Correlated Quantization for Faster Nonconvex Distributed Optimization

Andrei Panferov[1] **Yury Demidovich**[1] **Ahmad Rammal**[1] **Peter Richtárik**[1]

[1]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Abstract

Quantization [Alistarh et al., 2017] is an important (stochastic) compression technique that reduces the volume of transmitted bits during each communication round in distributed model training. Suresh et al. [2022] introduce correlated quantizers and show their advantages over independent counterparts by analyzing distributed SGD communication complexity. We analyze the forefront distributed non-convex optimization algorithm MARINA [Gorbunov et al., 2022] utilizing the proposed correlated quantizers and show that it outperforms the original MARINA and distributed SGD of Suresh et al. [2022] with regard to the communication complexity. We significantly refine the original analysis of MARINA without any additional assumptions using the weighted Hessian variance [Tyurin et al., 2022], and then we expand the theoretical framework of MARINA to accommodate a substantially broader range of potentially correlated and biased compressors, thus dilating the applicability of the method beyond the conventional independent unbiased compressor setup. Extensive experimental results corroborate our theoretical findings.

1 INTRODUCTION

Modern deep neural networks consist of numerous blocks comprising diverse layers that are arranged in a hierarchical structure [LeCun et al., 2015]. This complexity leads to a high demand for data in these networks [Vaswani et al., 2017, Brown et al., 2020]. Moreover, it is worth noting that such models exhibit a distinct nonconvex nature [Choromanska et al., 2015]. Hence, there is a requirement to distribute the data among various computing resources, giving rise to the challenge of effectively orchestrating distributed [Yang et al., 2019] model training. Another incentive for adopting distributed training emerges from the Federated Learning framework [Li et al., 2020a, Kairouz and et. al, 2019]. In this scenario, client-owned data is not readily shared among clients. Consequently, a centralized algorithm becomes responsible for overseeing the training of multiple clients. Given that contemporary Machine Learning models have grown substantially in size, during each round of gradient descent, every client is required to transmit a dense gradient vector often comprised of millions of parameters [Li et al., 2020b]. This places an overwhelming strain on the communication network. Therefore, it becomes compulsory to explore techniques capable of diminishing the volume of bits transmitted over communication channels while preserving the algorithm's convergence.

There exist various approaches of addressing this problem. The concept of *acceleration* or *momentum* [Nesterov, 1983, 2004] in gradient-type methods has received extensive attention in conventional optimization problems. It aims to attain quicker convergence rates, thereby reducing the number of communication rounds [Beck and Teboulle, 2009, Zhu, 2017, Lan et al., 2019, Kovalev et al., 2019, Li et al., 2020c]. Deep Learning practitioners commonly rely on Adam [Kingma and Ba, 2015] or one of its numerous variants, which, among other techniques, also employ momentum. *Local training*, which involves having each participating client perform multiple local optimization steps on their data before engaging in communication-intensive parameter synchronization, stands as one of the most practically valuable algorithmic components in Federated Learning model training [Povey et al., 2014, Moritz et al., 2016, McMahan et al., 2017, Mishchenko et al., 2022, Condat et al., 2024, Grudzień et al., 2023, Demidovich et al., 2024b].

Driven by the necessity to create distributed stochastic gradient communication-efficient methods in nonconvex scenarios, in this paper we consider the optimization problem

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

where n is the number of clients working in parallel, and $f_i : \mathbb{R}^d \to \mathbb{R}$ is a (potentially nonconvex) function representing the loss of the model parameterized by weights $x \in \mathbb{R}^d$ with respect to the training data stored on client i. We require the problem (1) to be well-posed:

Assumption 1. *The functions $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ are differentiable. Moreover, f is lower bounded, i.e., there exists $f^{\inf} \in \mathbb{R}$ such that $f(x) \geq f^{\inf}$ for all $x \in \mathbb{R}^d$.*

Assumption 2. *There exists a constant $L_+ \geq 0$ such that $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_+^2 \|x-y\|^2$, for all $x, y \in \mathbb{R}^d$. To avoid ambiguity, let L_+ be the smallest such number.*

Assumption 2 is stronger than L_--Lipschitz continuity of the gradient of f (by Jensen's inequality; also, $L_- \leq L_+$):

Assumption 3. *There exists a constant $L_- \geq 0$ such that $\|\nabla f(x) - \nabla f(y)\| \leq L_- \|x-y\|$, for all $x, y \in \mathbb{R}^d$.*

We are interested in finding an approximately stationary point of the nonconvex problem (1). In other words, our objective is to find a (random) vector $\widehat{x} \in \mathbb{R}^d$ such that $\mathbb{E}\left[\|\nabla f(\widehat{x})\|^2\right] \leq \varepsilon^2$, all while minimizing the amount of communication between the n clients and the server.

