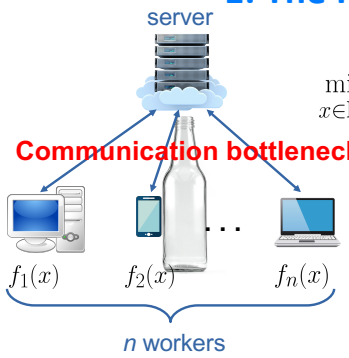
Linearly Converging Error Compensated SGD

Eduard Gorbunov[1,2,3,4], Dmitry Kovalev[4], Dmitry Makarenko[1], Peter Richtárik[4]

[1]MIPT (Russia), [2]Yandex (Russia), [3]Sirius (Russia), [4]KAUST (Saudi Arabia)

1. The Problem



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

Options for $f_i(x)$:

- $f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right]$
- $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$

Problem: Distributed optimization / training, where *n* workers (devices / clients) jointly solve a problem by communicating with a central server.

Assumptions: Smoothness and convexity of local loss functions and quasi-strong convexity of their average:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\| \qquad f_i(x) - f_i(y) \ge \langle \nabla f_i(y), x-y\rangle$$

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

The key bottleneck comes from the high cost of communication.
However, one can handle this using compression of messages.

- There is good theory for methods with unbiased compressors: there exist linearly converging SGD-like methods [1, 2].
- However, Parallel-SGD with biased compressors diverges exponentially fast on some problems [3].
- One can fix this using error compensation (EC-SGD) [4]
- but **EC-SGD fails to converge linearly** even when workers compute full gradients.

2. Our Contributions

- **The first linearly converging error compensated SGD method**
- **General theoretical framework** covering error compensated methods, methods with delayed updates and non-distributed methods
- In **one theorem**, we recover **the sharpest rates** for all known special cases
- **16 new methods** fitting our framework: 8 error-compensated methods, 7 methods with delayed updates and DIANA with bi-directional compression

3. Compression Operators

Unbiased compressors

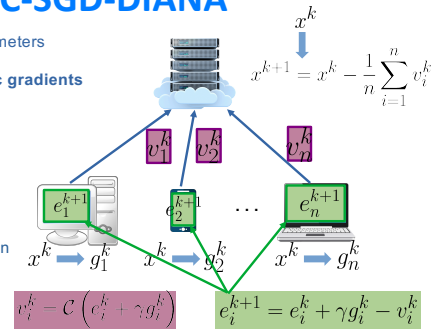
$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$
$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \le \omega\|x\|^2$$

Biased compressors

$$x \to \mathcal{C}(x)$$
$$\mathbb{E}\|\mathcal{C}(x) - x\|^2 \le (1-\delta)\|x\|^2$$

4. EC-SGD-DIANA

1. Server broadcasts new parameters
2. Workers compute **stochastic gradients** in parallel
3. Compression
4. Devices send compressed vectors and update unsent information
5. Server gathers the information and updates the parameters
6. Repeat steps 1 – 5



The main innovation is reflected in the definition of g_i^k :

$$g_i^k = \hat{g}_i^k - h_i^k + h^k \qquad \mathbb{E}\left[\hat{g}_i^k \mid x^k\right] = \nabla f_i(x^k)$$

"Learned shift vectors" are the key to get linear convergence:

$$h_i^{k+1} = h_i^k + \alpha\mathcal{Q}\left(\hat{g}_i^k - h_i^k\right) \qquad h^k = \frac{1}{n}\sum_{i=1}^n h_i^k$$

5. General Framework

The assumption below covers a very broad class of methods:

$$g^k = \frac{1}{n}\sum_{i=1}^n g_i^k, \quad \mathbb{E}\left[g^k \mid x^k\right] = \nabla f\left(x^k\right) \quad \bar{g}_i^k = \mathbb{E}\left[g_i^k \mid x^k\right]$$

$$\frac{1}{n}\sum_{i=1}^n \left\|\bar{g}_i^k\right\|^2 \le 2A\left(f\left(x^k\right) - f\left(x^*\right)\right) + B_1\sigma_{1,k}^2 + B_2\sigma_{2,k}^2 + D_1$$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|g_i^k - \bar{g}_i^k\right\|^2 \mid x^k\right] \le 2\widetilde{A}\left(f\left(x^k\right) - f\left(x^*\right)\right) + \widetilde{B}_1\sigma_{1,k}^2 + \widetilde{B}_2\sigma_{2,k}^2 + \widetilde{D}_1$$

$$\mathbb{E}\left[\left\|g^k\right\|^2 \mid x^k\right] \le 2A'\left(f\left(x^k\right) - f\left(x^*\right)\right) + B_1'\sigma_{1,k}^2 + B_2'\sigma_{2,k}^2 + D_1'$$

$$\mathbb{E}\left[\sigma_{1,k+1}^2 \mid \sigma_{1,k}^2, \sigma_{2,k}^2\right] \le (1-\rho_1)\sigma_{1,k}^2 + 2C_1\left(f\left(x^k\right) - f\left(x^*\right)\right) + G\rho_1\sigma_{2,k}^2 + D_2$$

$$\mathbb{E}\left[\sigma_{2,k+1}^2 \mid \sigma_{2,k}^2\right] \le (1-\rho_2)\sigma_{2,k}^2 + 2C_2\left(f\left(x^k\right) - f\left(x^*\right)\right)$$

