Safe-EF: Error Feedback for Nonsmooth Constrained Optimization

Rustem Islamov[1] Yarden As[2,3] Ilyas Fatkhullin[2,3]

[1]University of Basel [2]ETH Zürich [3]ETH AI Center

Problem Formulation

We want to solve a distributed (stochastic) problem:

$$\min_{x\in\mathcal{X}}\left\{f(x):=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right\}$$

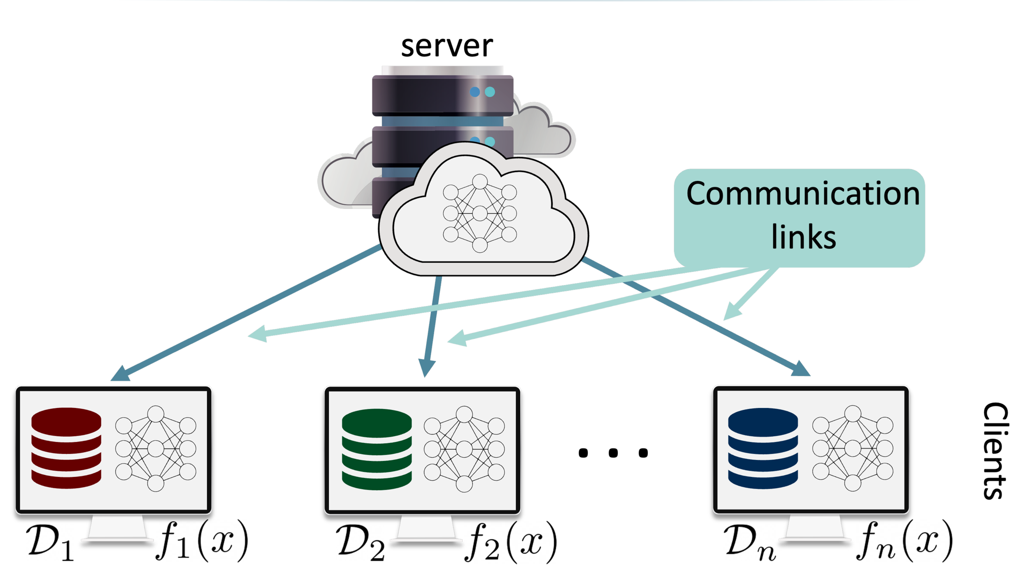
convex non-smooth — # clients/devices — model parameters — Empirical risk/loss — Local training data — Local loss function $f_i(x)=\mathbb{E}_{\xi\sim\mathcal{D}_i}[f_\xi(x)]$

where the constraint set is

$$\mathcal{X}:=\left\{x\in\mathbb{R}^d \mid g(x):=\frac{1}{n}\sum_{i=1}^{n}g_i(x)\le 0\right\}$$

- This problem has many applications in machine learning, data science and engineering.
- Safety constraints play a critical role in real-world applications such as federated reinforcement learning.

Federated Training



- Federated learning faces severe communication bottlenecks due to the high dimensionality of model updates

Contractive Compression

(C) We say that a (possibly randomized) mapping $\mathcal{C}\colon\mathbb{R}^d\to\mathbb{R}^d$ is a contractive compression operator if for some constant $0<\delta\le 1$ and all $x\in\mathbb{R}^d$ it holds

$$\mathbb{E}\left[\|\mathcal{C}(x)-x\|^2\right]\le(1-\delta)\|x\|^2.$$

We denote the class of δ-contractive compressors as $\mathbb{C}(\delta)$. A classic example of contractive compression is Top-K compressor.

$$(-2,1,1.5)^\top \overset{\text{Top-1}}{\to} (-2,0,0)^\top.$$

It preserves top K entries in magnitude. It is contractive with $\alpha = K/d$.

