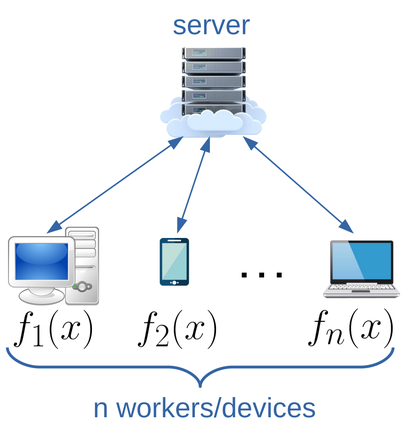
MARINA: Faster Non-Convex Distributed Learning with Compression

Eduard Gorbunov[1,2,3], Konstantin Burlachenko[3], Zhize Li[3], Peter Richtárik[3]
[1]MIPT (Russia), [2]Yandex (Russia), [3]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: Non-convex distributed optimization / training, where *n* workers (devices/clients) jointly solve a problem by communicating with a central server

Assumptions: smoothness of local loss functions and lower-boundedness

$$\forall x,y\in\mathbb{R}^d \quad \|\nabla f_i(x) - \nabla f_i(y)\| \le L_i\|x-y\|$$

$$\exists f_*\in\mathbb{R}:\ \forall x\in\mathbb{R}^d \quad f(x)\ge f_*$$

Distributed methods often suffer from **communication bottleneck**

One can handle this issue via **communication compression**

2. Unbiased Compression

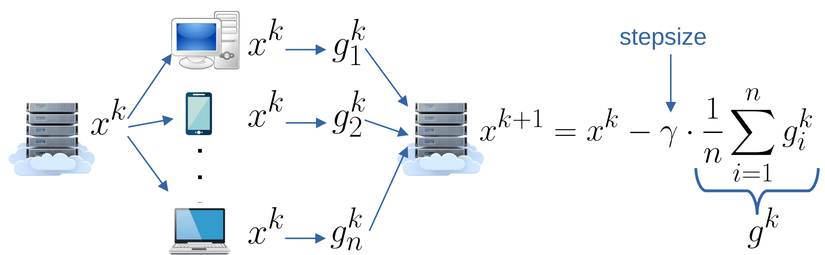
$$x\to\mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x \qquad \mathbb{E}\|\mathcal{Q}(x)-x\|^2 \le \omega\|x\|^2$$

Example: *RandK* compression operator picks *K* components uniformly at random

$$d=5,\ K=2 \qquad x=\begin{pmatrix}1\\-15\\0.2\\-7\\10\end{pmatrix} \longrightarrow \mathcal{Q}(x) = \frac{5}{2}\cdot\begin{pmatrix}1\\0\\0\\-7\\0\end{pmatrix} \qquad \omega = \frac{d}{K}-1$$

References: **[1]** Alistarh, Dan, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. "QSGD: Communication-efficient SGD via gradient quantization and encoding." In Advances in Neural Information Processing Systems, pp. 1709-1720, (2017).
[2] Mishchenko, Konstantin, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. "Distributed learning with compressed gradient differences." arXiv:1901.09269 (2019).
[3] Horváth, Samuel, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. "Stochastic distributed learning with gradient quantization and variance reduction." arXiv:1904.05115 (2019).

3. QGD and DIANA



QGD [1]: $g_i^k = \mathcal{Q}\left(\nabla f_i(x^k)\right)$

DIANA [2]: $g_i^k = h_i^k + \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$

learnable local shifts: $h_i^{k+1} = h_i^k + \alpha\mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$

# of communication rounds to find $\hat{x}$ such that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right]\le\varepsilon^2$

(ε-stationary point)

QGD: $\mathcal{O}\left(\frac{1}{\varepsilon^2} + \frac{1+\omega}{\varepsilon^4 n}\right)$

DIANA: $\mathcal{O}\left(\frac{1+(1+\omega)\sqrt{\omega/n}}{\varepsilon^2}\right)$ ← **previous SOTA complexity**

dependencies on numerical factors, smoothness constants and initial suboptimality are omitted

4. New Method: MARINA

typically small

$$g_i^k = \begin{cases}\nabla f_i(x^k) & \text{w.p. } p\\ g^{k-1} + \mathcal{Q}\left(\nabla f_i(x^k) - \nabla f_i(x^{k-1})\right) & \text{w.p. } 1-p\end{cases}$$

# of communication rounds to find an ε-stationary point

$$\mathcal{O}\left(\frac{1+\omega/\sqrt{n}}{\varepsilon^2}\right)$$ ← **current SOTA complexity**

$$p = \frac{1}{\omega+1} = \Theta\left(\frac{\zeta_{\mathcal{Q}}}{d}\right)$$

assumption (holds for RandK, l2-quantization)

$$\zeta_{\mathcal{Q}} = \sup_{x\in\mathbb{R}^d}\mathbb{E}\left[\|\mathcal{Q}(x)\|_0\right]$$

expected density

5. New Method: VR-MARINA

Assumption: $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$

typically small

$$g_i^k = \begin{cases}\nabla f_i(x^k) & \text{w.p. } p\\ g^{k-1} + \mathcal{Q}\left(\nabla f_{ij}(x^k) - \nabla f_{ij}(x^{k-1})\right) & \text{w.p. } 1-p\end{cases}$$

$j\sim\{1,\dots,m\}$ uniformly at random

# of communication rounds/oracle calls to find an ε-stationary point

$$p = \min\left\{\frac{1}{\omega+1}, \frac{1}{m+1}\right\}$$

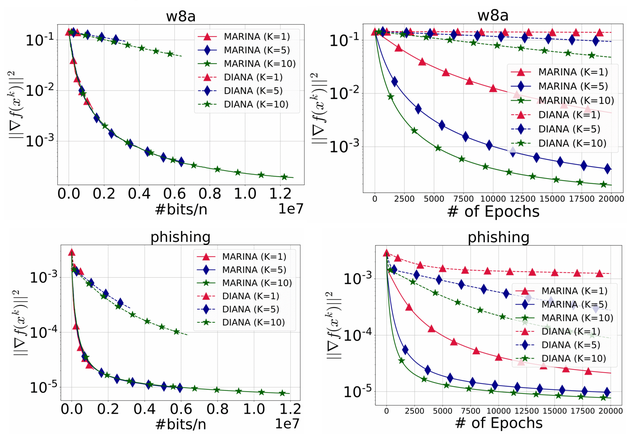
$$\mathcal{O}\left(\frac{1+\max\{\omega, \sqrt{(1+\omega)m}\}/\sqrt{n}}{\varepsilon^2}\right)$$

VR-DIANA [3]: $\mathcal{O}\left(\frac{(m^{2/3}+\omega)\sqrt{1+\omega/n}}{\varepsilon^2}\right)$

(previous SOTA complexity)

6. Numerical Experiments

Binary Classification with Non-Convex Loss



7. In the paper, we also have

(1) Mini-batch VR-MARINA, **(2)** VR-MARINA for expectation minimization, **(3)** MARINA with partial partcipation of clients, **(4)** analysis under Polyak-Lojasiewicz condition, **(5)** explicit dependencies on smoothness constants, non-uniform sampling, **(6)** experiments with neural networks