OPT2021: 13th Annual Workshop on Optimization for Machine Learning

Shifted Compression Framework: Generalizations and Improvements

Egor Shulgin EGOR.SHULGIN@KAUST.EDU.SA
Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract

Communication is one of the key bottlenecks in the distributed training of large-scale machine learning models, and lossy compression of exchanged information, such as stochastic gradients or models, is one of the most effective instruments to alleviate this issue. Among the most studied compression techniques is the class of unbiased compression operators with variance bounded by a multiple of the square norm of the vector we wish to compress. By design, this variance may remain high, and only diminishes if the input vector approaches zero. However, unless the model being trained is overparameterized, there is no a-priori reason for the vectors we wish to compress to approach zero during the iterations of classical methods such as distributed compressed SGD, which has adverse effects on the convergence speed. Due to this issue, several more elaborate and seemingly very different algorithms have been proposed recently, with the goal of circumventing this issue. These methods are based on the idea of compressing the *difference* between the vector we would normally wish to compress and some auxiliary vector which changes throughout the iterative process. In this work we take a step back, and develop a unified framework for studying such methods, conceptually, and theoretically. Our framework incorporates methods compressing both gradients and models, using unbiased and biased compressors, and sheds light on the construction of the auxiliary vectors. Furthermore, our general framework can lead to the improvement of several existing algorithms, and can produce new algorithms. Finally, we performed several numerical experiments which illustrate and support our theoretical findings.

1. Introduction

We consider distributed optimization problem

$$\min_{x\in\mathbb{R}^d} \left[f(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right], \tag{$\star$}$$

where n is the number of workers/clients and $f_i : \mathbb{R}^d \to \mathbb{R}$ is a smooth function representing the loss of the model parametrized by $x \in \mathbb{R}^d$ for data stored on node i. Such formulation has become very popular in recent years due to the need for training large-scale machine learning models [13].

Communication bottleneck. Compute nodes have to exchange information in a distributed learning process. The size of the sent messages (usually gradients or model updates) can be very large which forms a significant bottleneck [23, 29, 36] of the whole training procedure. One of the main practical solutions to this problem is lossy *communication compression* [2, 20, 37]. It suggests applying a (possibly randomized) mapping $\mathcal{C}$ to a vector/matrix/tensor x before it is transmitted in order to produce a less accurate estimate $\mathcal{C}(x) : \mathbb{R}^d \to \mathbb{R}^d$ and thus save bits sent per every communication round.

Summary of contributions The main results of this work include:

1. Generalizations of existing methods. We introduce the concept of *Shifted Compressor* which generalizes a common definition of compression operators used in distributed learning. It allows to study various strategies for updating the shifts using both biased and unbiased compressors; to recover and improve such previously known methods as DCGD and DIANA. As a byproduct a new algorithm is also obtained: DCGD-STAR, which enjoys linear convergence to the exact solution if we know the local gradients at the optimum.

2. Improved rates. The notion of a shifted compressor allows us to revisit existing analysis of distributed methods with *compressed iterates* and improve guarantees in both cases: with and without variance-reduction. Obtained results indicate that algorithms with model compression can have the same complexity as compressed gradient methods.

3. New algorithm. We present a novel distributed algorithm with compression, called Randomized DIANA, with linear convergence rate to the exact optimum. It has a significantly *simpler analysis* than the original DIANA method. Via examination of its experimental performance we highlight the cases when it can outperform DIANA in practice.

Obtained theoretical results are summarized in Table 1 with highlighted improvements over the previous works. Notation: κ - condition number, ω/δ - variance of the un/biased compressor. More methods covered by our framework can be found in Table 2. Due to space limitations Related works, Experiments sections and some of the theoretical results are left for the Appendix.

Table 1: Iteration complexities are presented in $\tilde{\mathcal{O}}$-notation to omit $\log 1/\varepsilon$ factors and for the simplified case $\omega_i \equiv \omega, \delta_i \equiv \delta, L_i \equiv L$, $p_i \equiv p$. More refined results are in Theorems with links in the second column. The last two rows refer to the methods with compressed iterates. Complexities for DCGD-SHIFT and GDCI are in the interpolation regimes: $\nabla f_i(x^\star) = 0 = x^\star - \gamma \nabla f_i(x^\star)$.

ALGORITHM	REF	PREVIOUS	OUR RESULT
DCGD-SHIFT	5	–	$\kappa\left(1+\frac{\omega}{n}\right)$
Rand-DIANA	6	–	$\max\left\{\kappa\left(1+\frac{\omega}{n}(1-\delta)\right), \frac{1}{p}\right\}$
DIANA	11	$\max\left\{\kappa\left(1+\frac{\omega}{n}\right), \omega\right\}$	$\max\left\{\kappa\left(1+\frac{\omega}{n}(1-\delta)\right), \omega(1-\delta)\right\}$
GDCI	14	$\kappa^2\left(1+\frac{\omega}{n}\right)$	$\kappa\left(1+\frac{\omega}{n}\right)$
VR-GDCI	D.2	$\max\left\{\kappa^2\left(1+\frac{\omega}{n}\right), \omega\right\}$	$\max\left\{\kappa\left(1+\frac{\omega}{n}\right), \omega\right\}$

2. General Framework

In this section we introduce compression operators and the framework of shifted compressors.

2.1. Standard Compression

At first recall some basic definitions.

Definition 1 (General contractive compressor) *A (possibly) randomized mapping $\mathcal{C}:\mathbb{R}^d\to\mathbb{R}^d$ is a* **compression operator** *($\mathcal{C}\in\mathbb{B}(\delta)$ for brevity) if for some $\delta\in[0,1]$ and $\forall x\in\mathbb{R}^d$*

$$\mathbf{E}\,\|\mathcal{C}(x)-x\|^2\leq(1-\delta)\|x\|^2,$$

where the expectation is taken w.r.t. (possible) randomness of operator $\mathcal{C}$.

One of the most known operators from this class is *greedy sparsification* (`Top-K`):

$$\mathcal{C}_{\texttt{Top-K}}(x)\coloneqq\textstyle\sum_{i=d-K+1}^{d}x_{(i)}e_{(i)},$$

where $K\in[d]\coloneqq\{1,\dots,d\}$, coordinates are ordered by their magnitudes so that $|x_{(1)}|\leq|x_{(2)}|\leq\cdots\leq|x_{(d)}|$, and $e_i\in\mathbb{R}^d$ are the standard unit basis vectors. This compressor belongs to $\mathbb{B}\left(d/K\right)$.

Definition 2 (Unbiased compressor) *A randomized mapping $\mathcal{Q}:\mathbb{R}^d\to\mathbb{R}^d$ is an* **unbiased compression operator** *($\mathcal{Q}\in\mathbb{U}(\omega)$ for brevity) if for some $\omega\geq0$ and $\forall x\in\mathbb{R}^d$*

$$\begin{aligned}&(a)\ \mathbf{E}\,\mathcal{Q}(x)=x, && \textit{(Unbiasedness)}\\ &(b)\ \mathbf{E}\,\|\mathcal{Q}(x)-x\|^2\leq\omega\|x\|^2 && \textit{(Bounded variance)}\end{aligned}$$

A notable example from this class is the *random sparsification* (`Rand-K` for $K\in[d]$) operator:

$$\mathcal{Q}_{\texttt{Rand-K}}(x)\coloneqq\tfrac{d}{K}\textstyle\sum_{i\in S}x_ie_i,\tag{1}$$

where S is a random subset of $[d]$ sampled from the uniform distribution on the all subsets of $[d]$ with cardinality K. `Rand-K` belongs to $\mathbb{U}\left(d/K-1\right)$. Notice that property (a) from Definition 2 is "uniform" across all vectors x, while property (b) is not. Namely, vector $x=0$ is treated *in a special way* because $\mathbf{E}\,\|\mathcal{Q}(0)-0\|^2=0$, which means that the compressed zero vector has *zero variance*. In other words, zero is mapped to itself with probability 1.

2.2. Compression with shift

We can generalize the class of unbiased compressors $\mathbb{U}(\omega)$ to a class of operators with other (not only 0) "special" vectors. Specifically, this class allows for **shifts** away from the origin, which is formalized in the following definition.

Definition 3 (Shifted compressor) *Let $h\in\mathbb{R}^d$. A randomized mapping $\mathcal{Q}_h:\mathbb{R}^d\to\mathbb{R}^d$ is a* ***shifted compression operator*** *($\mathcal{Q}_h\in\mathbb{U}(\omega;h)$ in short) if exists $\omega\geq0$ such that $\forall x\in\mathbb{R}^d$*

$$\begin{aligned}&(a)\ \mathbf{E}\,\mathcal{Q}_h(x)=x\\ &(b)\ \mathbf{E}\,\|\mathcal{Q}_h(x)-x\|^2\leq\omega\|x-h\|^2.\end{aligned}$$

Vector h is called a ***shift****. Note that class of unbiased compressors $\mathbb{U}(\omega)$ is equivalent to $\mathbb{U}(\omega;0)$.*

Next lemma shows that shifts add up and all shifted compression operators $\mathcal{Q}_h\in\mathbb{U}(\omega;h)$ arise by a shift of some operator $\mathcal{Q}_0$ from $\mathbb{U}(\omega;0)$.

Lemma 4 (Shifting a Shifted Compressor) *Let $\mathcal{Q}_h\in\mathbb{U}(\omega;h)$ and $v\in\mathbb{R}^d$. Then the (possibly) randomized mapping $\mathcal{Q}$ defined by $\mathcal{Q}(x)\coloneqq v+\mathcal{Q}_h(x-v)$ satisfies $\mathcal{Q}\in\mathbb{U}(\omega;h+v)$.*

Shifted compressor concept allows to construct a shifted **gradient estimator** $\mathcal{Q}_h \in \mathbb{U}(\omega; h)$

$$g_h(x) = \mathcal{Q}_h\left(\nabla f(x)\right) = h + \mathcal{Q}(\nabla f(x) - h), \tag{2}$$

which is the main focus of this work. In particular, we are going to study different mechanisms for choosing this shift vector throughout the optimization process.

Note: Estimator (2) is clearly unbiased, as soon as operator $\mathcal{Q}$ satisfies $\mathbf{E}\,\mathcal{Q}(x) = x$.

