NIPS, December 2-8, 2018, Montréal

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Sparsified SGD with Memory

{sebastian.stich, jean-baptiste.cordonnier, martin.jaggi}@epfl.ch

Quantized SGD: cheaper communication, but slower convergence

Problem:

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

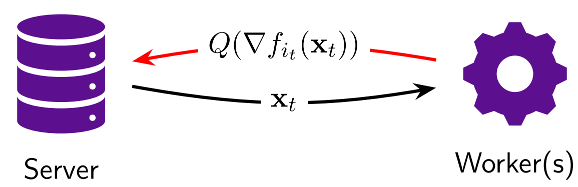
L-smooth $f_i\colon \mathbb{R}^d \to \mathbb{R}$, μ-strongly convex $f\colon \mathbb{R}^d \to \mathbb{R}$

Setting: Data-parallel SGD with parameter server

Quantization operator $Q\colon \mathbb{R}^d \to \mathbb{R}^d$

- **unbiased** $\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}, \forall \mathbf{x}\in\mathbb{R}^d$
- **bounded variance** $\mathbb{E}\left\|Q(\mathbf{x})\right\|^2 \leq \rho \left\|\mathbf{x}\right\|^2, \forall \mathbf{x}\in\mathbb{R}^d$

Gradient Quantization Reduces Communication Cost per Iteration



Example 1: Ternary Quantization

$$Q(\mathbf{x}) = \operatorname{sign}(\mathbf{x})\cdot\|\mathbf{x}\|\cdot\xi(\mathbf{x})$$

where $\xi(\mathbf{x})_i = 1$ with probability $\frac{x_i}{\|\mathbf{x}\|}$, $\xi(\mathbf{x})_i = 0$ otherwise.
$\mathbb{E}\left\|Q(\mathbf{x})-\mathbf{x}\right\|^2 \leq \sqrt{d}\left\|\mathbf{x}\right\|^2$, **sparsity** $\mathbb{E}\left\|Q(\mathbf{x})\right\|_0 \leq 1+\sqrt{d}$

Example 2: Quantization with s levels (QSGD)

$$Q(\mathbf{x}) = \operatorname{sign}(\mathbf{x})\cdot\|\mathbf{x}\|\cdot\xi(\mathbf{x},s)$$

where $\xi(\mathbf{x},s)_i = \frac{\ell+1}{s}$ with probability $s\frac{x_i}{\|\mathbf{x}\|}-\ell$, $\xi(\mathbf{x},s)_i = \frac{\ell}{s}$ otherwise. Here $\frac{\ell}{s} \leq \frac{x_i}{\|\mathbf{x}\|} \leq \frac{\ell+1}{s}$ for integers $\ell \leq s$.
$\mathbb{E}\left\|Q(\mathbf{x})-\mathbf{x}\right\|^2 \leq \frac{\sqrt{d}}{s}\left\|\mathbf{x}\right\|^2$, **sparsity** $\mathbb{E}\left\|Q(\mathbf{x})\right\|_0 \leq s(s+\sqrt{d})$

Previous results suffer from multiplicative slowdown:

quantization	$Q(\nabla f_{i_t})$ sparsity	convergence rate
general	—	$\mathcal{O}\left(\frac{G^2\rho}{T}\right)$
1 level (Ternary)	$\sqrt{d}$	$\mathcal{O}\left(\frac{G^2\sqrt{d}}{T}\right)$
s levels (QSGD)	$s(s+\sqrt{d})$	$\mathcal{O}\left(\frac{G^2 s(s+\sqrt{d})}{T}\right)$

Increasing the number of levels does not help:

$\sqrt{d}$ levels (QSGD)	d	$\mathcal{O}\left(\frac{G^2}{T}\right)$

This Paper: Better sparsity and faster rate:

1 compression	1	$\mathcal{O}\left(\frac{G^2+d}{T}\right)$
k compression	k	$\mathcal{O}\left(\frac{G^2+d/k}{T}\right)$

Mem-SGD: cheaper communication and faster convergence

Compression operator $\operatorname{comp}_k\colon \mathbb{R}^d \to \mathbb{R}^d$

$$\mathbb{E}\left\|\operatorname{comp}_k(\mathbf{x})-\mathbf{x}\right\|^2 \leq \left(1-\frac{k}{d}\right)\left\|\mathbf{x}\right\|^2, \forall \mathbf{x}\in\mathbb{R}^d$$

Example 1: Random-k Compression

$$\operatorname{comp}_k(\mathbf{x})_i = \begin{cases} x_i & \text{with probability } \frac{k}{d} \\ 0 & \text{otherwise} \end{cases}$$

Example 2: Top-k Compression

$$\operatorname{comp}_k(\mathbf{x})_i = \begin{cases} x_i & \text{if } |x_i| \in \left\{|\mathbf{x}|_{(1)},\dots,|\mathbf{x}|_{(k)}\right\} \\ 0 & \text{otherwise} \end{cases}$$

Example 3: Rescaled Random Quantization

$$\operatorname{comp}_{\sqrt{d}}(\mathbf{x}) = \frac{1}{1+\sqrt{d}}Q(\mathbf{x})$$

for ternary quantizer Q (and analogous for s-level quant.)

Main Principle:

Error compensation through auxiliary **memory** $\mathbf{m}\in\mathbb{R}^d$.

(Similar mechanism as e.g. in **1Bit-SGD**.)

