EControl: Fast Distributed Optimization with Compression and Error Control

Yuan Gao*[1,3] Rustem Islamov*[2] Sebastian Stich[3]

*Equal Contribution [1]Universität des Saarlandes [2]Universität Basel [3]CISPA Helmholtz Center for Information Security

Problem Formulation

We want to solve the finite-sum optimization problem

$$\min_{x\in\mathbb{R}^d}\left\{ f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}$$

workers/devices

model parameters

Empirical risk/loss

Local training data

Local loss function $f_i(x) = \mathbb{E}_{\xi\sim\mathcal{D}_i}[f_\xi(x)]$

- This problem has many applications in machine learning, data science and engineering.
- We focus on the regime when n **and** d **are very large**. This is typically the case in the big data settings (e.g., massively distributed and federated learning).

Assumptions

(A1) Let $f^\star := \operatorname{argmin}_{x\in\mathbb{R}^d} f(x) > -\infty$. Let f and each f_i be smooth, i.e. for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(x)\| \le L\|x-y\|,$$

$$\|\nabla f_i(x) - \nabla f_i(x)\| \le L_i\|x-y\|, \quad \widetilde{L}^2 = \frac{1}{n}\sum_{i=1}^{n} L_i^2.$$

(A2) Let f be μ-strongly quasi-convex for some $\mu \ge 0$, i.e. for all $\mathbf{x}\in\mathbb{R}^d$

$$f(\mathbf{x}^\star) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{x}^\star\|^2.$$

(A3) Let stochastic gradient oracles $\mathbf{g}^i(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}^d$ for each f_i be unbiased and have bounded variance, i.e. for all $\mathbf{x}\in\mathbb{R}^d$

$$\mathbb{E}\left[\mathbf{g}^i(\mathbf{x})\right] = \nabla f_i(\mathbf{x}), \quad \mathbb{E}\left[\|\mathbf{g}^i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2\right] \le \sigma^2.$$

Contractive Compression

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 $\mathbf{x}\in\mathbb{R}^d$ it holds

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

Motivation

There is no Error Compensation (EC) mechanism that is able to handle an error coming from stochastic gradients and contractive compression simultaneously in all standard regimes.

Existing Problems

Existing problems with Error Compensation include

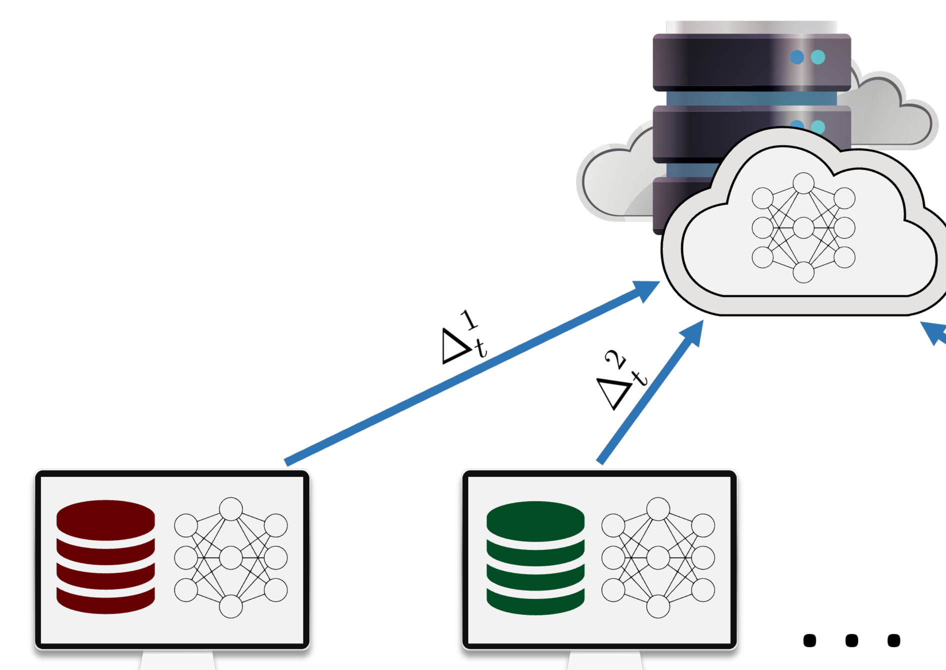
- **Additional Communication:** methods require two or more compressed messages instead of one per iteration;
- **Strong Assumptions:** the analysis is done under strong assumptions (e.g., bounded gradients or bounded heterogeneity);
- **Large Batches:** the analysis requires access to stochastic gradients with large batches;
- **Suboptimal Rates:** existing convergence rates do not match the lower bounds;
- **No Method in Convex Regime:** there is not an algorithm that provably works in convex regime.