Estimator (2) uses operator $\mathcal{Q}$ from class of unbiased compressors $\mathbb{U}(\omega)$, which are usually easier to analyse but have higher empirical variance than biased counterparts [6]. In an attempt to kill two birds with one stone we can incorporate biased compressor $\mathcal{C} \in \mathbb{B}(\delta)$ into h using a similar trick:

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

as $g_h(x)$ allows for virtually any shift vector. This leads to the following estimator[1]

$$\boxed{\begin{aligned} g_h(x) &= h + \mathcal{Q}\left(\nabla f(x) - h\right) \\ &= s + \mathcal{C}(\nabla f(x) - s) + \mathcal{Q}\left(\nabla f(x) - s - \mathcal{C}(\nabla f(x) - s)\right). \end{aligned}} \tag{4}$$

2.3. The meta-algorithm

Now we are ready to present the general distributed optimization algorithm for solving ($\star$) employing shifted gradient estimators

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

Algorithm 1: Distributed Compressed Gradient Descent with Shift (DCGD-SHIFT)

Input: learning rate $\gamma > 0$; unbiased compressors $\mathcal{Q}_1, \ldots, \mathcal{Q}_n$; initial iterate $x^0 \in \mathbb{R}^d$, initial local shifts $h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$ (stored on the n nodes)

Initialize: $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$ (stored on the master)

for $k = 0, 1, 2 \ldots$ **do**

  Broadcast x^k to all workers

  for $i = 1, \ldots n$ **do**

    Compute local gradient: $\nabla f_i(x^k)$

    Compress shifted local gradient: $m_i^k = \mathcal{Q}_i(\nabla f_i(x^k) - h_i^k)$

    Update the local shift: h_i^{k+1}

    Send message m_i^k and (possibly) the shifts h_i^{k+1} to the master

  end

  Aggregate received messages: $m^k = \frac{1}{n}\sum_{i=1}^n m_i^k$

  Compute global gradient estimator: $g^k = h^k + m^k$

  Take Gradient Descent step: $x^{k+1} = x^k - \gamma g^k$

  Update aggregated shift: $h^{k+1} = \frac{1}{n}\sum_{i=1}^n h_i^{k+1}$

end

[1] The resulting estimator is closely related to Induced compressor [14] $\mathcal{Q}_{ind}(x) = \mathcal{C}(x) + \mathcal{Q}\left(x - \mathcal{C}(x)\right)$, which belongs to $\mathbb{U}(\omega(1-\delta))$ class for $\mathcal{C} \in \mathbb{B}(\delta)$ and $\mathcal{Q} \in \mathbb{U}(\omega)$.

In Algorithm 1, each worker $i = 1, \ldots, n$ queries the gradient oracle $\nabla f_i(x^k)$ in iteration k. Then compression operator is applied to the difference between the local gradient and shift, and the result is sent to the master (with possibly the new shift as well). The shift is updated on both the server and workers. After receiving the messages m_i^k a global gradient estimator is formed on the server, and a gradient step is performed.

Please note that this method is not fully defined because it requires a description of the mechanism for updating the shifts h_i^{k+1} (highlighted in red) throughout the iteration process on both workers and master. Next, we will present a general convergence guarantee for this algorithm with fixed shifts $h_k^i \equiv h^i$ and later discuss how it can be practically fixed.

3. Convergence theory

We will use the following standard assumptions for the function $f : \mathbb{R}^d \to \mathbb{R}$

- μ-**Strong convexity**: $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$.
  If $\mu = 0$ then the function is **convex**.
- L-**Smoothness**: $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2$

Theorem 5 (DCGD **with** `fixed` SHIFT) *Assume each f_i is convex and L_i-smooth, and f is L-smooth and μ-strongly convex. Let $\mathcal{Q}_i \in \mathbb{U}(\omega_i)$ - independent unbiased compression operators. If the step-size satisfies*

$$\gamma \leq \frac{1}{L + 2\max_i (L_i \omega_i / n)}.$$

Then the iterates of Algorithm 1 with fixed shifts $h_i^k \equiv h_i$ satisfy

$$\mathbf{E}\left\|x^k - x^\star\right\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^\star\|^2 + \frac{2\gamma}{\mu} \frac{1}{n} \sum_{i=1}^n \frac{\omega_i}{n} \left\|\nabla f_i(x^\star) - h_i\right\|^2. \tag{5}$$

This theorem establishes a linear convergence rate up to a certain oscillation radius, controlled by the average distance of shift vectors h_i to the optimal local gradients $\nabla f_i(x^\star)$ multiplied by the step-size γ. It means that in the interpolation/**overparameterized regime** ($\nabla f_i(x^\star) = 0$ for all i) method reaches **exact solution** with zero shifts $h_i^0 = 0$.

3.1. Learning the optimal shifts

We need to design the sequences $\{h_1^k\}_{k\geq 0}, \ldots, \{h_n^k\}_{k\geq 0}$ in such a way that all of them converge to the optimal shifts:

$$h_i^k \to \nabla f_i(x^\star) \quad \text{as} \quad k \to \infty,$$

but at the same time we do not want to send uncompressed vectors from workers to the master. So, the challenge is not only in learning the shifts, but doing this in a communication-efficient way. In this work we present two different solutions (one left for Appendix) to this problem.

Randomized DIANA (Rand-DIANA) The simplest possible solution to the issue raised above would be just to set h_i^k to $\nabla f_i(x^k)$ because if $x^k \to x^\star$ in the optimization process then $\nabla f_i(x^k)$ converges to the optimal local shift. But this approach is not efficient, as workers have to transfer full (uncompressed) vectors $h_i^k = \nabla f_i(x^k)$. Our new solution is to update a reference point w_i^k for calculating the shift $h_i^k = \nabla f_i(w_i^k)$ infrequently (with small probability $p_i \in (0, 1]$) so that it needs to be communicated very rare:

$$h_i^k = \nabla f_i(w_i^k), \qquad w_i^{k+1} = \begin{cases} x^k & \text{with probability } p_i \\ w_i^k & \text{with probability } 1 - p_i \end{cases} \tag{6}$$

This method has a remarkably simpler analysis than DIANA, but can solve the original problem of eliminating the variance introduced by gradient compression. Next we state the convergence result of DCGD with shifts updated in a `randomized` fashion (6). We give it a Randomized-DIANA name (Rand-DIANA in short) to acknowledge the original method [25] first solving such problem.

Theorem 6 (Rand-DIANA) *Assume that f_i are convex, L_i-smooth for all i and f is μ-convex. If the stepsize satisfies*

$$\gamma \leq \frac{1}{\left(1 + \frac{2\omega}{n}\right) L_{\max} + M \max_i (p_i L_i)},$$

where $M > \frac{2\omega}{np_{\min}}$ and $p_{\min} = \min_i p_i$. Then the iterates of DCGD *with* Randomized-DIANA *shift update* (6) *satisfy*

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

where the Lyapunov function V^k is defined by

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

By appropriate choice of this method's parameters $M = \frac{4\omega}{np_{\min}}$ and $p_i \equiv p = \frac{1}{\omega+1}$ for every i, we can obtain basically the same iteration complexity as for the original DIANA [16]

$$\max\left\{\frac{1}{\gamma\mu}, \frac{1}{p_{\min} - \frac{2\omega}{nM}}\right\} \log\frac{1}{\varepsilon} = \max\left\{\frac{L_{\max}}{\mu}\left(1 + \frac{\omega}{n}\right), \omega + 1\right\} \log\frac{1}{\varepsilon}.$$

In the Appendix one can find proofs for the presented results, generalized and improved analysis of DIANA-like method, application of the shifted compressor to distributed methods with compressed iterates (with and without variance reduction) along with experimental results.

References

[1] Alyazeed Albasyoni, Mher Safaryan, Laurent Condat, and Peter Richtárik. Optimal gradient compression for distributed and federated learning. *arXiv preprint arXiv:2010.03246*, 2020.

[2] 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*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6c340f25839e6acdc73414517203f5f0-Paper.pdf.

[3] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5973–5983, 2018.

[4] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pages 14668–14679, 2019.

[5] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 2018.

[6] Aleksandr Beznosikov, Samuel Horvath, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

[7] Edited by: Peter Kairouz and H. Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1), 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL http://dx.doi.org/10.1561/2200000083.

[8] Sélim Chraibi, Ahmed Khaled, Dmitry Kovalev, Peter Richtárik, Adil Salim, and Martin Takáč. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:2102.07245*, 2019.

[9] Venkata Gandikota, Daniel Kane, Raj Kumar Maity, and Arya Mazumdar. vqsgd: Vector quantized stochastic gradient descent. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2197–2205. PMLR, 2021. URL http://proceedings.mlr.press/v130/gandikota21a.html.

[10] WM Goodall. Television by pulse code modulation. *Bell System Technical Journal*, 30(1): 33–49, 1951.

[11] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[12] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtarik. Linearly converging error compensated sgd. In *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ef9280fbc5317f17d480e4d4f61b3751-Paper.pdf.

[13] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2018.

[14] Samuel Horváth and Peter Richtarik. A better alternative to error feedback for communication-efficient distributed learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=vYVI1CHPaQg.

[15] Samuel Horváth, Chen-Yu Ho, L'udovit Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.

[16] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

[17] Ahmed Khaled and Peter Richtárik. Gradient descent with compressed iterates. *NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

[18] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

[19] Jakub Konečný and Peter Richtárik. Randomized distributed mean estimation: Accuracy vs. communication. *Frontiers in Applied Mathematics and Statistics*, 4:62, 2018.

[20] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

[21] Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 451–467. PMLR, 2020. URL http://proceedings.mlr.press/v117/kovalev20a.html.

[22] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[23] Liang Luo, Jacob Nelson, Luis Ceze, Amar Phanishayee, and Arvind Krishnamurthy. Parameter hub: a rack-scale parameter server for distributed deep neural network training. In *Proceedings of the ACM Symposium on Cloud Computing, SoCC 2018*, pages 41–54, 2018. doi: 10.1145/3267809.3267840. URL https://doi.org/10.1145/3267809.3267840.

[24] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, 20–22 Apr 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

[25] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[26] Konstantin Mishchenko, Bokun Wang, Dmitry Kovalev, and Peter Richtárik. Intsgd: Floatless compression of stochastic gradients. *arXiv preprint arXiv:2102.08374*, 2021.

[27] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk USSR*, 269(3):543–547, 1983.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] Yanghua Peng, Yibo Zhu, Yangrui Chen, Yixin Bao, Bairen Yi, Chang Lan, Chuan Wu, and Chuanxiong Guo. A generic communication scheduler for distributed DNN training acceleration. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles, SOSP 2019*, pages 16–29, 2019. doi: 10.1145/3341301.3359642. URL https://doi.org/10.1145/3341301.3359642.

[30] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2021–2031. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/reisizadeh20a.html.

[31] Lawrence Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, 1962.

[32] Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better theory. *arXiv preprint arXiv:1905.12938*, 2021.

[33] Mher Safaryan, Filip Hanzely, and Peter Richtárik. Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization. *arXiv preprint arXiv:2102.07245*, 2021.

[34] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021.

[35] Mher Safaryan, Egor Shulgin, and Peter Richtárik. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *Information and Inference: A Journal of the IMA*, 04 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaab006. URL https://doi.org/10.1093/imaiai/iaab006. iaab006.

[36] 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. In *18th USENIX Symposium*

on Networked Systems Design and Implementation, NSDI 2021, April 12-14, pages 785–808. USENIX Association, 2021. URL https://www.usenix.org/conference/nsdi21/presentation/sapio.

[37] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[38] Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: Sgd with delayed gradients. *Journal of Machine Learning Research*, 21(237):1–36, 2020. URL http://jmlr.org/papers/v21/19-748.html.