A typical approach for solving the optimization problem (1) involves employing Distributed Gradient Descent (DGD). Starting with an initial iterate $x^0 \in \mathbb{R}^d$ and a learning rate $\gamma > 0$, at each iteration t, the server broadcasts the current iterate $x^t \in \mathbb{R}^d$ to the clients. Subsequently, each client computes its gradient $\nabla f_i(x^t)$ and sends it back to the server. Finally, the server aggregates all the gradients and utilizes them to perform the gradient descent step $x^{t+1} = x^t - \frac{\gamma}{n}\sum_{i=1}^n \nabla f_i(x^t)$, updating the iterate to x^{t+1}. This process is then repeated. Although DGD is widely acknowledged as an optimal algorithm for attaining a stationary point with minimal iterations in smooth nonconvex problems [Nesterov, 2004], it also places a substantial burden on the communication network. During each communication round, DGD sends dense gradients to the server. As mentioned earlier, this level of communication load is deemed impractical in numerous scenarios. One approach to address this issue is to employ an unbiased compressor on the transmitted data [Seide et al., 2014, Alistarh et al., 2017, Lin et al., 2017, Zhang et al., 2017, Lim et al., 2018, Alistarh et al., 2018, Wang et al., 2018].

Definition 1. *A (possibly randomized) mapping $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ is called an unbiased compressor if $\mathbb{E}[\mathcal{Q}(a)] = a$ and there exists a constant $\omega \geq 0$ such that*

$$\mathbb{E}\left[\|\mathcal{Q}(a) - a\|^2\right] \leq \omega \|a\|^2, \quad \forall x \in \mathbb{R}^d.$$

If this condition is satisfied for a compressor $\mathcal{Q}$, we shall write $\mathcal{Q} \in \mathbb{U}(\omega)$.

The subsequent phase in reducing the communication burden within DGD involves the implementation of client-to-server communication compression. This modification of DGD is referred to as Distributed Compressed Gradient Descent (DCGD), and it conducts iterations of the form

$$x^{t+1} = x^t - \frac{\gamma}{n}\sum_{i=1}^n \mathcal{Q}_i^t\left(\nabla f_i(x^t)\right),$$

where $\mathcal{Q}_i^t$ is the compressor used by the client i at iteration t. DCGD stands out as one of the simplest distributed methods that employ compression. More advanced methods include DIANA [Mishchenko et al., 2019], MARINA [Gorbunov et al., 2022, Sokolov and Richtárik, 2024, Demidovich et al., 2024a].

Most of the common compression techniques can be attributed to one of the two classes: *sparsification* or *quantization*. Sparsification [Alistarh et al., 2018, Szlendak et al., 2021, Demidovich et al., 2023a] methods reduce communication by only selecting an important sparse subset of the vectors to broadcast at each step. Highly popular sparsifiers are TopK and RandK [Beznosikov et al., 2020]. More examples can be found in the survey Demidovich et al. [2023b]. In the present work we focus on the quantization compression technique [Alistarh et al., 2017]. When provided with the gradient vector at a client, we quantize each component through randomized rounding to a discrete set of values, preserving the statistical properties of the original vector. Below we define several widely used quantizers.

Definition 2. *Let $1 \leq q \leq +\infty$, $a \in \mathbb{R}^d$. Standard dithering operator $\mathcal{D}_{sta}^{q,k}$ with k levels $0 = l_k < l_{k-1} = \frac{1}{k} < \ldots < l_1 = \frac{k-1}{k} < l_0 = 1$, is defined as follows. If $a = 0$, then $\mathcal{D}_{sta}^{q,k} = 0$. If $a \neq 0$, let $y_i \overset{def}{=} \frac{|a_i|}{\|a\|_q}$, for all $i \in [d]$. Fix i, let $u \in \{0, 1, \ldots, k-1\}$ be such that $l_{u+1} \leq y_i \leq l_u$. Then $\left(\mathcal{D}_{sta}^{q,k}\right)_i = \|a\|_q \times sign(a_i) \times \xi(y_i)$, where $\xi(y_i) = l_u$ with probability $\frac{y_i - l_{u+1}}{l_u - l_{u+1}}$ or $\xi(y_i) = l_{u+1}$ otherwise.*

Definition 3 (Natural dithering)**.** *Natural dithering operator $\mathcal{D}_{nat}^{q,k}$ with k levels is defined in the same way, but with $l_k = 0$, $l_{k-1} = \frac{1}{2^{k-1}}$, $\ldots$, $l_1 = \frac{1}{2}$, $l_0 = \frac{1}{2^0} = 1$.*

In particular, Alistarh et al. [2017] consider QSGD with independent standard dithering quantizers $\mathcal{D}_{sta}^{2,k}$.

