On Biased Compression for Distributed Learning

Aleksandr Beznosikov[1,2] Samuel Horváth[3] Peter Richtárik[3] Mher Safaryan[3]

[1]MIPT [2]Sirius University [3]KAUST

The Problem

workers/devices

L-smooth μ-strongly convex

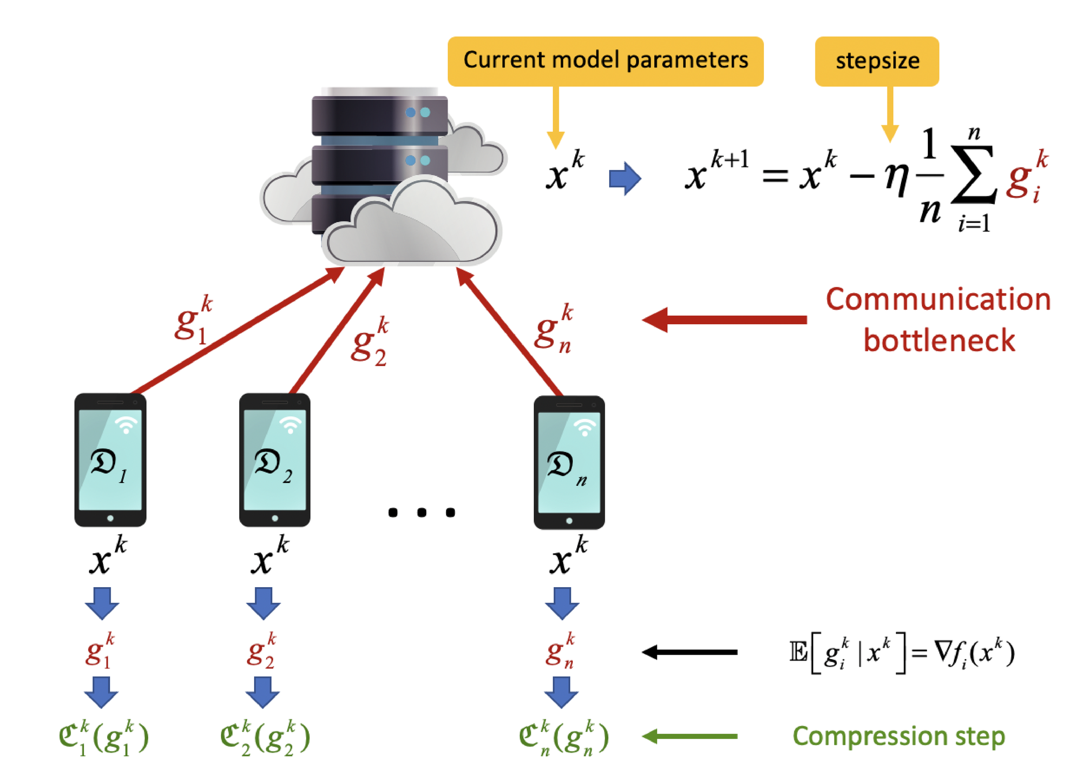
$$\min_{x\in\mathbb{R}^d} f(x) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