Table 1:Summary of theoretical results on error compensated algorithms using only contractive compressors. nCVX = supports nonconvex functions; CVX = supports covnex functions; sCVX = supports strongly convex functions.

Method	Nonconvex(a)	Convex(b)	Strongly Convex(b)	Extra Assumptions
EC [1]	$\frac{\sigma^2}{n\varepsilon^2} + \frac{\sigma+\zeta/\sqrt{\delta}}{\sqrt{\delta}\varepsilon^{3/2}} + \frac{1}{\delta\varepsilon}$ (c)	$\frac{\sigma^2}{n\varepsilon} + \frac{\sigma+\zeta/\sqrt{\delta}}{\sqrt{\delta\varepsilon}} + \frac{1}{\delta}$ (c)	$\frac{\sigma^2}{n\varepsilon} + \frac{\sigma+\zeta/\sqrt{\delta}}{\sqrt{\delta\varepsilon}} + \frac{1}{\delta}$ (c)	Bounded Heterogeneity
Choco-SGD [2]	$\frac{\sigma^2}{n\varepsilon^2} + \frac{G}{\varepsilon^{3/2}} + \frac{1}{\delta\varepsilon}$ (d)	✗	$\frac{\sigma^2}{n\varepsilon} + \frac{G}{\delta\sqrt{\varepsilon}} + \frac{1}{\delta}$ (d)	Bounded gradients $\mathbb{E}\left[\|\mathbf{g}^i(\mathbf{x})\|^2\right] \le G^2$.
EF21-SGD [3]	$\frac{\sigma^2}{\delta^3\varepsilon^2} + \frac{1}{\delta\varepsilon}$	✗	$\frac{\sigma^2}{\delta^3\varepsilon} + \frac{1}{\delta}$	Large batches of order $\frac{\sigma^2}{\delta^2\varepsilon}$
EF21-SGD2M [4]	$\frac{\sigma^2}{n\varepsilon^2} + \frac{\sigma^{2/3}}{\delta^{2/3}\varepsilon^{4/3}} + \frac{1+\sigma}{\delta\varepsilon}$ (c)	✗	✗	✗
EControl **This work**	$\frac{\sigma^2}{n\varepsilon^2} + \frac{\sigma}{\delta^2\varepsilon^{3/2}} + \frac{1+\sigma}{\delta\varepsilon}$	$\frac{\sigma^2}{n\varepsilon} + \frac{\sigma}{\delta^2\sqrt{\varepsilon}} + \frac{1}{\delta\varepsilon}$	$\frac{\sigma^2}{n\varepsilon} + \frac{\sigma}{\delta^2\sqrt{\varepsilon}} + \frac{1}{\delta}$	✗

(a) The convergence in terms of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\text{out}})\|^2\right] \le \varepsilon$. (b) The convergence in terms of $\mathbb{E}\left[f(\mathbf{x}_{\text{out}}) - f^\star\right] \le \varepsilon$.
(c) The last term becomes $\frac{\sigma}{\delta\varepsilon}$ if the initial batch size is of order σ^2.