[39] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[40] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Practical low-rank communication compression in decentralized deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 14171–14181. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a376802c0811f1b9088828288eb0d3f0-Paper.pdf.

[41] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/33b3214d792caf311e1f00fd22b392c5-Paper.pdf.

[42] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1299–1309, 2018.

[43] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

[44] Hang Xu, Chen-Yu Ho, Ahmed M. Abdelmoniem, Aritra Dutta, El Houcine Bergou, Konstantinos Karatsenidis, Marco Canini, and Panos Kalnis. Compressed communication for distributed deep learning: Survey and quantitative evaluation. *Technical report*, 2020. URL http://hdl.handle.net/10754/662495.

Appendix

Contents

Appendix A. Related work

A.1. Compression operators.

The topic of gradient compression in distributed learning has been studied extensively over the last years form both practical [44] and theoretical [1, 6, 35] sides. Compression operators are typically divided into 2 large groups: *unbiased* and *biased* ones. The first group includes methods based on some sort of rounding or *quantization*: Random Dithering [10, 31], Ternary quantization [43], Natural [15] and Integer [26] compression. Another popular example is random *sparsification* – `Rand-K` [19, 39, 42], which preserves only a subset of the original vector coordinates. These two approaches can also be combined [4] for even more aggressive compression. There are also many other approaches based on low-rank approximation [34, 40, 41], vector quantization [9], etc. The second group of biased compressors mainly includes greedy sparsification – `Top-K` [3, 39] and various sign-based quantization methods [5, 32, 37]. For a more complete review of compression operators, one can refer to the survey [44] and [6, 35].

A.2. Optimization algorithms.

Compression operators on their own are not sufficient for building a distributed learning system because they always go along with optimization algorithms. One of the first theoretically analyzed methods is Distributed Compressed Gradient Descent (DCGD) [18] which considered arbitrary unbiased compressors. The issue with it is that it was proven to converge linearly only to a neighborhood of the optimal point with constant step-size. DIANA [25] fixed this problem by compressing specially designed gradient differences. Later it was combined with variance reduction [16], accelerated [22] in Nesterov's sense [27] and by using smoothness matrices [33] with a properly designed sparsification technique.

On the other side are methods working with biased compressors, which require the use of the error-feedback (EF) mechanism [3, 37, 38]. Such algorithms were considered to be often better in practice due to smaller variance of biased updates [6]. But recently it was demonstrated that biased compressors can be incorporated into specially designed unbiased operators and show superior to error-feedback results [14]. In addition, error-feedback was recently combined with the DIANA trick [12], which led to the first linearly converging method with EF.

A.3. Compressed iterates.

Most of the existing literature (including all methods described above) focuses on compression of the gradients, while in applications like Federated Learning [7, 20, 24], it is vital to reduce the size of the broadcasted model parameters [30]. This demand gives rise to optimization algorithms with compressed iterates. The first attempt to analyse such methods was done in [17] for a single node set up. Later it was combined with variance-reduction for noise introduced by compression and generalized to a much more general setting of distributed fixed point methods [8].

In Table 2, we show the generality of our approach by presenting some of the existing and new distributed methods falling into our framework of DCGD-SHIFT with shift updates of the form (3).

[3] $\mathcal{B}e_p(x) := \begin{cases} x & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

Table 2: List of existing and new algorithms which fit our general framework. **VR** – variance reduced method. $\mathcal{O}/\mathcal{I}$ – zero/identity operator, $\mathcal{B}e_{p_i}$ – Bernoulli[3] compressor. DGD refers to Distributed Gradient Descent.

			SHIFT $h_i^{k+1} = s_i^k + \mathcal{C}_i\left(\nabla f_i(x^k) - s_i^k\right)$	
METHOD	**REFERENCE**	**VR?**	s_i^k	$\mathcal{C}_i$
DCGD	[18]	✗	0	$\mathcal{O}$
DCGD-SHIFT	[**New**]	✗	s_i^0	$\mathcal{O}$
DGD		✓	s_i^0	$\mathcal{I}$
DCGD-STAR	[**New**]	✓	$\nabla f_i(x^\star)$	any $\mathcal{C}_i$
DIANA	[25]	✓	h_i^k	$\alpha\mathcal{Q}_i$
RAND-DIANA	[**New**]	✓	h_i^k	$\mathcal{B}e_{p_i}$
GDCI	[8]	✗	x^k/γ	$\mathcal{O}$

Appendix B. Basic Facts

Bregman Divergence associated with continuously-differentiable, strictly convex function f is defined as

$$D_f(x,y) \coloneqq f(x) - f(y) - \langle \nabla f(y), x - y\rangle.$$

Bregman Divergence and Strong Convexity inequality

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq D_f(x,y) + \frac{\mu}{2}\|x-y\|^2 \tag{7}$$

Bregman Divergence and L-smoothness inequality

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(x,y) \tag{8}$$

Basic Inequalities For all $a, b \in \mathbb{R}^d$

$$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \tag{9}$$

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2 \tag{10}$$

Appendix C. Proofs

For brevity we can use notation $\mathbf{E}\,\|x\|^2$ instead of $\mathbf{E}\left[\|x\|^2\right]$.

C.1. Shifted Compression

C.1.1. PROOF OF LEMMA 4

Proof For proof we need to show unbiasedness and (shifted) bounded variance property for operator $\mathcal{Q} \coloneqq v + \mathcal{Q}_h(x - v)$ with $\mathcal{Q}_h \in \mathbb{U}(\omega; h)$.
(a) Unbiasdeness:

$$\mathbf{E}\,\mathcal{Q}(x) = \mathbf{E}\left[v + \mathcal{Q}_h(x - v)\right] = v + \mathbf{E}\,\mathcal{Q}_h(x - v) = v + (x - v) = x.$$

(b) Variance:

$$\begin{aligned}\mathbf{E}\left\|\mathcal{Q}(x) - x\right\|^2 &= \mathbf{E}\left\|v + \mathcal{Q}_h(x - v) - x\right\|^2 \\ &= \mathbf{E}\left\|\mathcal{Q}_h(x - v) - (x - v)\right\|^2 \\ &\leq \omega\left\|x - v - h\right\|^2 = \omega\left\|x - (v + h)\right\|^2.\end{aligned}$$

■

C.1.2. COMPRESSION OF THE ITERATES

Lemma 7 *Let $\mathcal{Q} \in \mathbb{U}(\omega)$, then for*

$$\hat{\mathcal{Q}}(z) \coloneqq -t \cdot \mathcal{Q}\left(-\frac{z}{t}\right), \qquad t \neq 0 \tag{11}$$

we have $\hat{\mathcal{Q}}(z) \in \mathbb{U}(\omega)$.

Proof We consequentially show the unbiasedness (1) and bounded variance (2) properties according to Definition 2

1) $\mathbf{E}\,\hat{\mathcal{Q}}(z) = -t \cdot \mathbf{E}\,\mathcal{Q}\left(-\frac{z}{t}\right) = -t \cdot \left(-\frac{z}{t}\right) = z$;

2) $\mathbf{E}\left\|\hat{(z)}\right\|^2 = \mathbf{E}\left\|-t \cdot \mathcal{Q}\left(-\frac{z}{t}\right)\right\|^2 = t^2\,\mathbf{E}\left\|\mathcal{Q}\left(-\frac{z}{t}\right)\right\|^2 \leq t^2(\omega + 1)\left\|-\frac{z}{t}\right\|^2 = (\omega + 1)\|z\|^2$

■

Now we prove that shifted compressor $\tilde{\mathcal{Q}}(z) \coloneqq \frac{1}{\gamma}\left[x - \mathcal{Q}(x - \gamma z)\right]$, obtained from (11) by procedure described in Section D, belongs to class $\mathbb{U}\left(\omega; x/\gamma\right)$ for $\mathcal{Q} \in \mathbb{U}(\omega)$.
Proof 1) **Unbiasedness** follows from $\mathbf{E}\,\mathcal{Q}(x) = x$.

2) Computation of the **variance**:
Expectation conditional on x

$$\begin{aligned}
\mathbf{E}\left\|\tilde{\mathcal{Q}}(z)\right\|^2 &= \mathbf{E}\left\|\frac{1}{\gamma}\left[x-\mathcal{Q}(x-\gamma z)\right]\right\|^2 \\
&= \frac{1}{\gamma^2}\mathbf{E}\left\|x-\mathcal{Q}(x-\gamma z)\right\|^2 \\
&= \frac{1}{\gamma^2}\mathbf{E}\left\|x-\gamma z-\mathcal{Q}(x-\gamma z)+\gamma z\right\|^2 \\
&= \frac{1}{\gamma^2}\left[\mathbf{E}\left\|x-\gamma z-\mathcal{Q}(x-\gamma z)\right\|^2+\left\|\gamma z\right\|^2\right] \\
&\leq \frac{1}{\gamma^2}\left[\omega\left\|x-\gamma z\right\|^2+\left\|\gamma z\right\|^2\right] \\
&= \omega\left\|\frac{x}{\gamma}-z\right\|^2+\|z\|^2,
\end{aligned}$$

which implies

$$\mathbf{E}\left\|\tilde{\mathcal{Q}}(z)-z\right\|^2 \leq \omega\left\|z-\frac{x}{\gamma}\right\|^2.$$

■

C.1.3. INDUCED COMPRESSOR

Definition 8 (Induced Compression Operator [14]) *For $\mathcal{C}\in\mathbb{B}(\delta)$, choose $\mathcal{Q}\in\mathbb{U}(\omega)$ and define the induced compressor via*

$$\mathcal{C}_{ind}(x) \coloneqq \mathcal{C}(x)+\mathcal{Q}(x-\mathcal{C}(x)). \tag{12}$$

Lemma 9 *The induced operator satisfies $\mathcal{C}_{\rm ind}\in\mathbb{U}(\tilde{\omega})$ for*

$$\tilde{\omega} = \omega(1-\delta) \tag{13}$$

Proof

$$\begin{aligned}
\mathbf{E}\left\|\mathcal{C}_{\rm ind}(x)-x\right\|^2 &= \mathbf{E}\left\|\mathcal{Q}(x-\mathcal{C}(x))-(x-\mathcal{C}(x))\right\|^2 \\
&\leq \omega\,\mathbf{E}\left\|x-\mathcal{C}(x)\right\|^2 \\
&\leq \underbrace{\omega(1-\delta)}_{\tilde{\omega}}\left\|x\right\|^2
\end{aligned}$$

■

C.2. Proof of Theorem 5 (DCGD-SHIFT)

According to Algorithm 1 gradient estimator always has the following form

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

Obvious unbiasedness (as $\mathcal{Q}_i \in \mathbb{U}(\omega_i)$) of this estimator for any h_i will be used in all further proofs.

Decomposition

$$\mathbf{E}\left\|g_h(x^k) - \nabla f(x^\star)\right\|^2 = \mathbf{E}\left\|g_h(x^k) - \nabla f(x^k)\right\|^2 + \left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2 \tag{14}$$

Next we upper-bound the first term from (14).

