Error Feedback under (L_0, L_1)-Smoothness: Normalization and Momentum

Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabin,
Eduard Gorbunov*, Peter Richtárik
KAUST, MBZUAI (*)

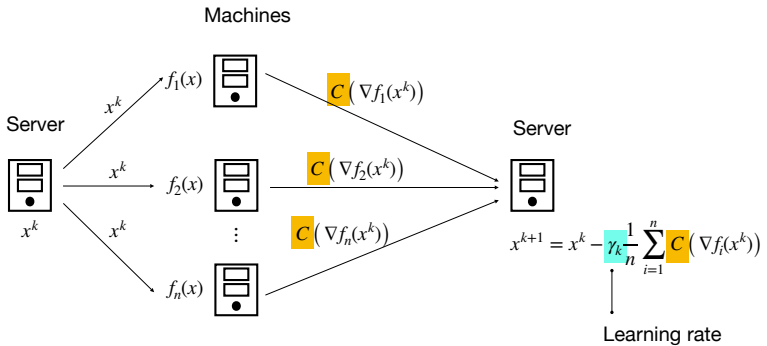
November 7, 2025

Distributed Learning Challenge

Distributed optimization

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

Distributed methods with gradient compression

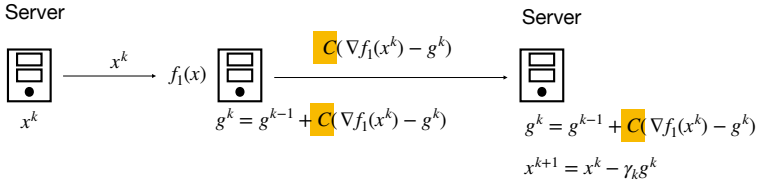


Distributed Compression Methods

Problem: Diverge for biased, contractive compressors (Beznosikov et. al, 2024).

$$\mathbb{E}\|g - \mathcal{C}(g)\|^2 \leq (1-\alpha)\|g\|^2 \quad \text{for} \quad \alpha \in (0,1].$$

Solution: Apply **error feedback**, e.g. **EF21** (Richtárik et. al, 2021).



EF21 Convergence under Standard L-smoothness

Assume

(1) $\frac{\|\nabla f(x)-\nabla f(y)\|}{\|x-y\|} \leq L$.

(2) $\gamma_k = \mathcal{O}(1/L)$.

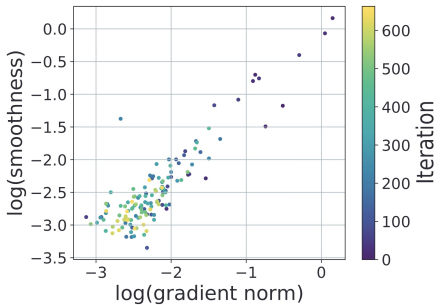
Then,

$$\min_{k=0,1,\ldots,K} \|\nabla f(x^k)\| \leq \mathcal{O}\left(\frac{1}{\sqrt{K+1}}\right)$$

Problem: L-smoothness limits applicability of EF21 for several ML problems.

Generalized (L_0, L_1)-smoothness

$$\frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} \leq L_0 + L_1 \sup_{u \in [x,y]} \|\nabla f(u)\|.$$



LSTM by Zhang et. al (2020).

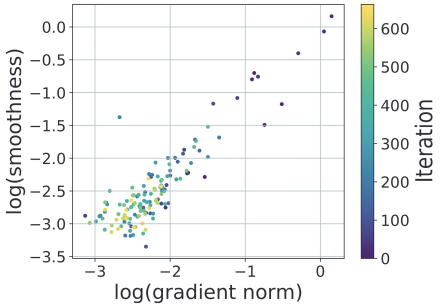
- Recovers $\frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} \leq L_0$ when $L_1 = 0$.
- Captures problems standard smoothness does not hold, e.g. training DNNs like LSTMs.

EF21 Convergence under Generalized (L_0, L_1)-smoothness

Assume

(1) $\frac{\|\nabla f(x)-\nabla f(y)\|}{\|x-y\|} \leq L_0 + L_1 \sup_{u\in[x,y]} \|\nabla f(u)\|.$

(2) $\gamma_k = \mathcal{O}\left(\frac{1}{\max_{k=0,1,\ldots,K} \|\nabla f(x^k)\|}\right).$

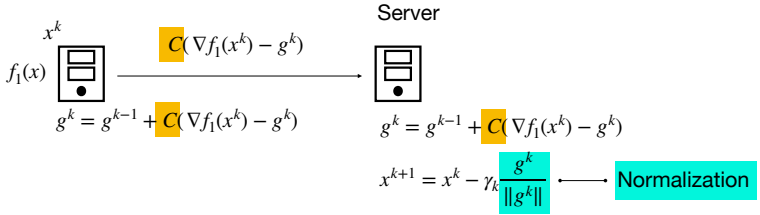


LSTM by Zhang et. al (2020).

Then,

$$\min_{k=0,1,\ldots,K} \|\nabla f(x^k)\| \leq \mathcal{O}\left(\frac{\max_{k=0,1,\ldots,K} \|\nabla f(x^k)\|}{\sqrt{K+1}}\right) \implies \text{Vacuous bounds}$$

‖EF21‖: Error Feedback with Normalization



	EF21	‖**EF21**‖ (NEW)
Normalization	No	Yes
Smoothness	Standard L	Generalized (L_0, L_1)
Rate in $\min_{k=0,1,\ldots,K} \|\nabla f(x^k)\|$	$\mathcal{O}(1/\sqrt{K+1})$	$\mathcal{O}(1/\sqrt{K+1})$
Learning rate	$1/L$	$\mathcal{O}(1/\sqrt{K+1})$

Conclusion

- Normalization stabilizes EF21 under (L_0, L_1)-smoothness.
- $\|\text{EF21}\|$ under (L_0, L_1)-smoothness with the same rate as EF21 under L-smoothness.
- Extension to stochastic optimization: $\|\text{EF21}\|$ with momentum and stochastic gradients.

Q & A at the poster session!