Algorithms



$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\mathbf{h}_t - \frac{\gamma}{n}\sum_{i=1}^{n}\Delta_t^i$$

$\Delta_t^1 = \mathcal{C}_\delta(\eta\mathbf{e}_t^1 + \mathbf{g}_t^1 - \mathbf{h}_t^1)$ $\Delta_t^2 = \mathcal{C}_\delta(\eta\mathbf{e}_t^2 + \mathbf{g}_t^2 - \mathbf{h}_t^2)$ $\Delta_t^n = \mathcal{C}_\delta(\eta\mathbf{e}_t^n + \mathbf{g}_t^n - \mathbf{h}_t^n)$

Algorithm 1: EC-Ideal

Input: $\mathbf{x}_0, \mathbf{e}_0^i = \mathbf{0}_d$, $\mathbf{h}_\star^i = \nabla f_i(\mathbf{x}^\star)$, γ, η, $\mathcal{C}_\delta$, $\mathbf{h}_\star = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_\star^i$
for $t = 0, 1, \dots T-1$ **do**
  client side:
  compute $\mathbf{g}_t^i = \mathbf{g}^i(\mathbf{x}_t)$ and $\Delta_t^i = \mathcal{C}_\delta(\mathbf{e}_t^i + \mathbf{g}_t^i - \mathbf{h}_\star^i)$
  update $\mathbf{e}_{t+1}^i = \mathbf{e}_t^i + \mathbf{g}_t^i - \mathbf{h}_\star^i - \Delta_t^i$ and $\mathbf{h}_{t+1}^i = \mathbf{h}_t^i + \Delta_t^i$
  send to server Δ_t^i
  server side:
  update $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma\mathbf{h}_\star - \frac{\gamma}{n}\sum_{i=1}^{n}\Delta_t^i$
end

Algorithm 2: EControl

Input: $\mathbf{x}_0, \mathbf{e}_0^i = \mathbf{0}_d$, $\mathbf{h}_0^i = \mathbf{g}_0^i$, γ, η, $\mathcal{C}_\delta$, $\mathbf{h}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_0^i$
for $t = 0, 1, \dots T-1$ **do**
  client side:
  compute $\mathbf{g}_t^i = \mathbf{g}^i(\mathbf{x}_t)$ and $\Delta_t^i = \mathcal{C}_\delta(\eta\mathbf{e}_t^i + \mathbf{g}_t^i - \mathbf{h}_t^i)$
  update $\mathbf{e}_{t+1}^i = \mathbf{e}_t^i + \mathbf{g}_t^i - \mathbf{h}_t^i - \Delta_t^i$ and $\mathbf{h}_{t+1}^i = \mathbf{h}_t^i + \Delta_t^i$
  send to server Δ_t^i
  server side:
  update $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma\mathbf{h}_t - \frac{\gamma}{n}\sum_{i=1}^{n}\Delta_t^i$
  and $\mathbf{h}_{t+1} = \mathbf{h}_t + \frac{1}{n}\sum_{i=1}^{n}\Delta_t^i$
end

Convergence Theory

EControl in Strongly Convex Regime

Assume (A1), (A2) with $\mu > 0$ and (A3) hold. Then there exist stepsizes $\eta = \mathcal{O}(\delta)$ and $\gamma = \mathcal{O}(\delta/\widetilde{L})$ such that $\mathbb{E}\left[f(\mathbf{x}_{\text{out}}) - f^\star\right] \le \varepsilon$ after

$$T = \mathcal{O}\left(\frac{\sigma^2}{\mu n\sigma^2} + \frac{\sqrt{L}\sigma}{\mu\delta^2\varepsilon^{1/2}} + \frac{\widetilde{L}}{\mu\delta}\right)$$

iterations of Algorithm 2.

EControl in Nonconvex Regime

Assume (A1) and (A3) hold. Then there exist stepsizes $\eta = \mathcal{O}(\delta)$ and $\gamma = \mathcal{O}(\delta/\widetilde{L})$ such that $\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\text{out}})\|^2\right] \le \varepsilon$ after

$$T = \mathcal{O}\left(\frac{LF_0\sigma^2}{\mu n\sigma^2} + \frac{LF_0\sigma}{\delta^2\varepsilon^{3/2}} + \frac{\widetilde{L}F_0}{\mu\delta}\right)$$

iterations of Algorithm 2 where $F_0 := f(\mathbf{x}_0) - f^\star$.