Failure of Algorithms from the Smooth Setting

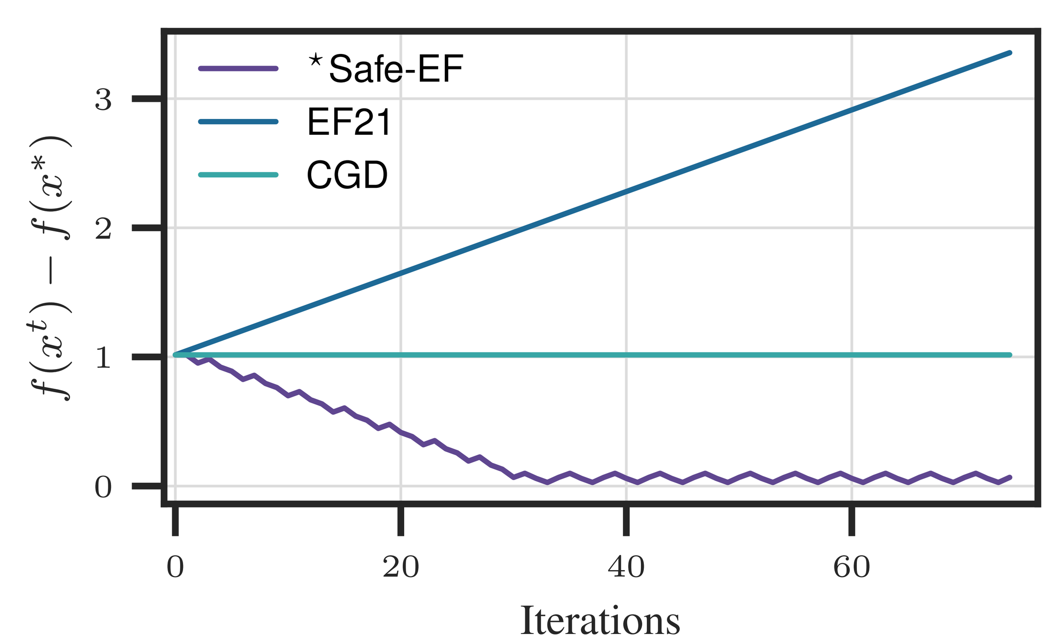


Figure 1:Non-convergence of CGD, divergence of EF21 [2] and convergence of Safe-EF for the problem $f(x)=\|x\|_1$. *Safe-EF coincides with EF14 [1] in this example.

$$\texttt{CGD}:\quad x^{t+1}=x^t-\frac{\gamma}{n}\sum_{i=1}^{n}\mathcal{C}(f_i'(x^t))$$

$$\texttt{EF21}:\quad x^{t+1}=x^t-\gamma\, v^t,\qquad v^t=\frac{1}{n}\sum_{i=1}^{n}v_i^t,$$

$$v_i^{t+1}=v_i^t+\mathcal{C}(f_i'(x^{t+1})-v_i^t)$$

Example. Consider $f(x)=\|x\|_1$ with $\mathcal{X}=\mathbb{R}^2$. For any $n\ge 1$, CGD and EF21 *do not converge*, i.e., for any $\gamma>0$, $t\ge 0$

$$\texttt{CGD}:\quad f(x^t)-\min_x f(x)=1+\gamma/2$$

$$\texttt{EF21}:\quad f(x^t)-\min_x f(x)=1+\gamma/2+t\,\gamma.$$

Takeaway: "Smooth" algorithms not suitable for non-smooth. It leads to extra challenges for federated learning.

Our Goals

Question 1: What are the limits of compressed gradient methods in the non-smooth regime?

Question 2: Can we design a provably convergent compressed gradient method with a Top-K compressor for "non-smooth"?

Communication Protocol and Main Results

Algorithm Class:

- *Centralized communication:* Workers are restricted to communicating directly with a central server only;
- *Synchronous Communication:* All workers begin each iteration simultaneously;
- *"Zero-respecting" Property:* Non-zero entries appear through local subgradient queries or synchronization with the server;
- *Output of Algorithm:* The output $\hat{x}_{A,t}$ of the algorithm A after t iterations can be expressed as any linear combination of all previous local models.

Lower Bound

Let f_i, g_i are convex and $\|f_i'(x,\xi^i)\|,\ \|g_i'(x,\xi^i)\|\le M$. Then there exists an instance of such problem that

$$\mathbb{E}[f(\hat{x}_{A,T})-f(x^*)]\ge\Omega\left(\frac{M\|x^0-x^*\|}{\sqrt{\delta T}}\right),\quad\text{and}$$

$$\mathbb{E}[g(\hat{x}_{A,T})]\ge\Omega\left(\frac{M\|x^0-x^*\|}{\sqrt{\delta T}}\right).$$

Key idea: Construct a "worst-case" function and account for compression in the distributed setting. We use for all $i\in[n]$

$$f_i(x):=C\cdot\max_{1\le j\le T}x_j+\frac{\mu}{2}\|x\|\cdot\max\left\{\|x\|;\frac{\|x^*\|}{2}\right\},$$

$$g_i(x):=f_i(x)-\min_{x\in\mathbb{R}^d}f_i(x),$$

where $C,\mu>0$ are some constants.