The major measure of the effectiveness of the distributed training method is its *communication complexity*. It is the result of multiplying the number of communication rounds required to find $\widehat{x}$ by a properly defined measure of the amount of communication carried out in each round. Consistent with the standard practice in the literature, we make the assumption that client-to-server communication constitutes the primary bottleneck, and therefore, we do not include server-to-client communication in our calculations.

Table 1: Comparison of communication complexities of different distributed methods combined with different quantizers in the nonconvex regime with homogeneous clients (see Section 3.2), when $d \leq n$. In the homogeneous scenario, $L_- = L_+ = L$ and $L_\pm = 0$. Notation: $\Delta^0 = f(x^0) - f^*$. Abbreviations: CQ = "Correlated Quantizers", ISCC = "Importance Sampling Combinatorial Compressors", IQ = "Independent Quantizers".

Method	Quantizer	Communication Complexity	Correlated Compressors	Reference
DCGD	IQ, Def. 6	$\mathcal{O}\left(\frac{\Delta^0 dL}{\varepsilon^2}\right)$	✗	Suresh et al. [2022]
DCGD	CQ, Def. 7	$\mathcal{O}\left(\frac{\Delta^0 dL}{\varepsilon^2}\right)$	✓	Suresh et al. [2022]
MARINA	$\mathcal{D}^{q,k}_{nat}$, Def. 3	$\mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1+\frac{d}{\sqrt{n}}\right\}\right)$	✗	Gorbunov et al. [2022]
MARINA	IQ, Def. 6	$\mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1+\frac{d}{\sqrt{n}}\right\}\right)$	✗	Gorbunov et al. [2022]
MARINA	ISCC, Asm. 6	$\mathcal{O}\left(\frac{\Delta^0 d}{\varepsilon^2}\min\left\{L, \frac{L}{n}+\frac{\sqrt{\omega+1}L_{avg}}{\sqrt{n}}\right\}\right)$	✗	Corollary 4, this work
MARINA	CQ, Def. 7	$\mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1+\frac{d}{n}\right\}\right)$	✓	Proposition 4, this work

Table 2: Comparison of important characteristics of different quantizers in the nonconvex zero-Hessian-variance regime and when $d \leq n$: bits sent per client and MSE (Mean Square Error, Section 3.1). Notation: $\mathcal{D}^{2,k}_{sta}$ – Standard Dithering, $\mathcal{D}^{\infty,1}_{sta}$ – Ternary Quantization, $\mathcal{D}^{q,k}_{nat}$ – Natural Dithering.

Quantizer	Bits Sent	MSE	Correlated?	Reference
$\mathcal{D}^{2,k}_{sta}$, Def. 2	$\mathcal{O}\left(k(k+\sqrt{d})\right)$	$\frac{\sqrt{d}}{nk}$	✗	Alistarh et al. [2017]
$\mathcal{D}^{\infty,1}_{sta}$, Def. 2	$31 + d\log_2 3$	$\frac{\sqrt{d}-1}{n}$	✗	Wen et al. [2017]
$\mathcal{D}^{q,k}_{nat}$, Def. 3	$31 + d\log_2(2k+1)$	$\frac{\sqrt{d}}{n2^{k-1}}$	✗	Gorbunov et al. [2022]
IQ, Def. 6	$32 + d$	$\frac{d\|a\|^2}{n}$, Cor. 1	✗	Gorbunov et al. [2022]
CQ, Def. 7	$32 + d$	$\frac{d\|a\|^2}{n^2}$, Cor. 2	✓	Suresh et al. [2022]
ISCC, Asm. 6	$\frac{\mathcal{O}(d)}{n}$	$\left(\frac{A}{n^2}\sum_{i=1}^n \frac{1}{w_i} - B\right)\|a\|^2$, Asm. 6	✗	Corollary 4, this work

