Shifted Compression Framework: Generalizations and Improvements

Egor Shulgin Peter Richtárik

King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

The Problem: Distributed Optimization

$$\text{Find } x^\star = \arg\min_{x\in\mathbb{R}^d} \left[f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) \right], \qquad (\star)$$

where x represents the parameters of a machine learning model we wish to train, n is the number of workers/clients, and each $f_i : \mathbb{R}^d \to \mathbb{R}$ is an L_i-smooth loss and f is μ-strongly convex.

Communication as the Bottleneck

Problem: In distributed systems, communication from workers to the server can take much more time than computation.
Possible Solution: Lossy Compression $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$

Server: $x^{k+1} = x^k - \frac{\gamma}{n}\sum_{i=1}^n \mathcal{C}_i(\nabla f_i(x^k))$

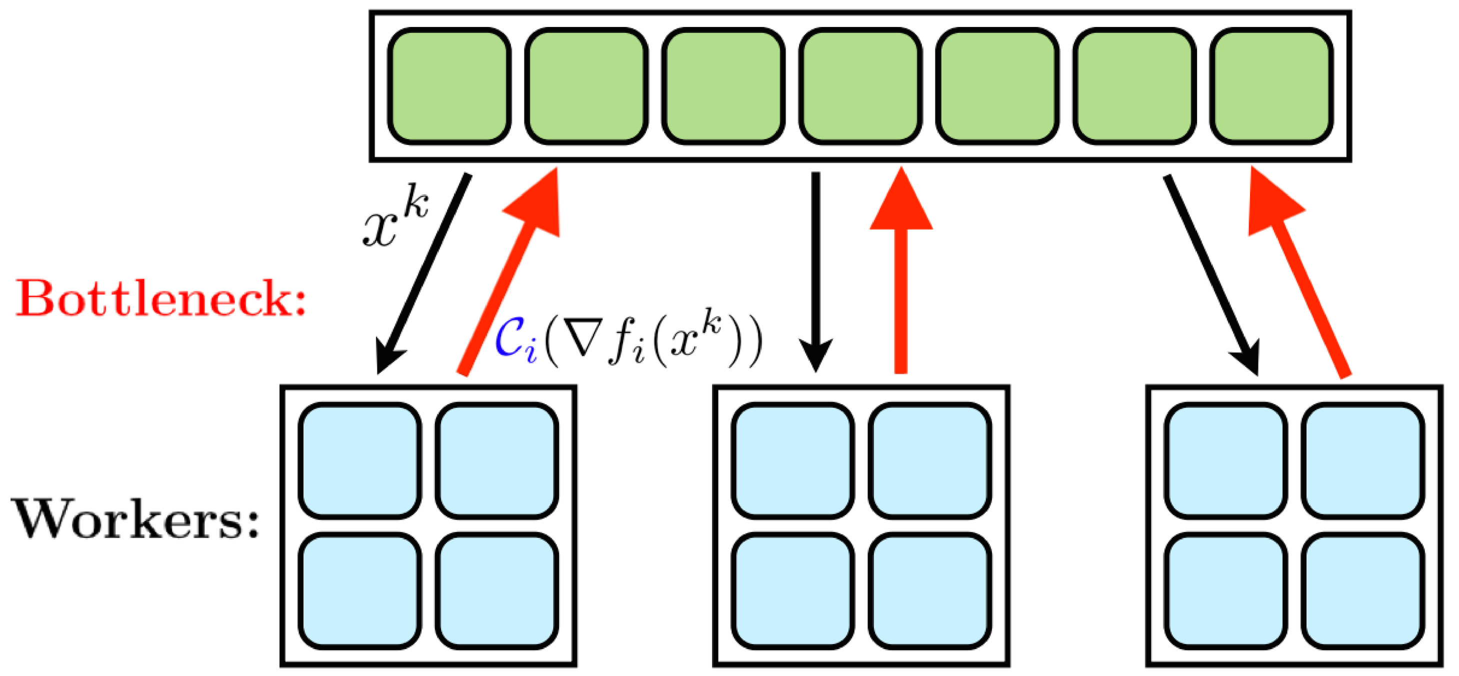

Figure 1: Distributed Compressed Gradient Descent (DCGD) scheme [3]

Compression Operators

Contractive ($\mathcal{C} \in \mathbb{B}(\delta)$, $\delta \in (0,1]$):

$$\mathbf{E}\,\|\mathcal{C}(x) - x\|^2 \le (1-\delta)\|x\|^2, \qquad \forall x \in \mathbb{R}^d$$