Convergence Theorem

Let $f_i(x,\xi^i)$, $g_i(x,\xi^i)$ are convex, $\|f_i'(x)\|,\ \|g_i'(x)\|\le M$, and $\mathbb{E}\left[\frac{(g_i(x,\xi^i)-g_i(x))^2}{\sigma_{\rm fv}^2/N_{\rm fv}}\right]\le\exp(1)$. Then the iterates of Safe-EF satisfy with probability $1-2\beta$ for any $\beta<1/2$

$$f(\overline{x}^T)-f(x^*)\le\mathcal{O}\left(\frac{(M\|x^0-x^*\|+\sigma_{\rm fv}/\sqrt{N_{\rm fv}})(1+\log 1/\beta)}{\sqrt{\delta_{\rm s}\delta T}}\right),$$

$$\mathbb{E}g(\overline{x}^T)\le\mathcal{O}\left(\frac{(M\|x^0-x^*\|+\sigma_{\rm fv}/\sqrt{N_{\rm fv}})(1+\log 1/\beta)}{\sqrt{\delta_{\rm s}\delta T}}\right),$$

where $\overline{x}^T:=\frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}}x^t$, $\mathcal{B}:=\{t\in[T-1]\mid g(x^t)\le c\}$.

Main implications:

- *General Claim:* Design a provably convergent compressed gradient method for distributed non-smooth optimization. Extend it to practically relevant settings with safety constraints and noise;
- *Single-node Training:* Recover the optimal rate known in the literature extending EF14 [3];
- *High-probability Analysis:* The dependency on the failure probability β is logarithmic → optimal;
- *Unidirectional Compression:* The rate of Safe-EF matches established lower bound → dependency on the compression level δ is optimal;
- *Bidirectional Compression:* Safe-EF — the first algorithm provably convergent when contractive compression **(C)** used in both server-to-worker and worker-to-server directions.

Algorithm 1: `Safe-EF`: Safe Error Feedback

Input: $w^0=x^0$, $\{\mathcal{C}_i\}_{i=0}^n$, $\gamma, c>0$, $e_i^0=0$
For $t=0,\dots,T-1$ **do**
  For $i=1,\dots,n$ **in parallel do**
    Send $g_i(x^t,\xi_i^t)$ to server
  end for
  Send $g(x^t,\xi^t)=\frac{1}{n}\sum_{i=1}^n g_i(x^t,\xi_i^t)$ to workers
  For $i=1,\dots,n$ **in parallel do**
    Compute $h_i^t=f_i'(x^t,\xi_i^t)$ **if** $g(x^t,\xi^t)\le c$ **else** $g_i'(x^t,\xi_i^t)$
    Send $v_i^t=\mathcal{C}_i(e_i^t+h_i^t)$ to server
    Compute $e_i^{t+1}=e_i^t+h_i^t-v_i^t$
  end for
  Compute $v^t=\frac{1}{n}\sum_{i=1}^n v_i^t$ and $w^{t+1}=w^t-\gamma v^t$
  Compute $x^{t+1}=x^t+\mathcal{C}_0(w^{t+1}-x^t)$
  Send $\mathcal{C}_0(w^{t+1}-x^t)$ to workers
end for

