Natural Compression for Distributed Deep Learning[2]

Samuel Horváth[1] Chen-Yu Ho[1] Ludovít Horváth[2] Atal Sahu[1]
Marco Canini[1] Peter Richtárik[1,3]

[1] KAUST [2] Comenius University [3] Moscow Institute of Physics and Technology

The Problem

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

- $f_i(x) := \mathrm{E}_{\zeta\sim\mathcal{D}_i}\left[f_\zeta(x)\right]$, where $\mathcal{D}_i$ is the distribution of data stored on worker i
- f is smooth, f_i's have bounded variance σ_i^2, $\mathrm{E}\left[\left\|\nabla f_{\zeta_i(x)} - \nabla f_i(x)\right\|^2\right] \leq \sigma_i^2$, $\sigma^2 = {}^1\!/_n \sum \sigma_i^2$,
- n is number of nodes

Communication as the Bottleneck

- A key bottleneck of distributed SGD is the cost of communication of the typically dense gradient vectors $g_i(x^k)$.
- In typical distributed computing environments, communication takes more time than computation.
- Two orthogonal types of remedies:
  - Local iterations: give each worker "more useful work" to do before any communication takes place
  - Gradient compression: communicate compressed gradients instead of full gradients

Main Contributions

- New Compression Operators. We construct a new *"natural"* operators based on a randomized binary rounding scheme.
- Computation-Free Simple Low-Level Implementation. "Natural" compatibility with binary floating point types.
- Post-compression Mechanism. Provable theoretical and practical speedup improvements through composition ($\circ$) with previous methods.
- Proof-of-Concept System with In-Network Aggregation. Our mechanisms are the first mechanism that are provably able to operate in the SwitchML [2] framework.
- Theory of general quantized SGD. (Algorithm 1)

Compression Operators

$\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ is called an unbiased bounded-second moment compression operator (not. $\mathcal{C} \in \mathbb{B}(\omega)$) if

$$\mathrm{E}\left[\mathcal{C}(x)\right] = x, \quad \mathrm{E}\left\|\mathcal{C}(x)\right\|^2 \leq (\omega+1)\left\|x\right\|^2, \quad \forall x \in \mathbb{R}^d.$$

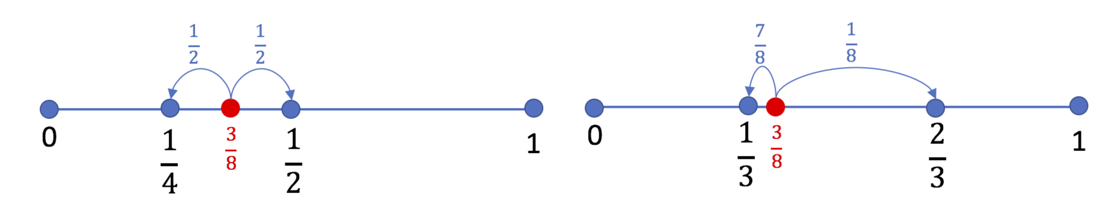
Relative Iterations Slowdown

$$(\omega_M+1)\left({}^{(\omega_W+1)\sigma^2}\!/_n + (1+{}^{\omega_W}\!/_n)\varepsilon\right)\Big/\left({}^{\sigma^2}\!/_n+\varepsilon\right),$$

with respect to non-compressed SGD. (ϵ-precision, worker's $\mathcal{C}_W \in \mathbb{B}(\omega_W)$, master's $\mathcal{C}_M \in \mathbb{B}(\omega_M)$ compression).

Natural Dithering $\mathcal{D}_{\mathrm{nat}}^{p,s}$

- Inexact version of natural compression.
- Fixing the number of level s, natural dithering $\mathcal{D}_{\mathrm{nat}}^{p,s}$ has $\mathcal{O}(2^{s-1}/s)$ times smaller variance than standard dithering $\mathcal{D}_{\mathrm{sta}}^{p,s}$.



