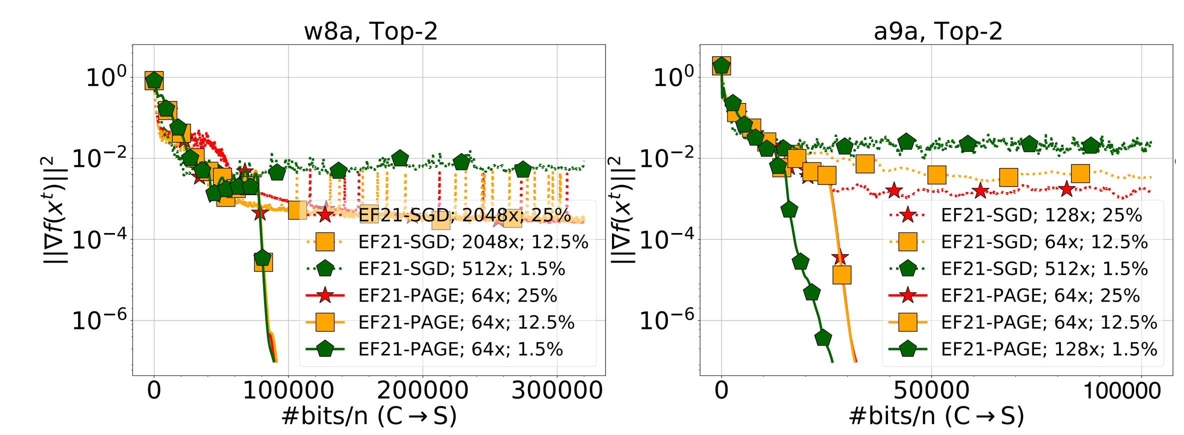
- constrained optimization
- better generalization and sparsity of the solution
- first analysis with error feedback

Experiments

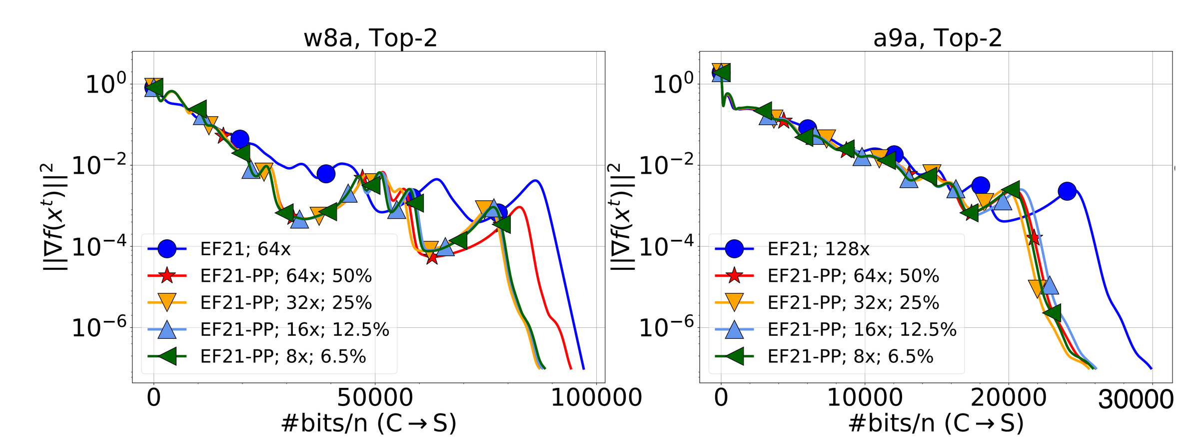
Logistic regression problem with a non-convex regularizer

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp\left(-y_i a_i^\top x\right)\right) + \lambda\sum_{j=1}^{d}\frac{x_j^2}{1+x_j^2}.$$

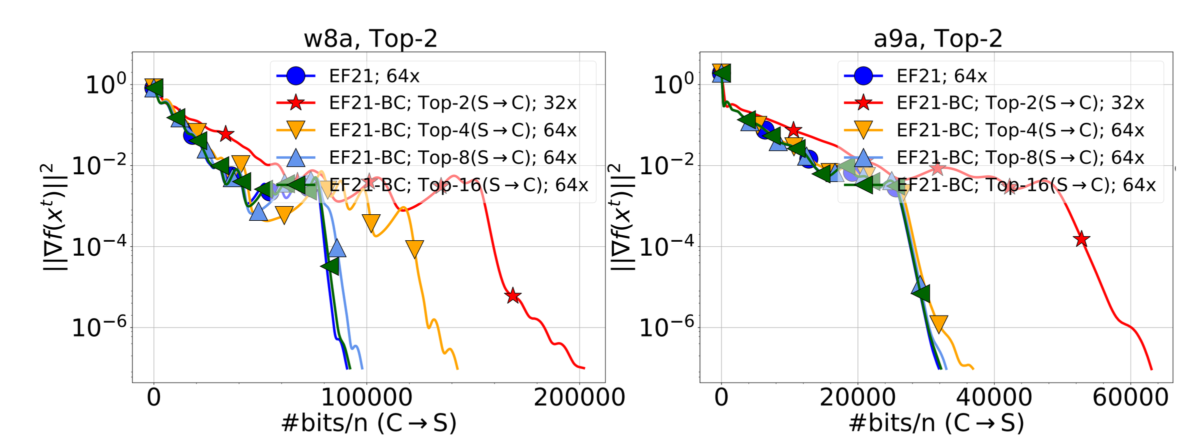
EF21-SGD vs EF21-PAGE



EF21-PP vs EF21



EF21-BC vs EF21



By $1\times, 2\times, 4\times$ (and so on) it is indicated that the stepsize was set to a multiple of the largest stepsize predicted by our theory. $k = 1$ means that Top-1 compressor was used in the experiment. Stepsizes were fine-tuned in all experiments.

References

[1] A. Beznosikov, S. Horváth, P. Richtárik, M. Safaryan. On biased compression for distributed learning. arXiv:2002.12410, 2020.

[2] P. Richtárik, I. Sokolov, I. Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. Oral NeurIPS'21, arXiv:2106.05203, 2021

[3] A. Koloskova, T. Lin, S. Stich, M. Jaggi. Decentralized deep learning with arbitrary communication compression. ICLR'20.

[4] H. Tang, X. Lian, C. Yu, T. Zhang, J. Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. ICML'20, arXiv:1905.05957, 2019.

[5] C. Xie, S. Zheng,O. Koyejo, I. Gupta, M. Li, H. Lin. CSER: Communication-efficient SGD with error reset. NeurIPS'20.