Algorithm 1 MEM-SGD

1: Initialize variables $\mathbf{x}_0$ and $\mathbf{m}_0 = \mathbf{0}$
2: **for** t **in** $0\dots T-1$ **do**
3: Sample i_t uniformly in $[n]$
4: $\mathbf{g}_t \leftarrow \operatorname{comp}_k(\mathbf{m}_t + \eta_t\nabla f_{i_t}(\mathbf{x}_t))$ ▷ on worker
5: $\mathbf{m}_{t+1} \leftarrow \mathbf{m}_t + \eta_t\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{g}_t$
6: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \mathbf{g}_t$ ▷ on server
7: **end for**

Theorem:
For stepsizes $\eta_t = \frac{8}{\mu(5\frac{d}{k}+t)}$, $G^2 \geq \mathbb{E}\left\|\nabla f_{i_t}(x_t)\right\|^2$ it holds

$$\mathbb{E}f(\bar{x}_T) - f^\star = \mathcal{O}\left(\frac{G^2 + \frac{d}{k}\sqrt{\kappa}}{\mu T}\right)$$

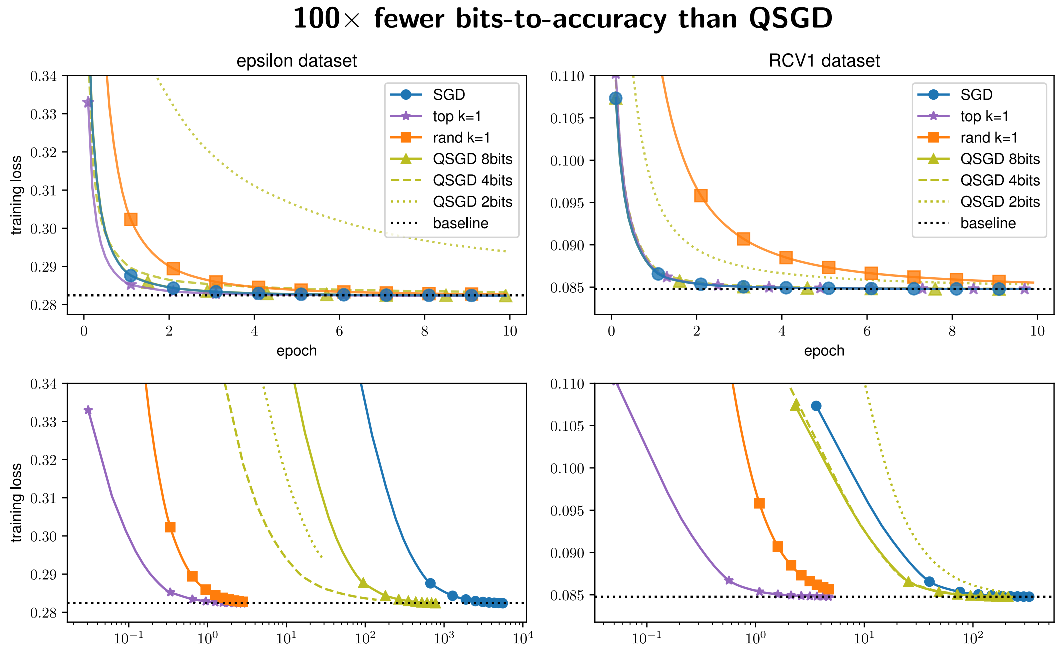
with $\bar{\mathbf{x}}_T := \frac{1}{\sum_{t=0}^{T-1} w_t}\sum_{i=0}^{T-1} w_t\mathbf{x}_t$, $w_t = \left(5\frac{d}{k}+t\right)^2$, $\kappa = \frac{L}{\mu}$.

Remarks:

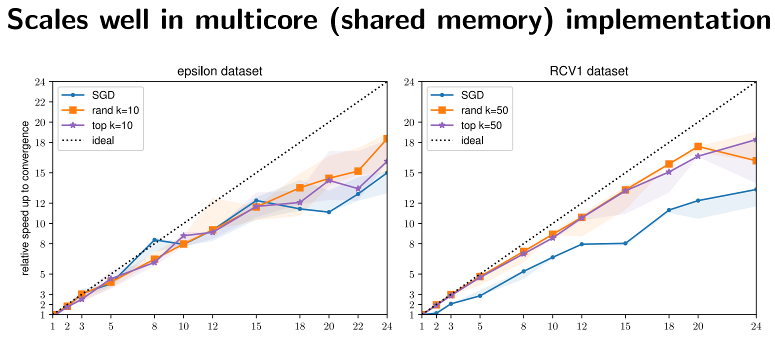
- Previous methods required $\mathcal{O}\left(G^2\cdot\frac{d}{k}\right)$ steps to converge, we need $\mathcal{O}\left(G^2+\frac{d}{k}\sqrt{\kappa}\right)$ instead.
- Theory holds for arbitrary compression operators.
- Previous analyses were often limited to *unbiased* updates. Our analysis avoids this limitation which allows—together with the memory variable—to obtain faster rates.

Experiments

100× fewer bits-to-accuracy than QSGD



Scales well in multicore (shared memory) implementation



Logistic Regression:

$$\frac{1}{n}\sum_{i=1}^n \log\left(1+\exp(-b_i\mathbf{a}_i^\top\mathbf{x})\right) + \frac{\lambda}{2}\|\mathbf{x}\|^2$$

Datasets:

	n	d	density
epsilon	400'000	2'000	100%
RCV1-test	677'399	47'236	0.15%

Open Problems and Future Work

- ✓ Theoretical analysis for $W > 1$ workers, also with compression of the state $\mathbf{x}_t$ communication.
- ✓ Asynchronous updates.
- Extension of the theory to non-convex objectives.

Code
github.com/epfml/sparsifiedSGD