The literature has introduced several distributed methods more advanced than DCGD, often in conjunction with various quantization techniques. Horváth et al. [2022] consider DCGD with independent natural dithering quantizers. Mishchenko et al. [2019], Horváth et al. [2023] study DIANA with arbitrary unbiased independent quantizers. Wen et al. [2017] examine distributed SGD algorithm with $\mathcal{D}^{\infty,1}_{sta}$. To the best of our knowledge, there exists only one paper [Suresh et al., 2022] that provides an analysis of unbiased correlated quantizers. However, these quantizers are integrated with the basic DCGD algorithm, and the optimization problem is exclusively examined in the convex setting. In Table 1 we compare communication complexities of the state-of-the-art method MARINA with correlated quantizers against other proposed combinations of algorithms and quantizers. In fact, DCGD with any quantizers has a communication complexity of $\mathcal{O}\left(\frac{\Delta^0 dL}{\varepsilon^2}\right)$, and it is hard to see theoretical advantages of CQ. Our results are better: MARINA communication complexity is lower than of DCGD, the proposed combinations with quantizers ISCC and CQ allow to reduce it even further. Since L_{avg} can be $\sqrt{n}$ times smaller than L_+, MARINA with ISCC can converge up to $\sqrt{n}$ times faster than the original method. This suggests that we can develop more effective assumptions for the framework of dependent compressors. In Table 2 we compare the number of bits sent and MSE of different quantizers. CQ send roughly the same amount of bits as its competitors yet they have lower MSE, which allows MARINA+CQ to achieve lower communication complexity (as shown in Table 1).

2 CONTRIBUTIONS

First, we highlight our main contributions.

$\diamond$ We extend the analysis of the state-of-the-art distributed optimization method MARINA [Gorbunov et al., 2022] (see Algorithm 1) beyond the use of independent quantizers. We rigorously demonstrate that MARINA achieves faster convergence when employing Correlated Quantizers (CQ) proposed by Suresh et al. [2022] in the zero-Hessian-variance regime [Szlendak et al., 2021]. Specifically, we establish the communication complexity of MARINA with CQ and showcase a significant enhancement compared to MARINA integrated with strong baseline independent quantizers (see Table 1; Proposition 4). Our experiments confirm the validity of our theoretical insights.

$\diamond$ We compare two distributed algorithms that utilize corre-

lated quantizers: MARINA and DCGD (see Table 1; Section 4.1). Our analysis reveals that in the zero-Hessian-variance regime, MARINA exhibits substantially lower communication complexity, making it a superior algorithm. Our experimental results corroborate the validity of our theoretical discoveries.

$\diamond$ We demonstrate that CQ from Suresh et al. [2022] exhibit significantly lower (by a factor of n) MSE compared to their independent counterparts when applied to homogeneous data (see Table 2; Corollaries 1 and 2). Furthermore, we provide insights into why these compressors are particularly effective when used with MARINA in the zero-Hessian-variance regime (see Section 3.3).

Our further contributions can be summarized as follows.

We propose a new way to combine CQ with correlated sparsifiers [Szlendak et al., 2021], allowing for even stronger compression (see Algorithm 3; Corollary 3).

We expand the scope of our findings by demonstrating through experiments that they remain applicable beyond the zero-Hessian-variance regime (see Section 4; Appendix D).

The initial analysis of MARINA was conducted under the assumption of individual unbiasedness of compressors. We revise it and demonstrate that equivalent convergence results can be achieved for a much wider range of Distributed Mean Estimation algorithms. Additionally, under the weighted AB-inequality Assumption 6 [Tyurin et al., 2022], we enhance the analysis of MARINA (see Algorithm 4) by investigating its convergence guarantees (see Theorem 2). We propose an Importance Sampling Combinatorial Compressor which in combination with MARINA allows for an up to $\sqrt{n}$ times faster convergence than the original method (see Corollary 4; Table 1). Our findings are corroborated by experiments (see Appendix C.3).

3 MAIN RESULTS

Let us introduce an assumption for the set of compressors that is used in many results of the paper.

Assumption 4 (Individual Unbiasedness)**.** *The random operators $\mathcal{Q}_1, \ldots, \mathcal{Q}_n : \mathbb{R}^d \to \mathbb{R}^d$ are unbiased, i.e., $\mathbb{E}[\mathcal{Q}_i(a)] = a$ for all $i \in \{1, 2, \ldots, n\}$ and all $a \in \mathbb{R}^d$. If these conditions are satisfied, we write $\{\mathcal{Q}_i\}_{i=1}^n \in \mathbb{U}_{\text{ind}}$.*