Experiments

First, we test the performance of EControl on toy problem where $f_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\|^2$ with $\mathbf{A}_i = \frac{i^2}{n}\mathbf{I}_d$ and $\mathbf{b}_i \sim \mathcal{N}(0, \frac{\zeta^2}{i^2}\mathbf{I}_d)$.

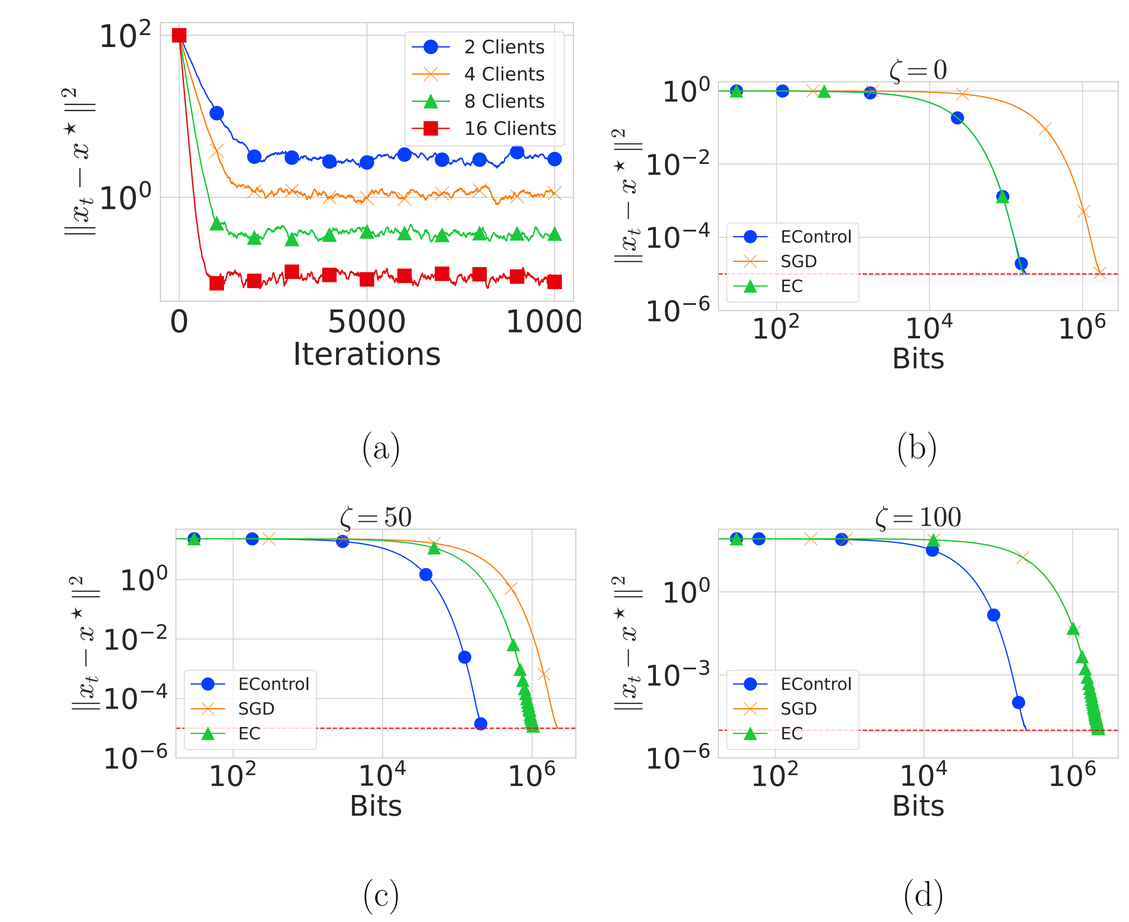


Figure 1:(a): behaviour of EControl changing the numberof workers. (b-d): comparison of EControl, SGD, and EC changing the heterogeneity of the problem.