model parameters

Local training data

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

Communication Bottleneck in Distributed Systems



Biased vs. Unbiased Compression

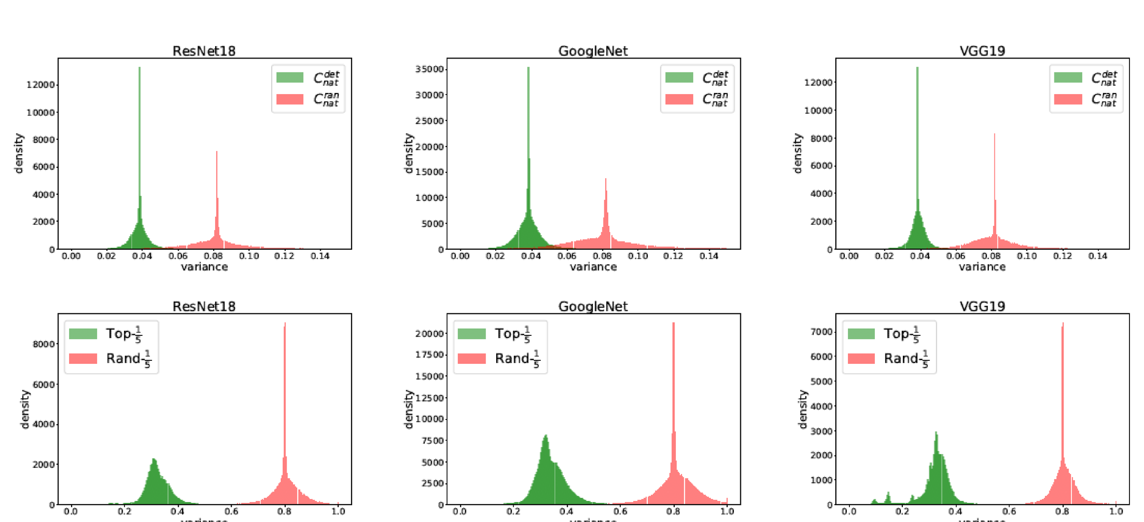


Figure 1: Comparison of empirical variance $\|\mathcal{C}(x) - x\|^2/\|x\|^2$ during training procedure of ResNet18, GoogleNet, and VGG19 on CIFAR10 dataset for two pairs of methods–deterministic biased against unbiased $\mathcal{C}_{\text{nat}}$, and Top-k against Rand-k, where $k = d/5$.

Classes of Compression Operators

Definition 1 *[1].* $\mathcal{C} \in \mathbb{U}(\zeta)$ for some $\zeta \geq 1$ if $\mathbb{E}\,\mathcal{C}(x) = x$, $\|\mathcal{C}(x)\|_2^2 \leq \zeta\|x\|_2^2$, $\forall x \in \mathbb{R}^d$.

Definition 2. $\mathcal{C} \in \mathbb{B}^1(\alpha, \beta)$ for some $\alpha, \beta > 0$ if $\alpha\|x\|_2^2 \leq \|\mathcal{C}(x)\|_2^2 \leq \beta\langle\mathcal{C}(x), x\rangle$, $\forall x \in \mathbb{R}^d$.

Definition 3. $\mathcal{C} \in \mathbb{B}^2(\gamma, \beta)$ for some $\gamma, \beta > 0$ if $\max\left\{\gamma\|x\|_2^2, \frac{1}{\beta}\|\mathcal{C}(x)\|_2^2\right\} \leq \langle\mathcal{C}(x), x\rangle$, $\forall x \in \mathbb{R}^d$.

Definition 4 *[2].* $\mathcal{C} \in \mathbb{B}^3(\delta)$ for some $\delta > 1$ if $\|\mathcal{C}(x) - x\|_2^2 \leq (1 - 1/\delta)\|x\|_2^2$, $\forall x \in \mathbb{R}^d$.

- Rand-$k \in \mathbb{U}(\frac{d}{k})$, Top-$k \in \mathbb{B}^1(\frac{k}{d}, 1), \mathbb{B}^2(\frac{k}{d}, 1), \mathbb{B}^3(\frac{d}{k})$.

Lyapunov function $\mathbb{E}\left[f(x^k)\right] - f(x^*)$

$$x^{k+1} = x^k - \eta\mathfrak{C}^k(\nabla f(x^k)) \qquad (CGD)$$

Compressor	$\mathfrak{C} \in \mathbb{U}(\zeta)$	$\mathfrak{C} \in \mathbb{B}^1(\alpha,\beta)$	$\mathfrak{C} \in \mathbb{B}^2(\gamma,\beta)$	$\mathfrak{C} \in \mathbb{B}^3(\delta)$
Complexity	$O\left(\zeta\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	$O\left(\frac{\beta^2}{\alpha}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	$O\left(\frac{\beta}{\gamma}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	$O\left(\delta\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$

scaling parameter $\lambda > 0$	$\mathbb{B}^1$	$\mathbb{B}^2$	$\mathbb{B}^3$
$\mathfrak{C} \in \mathbb{B}^1(\alpha,\beta)$, $\beta^2 \geq \alpha \geq 0$	$\lambda\mathfrak{C} \in \mathbb{B}^1(\lambda^2\alpha, \lambda\beta)$	$\mathfrak{C} \in \mathbb{B}^2(\alpha, \beta^2)$	$\frac{1}{\beta}\mathfrak{C} \in \mathbb{B}^3\left(\frac{\beta^2}{\alpha}\right)$
$\mathfrak{C} \in \mathbb{B}^2(\gamma,\beta)$, $\beta \geq \gamma \geq 0$	$\mathfrak{C} \in \mathbb{B}^1(\gamma^2, \beta)$	$\lambda\mathfrak{C} \in \mathbb{B}^2(\lambda\gamma, \lambda\beta)$	$\frac{1}{\beta}\mathfrak{C} \in \mathbb{B}^3\left(\frac{\beta}{\gamma}\right)$
$\mathfrak{C} \in \mathbb{B}^3(\delta)$, $\delta \geq 1$	$\mathfrak{C} \in \mathbb{B}^1\left(\frac{1}{4\delta^2}, 2\right)$	$\mathfrak{C} \in \mathbb{B}^2\left(\frac{1}{2\delta}, 2\right)$	-
$\mathfrak{C} \in \mathbb{U}(\zeta)$, $\zeta \geq 1$	$\lambda\mathfrak{C} \in \mathbb{B}^1(\lambda^2, \lambda\zeta)$	$\lambda\mathfrak{C} \in \mathbb{B}^2(\lambda, \lambda\zeta)$	$\frac{1}{\zeta}\mathfrak{C} \in \mathbb{B}^3\left(\frac{1}{\zeta}\right)$