Expectation conditional on x^k and $h = (h_1, \ldots, h_n)$ for brevity

$$\begin{aligned}
\mathbf{E}\left\|g_h(x^k) - \nabla f(x^k)\right\|^2 &= \mathbf{E}\left\|g_h(x^k) - \nabla f(x^k)\right\|^2 \\
&\le \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \underbrace{\mathcal{Q}_i\left(\nabla f_i(x^k) - h_i\right)}_{b_i^k} + h_i - \nabla f_i(x^k)\right\|^2 \\
&= \frac{1}{n^2}\mathbf{E}\left[\sum_{i=1}^n \|b_i^k\|^2 + \sum_{i\neq j}\langle b_i^k, b_j^k\rangle\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\,\|b_i^k\|^2 + \frac{1}{n^2}\sum_{i\neq j}\underbrace{\langle \mathbf{E}\, b_i^k, \mathbf{E}\, b_j^k\rangle}_{=0\ (\mathbf{E}\,\mathcal{Q}_i(x)=x)} \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\,\|\mathcal{Q}_i\left(\nabla f_i(x^k) - h_i\right) + h_i - \nabla f_i(x^k)\|^2 \\
&\overset{(2)}{\le} \frac{1}{n^2}\sum_{i=1}^n \omega_i \|\nabla f_i(x^k) - h_i\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n \omega_i \|\nabla f_i(x^k) - \nabla f_i(x^\star) - (h_i - \nabla f_i(x^\star))\,\|^2 \\
&\overset{(9)}{\le} \frac{2}{n^2}\sum_{i=1}^n \omega_i \left[\left\|\nabla f_i(x^k) - \nabla f_i(x^\star)\right\|^2 + \|h_i - \nabla f_i(x^\star)\|^2\right] \\
&\overset{(8)}{\le} \frac{2}{n^2}\sum_{i=1}^n \omega_i \cdot 2L_i D_{f_i}(x^k, x^\star) + \frac{2}{n^2}\sum_{i=1}^n \omega_i \left\|h_i - \nabla f_i(x^\star)\right\|^2 \\
&\le \frac{4}{n}\max\left(\omega_i L_i\right)\frac{1}{n}\sum_{i=1}^n D_{f_i}(x^k, x^\star) + \frac{2}{n^2}\sum_{i=1}^n \omega_i \left\|h_i - \nabla f_i(x^\star)\right\|^2 \\
&\le \frac{4}{n}\max\left(\omega_i L_i\right) D_f(x^k, x^\star) + \frac{2}{n^2}\sum_{i=1}^n \omega_i \left\|h_i - \nabla f_i(x^\star)\right\|^2.
\end{aligned}$$

Combined with (14) we obtain

$$\mathbf{E}\left[\left\|g_h(x^k)-\nabla f(x^\star)\right\|^2 \mid x^k,h\right] \leq 2\left[\frac{2}{n}\max\left(\omega_i L_i\right)+L\right] D_f(x^k,x^\star)+\frac{2}{n^2}\sum_{i=1}^n \omega_i \left\|h_i-\nabla f_i(x^\star)\right\|^2$$

Now from [11, Theorem 4.1] we get statement of Theorem 11.

Next subsection we start by introducing practically useless, but theoretically insightful DCGD-SHIFT* and then move on to implementable Algorithms.

C.3. Optimal shifts

Assume, for the sake of argument, that we know the values $\nabla f_i(x^\star)$ for every $i \in [n]$. Then we can construct optimally shifted compressed shift updates sequence using the form (3)

$$h_i^{k+1} = \nabla f_i(x^\star) + \mathcal{C}_i(\nabla f_i(x^k) - \nabla f_i(x^\star)). \tag{15}$$

This is enough to fully characterize the Algorithm 1 and obtain the following convergence guarantee.

Theorem 10 (DCGD-SHIFT*/STAR) *Assume each f_i is convex and L_i-smooth, and f is L-smooth and μ-strongly convex. Let $\mathcal{Q}_i \in \mathbb{U}(\omega_i), \mathcal{C}_i \in \mathbb{U}(\delta_i)$ - independent compression operators. If the step-size satisfies*

$$\gamma \leq 1/\left[L+\max_i\left(L_i\omega_i(1-\delta_i)/n\right)\right], \tag{16}$$

Then the iterates of DCGD *with* ***optimally shifted compressed shift*** *update* (15) *satisfy*

$$\mathbf{E}\left\|x^k-x^\star\right\|^2 \leq (1-\gamma\mu)^k\|x^0-x^\star\|^2.$$

This is the first presented Algorithm with linear convergence to the exact solution for the general not-overparameterized case. Notice that for zero-identity operators $\mathcal{C}_i \equiv 0$ we obtain the simplest optimal shift $h_i = \nabla f_i(x^\star)$ and term δ_i in (16) should be interpreted as zero.

The issue with the described method is that in general, we do not know the values $h_i^\star := \nabla f_i(x^\star)$, which makes it impractical.

C.3.1. PROOF OF THEOREM 10 (DCGD-STAR)

Proof At first compute the **variance of the estimator**.

Expectation conditional on x^k and $h^k = (h_1^k, \ldots, h_n^k)$ for brevity

$$
\begin{aligned}
\mathbf{Var}[g^k] &= \mathbf{E}\left\|g^k - \mathbf{E}\left[g^k\right]\right\|^2 \\
&= \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \underbrace{\mathcal{Q}_i\left(\nabla f_i(x^k) - h_i^k\right)}_{b_i^k} + h_i^k - \nabla f_i(x^k)\right\|^2 \\
&= \frac{1}{n^2}\mathbf{E}\left[\sum_{i=1}^n \|b_i^k\|^2 + \sum_{i\neq j}\langle b_i^k, b_j^k\rangle\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\,\|b_i^k\|^2 + \frac{1}{n^2}\sum_{i\neq j}\underbrace{\langle \mathbf{E}\, b_i^k, \mathbf{E}\, b_j^k\rangle}_{=0\ (\mathbf{E}\,\mathcal{Q}_i(x)=x)} \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\,\|\mathcal{Q}_i\left(\nabla f_i(x^k) - h_i^k\right) + h_i^k - \nabla f_i(x^k)\|^2 \\
&\overset{(2)}{\leq} \frac{1}{n^2}\sum_{i=1}^n \omega_i\,\mathbf{E}\,\|\nabla f_i(x^k) - h_i^k\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n \omega_i\,\mathbf{E}\,\|\nabla f_i(x^k) - h_i^\star - \mathcal{C}_i\left(\nabla f_i(x^k) - h_i^\star\right)\|^2 \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \omega_i(1-\delta_i)\|\nabla f_i(x^k) - \nabla f_i(x^\star)\|^2 \\
&\overset{(8)}{\leq} \frac{1}{n^2}\sum_{i=1}^n \omega_i(1-\delta_i)\cdot 2L_i D_{f_i}(x^k, x^\star) \\
&\leq \frac{2}{n}\max\{\omega_i(1-\delta_i)L_i\}\frac{1}{n}\sum_{i=1}^n D_{f_i}(x^k, x^\star) \\
&\leq \frac{2}{n}\max\{\omega_i(1-\delta_i)L_i\}D_f(x^k, x^\star).
\end{aligned}
\tag{17}
$$

Now we can move to **convergence proof**.
Expectation conditional on x^k and h^k

$$
\begin{aligned}
\mathbf{E}\,\|r^{k+1}\|^2 &= \mathbf{E}\,\|x^{k+1} - x^\star\|^2 \\
&= \mathbf{E}\,\|x^k - \gamma g^k - (x^\star - \gamma\nabla f(x^\star))\|^2 \\
&= \mathbf{E}\,\|x^k - x^\star - \gamma(g^k - \nabla f(x^\star))\|^2 \\
&= \|r^k\|^2 - 2\gamma\langle x^k - x^\star, \mathbf{E}\, g^k - \nabla f(x^\star)\rangle + \gamma^2\,\mathbf{E}\left\|g^k - \nabla f(x^\star))\right\|^2 \\
&= \|r^k\|^2 - 2\gamma\langle x^k - x^\star, \nabla f(x^k) - \nabla f(x^\star)\rangle \\
&\quad + \gamma^2\left[\mathbf{E}\left\|g^k - \nabla f(x^k))\right\|^2 + \left\|\nabla f(x^k) - \nabla f(x^\star))\right\|^2\right] \\
&\overset{(7)}{\leq} \|r^k\|^2 - 2\gamma\left[D_f(x^k, x^\star) + \frac{\mu}{2}\|x^k - x^\star\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\gamma^2\left[\mathbf{Var}[g^k]+\left\|\nabla f(x^k)-\nabla f(x^\star))\right\|^2\right]\\
&\overset{(17)}{\leq}(1-\gamma\mu)\|r^k\|^2-2\gamma D_f(x^k,x^\star)\\
&\quad+\gamma^2\left[\frac{2}{n}\max\{\omega_i(1-\delta_i)L_i\}D_f(x^k,x^\star)+2LD_f(x^k,x^\star)\right]\\
&=(1-\gamma\mu)\|r^k\|^2-2\gamma\left[1-\gamma\left(L+\max\{L_i\omega_i(1-\delta_i)/n\}\right)\right]D_f(x^k,x^\star). \qquad (18)
\end{aligned}
$$

By choosing step-size

$$\gamma\leq\frac{1}{L+\max_i\left(L_i\omega_i(1-\delta_i)/n\right)},$$

it is guaranteed that the second term in (18) is positive. Therefore,

$$
\begin{aligned}
\mathbf{E}\left\|x^{k+1}-x^\star\right\|^2&=\mathbf{E}\left[\mathbf{E}\left[\left\|x^{k+1}-x^\star\right\|^2\mid x^k,h^k\right]\right]\\
&\leq(1-\gamma\mu)\,\mathbf{E}\left\|x^k-x^\star\right\|^2\\
&\leq(1-\gamma\mu)^{k+1}\,\mathbf{E}\left\|x^0-x^\star\right\|^2,
\end{aligned}
$$

which concludes the proof. ■

C.4. DIANA-like trick

Next we present approach to issue raised in 3.1 based on the celebrated DIANA [16, 25] Algorithm:

$$h_i^{k+1}=h_i^k+\alpha\left[\mathcal{C}_i(\nabla f_i(x^k)-h_i^k)+\mathcal{Q}_i\left(\nabla f_i(x^k)-h_i^k-\mathcal{C}_i(\nabla f_i(x^k)-h_i^k)\right)\right], \qquad (19)$$

where α is a suitably chosen step-size. For $\mathcal{C}_i\equiv 0$ it takes the simplified form

$$h_i^{k+1}=h_i^k+\alpha\mathcal{Q}_i\left(\nabla f_i(x^k)-h_i^k\right). \qquad (20)$$

This recursion resolves both of the earlier raised issues. Firstly, this sequence of h_i^k indeed converges to the optimal shifts $\nabla f_i(x^\star)$, which is formalized in the Theorem 11 presented later. Besides, shift on the master $h^{k+1}=\frac{1}{n}\sum_{i=1}^n h_i^{k+1}$ is updated this way

$$
\begin{aligned}
h^{k+1}&=\frac{1}{n}\sum_{i=1}^n\left\{h_i^k+\alpha\left[\mathcal{C}_i(\nabla f_i(x^k)-h_i^k)+\mathcal{Q}_i\left(\nabla f_i(x^k)-h_i^k-\mathcal{C}_i(\nabla f_i(x^k)-h_i^k)\right)\right]\right\}\\
&=\frac{1}{n}\sum_{i=1}^n h_i^k+\alpha\frac{1}{n}\sum_{i=1}^n\left\{c_i^k+m_i^k\right\}=h^k+\alpha\left(c^k+m^k\right),
\end{aligned}
$$

which requires aggregation of only compressed vectors $c_i^k:=\mathcal{C}_i(\nabla f_i(x^k)-h_i^k)$ and $m_i^k:=\mathcal{Q}_i\left(\nabla f_i(x^k)-h_i^k-c_i^k\right)$ from the workers. In case of update (20) it is even not needed to send anything in addition to messages m_i^k required by default in DCGD-SHIFT (1).