Randomized rounding for natural (left) and standard (right) dithering ($s = 3$ levels).

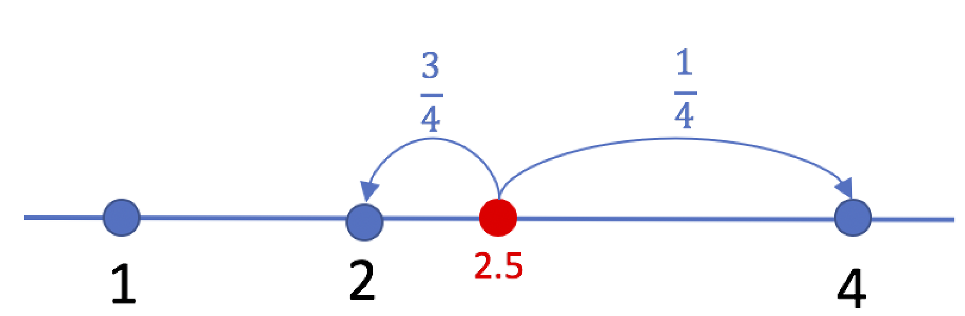
Relative Iteration Complexity to Achieve $\mathrm{E}\left[\|\nabla f(x)\|^2\right] \leq \epsilon$ ($\omega_M = 0$)

Approach	$\mathcal{C}_{W_i}$	Relative # Iterations $\theta(n) \in (0,1]$, decreasing in n	Bits per 1 iter. $W_i \mapsto M$	Speedup Factor
Baseline	identity	1	$32d$	1
New	$\mathcal{C}_{\mathrm{nat}}$	$({}^9\!/_8)^\theta$	$9d$	3.2×–3.6×
Sparsification (q non-zeros)	$\mathcal{S}^q$	$({}^d\!/_q)^\theta$	$(33+\log_2 d)q$	0.6×–6.0×
New	$\mathcal{C}_{\mathrm{nat}} \circ \mathcal{S}^q$	$({}^{9d}\!/_{8q})^\theta$	$(10+\log_2 d)q$	1.0×–10.7×
Dithering	$\mathcal{D}_{\mathrm{sta}}^{p,2^{s-1}}$	$(1+\min\{1,\sqrt{d}2^{1-s}\}d^{\frac{1}{\min\{r,2\}}}2^{1-s})^\theta$	$31+d(2+s)$	1.8×–15.9×
New	$\mathcal{D}_{\mathrm{nat}}^{p,s}$	$({}^{81}\!/_{64} + {}^9\!/_8\min\{1,\sqrt{d}2^{1-s}\}d^{\frac{1}{\min\{r,2\}}}2^{1-s})^\theta$	$8+d(2+\log_2 s)$	4.1×–16.0×

Natural Compression $\mathcal{C}_{\mathrm{nat}}$

- New (randomized) compression technique, which performs an element-wise randomized binary rounding of its input $t \in \mathbb{R}$. $\mathcal{C}_{\mathrm{nat}} \in \mathbb{B}(1/8)$

$$\mathcal{C}_{\mathrm{nat}}(t) \stackrel{\text{def}}{=} \begin{cases} \mathrm{sign}(t)\cdot 2^{\lfloor \log_2|t| \rfloor}, & \text{Prob. } p(t) \stackrel{\text{def}}{=} \frac{2^{\lceil \log_2|t| \rceil}-|t|}{2^{\lfloor \log_2|t| \rfloor}}, \\ \mathrm{sign}(t)\cdot 2^{\lceil \log_2|t| \rceil}, & \text{Prob. } 1-p(t), \end{cases}$$