$$\bar{x}^K = \frac{1}{W^K}\sum_{k=0}^{K} w^k x^k, \quad W^K = \sum_{k=0}^{K} w^k$$

$$w^k = \left(1 - \frac{\mu\eta}{2}\right)^{-(k+1)}, \quad \eta \leq \frac{1}{14(2\delta + B)L}$$

assumption on stochasticity $(B, C \geq 0)$

$$\mathbb{E}\left[\|g_i^k - \nabla f_i(x^k)\|^2\right] \leq B\|\nabla f_i(x^k)\|^2 + C$$

stochastic gradient

$$D = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|^2$$

$$\mathbb{E}\left[f(\bar{x}^K)\right] - f(x^*) = O\left(\delta L\exp\left[-\frac{\mu K}{\delta L}\right] + \frac{C + \delta D}{K\mu}\right)$$

total # iterations

the optimal weights

compression level $\mathfrak{C} \in \mathbb{B}^3(\delta)$

Exponential Divergence with Biased Compressor

Consider $n = d = 3$ and $x^0 = (t, t, t)$. Define

$$f_1(x) = \langle a, x\rangle^2 + \frac{1}{4}\|x\|_2^2, \quad a = (-3, 2, 2)$$

$$f_2(x) = \langle b, x\rangle^2 + \frac{1}{4}\|x\|_2^2, \quad b = (2, -3, 2)$$

$$f_3(x) = \langle c, x\rangle^2 + \frac{1}{4}\|x\|_2^2, \quad c = (2, 2, -3).$$

After k iterations with Top-1 compression, we get

$$x^k = (1 + {}^{11\eta}/_{6})^k x^0 \to \infty.$$

A Fix: Error Feedback *[2,3]*

$$x^{k+1} = x^k - \eta\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i^k$$