- Pro: low empirical variance
- Con: may not converge without Error-Feedback [1]

Unbiased ($\mathcal{Q} \in \mathbb{U}(\omega)$, $\omega \ge 0$):

$$\mathbf{E}\,\mathcal{Q}(x) = x, \qquad \mathbf{E}\,\|\mathcal{Q}(x) - x\|^2 \le \omega\|x\|^2, \qquad \forall x \in \mathbb{R}^d$$

- Pro: have better guarantees (variance decreases with n)
- Con: can have higher empirical variance

Issue: DCGD with unbiased compressors $\mathcal{Q}_i \in \mathbb{U}(\omega)$ and a constant step-size converges (linearly) to a neighbourhood:

$$\mathbf{E}\left\|x^k - x^\star\right\|^2 \le (1-\gamma\mu)^k\|x^0 - x^\star\|^2 + \frac{2\gamma\omega}{\mu n}\cdot\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2 \qquad (1)$$

Fix: Shifted Compressor

Randomized mapping $\mathcal{Q}_h$ is a **shifted compression operator** ($\mathcal{Q}_h \in \mathbb{U}(\omega; h)$) if

$$\mathbf{E}\,\mathcal{Q}_h(x) = x, \qquad \mathbf{E}\,\|\mathcal{Q}_h(x) - x\|^2 \le \omega\|x - h\|^2 \qquad \forall x \in \mathbb{R}^d. \qquad (2)$$

Lemma. All shifted compressors arise by a shift of unbiased operator $\mathcal{Q} \in \mathbb{U}(\omega)$

$$\mathcal{Q}_h(x) = h + \mathcal{Q}(x - h).$$

This gives rise to a shifted **gradient estimator**: $g_h(x) = \mathcal{Q}_h(\nabla f(x))$ and method

$$x^{k+1} = x^k - \gamma\frac{1}{n}\sum_{i=1}^n g_{h_i}(x) = x^k - \gamma\frac{1}{n}\sum_{i=1}^n \left[h_i^k + \mathcal{Q}_i\left(\nabla f_i(x) - h_i^k\right)\right]. \qquad \text{(DCGD-SHIFT)}$$

The same trick can be applied using a (possibly biased) compressor $\mathcal{C}$ for the **shift** h:

$$h = s + \mathcal{C}(\nabla f(x) - s). \qquad (3)$$

General Framework: Choosing the Shifts

			SHIFT $h_i^{k+1} = s_i^k + \mathcal{C}_i\left(\nabla f_i(x^k) - s_i^k\right)$	
METHOD	**REF**	**VR?**	s_i^k	$\mathcal{C}_i$
DCGD	[3]	✗	0	$\mathcal{O}$
DCGD-SHIFT	[New]	✗	s_i^0	$\mathcal{O}$
DCGD-STAR	[New]	✓	$\nabla f_i(x^\star)$	any $\mathcal{C}_i \in \mathbb{B}(\delta)$
DIANA	[4]	✓	h_i^k	$\alpha\mathcal{Q}_i,\ \mathcal{Q}_i \in \mathbb{U}(\omega_i)$
Rand-DIANA	[New]	✓	h_i^k	$\mathcal{B}e_{p_i}$
GDCI	[2]	✗	x^k/γ	$\mathcal{O}$

Table 1: List of existing and new algorithms which fit our framework. **VR** – variance reduced method. $\mathcal{O}/\mathcal{I}$ – zero/identity, $\mathcal{B}e_p = \{x/0 \text{ with prob. } p/(1-p)\}$ – Bernoulli compressor.

ALGORITHM	**PREVIOUS**	**OUR RESULT**
DCGD-SHIFT	–	$\kappa\left(1+\frac{\omega}{n}\right)$
DIANA	$\max\left\{\kappa\left(1+\frac{\omega}{n}\right), \omega\right\}$	$\max\left\{\kappa\left(1+\frac{\omega}{n}(1-\delta)\right), \omega(1-\delta)\right\}$
Rand-DIANA	–	$\max\left\{\kappa\left(1+\frac{\omega}{n}(1-\delta)\right), \frac{1}{p}\right\}$
GDCI	$\kappa^2\left(1+\frac{\omega}{n}\right)$	$\kappa\left(1+\frac{\omega}{n}\right)$
VR-GDCI	$\max\left\{\kappa^2\left(1+\frac{\omega}{n}\right), \omega\right\}$	$\max\left\{\kappa\left(1+\frac{\omega}{n}\right), \omega\right\}$

Table 2: Summary of iteration complexity results (without $\log 1/\varepsilon$ factors) with highlighted improvements over the previous works. Results for non VR methods are in the interpolation regimes: $\nabla f_i(x^\star) = 0 = x^\star - \gamma\nabla f_i(x^\star)$. Last two rows: methods with compressed iterates.