Furthermore, simplified recursion (20) can be interpreted as 1 step of Compressed Gradient Descent (CGD) with stepsize α applied to such optimization problem:

$$\max_{h_i\in\mathbb{R}^d}\left[\phi_i^k(h_i):=-\frac{1}{2}\left\|h_i-\nabla f_i(x^k)\right\|^2\right],$$

which is in fact a 1-smooth and 1-strongly concave function. In this way h_i^{k+1} keeps track of the latest local gradient and produces a better estimate than the previous shift h_i^k offered.

Now we present the convergence result for the Algorithm 1 with described shift learning procedure.

Theorem 11 (DIANA) *Assume each f_i is convex and L_i-smooth, and f is μ-strongly convex. Let $\mathcal{Q}_i \in \mathbb{U}(\omega_i), \mathcal{C}_i \in \mathbb{U}(\delta_i)$ - independent compression operators. If the step-sizes satisfy*

$$\alpha \leq \frac{1}{1+\omega_i(1-\delta_i)} \text{ (for all } i), \qquad \gamma \leq \frac{1}{\frac{2}{n}\max_i(\omega_i L_i) + (1+\alpha M)L_{\max}},$$

where $L_{\max} := \max_i L_i$, $M > 2/(n\alpha)$ and δ_i should be interpreted as zero for $\mathcal{C}_i \equiv 0$. Then the iterates of DCGD *with* DIANA*-like shift update (19) satisfy*

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

where the Lyapunov function V^k is defined by

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

Our result improves over original DIANA in several ways. Firstly, it uses much more general shift updates involving $\mathcal{C}_i$, which allows using biased operators for learning the optimal shifts. Secondly, one can use different compressors $\mathcal{Q}_i$, which can be particularly beneficial when different workers have various bandwidths/connection speeds to the master. So, the slower processors can compress more, and therefore use operators with higher ω_i. At the same, time the opposite makes sense for "faster" workers.

C.4.1. PROOF OF THEOREM 11 (DIANA-LIKE)

DIANA-like shift update has the following form

$$h_i^{k+1} = h_i^k + \alpha\left[\mathcal{C}_i(\nabla f_i(x^k) - h_i^k) + \mathcal{Q}_i\left(\nabla f_i(x^k) - h_i^k - \mathcal{C}_i(\nabla f_i(x^k) - h_i^k)\right)\right],$$

which is in fact equivalent to the standard DIANA shift update with induced compressor $\mathcal{Q}_i^{\text{ind}} \in \mathbb{U}\left(\omega_i(1-\delta_i)\right)$ defined before (8):

$$h_i^{k+1} = h_i^k + \alpha\mathcal{Q}_i^{\text{ind}}\left(\nabla f_i(x^k) - h_i^k\right),$$

The proof of Theorem 11 is mainly a generalization of the original DIANA analysis. Our approach is based on [11, Theorem 4.1], which requires the following Lemma (referred as Assumption 4.1 in [11]).

Lemma 12 *Assume that functions f_i are convex and L_i-smooth for all i, and $\mathcal{Q}_i \in \mathbb{U}(\omega_i), \mathcal{C}_i \in \mathbb{B}(\delta_i)$. Let $h = (h_1, h_2, \ldots, h_n) \in \mathbb{R}^d \times \mathbb{R}^d \ldots \times \mathbb{R}^d = \mathbb{R}^{nd}$ and define $\sigma : \mathbb{R}^{nd} \to [0,\infty)$ and σ^k by*

$$\sigma(h) = \frac{1}{n}\sum_{i=1}^n \omega_i \left\|h_i - \nabla f_i(x^\star)\right\|^2, \qquad \sigma^k := \sigma(h^k) = \frac{1}{n}\sum_{i=1}^n \omega_i \left\|h_i^k - \nabla f_i(x^\star)\right\|^2.$$

Then for all iterations of DCGD *with* DIANA*-like shift update we have*

$$\begin{aligned}
\textbf{\textit{1)}}\ \mathbf{E}\left[g^k \mid x^k, h^k\right] &= \nabla f(x^k), \\
\textbf{\textit{2)}}\ \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, h^k\right] &\leq 2\left(2\max_i(L_i\omega_i/n) + L_{\max}\right) D_f(x^k, x^\star) + \frac{2}{n}\sigma^k, \\
\textbf{\textit{3)}}\ \mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &\leq (1-\alpha)\sigma^k + 2\alpha\max_i(L_i\omega_i) D_f(x^k, x^\star).
\end{aligned}$$

Proof The Lemma consists of three points.

1) Unbiasedness of the shifted gradient estimator was already shown in C.2.

2) Expected smoothness:

$$\begin{aligned}
G^k &:= \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, h^k\right] \\
&= \mathbf{E}\left[\|g^k - \nabla f(x^k)\|^2 \mid x^k, h^k\right] + \left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2 \\
&\leq \mathbf{E}\left[\|g^k - \nabla f(x^k)\|^2 \mid x^k, h^k\right] + 2LD_f(x^k, x^\star) \\
&= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^k + \mathcal{Q}_i\left(\nabla f_i(x^k) - h_i^k\right) - \nabla f_i(x^k)\right)\right\|^2 \mid x^k, h^k\right] + 2LD_f(x^k, x^\star) \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\left\|\mathcal{Q}_i\left(\nabla f_i(x^k) - h_i^k\right) - (\nabla f_i(x^k) - h_i^k)\right\|^2 \mid x^k, h^k\right] + 2LD_f(x^k, x^\star) \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \omega_i \left\|\nabla f_i(x^k) - h_i^k)\right\|^2 + 2LD_f(x^k, x^\star) \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \omega_i \left[2\left\|\nabla f_i(x^k) - \nabla f_i(x^\star))\right\|^2 + 2\left\|\nabla f_i(x^\star) - h_i^k)\right\|^2\right] + 2LD_f(x^k, x^\star) \\
&\leq 2\frac{2}{n}\max(\omega_i L_i)\frac{1}{n}\sum_{i=1}^n D_{f_i}(x^k, x^\star) + \frac{2}{n}\frac{1}{n}\sum_{i=1}^n \omega_i \left\|\nabla f_i(x^\star) - h_i^k)\right\|^2 + 2LD_f(x^k, x^\star) \\
&\leq 2\left[\frac{2}{n}\max(\omega_i L_i) + L_{\max}\right] D_f(x^k, x^\star) + \frac{2}{n}\sigma^k.
\end{aligned}$$

Denote $m_i^k := \mathcal{Q}_i^{\text{ind}}\left(\nabla f_i(x^k) - h_i^k\right)$.

3) Sigma-k recursion:

$$\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n \omega_i \left\|h_i^{k+1} - \nabla f_i(x^\star)\right\|^2 \mid x^k, h^k\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n \omega_i \left(\left\|h_i^k - h_i^\star\right\|^2 + 2\alpha\left\langle\nabla f_i(x^k) - h_i^k, h_i^k - h_i^\star\right\rangle + \alpha^2\,\mathbf{E}\left[\left\|m_i^k\right\|^2 \mid x^k, h^k\right]\right) \\
&\leq \sigma^k + \frac{1}{n}\sum_{i=1}^n \omega_i \left(2\alpha\left\langle\nabla f_i(x^k) - h_i^k, h_i^k - h_i^\star\right\rangle + \alpha^2\left(1 + \omega_i(1-\delta_i)\right)\left\|\nabla f_i(x^k) - h_i^k\right\|^2\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sigma^k + \frac{1}{n}\sum_{i=1}^n \omega_i \alpha \left(2\left\langle \nabla f_i(x^k) - h_i^k, h_i^k - h_i^\star \right\rangle + \left\|\nabla f_i(x^k) - h_i^k\right\|^2\right) \\
&= \sigma^k + \frac{1}{n}\sum_{i=1}^n \alpha\omega_i \left(\left\|\nabla f_i(x^k) - h_i^\star\right\|^2 - \left\|h_i^k - h_i^\star\right\|^2\right) \\
&= \frac{1}{n}\sum_{i=1}^n (1-\alpha)\sigma^k + \frac{1}{n}\sum_{i=1}^n \alpha\omega_i \left\|\nabla f_i(x^k) - \nabla f_i(x^\star)\right\|^2 \\
&\leq (1-\alpha)\sigma^k + \frac{1}{n}\sum_{i=1}^n \alpha 2\omega_i L_i D_{f_i}(x^k, x^\star) \\
&\leq (1-\alpha)\sigma^k + 2\max_i(L_i\omega_i) D_f(x^k, x^\star),
\end{aligned}
$$

which concludes the proof. ■

Now by direct application of [11, Theorem 4.1] we get statement of Theorem 11.

C.5. Proof of Theorem 6 (Randomized-DIANA)

In short Rand-DIANA is defined by the shift update

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

which is similar to gradient estimator structure of Loopless-SVRG [21].

The proof of Theorem 6 is also based on [11, Theorem 4.1], which requires a modified version of Lemma 12.

Lemma 13 *Assume that functions f_i are convex and L_i-smooth for all i, and $\mathcal{Q}_i \in \mathbb{U}(\omega)$ for all i. Let $h = (h_1, h_2, \ldots, h_n) \in \mathbb{R}^d \times \mathbb{R}^d \ldots \times \mathbb{R}^d = \mathbb{R}^{nd}$ and define $\sigma : \mathbb{R}^{nd} \to [0, \infty)$ and σ^k by*

$$
\sigma(h) = \frac{1}{n}\sum_{i=1}^n \|h_i - \nabla f_i(x^\star)\|^2 \quad \sigma^k := \sigma(h^k) = \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^\star)\right\|^2.
$$

Then for all iterations of Rand-DIANA *we have*

$$
\begin{aligned}
\textbf{1)}\ \mathbf{E}\left[g^k \mid x^k, h^k\right] &= \nabla f(x^k), \\
\textbf{2)}\ \mathbf{E}\left[\left\|g^k - \nabla f(x^\star)\right\|^2 \mid x^k, h^k\right] &\leq 2\left(\frac{2\omega}{n} + 1\right) L_{\max} D_f(x^k, x^\star) + \frac{2\omega}{n}\sigma^k, \\
\textbf{3)}\ \mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &\leq 2\max_i(p_i L_i) D_f(x^k, x^\star) + (1 - \min_i p_i)\sigma^k.
\end{aligned}
$$

Proof **1) Unbiasedness** of the shifted gradient estimator was already shown in C.2.