$$\tilde{g}_i^k = \mathfrak{C}_i^k(e_i^k + \eta g_i^k)$$

$$e_i^{k+1} = e_i^k + \eta g_i^k - \tilde{g}_i^k$$

Experiments

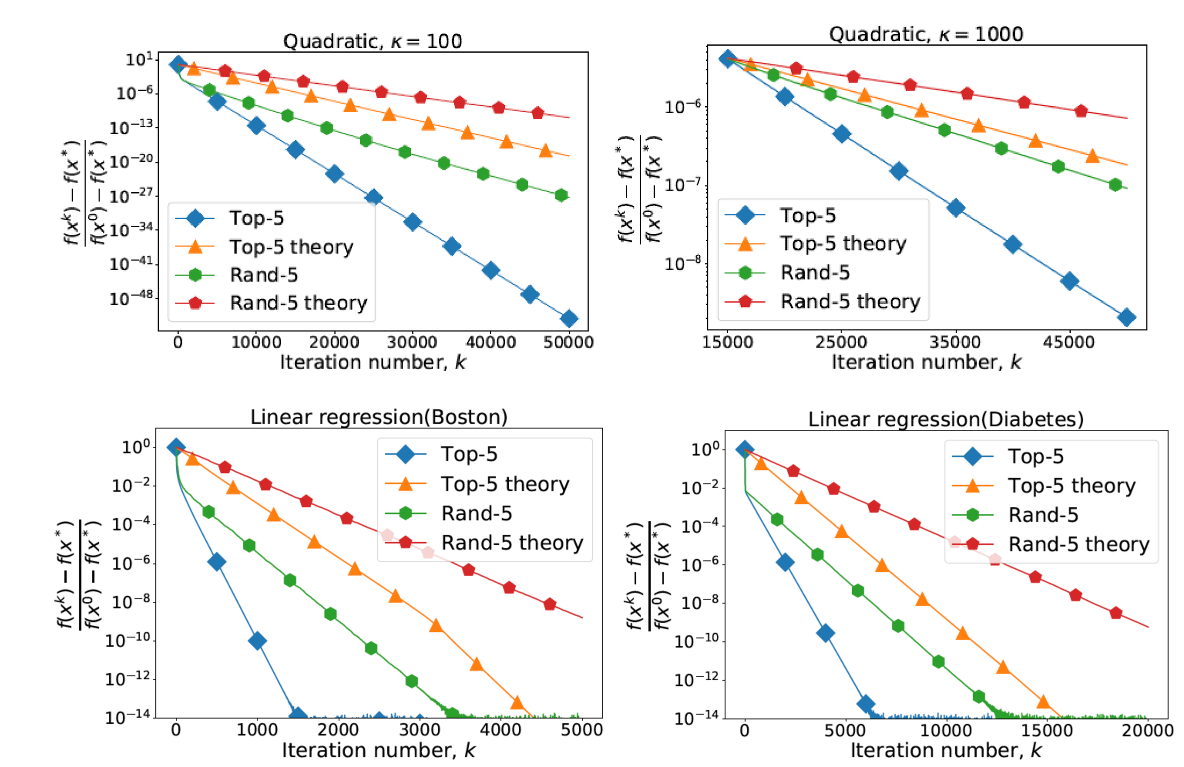


Figure 2:CGD on Quadratic problems (1st row) and Linear Regression (2nd row) with Top-5 and Rand-5 compression.

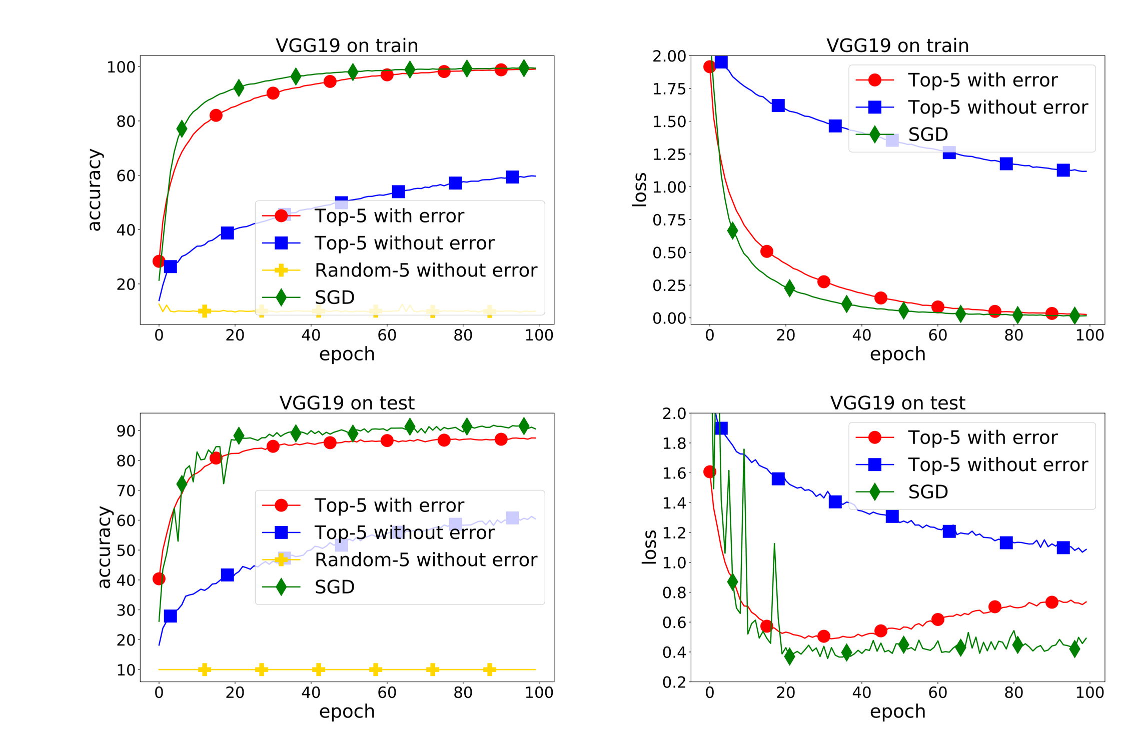


Figure 3:Training/Test loss and accuracy for VGG19 on CIFAR10 distributed among 4 nodes for 4 compression operators.

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *NeurIPS, 2017.*

[2] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. *ICML, 2019.*

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