New Method: Rand-DIANA

Learns the shift in a **randomized** (loop-less) way:

$$h_i^k = \nabla f_i(w_i^k)$$
$$w_i^{k+1} = \begin{cases} x^k & \text{with probability } p_i \\ w_i^k & \text{with probability } 1 - p_i \end{cases} \qquad (4)$$

Convergence of Rand-DIANA

Assume f_i are convex and L_i-smooth, f is μ-convex and step size

$$\gamma \le \left[\left(1+\frac{2\omega}{n}\right)L_{\max} + M\max_i(p_i L_i)\right]^{-1}$$

where $M > 2\omega/(np_m)$, $L_{\max} = \max_i L_i$, $p_m := \min_i p_i$. Then the iterates of DCGD-SHIFT with Rand-DIANA shift update (4) satisfy

$$\mathbf{E}\left[V^k\right] \le \max\left\{(1-\gamma\mu)^k, \left(1 - p_m + \frac{2\omega}{nM}\right)^k\right\} V^0,$$

where the Lyapunov function V^k is defined by

$$V^k := \left\|x^k - x^\star\right\|^2 + M\gamma^2\cdot\frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^\star)\right\|^2.$$

Experiments

ℓ_2-regularized logistic regression problem with *w2a* LibSVM dataset.

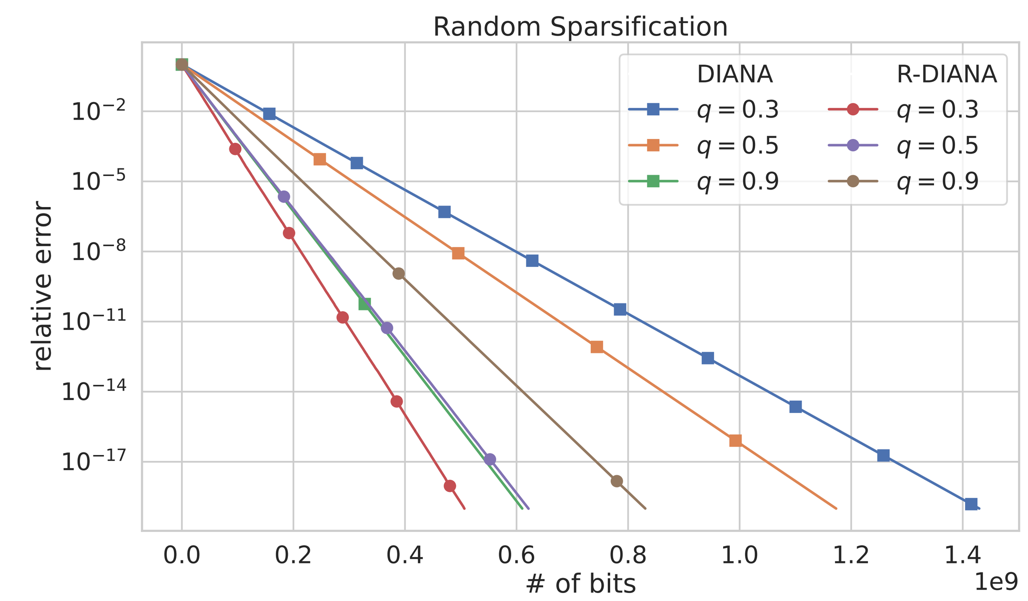

Figure 2: Comparison of DIANA and Rand-DIANA with varying parameter of Rand-K sparsification compressor.

Shifted compressor can also be used for **model compression**:

$$x^{k+1} = x^k - (\eta\gamma)\left[x^k - \mathcal{Q}\left(x^k - \gamma\nabla f(x^k)\right)\right]/\gamma \qquad \text{(GDCI)}$$

References:

[1] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

[2] S. Chraibi, A. Khaled, D. Kovalev, P. Richtárik, A. Salim, and M. Takáč. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:2102.07245*, 2019.

[3] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

[4] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.