- Reflects smoothness properties of the problem and noises introduced by compressions and stochastic gradients
- Describes the process of variance reduction of the variance coming from compressions
- Describes the process of variance reduction of the variance coming from stochastic gradients

Table 1: Complexity of Error-Compensated SGD methods established in this paper. Symbols: ε = error tolerance; δ = contraction factor of compressor $\mathcal{C}$; ω = variance parameter of compressor $\mathcal{Q}$; $\kappa = L/\mu$; $\mathcal{L}$ = expected smoothness constant; σ_*^2 = variance of the stochastic gradients in the solution; ζ_*^2 = average of $\|\nabla f_i(x^*)\|^2$; σ^2 = average of the uniform bounds for the variances of stochastic gradients of workers. EC-GDstar, EC-LSVRGstar and

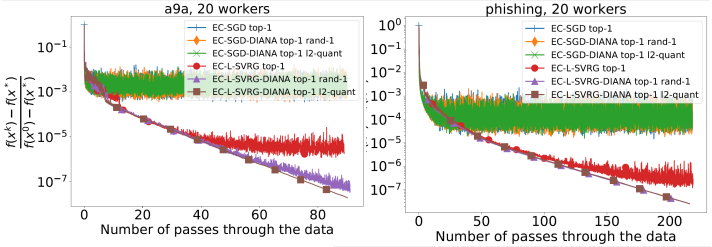
Probl.	Method	Citation	Rate (constants ignored)
Σ	EC-SGDsr	new	$\widetilde{\mathcal{O}}\left(\frac{\mathcal{L}}{\mu} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L(\sigma_*^2+\zeta_*^2/\delta)}}{\mu\sqrt{\delta\varepsilon}}\right)$
E	EC-SGD	[4]	$\widetilde{\mathcal{O}}\left(\frac{\kappa}{\delta} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L(\sigma_*^2+\zeta_*^2/\delta)}}{\delta\mu\sqrt{\varepsilon}}\right)$
Σ	EC-GDstar	new	$\mathcal{O}\left(\frac{\kappa}{\delta}\log\frac{1}{\varepsilon}\right)$
E	EC-SGD-DIANA	new	Opt. I: $\widetilde{\mathcal{O}}\left(\omega + \frac{\kappa}{\delta} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma_*^2}}{\delta\mu\sqrt{\varepsilon}}\right)$ Opt. II: $\widetilde{\mathcal{O}}\left(\frac{1+\omega}{\delta} + \frac{\kappa}{\delta} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma_*^2}}{\mu\sqrt{\delta\varepsilon}}\right)$
Σ	EC-SGDsr-DIANA	new	Opt. I: $\widetilde{\mathcal{O}}\left(\omega + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma_*^2}}{\delta\mu\sqrt{\varepsilon}}\right)$ Opt. II: $\widetilde{\mathcal{O}}\left(\frac{1+\omega}{\delta} + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma_*^2}}{\mu\sqrt{\delta\varepsilon}}\right)$
E	EC-GD-DIANA[†]	new	$\mathcal{O}\left(\left(\omega + \frac{\kappa}{\delta}\right)\log\frac{1}{\varepsilon}\right)$
Σ	EC-LSVRG	new	$\widetilde{\mathcal{O}}\left(m + \frac{\kappa}{\delta} + \frac{\sqrt{L\zeta_*^2}}{\delta\mu\sqrt{\varepsilon}}\right)$
Σ	EC-LSVRGstar	new	$\mathcal{O}\left(\left(m + \frac{\kappa}{\delta}\right)\log\frac{1}{\varepsilon}\right)$
Σ	EC-LSVRG-DIANA	new	$\mathcal{O}\left(\left(\omega + m + \frac{\kappa}{\delta}\right)\log\frac{1}{\varepsilon}\right)$

See more examples of methods (with delayed updates and without error feedback) fitting our framework together with the rates in weakly convex case in our paper.

6. Numerical Experiments

We conducted several numerical experiments on logistic regression problem with L_2-regularization:

$$\min_{x\in\mathbb{R}^d}\left\{f(x) = \frac{1}{N}\sum_{i=1}^N \log\left(1 + \exp\left(-y_i\cdot(Ax)_i\right)\right) + \frac{\mu}{2}\|x\|^2\right\}$$



References

[1] Mishchenko, Konstantin, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. **"Distributed learning with compressed gradient differences."** arXiv preprint arXiv:1901.09269 (2019).

[2] Horváth, Samuel, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. **"Stochastic distributed learning with gradient quantization and variance reduction."** arXiv preprint arXiv:1904.05115 (2019).

[3] Beznosikov, Aleksandr, Samuel Horváth, Peter Richtárik, and Mher Safaryan. **"On biased compression for distributed learning."** arXiv preprint arXiv:2002.12410 (2020).

[4] Stich, Sebastian U. and Sai Praneeth Karimireddy. **"The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication."** arXiv preprint arXiv:1909.05350 (2019).