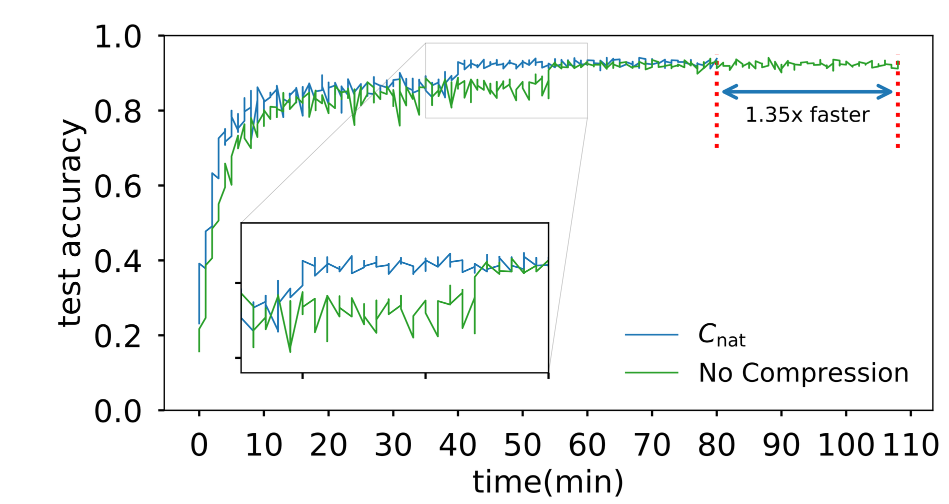


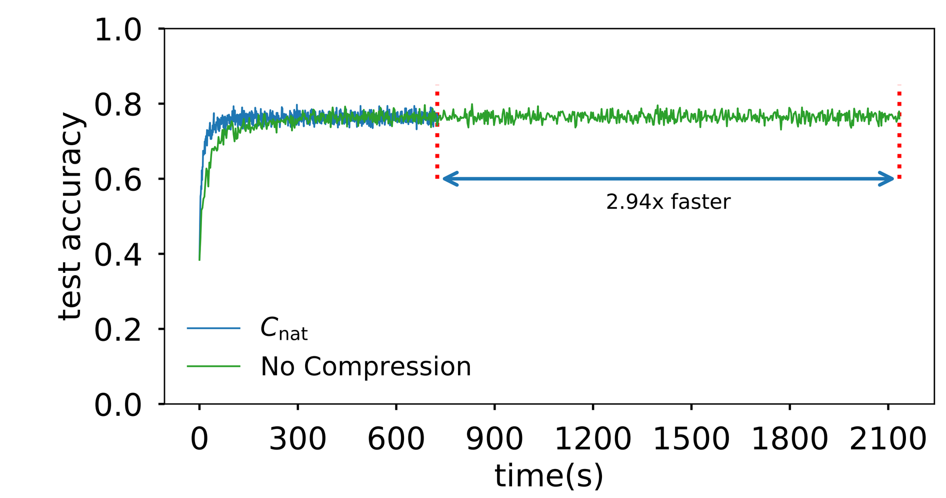
Bi-directional Compression

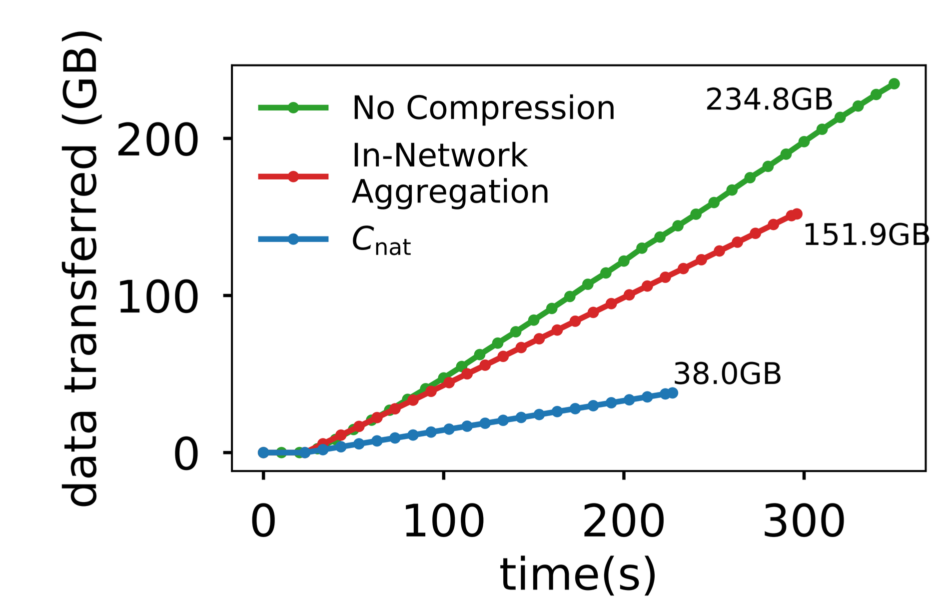
Algorithm 1: `General Quantized SGD`

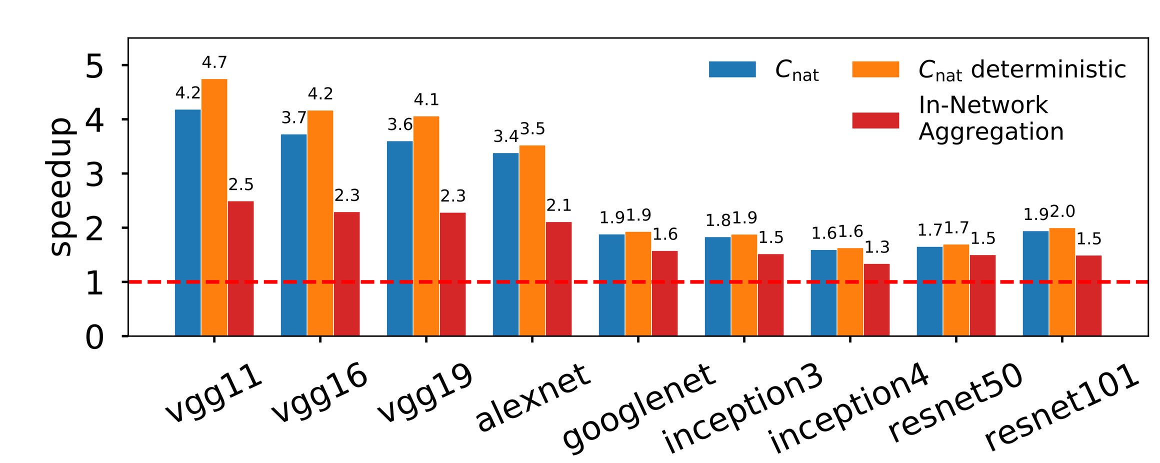
init. vector $x^0 \in \mathbb{R}^d$, step sizes $\{\eta^k\}_{k=0}^T > 0$;
for $k = 0$ *to* T **do**
  for $i = 1$ *to* n **do** in parallel ▷ *Worker side*
    compute stoch. gradient $g_i(x^k) \approx f_i(x^k)$
    compress stoch. gradient $\Delta_i^k = \mathcal{C}_{W_i}(g_i(x^k))$
  end
  aggregate compressed gradients $\Delta^k = \sum_{i=1}^n \Delta_i^k$
  compress aggregated vector $g^k = \mathcal{C}_M(\Delta^k)$
  broadcast g^k ▷ *Master side*
end
for $i = 1, \ldots, n$ **do** in parallel ▷ *Worker side*
  $x^{k+1} = x^k - \frac{\eta^k}{n} g^k$;
end

Numerical Results









[1] Dan Alistarh, 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*, pages 1709–1720, 2017.

[2] Amedeo Sapio, Marco Canini, Chen-Yu Ho, Jacob Nelson, Panos Kalnis, Changhoon Kim, Arvind Krishnamurthy, Masoud Moshref, Dan R. K. Ports, and Peter Richtárik. Scaling distributed machine learning with in-network aggregation. *CoRR*, abs/1903.06701, 2019.