2) Expected smoothness: Exactly the same as in Lemma 12 with simplified σ^k (without ω_i) and $\omega_i \equiv \omega$.

3) Sigma-k recursion:

$$\begin{aligned}
\mathbf{E}\left[\sigma^{k+1} \mid x^k, h^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|h_i^{k+1} - h_i^\star\|^2 \mid x^k, h^k\right] \\
&= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(w_i^{k+1}) - \nabla f_i(x^\star)\|^2 \mid x^k, h^k\right] \\
&= \frac{1}{n}\sum_{i=1}^n \left[p_i\|\nabla f_i(x^k) - \nabla f_i(x^\star)\|^2 + (1-p_i)\|\nabla f(w_i^k) - \nabla f_i(x^\star)\|^2\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n p_i \cdot 2L_i D_{f_i}(x_k, x^\star) + \max_i(1-p_i)\frac{1}{n}\sum_{i=1}^n \|\nabla f(w_i^k) - \nabla f_i(x^\star)\|^2 \\
&\leq 2\max_i(p_i L_i) D_f(x_k, x^\star) + (1 - \min_i p_i)\sigma^k
\end{aligned}$$

■

Appendix D. Compressing the iterates

In this section, we discuss how the shifted compression framework can be applied to the case where the iterates themselves need to be compressed and lead to the improved results.

Let $\mathcal{Q} \in \mathbb{U}(\omega)$. Consider such shifted by vector x/γ compressor

$$\hat{\mathcal{Q}}(z) \coloneqq \frac{x}{\gamma} + \mathcal{Q}\left(z - \frac{x}{\gamma}\right),$$

which clearly belongs to the class $\mathbb{U}(\omega; x/\gamma)$.

By using the fact that compressor $\bar{\mathcal{Q}}(z) \coloneqq -\frac{1}{\gamma}\cdot\mathcal{Q}(-\gamma z) \in \mathbb{U}(\omega)$ (for $\gamma \neq 0$) we can transform $\hat{\mathcal{Q}}$ to operator

$$\tilde{\mathcal{Q}}(z) \coloneqq \frac{x}{\gamma} + \bar{\mathcal{Q}}\left(z - \frac{x}{\gamma}\right) = \frac{x}{\gamma} - \frac{1}{\gamma}\cdot\mathcal{Q}\left(-\gamma\left[z - \frac{x}{\gamma}\right]\right) = \frac{1}{\gamma}\left[x - \mathcal{Q}(x - \gamma z)\right],$$

which also belongs to $\mathbb{U}(\omega; x/\gamma)$ and it is helpful for analysing algorithms with compressed iterates.

D.1. Distributed Gradient Descent with Compressed Iterates (GDCI)

GDCI was at first analyzed in [17] for single node and in short, was formulated in the relaxed form

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

by [8]. This algorithm can be reformulated using previously described shifted compressor $\tilde{\mathcal{Q}}$

$$x^{k+1} = x^k - (\eta\gamma)\frac{1}{\gamma}\left[x^k - \mathcal{Q}\left(x^k - \gamma\nabla f(x^k)\right)\right] = x^k - (\eta\gamma)\tilde{\mathcal{Q}}^k(\nabla f(x^k)),$$

which for distributed case takes the form

$$x^{k+1} = x^k - (\eta\gamma)\frac{1}{\gamma}\left[x^k - \frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i\left(x^k - \gamma\nabla f_i(x^k)\right)\right]. \tag{21}$$

The essence of this method is compression of the local workers' iterates $\mathcal{Q}_i\left(x^k - \gamma\nabla f_i(x^k)\right)$ and aggregation on the master (21). Established linear convergence up to a neighborhood introduced due to variance of compression operator (similarly to DCGD with fixed shifts result 5) is presented next.

Theorem 14 (GDCI) *Assume each f_i is convex and L_i-smooth, and f is L-smooth and μ-strongly convex. Let $\mathcal{Q}_i \in \mathbb{U}(\omega)$ - independent compression operators. If the step-sizes satisfy*

$$\eta \leq \left[\frac{L}{\mu} + \frac{2\omega}{n}\left(\frac{L_{\max}}{\mu} - 1\right)\right]^{-1}, \qquad \gamma \leq \frac{1 + 2\eta\omega/n}{\eta\left(L + 2L_{\max}\omega/n\right)}$$

then the iterates of Distributed GDCI (21) *satisfy*

$$\mathbf{E}\left\|x^k - x^\star\right\|^2 \leq (1-\eta)^k\|x^0 - x^\star\|^2 + \eta\frac{2\omega}{n}\frac{1}{n}\sum_{i=1}^{n}\|x^\star - \gamma\nabla f_i(x^\star)\|^2. \tag{22}$$

In the interpolation regime ($\nabla f_i(x^\star) = 0 = x^\star - \gamma\nabla f_i(x^\star)$ for every i) this result matches the complexity of DCGD with fixed shifts (5)

$$\tilde{\mathcal{O}}\left(\kappa\left(1 + \omega/n\right)\right)$$

and improves over the original rate of GDCI from [8] analyzed for fixed point problems and specialized for gradient mappings:

$$\tilde{\mathcal{O}}\left(\kappa\max\left\{1, \kappa\omega/n\right\}\right) \leq \tilde{\mathcal{O}}\left(\kappa^2\omega/n\right)$$

D.1.1. PROOF OF THEOREM 14 (GDCI)

This and further sections rely on Subsection C.1.2.

Distributed Gradient Descent with Compressed Iterates (GDCI) has the form

$$x^{k+1} = x^k - (\eta\gamma)\frac{1}{\gamma}\left[x^k - \frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i\left(x^k - \gamma\nabla f_i(x^k)\right)\right] = x^k - (\eta\gamma)\tilde{\mathcal{Q}}^k(\nabla f(x^k)).$$

where shifted compressor $\tilde{\mathcal{Q}}^k(\nabla f(x^k))$ belongs to class $\mathbb{U}\left(\omega;\, x^k/\gamma\right)$ for $\mathcal{Q}_i \in \mathbb{U}(\omega)$

Convergence analysis for the non-regularized case ($\nabla f(x^\star) = 0$).[4]

[4] Can be easily generalized to a proximal setup as the previous parts.

Proof Expectation conditional on x^k

$$\begin{aligned}
\mathbf{E}\left\|r^{k+1}\right\|^2 &:= \mathbf{E}\left\|x^{k+1}-x^\star\right\|^2 \\
&= \mathbf{E}\left\|x^k-\eta\gamma\tilde{\mathcal{Q}}^k(\nabla f(x^k))-x^\star\right\|^2 \\
&= \mathbf{E}\left\|x^k-\eta\gamma\tilde{\mathcal{Q}}^k(\nabla f(x^k))-(x^\star-\eta\gamma\nabla f(x^\star))\right\|^2 \\
&= \left\|r^k\right\|^2+(\eta\gamma)^2\,\mathbf{E}\left\|\tilde{\mathcal{Q}}^k(\nabla f(x^k))-\nabla f(x^\star)\right\|^2-2\eta\gamma\langle x^k-x^\star,\nabla f(x^k)-\nabla f(x^\star)\rangle \\
&= \left\|r^k\right\|^2+(\eta\gamma)^2\left[\mathbf{E}\left\|\tilde{\mathcal{Q}}^k(\nabla f(x^k))-\nabla f(x^k)\right\|^2+\left\|\nabla f(x^k)-\nabla f(x^\star)\right\|^2\right] \\
&\quad -2\eta\gamma\langle x^k-x^\star,\nabla f(x^k)-\nabla f(x^\star)\rangle.
\end{aligned} \tag{23}$$

Next we upper-bound the variance of $\tilde{\mathcal{Q}}^k(\nabla f(x^k))$

$$\begin{aligned}
\mathbf{E}\left\|\tilde{\mathcal{Q}}^k(\nabla f(x^k))-\nabla f(x^k)\right\|^2 &= \mathbf{E}\left\|\frac{1}{\gamma}\left[x^k-\frac{1}{n}\sum_{i=1}^n\mathcal{Q}_i\left(x^k-\gamma\nabla f_i(x^k)\right)\right]-\frac{1}{n}\sum_{i=1}^n\nabla f_i(x^k)\right\|^2 \\
&= \mathbf{E}\left\|\frac{1}{n^2}\sum_{i=1}^n\left(\nabla f_i(x^k)-\frac{1}{\gamma}\left[x^k-\mathcal{Q}_i\left(x^k-\gamma\nabla f_i(x^k)\right)\right]\right)\right\|^2 \\
[\text{independence of }\mathcal{Q}_i] &\le \frac{1}{n^2}\sum_{i=1}^n\mathbf{E}\left\|\nabla f_i(x^k)-\frac{1}{\gamma}\left[x^k-\mathcal{Q}_i\left(x^k-\gamma\nabla f_i(x^k)\right)\right]\right\|^2 \\
&\le \frac{1}{n^2}\sum_{i=1}^n\omega\left\|\nabla f_i(x^k)-\frac{1}{\gamma}x^k\right\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n\frac{\omega}{\gamma^2}\left\|x^k-\gamma\nabla f_i(x^k)\right\|^2 \\
&= \frac{\omega}{n^2\gamma^2}\sum_{i=1}^n\left\|x^k-\gamma\nabla f_i(x^k)\pm(x^\star-\gamma\nabla f_i(x^\star))\right\|^2 \\
&\le \frac{\omega}{n^2\gamma^2}\sum_{i=1}^n 2\left[\left\|x^k-\gamma\nabla f_i(x^k)-(x^\star-\gamma\nabla f_i(x^\star))\right\|^2+\left\|x^\star-\gamma\nabla f_i(x^\star)\right\|^2\right] \\
&= \frac{2\omega}{n^2\gamma^2}\sum_{i=1}^n\Big[\left\|x^k-x^\star\right\|^2-2\gamma\langle x^k-x^\star,\nabla f_i(x^k)-\nabla f_i(x^\star)\rangle \\
&\quad +\gamma^2\left\|\nabla f_i(x^k)-\nabla f_i(x^\star)\right\|^2+\left\|x^\star-\gamma\nabla f_i(x^\star)\right\|^2\Big] \\
&\le \frac{2\omega}{n^2\gamma^2}\sum_{i=1}^n\Big[\left\|r^k\right\|^2-2\gamma\left(D_{f_i}(x^k,x^\star)+\frac{\mu}{2}\|x^k-x^\star\|^2\right) \\
&\quad +\gamma^2 2L_iD_{f_i}(x^k,x^\star)+\left\|x^\star-\gamma\nabla f_i(x^\star)\right\|^2\Big] \\
&= \frac{2\omega}{n^2\gamma^2}\sum_{i=1}^n\left[(1-\gamma\mu)\left\|r^k\right\|^2-2\gamma\left(1-\gamma L_i\right)D_{f_i}(x^k,x^\star)+\left\|x^\star-\gamma\nabla f_i(x^\star)\right\|^2\right]
\end{aligned}$$

$$\leq \frac{2\omega}{n\gamma^2}(1-\gamma\mu)\left\|r^k\right\|^2 + \frac{2\omega}{n^2\gamma^2}\sum_{i=1}^n \|x^\star - \gamma\nabla f_i(x^\star)\|^2$$
$$- \frac{2\omega}{n\gamma^2}\cdot 2\gamma\left(1-\gamma L_{\max}\right)D_f(x^k,x^\star).$$