Next, we compare the performance of EControl, EF21, and EF21-SGDM on training deep networks such as Resnet18 and VGG13.

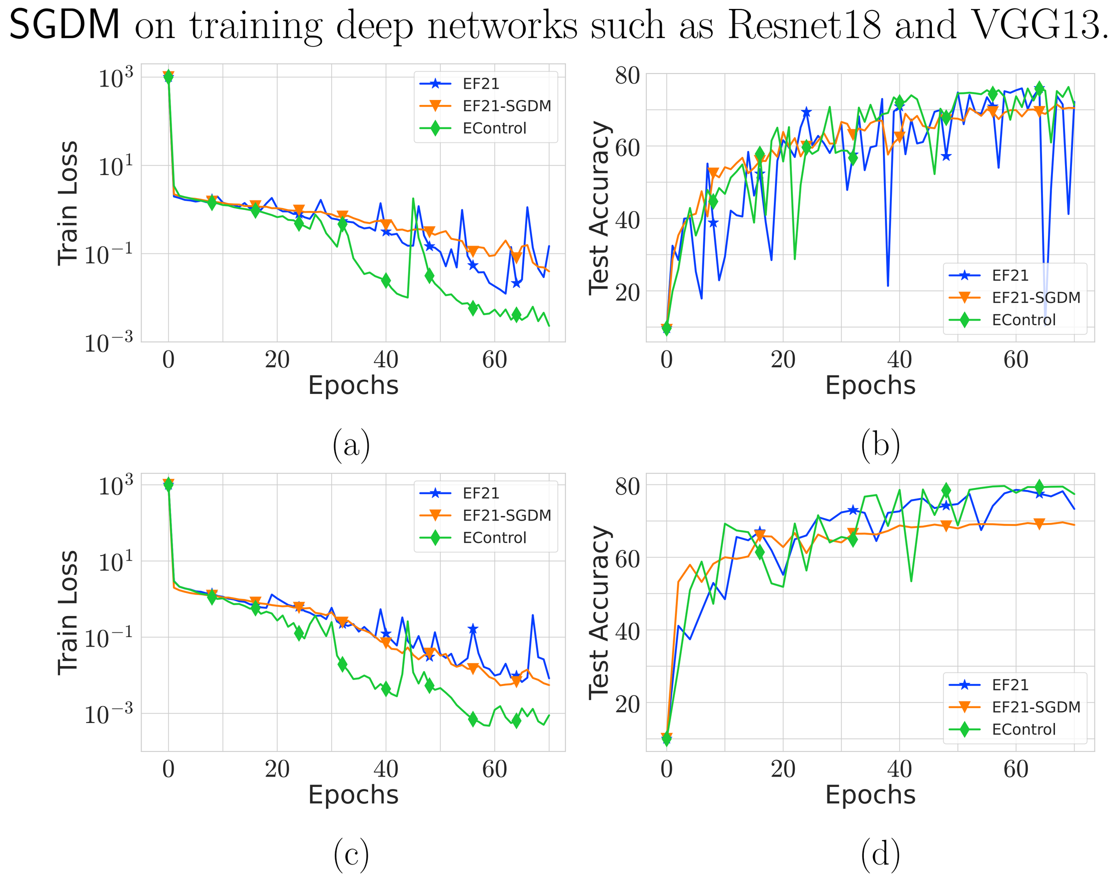


Figure 2:(a-b): comparison on Resnet 18; (c-d): comparison on VGG13.

References

[1] F. Seide, H. Fu, J. Droppo, G. Li, D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. Annual conference of the international speech communication association, 2014.

[2] A. Koloskova, T. Lin, S. U. Stich, M. Jaggi. Decentralized deep learning with arbitrary communication compression. International Conference on Learning Representations, 2020.

[3] I. Fatkhullin, I. Sokolov, E. Gorbunov, Z. Li, P. Richtarik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. arXiv preprint arXiv: 2110.03294, 2021.

[4] I. Fatkhullin, A. Tyurin, P. Richtarik. Momentum provably improves error feedback! Advances in Neural Information Processing Systems, 2023.