Experiments

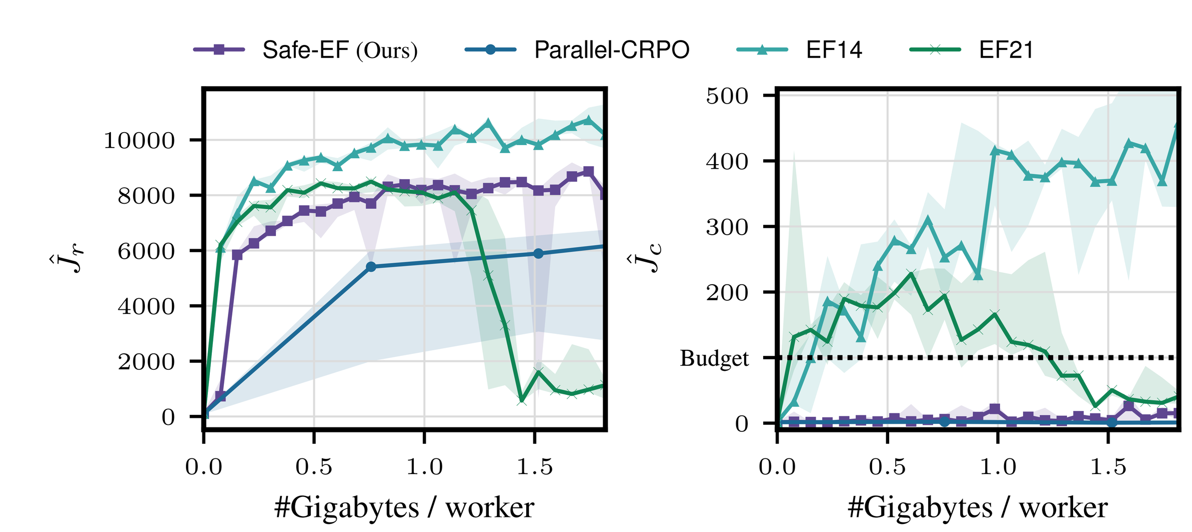


Figure 2:Objective and constraint during training Humanoid Robot Fleet. Budget denotes the level below which J_c must remain to satisfy the constraint.

Here, the objective f_i is

$$\mathop{\mathbb{E}}_{s,a\sim\bar{\pi}}\left[\min\left\{\frac{\pi_x(a|s)}{\bar{\pi}(a|s)}A_{p_i}^{\bar{\pi}}(s,a),\ \mathrm{clip}\left(\frac{\pi_x(a|s)}{\bar{\pi}(a|s)},1-\tilde{\epsilon},1+\tilde{\epsilon}\right)A_{p_i}^{\bar{\pi}}(s,a)\right\}\right],$$

where

- $A_{p_i}^{\bar{\pi}}$ denotes the advantage in terms of cumulative rewards
- Surrogate for the constraint $g_i(x)$ is given by replacing rewards with costs when computing the advantage
- $\mathrm{clip}(x,l,u):=\max\{l,\min\{x,u\}\}$

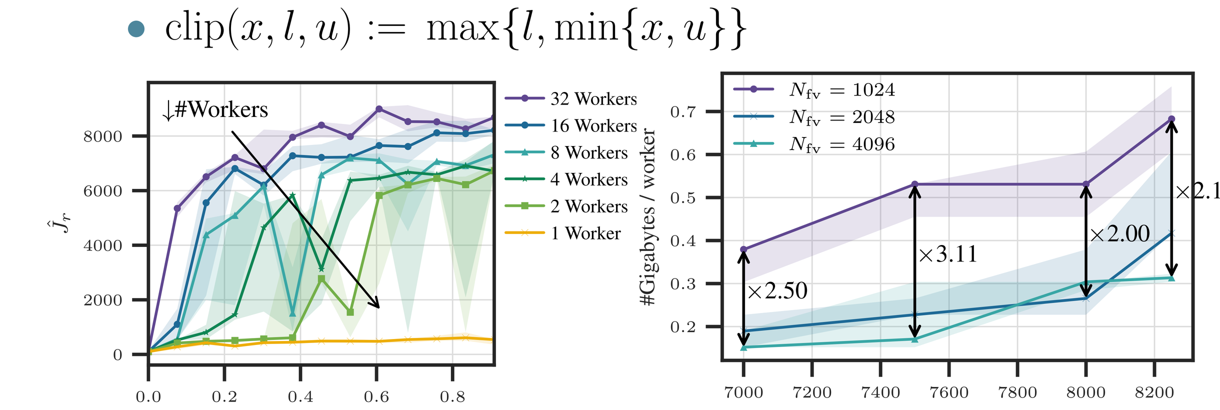


Figure 3:Left: Convergence for different number of workers. Right: Communication required to reach a desired performance for different batch samples $N_{\rm fv}$.

References

[1] Seide et al. "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs", Interspeech, 2014.
[2] Richtárik et al., "EF21: A new, simpler, theoretically better, and practically faster error feedback", NeurIPS 2021.
[3] Karimireddy et al. "Error feedback fixes SignSGD and other gradient compression schemes", ICML 2019.