Combining it with (23) and using notation $\mathcal{T}_i(x) \coloneqq x - \gamma\nabla f_i(x)$ we get

$$\begin{aligned}
\mathbf{E}\,\|r^{k+1}\|^2 &\leq \|r^k\|^2 - 2\eta\gamma\langle x^k - x^\star, \nabla f(x^k) - \nabla f(x^\star)\rangle + (\eta\gamma)^2\left\|\nabla f(x^k) - \nabla f(x^\star)\right\|^2 \\
&\quad + \frac{2\omega\eta^2}{n}\left[(1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma\left(1-\gamma L_{\max}\right)D_f(x^k,x^\star) + \frac{1}{n}\sum_{i=1}^n\|\mathcal{T}_i(x^\star)\|^2\right] \\
&\leq \|r^k\|^2 - 2\eta\gamma\left[D_f(x^k,x^\star) + \frac{\mu}{2}\|x^k - x^\star\|^2\right] + (\eta\gamma)^2\cdot 2LD_f(x^k,x^\star) \\
&\quad + \frac{2\omega\eta^2}{n}\left[(1-\gamma\mu)\left\|r^k\right\|^2 - 2\gamma\left(1-\gamma L_{\max}\right)D_f(x^k,x^\star) + \frac{1}{n}\sum_{i=1}^n\|\mathcal{T}_i(x^\star)\|^2\right] \\
&\leq \left[1 - \eta\gamma\mu + \frac{2\omega\eta^2}{n}(1-\gamma\mu)\right]\|r^k\|^2 + \frac{2\omega\eta^2}{n}\frac{1}{n}\sum_{i=1}^n\|\mathcal{T}_i(x^\star)\|^2 \\
&\quad - 2\eta\gamma\left[1 - \eta\gamma L + \frac{2\omega\eta}{n}\left(1-\gamma L_{\max}\right)\right]D_f(x^k,x^\star),
\end{aligned}$$

which after choosing the step-size

$$\gamma \leq \frac{1 + 2\eta\omega/n}{\eta\left(L + 2L_{\max}\omega/n\right)}$$

and after optimizing over γ and η (to maximize contraction term before $\|r^k\|^2$) leads to

$$\mathbf{E}\left[\|x^{k+1} - x^\star\|^2 \mid x^k\right] \leq (1-\eta)\|r^k\|^2 + \frac{2\omega\eta^2}{n}\frac{1}{n}\sum_{i=1}^n\|\mathcal{T}_i(x^\star)\|^2,$$

for

$$\eta \leq \left[\frac{L}{\mu} + \frac{2\omega}{n}\left(\frac{L_{\max}}{\mu} - 1\right)\right]^{-1},$$

And after unrolling the recursion and standard simplifications we obtain the desired result

$$\mathbf{E}\,\|x^k - x^\star\|^2 \leq (1-\eta)^k\left\|x^0 - x^\star\right\|^2 + \frac{2\omega\eta}{n}\frac{1}{n}\sum_{i=1}^n\|x^\star - \gamma\nabla f_i(x^\star)\|^2.$$

■

D.2. Variance-Reduced Gradient Descent with Compressed Iterates (VR-GDCI)

Algorithm 2: Variance-Reduced Gradient Descent with Compressed Iterates (VR-GDCI)

Input: learning rates $\alpha, \gamma, \eta > 0$; compressors $\mathcal{Q}_i$, initial iterate $x^0 \in \mathbb{R}^d$, initial local shifts $h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$ (stored on the n nodes)
Initialize: $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$ (stored on the master)
for $k = 0, 1, 2 \ldots$ **do**
  Broadcast x^k to all workers
  for $i = 1, \ldots n$ **do**
    Compute local gradient: $\nabla f_i(x^k)$
    Compress shifted local model $\delta_i^{k+1} = \mathcal{Q}_i\left(x^k - \gamma \nabla f_i(x^k) - h_i^k\right)$
    Update the local shift: $h_i^{k+1} = h_i^k + \alpha \delta_i^{k+1}$
    Send message δ_i^{k+1} to the master
  end
  Aggregate received messages: $\delta^{k+1} = \frac{1}{n}\sum_{i=1}^n \delta_i^{k+1}$
  Update aggregated shift: $h^{k+1} = h^k + \alpha \delta^{k+1}$
  Compute $\Delta^{k+1} = \delta^{k+1} + h^k$
  Take "model" step: $x^{k+1} = (1-\eta)x^k + \eta \Delta^{k+1}$
end

We can rewrite VR-GDCI (Algorithm 2) in the following equivalent way

$$\delta^{k+1} = \frac{1}{n}\sum_{i=1}^n \delta_i^{k+1} = \frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i\left(\mathcal{T}_i(x^k) - h_i^k\right)$$
$$h^{k+1} = h^k + \alpha \delta^k.$$

which leads to such update rule

$$\begin{aligned} x^{k+1} &= x^k - \eta\left(x^k - h^k - \delta^k\right) \\ &= x^k - (\eta\gamma)\frac{1}{\gamma}\left(x^k - h^k - \frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i\left(x^k - h_i^k - \gamma \nabla f_i(x^k)\right)\right) \\ &= x^k - (\eta\gamma)\tilde{\mathcal{Q}}^k(\nabla f(x^k)) \end{aligned}$$

Theorem 15 *Let Ψ^k be the following Lyapunov function:*

$$\Psi^k := \|x^k - x^\star\|^2 + \frac{4\eta^2\omega}{\alpha n}\sigma^k, \qquad \sigma^k := \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \mathcal{T}_i(x^\star)\right\|^2$$

Suppose that f is L-smooth and μ-strongly convex. Choose the stepsizes α, η, γ such that

$$\alpha \leq \frac{1}{\omega+1}, \qquad \eta = \left[\frac{L}{\mu} + \frac{6\omega}{n}\left(\frac{L_{\max}}{\mu} - 1\right)\right]^{-1}, \qquad \gamma \leq \frac{1 + 6\omega\eta/n}{\eta(L + 6L_{\max}\omega/n)}.$$

Then the iterates defined by Algorithm 2 satisfy

$$\mathbf{E}\,\Psi^k \leq \left(1 - \min\left\{\frac{\alpha}{2}, \eta\right\}\right)^k \|x^0 - x^\star\|^2 \Psi^0.$$

Proof Start of the analysis is exactly the same as in the previous section (23).
Expectation conditional on x^k and h^k

$$\begin{aligned}\mathbf{E}\left\|r^{k+1}\right\|^2 \coloneqq \mathbf{E}\left\|x^{k+1}-x^\star\right\|^2 &= \left\|r^k\right\|^2 - 2\eta\gamma\langle x^k-x^\star, \nabla f(x^k)-\nabla f(x^\star)\rangle \\ &\quad + (\eta\gamma)^2\Big[\underbrace{\mathbf{E}\left\|\tilde{\mathcal{Q}}^k(\nabla f(x^k))-\nabla f(x^k)\right\|^2}_{\tau^k} + \left\|\nabla f(x^k)-\nabla f(x^\star)\right\|^2\Big]. \end{aligned} \tag{24}$$

For term τ^k we employ similar upper bound using the fact that $\tilde{\mathcal{Q}}_i^k \in \mathbb{U}(\omega; \left(x^k - h_i^k\right)/\gamma)$

$$\begin{aligned}\mathbf{E}\left[\left\|\tilde{\mathcal{Q}}^k(\nabla f(x^k))-\nabla f(x^k)\right\|^2 \mid x^k, h^k\right] &\le \frac{\omega}{n^2}\sum_{i=1}^n \left\|\nabla f_i(x^k) - \left(x^k-h_i^k\right)/\gamma\right\|^2 \\ &= \frac{\omega}{n^2}\frac{1}{\gamma^2}\sum_{i=1}^n \left\|x^k - h_i^k - \gamma\nabla f_i(x^k) \pm \mathcal{T}_i(x^\star)\right\|^2 \\ &\le \frac{2\omega}{n^2\gamma^2}\sum_{i=1}^n \left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2 \\ &\quad + \frac{2\omega}{\gamma^2 n}\underbrace{\frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \mathcal{T}_i(x^\star)\right\|^2}_{\sigma^k}. \end{aligned} \tag{25}$$

Next we upper-bound the σ^{k+1} term (expectation conditional on x^k and h^k):

$$\begin{aligned}\mathbf{E}\,\sigma^{k+1} &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\|h_i^{k+1}-\mathcal{T}_i(x^\star)\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\|h_i^k + \alpha\mathcal{Q}_i\left(\mathcal{T}_i(x^k)-h_i^k\right) - (x^\star - \gamma\nabla f_i(x^\star))\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left[\left\|h_i^k-\mathcal{T}_i(x^\star)\right\|^2 + \alpha^2\,\mathbf{E}\left\|\mathcal{Q}_i\left(\mathcal{T}_i(x^k)-h_i^k\right)\right\|^2 + 2\alpha\langle h_i^k-\mathcal{T}_i(x^\star), \mathcal{T}_i(x^k)-h_i^k\rangle\right] \\ &\overset{(2)}{\le} \sigma^k + \frac{1}{n}\sum_{i=1}^n \left[\alpha^2(\omega+1)\left\|\mathcal{T}_i(x^k)-h_i^k\right\|^2 - 2\alpha\langle h_i^k-\mathcal{T}_i(x^\star), h_i^k-\mathcal{T}_i(x^k)\rangle\right] \\ &[\alpha \le 1/(\omega+1)] \le \sigma^k + \frac{1}{n}\sum_{i=1}^n \left[\alpha\left\|\mathcal{T}_i(x^k)-h_i^k\right\|^2 - 2\alpha\langle h_i^k-\mathcal{T}_i(x^\star), h_i^k-\mathcal{T}_i(x^k)\rangle\right] \\ &\overset{(10)}{=} \sigma^k + \frac{1}{n}\sum_{i=1}^n \alpha\left[\left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2 - \left\|h_i^k-\mathcal{T}_i(x^\star)\right\|^2\right] \\ &= (1-\alpha)\sigma^k + \alpha\frac{1}{n}\sum_{i=1}^n \left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2. \end{aligned}$$

Simplification of the second term (same as first term in (25))

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2 &= \frac{1}{n}\sum_{i=1}^n \left\|x^k-\gamma\nabla f_i(x^k)-(x^\star-\gamma\nabla f_i(x^\star))\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big[\left\|x^k-x^\star\right\|^2+\gamma^2\left\|\nabla f_i(x^k)-\nabla f_i(x^\star)\right\|^2 \\
&\quad -2\gamma\langle x^k-x^\star,\nabla f_i(x^k)-\nabla f_i(x^\star)\rangle\Big] \\
&\le \left\|r^k\right\|^2+\frac{1}{n}\sum_{i=1}^n\left[\gamma^2 2L_iD_{f_i}(x^k,x^\star)-2\gamma\left(D_{f_i}(x^k,x^\star)+\frac{\mu}{2}\|x^k-x^\star\|^2\right)\right] \\
&= (1-\gamma\mu)\left\|r^k\right\|^2-2\gamma\cdot\frac{1}{n}\sum_{i=1}^n(1-\gamma L_i)\,D_{f_i}(x^k,x^\star) \\
&\le (1-\gamma\mu)\left\|x^k-x^\star\right\|^2-2\gamma\cdot(1-\gamma L_{\max})\,D_f(x^k,x^\star).
\end{aligned} \tag{26}$$

Combining the obtained bounds for $\|x^{k+1}-x^\star\|^2$ and σ^{k+1} we get the Lyapunov function (expectation conditional on x^k,h^k):

$$\begin{aligned}
\mathbf{E}\,\Psi^{k+1} &:= \mathbf{E}\left\|x^{k+1}-x^\star\right\|^2+\frac{4\eta^2\omega}{\alpha n}\mathbf{E}\,\sigma^{k+1} \\
&\le \|r^k\|^2-2\eta\gamma\langle x^k-x^\star,\nabla f(x^k)-\nabla f(x^\star)\rangle+(\eta\gamma)^2\left\|\nabla f(x^k)-\nabla f(x^\star)\right\|^2 \\
&\quad+(\eta\gamma)^2\left[\frac{2\omega}{n^2\gamma^2}\sum_{i=1}^n\left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2+\frac{2\omega}{\gamma^2 n}\sigma^k\right] \\
&\quad+\frac{4\eta^2\omega}{\alpha n}\left[(1-\alpha)\sigma^k+\alpha\frac{1}{n}\sum_{i=1}^n\left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2\right] \\
&\overset{(7,8)}{\le} (1-\eta\gamma\mu)\|r^k\|^2-2\eta\gamma\,(1-L\eta\gamma)\,D_f(x^k,x^\star) \\
&\quad+\frac{4\eta^2\omega}{\alpha n}\left(1-\frac{\alpha}{2}\right)\sigma^k+\frac{6\omega\eta^2}{\alpha n}\frac{1}{n}\sum_{i=1}^n\left\|\mathcal{T}_i(x^k)-\mathcal{T}_i(x^\star)\right\|^2 \\
&\overset{(26)}{\le}\left[1-\mu\eta\gamma+\frac{6\omega\eta^2}{n}(1-\gamma\mu)\right]\|r^k\|^2+\frac{4\eta^2\omega}{\alpha n}\left(1-\frac{\alpha}{2}\right)\sigma^k \\
&\quad-2\eta\gamma\left[1-L\eta\gamma+\frac{6\eta\omega}{n}(1-\gamma L_{\max})\right]D_f(x^k,x^\star) \\
\left[\gamma\le\frac{1+\frac{6\omega}{n}\eta}{\eta\left(L+\frac{6\omega}{n}L_{\max}\right)}\right] &\le \left[1-\mu\eta\gamma+\frac{6\omega\eta^2}{n}(1-\gamma\mu)\right]\|r^k\|^2+\frac{4\eta^2\omega}{\alpha n}\left(1-\frac{\alpha}{2}\right)\sigma^k \\
&\le (1-\eta)\,\|x^k-x^\star\|^2+\frac{4\eta^2\omega}{\alpha n}\left(1-\frac{\alpha}{2}\right)\sigma^k
\end{aligned}$$

for

$$\eta=\mu\left[L+\frac{6\omega}{n}(L_{\max}-\mu)\right]^{-1}.$$

The last inequality was obtained via minimization of the term $\left[1-\mu\eta\gamma+\frac{6\omega\eta^2}{n}\left(1-\gamma\mu\right)\right]$ w.r.t. γ,η and using contraction inequality constraint. Using the definition of the Lyapunov function we obtain

$$\mathbf{E}\left[\Psi^{k+1}\mid x^k,h^k\right]\leq\left(1-\min\left\{\eta,\frac{\alpha}{2}\right\}\right)\|x^k-x^\star\|^2\Psi^k,$$

which by unrolling the recursion and taking full expectation leads to the statement of the Theorem 15. ■

Obtained Theorem gives rise to the following iteration complexity result

$$\max\left\{2(\omega+1),\frac{L}{\mu}+\frac{6\omega}{n}\left(\frac{L_{\max}}{\mu}-1\right)\right\}\log 1/\varepsilon,$$

which after simplification ($L_i\equiv L$) is equivalent to complexity of DIANA up to numerical constants

$$\max\{2(\omega+1),\left(1+6\frac{\omega}{n}\right)\kappa\}\log 1/\varepsilon,$$

and improves over the original rate of VR-GDCI from [8] analyzed for fixed point problems and specialized for gradient mappings:

$$2\max\left\{\omega+1,\kappa\min\left(1,12\frac{\omega}{n}\kappa\right)\right\}\log 1/\varepsilon.$$

Appendix E. Experiments

Consider a classical ridge-regression optimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) \coloneqq \frac{1}{2}\|\mathbf{A}x-y\|_2^2+\frac{\lambda}{2}\|x\|_2^2\right],$$

where $\lambda = 1/m$ and $\mathbf{A} \in \mathbb{R}^{m\times d}, y \in \mathbb{R}^m$ are generated using Scikit-learn library [28] method sklearn.datasets.make_regression with default parameters for $m = 100, d = 80$. Obtained data is uniformly at random distributed evenly among 10 workers. To compare selected Algorithms, we evaluate the logarithm of a relative argument error $\log\left(\|x^k - x^\star\|^2/\|x^0 - x^\star\|^2\right)$ on vertical axis, while for horizontal one we calculate number of communicated bits needed to reach certain error tolerance ε. The starting point $x^0 \in \mathbb{R}^d$ entries are sampled from the normal distribution $\mathcal{N}(0, 10)$.

E.1. Randomized-DIANA vs DIANA

In the first set of experiments we compare Rand-DIANA and DIANA with different compressors $\mathcal{Q}_i$ ($\mathcal{C}_i \equiv 0$) and varied operators' parameters. Obtained results are summarized Figure 1. Designation $q \coloneqq k/d$ is used for the share of non-zeroed coordinates of `Random sparsification` (`Rand-K`) operator and s corresponds to number of levels for `Natural Dithering` (`ND`) [15] compressor. Parameter p of Rand-DIANA was chosen as $1/(\omega+1)$ for every run.

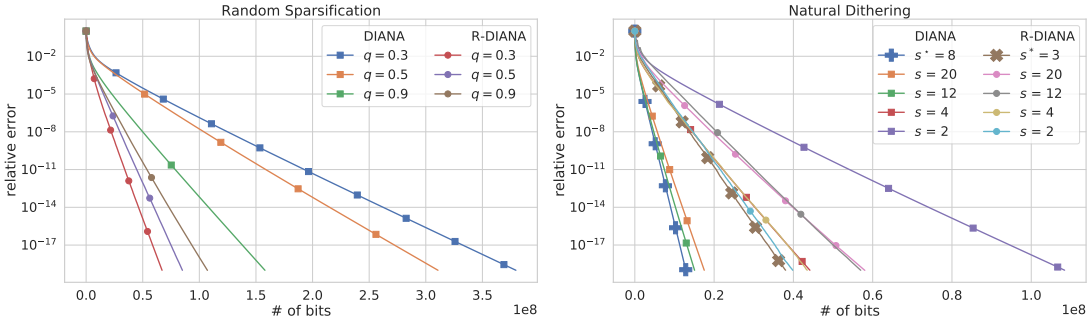

Figure 1: Comparison of DIANA and Rand-DIANA. **Left plot**: methods equipped with `Rand-K` for various q values. **Right plot**: selected results of grid search for `ND` parameter $s \in [20]$.

Left plot in Figure 1 clearly shows that Rand-DIANA performs better than original DIANA for every value of `Rand-K` compressor parameter. It is worth noting that DIANA performs better with higher q, while for Rand-DIANA the opposite holds. From the right plot one can find that DIANA with `ND` can be superior to Rand-DIANA for optimized parameter $s^\star$. Nevertheless, for very aggressive compression (e.g., $s = 2$) Rand-DIANA is preferable.

In the next experimental setup, we closer investigate the behavior of Rand-DIANA.

E.2. Randomized-DIANA study

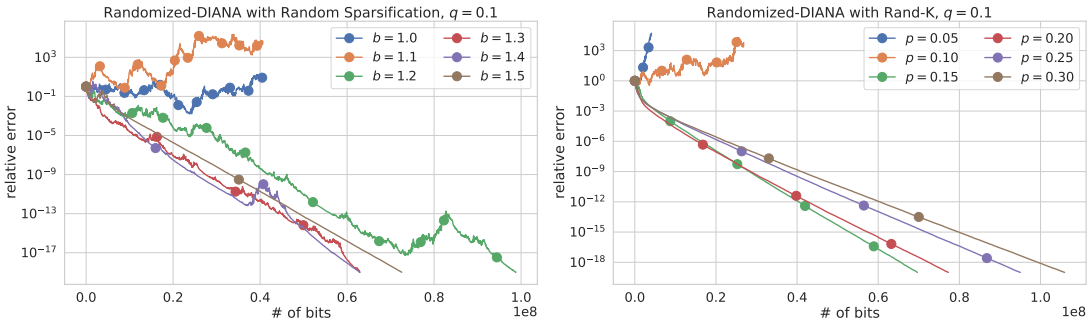

Figure 2: Study of Rand-DIANA stability and performance with varying parameters b and p.

According to the formulation of Theorem 6 constant M has to be chosen strictly greater than $M' := 2\omega/(np)$. In the left plot of Figure 2 we show that for smaller values of M (set to $M' \cdot b$) the method becomes less stable and can even diverge. But too big M (for $b = 1.5$) can lead to overall (stable) slowdown. The right plot examines how parameter p affects the convergence in high compression regime ($q = 0.1$). The conclusion is that for smaller p method converges faster, and after certain threshold it can diverge. So, there is a trade-off similar to the previous study of M.