3.1 AB-INEQUALITY: BETTER CONTROL OF MSE

Given n vectors $a_1, \ldots, a_n \in \mathbb{R}^d$, *compression variance* or *Mean Square Error (MSE)* associated with the set of randomized compressors $\{\mathcal{Q}_i\}_{i=1}^n$ is the quantity $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i(a_i) - \frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]$. In their works, Suresh et al. [2017, 2022] investigate the problem of distributed mean estimation under communication constraints

Algorithm 1 MARINA

Input: initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, probability $p \in (0, 1]$, number of iterations T
$g^0 = \nabla f(x^0)$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $c_t \sim \text{Bern}(p)$
  Broadcast g^t to all workers
  for $i = 1, \ldots, n$ in parallel **do**
    $x^{t+1} = x^t - \gamma g^t$
    $g_i^{t+1} = \nabla f_i(x^{t+1})$ if $c_t = 1$, and $g_i^{t+1} = g_i^t + \mathcal{Q}_i\left(\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right)$ otherwise
  end for
  $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$
end for
Output: $\hat{x}^T$ chosen uniformly at random from $\{x^t\}_{t=0}^{T-1}$

and mainly focus on the task of minimizing the MSE of quantizers. In fact, compression variance naturally emerges in the analysis of MARINA (see Algorithm 1), a cutting-edge distributed algorithm designed for solving nonconvex optimization problems, and the theoretical communication complexity of this method linearly depends on the square root of the compression variance. Therefore, it is crucial to identify compressors with low MSE when analyzing MARINA. Nevertheless, there exists a trade-off between MSE and communication cost. Typically, as MSE increases, compression becomes more aggressive, but concurrently, the number of communication rounds also increases. For this reason, recently, Szlendak et al. [2021] introduced the following tool for achieving a more precise control of compression variance.

Assumption 5 (AB-inequality)**.** *There exist constants $A, B \geq 0$, such that random operators $\mathcal{Q}_1, \ldots, \mathcal{Q}_n : \mathbb{R}^d \to \mathbb{R}^d$ satisfy the inequality*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i(a_i) - \frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right] \leq A\left(\frac{1}{n}\sum_{i=1}^n \|a_i\|^2\right) - B\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2,$$

for all $a_1, \ldots, a_n \in \mathbb{R}^d$. If these conditions are satisfied, we write $\{\mathcal{Q}_i\}_{i=1}^n \in \mathbb{U}(A, B)$ for the set of operators.

Note that the MSE of the estimate $\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i(a_i)$ for $\frac{1}{n}\sum_{i=1}^n a_i$ on the right-hand side can be viewed as a variance of the sum of the compressors. The question of interest here is how correlation between the random compression operators, or their independence, can affect the MSE. The following observations were made by Szlendak et al. [2021]. If compressors are unbiased (see Definition 1), then the AB-inequality holds without any assumption on their independence. Generally, requiring independence can lead to a significant improvement in the constant A. Formally:

Proposition 1. *If, for all $i \in [n]$, $\mathcal{Q}_i \in \mathbb{U}(\omega_i)$ and $\{\mathcal{Q}_i\}_{i=1}^n \in \mathbb{U}_{\text{ind}}$, then $\{\mathcal{Q}_i\}_{i=1}^n \in \mathbb{U}\left(\max_i\{\omega_i\}, 0\right)$. If we further assume that the compressors are independent, then $\{\mathcal{Q}_i\}_{i=1}^n \in \mathbb{U}\left(\frac{1}{n}\max_i\{\omega_i\}, 0\right)$.*

However, it is possible to design correlated unbiased quantizers with an even smaller constant A.

3.2 WHY CORRELATION MAY HELP

It could be feasible to decrease the compression variance by introducing dependencies between the compressors. The right-hand side of Assumption 5 can be rewritten as

$$A\left[\left(1-\frac{B}{A}\right)\left(\frac{1}{n}\sum_{i=1}^n \|a_i\|^2\right)+\frac{B}{A}\operatorname{Var}\left(a_1,\ldots,a_n\right)\right],$$

where $\operatorname{Var}\left(a_1,\ldots,a_n\right)=\frac{1}{n}\sum_{i=1}^n\left\|a_i-\sum_{i=1}^n a_i\right\|^2$ is the variance of the vectors $\{a_i\}_{i=1}^n$. It is preferable to design compressors with B as large as A, since $\operatorname{Var}\left(a_1,\ldots,a_n\right)$ can be much smaller than $\frac{1}{n}\sum_{i=1}^n\|a_i\|^2$, This result was obtained by Szlendak et al. [2021]: PermK sparsifiers introduced in this work are designed so that the sparsified vectors have zero scalar products, which enforces $A=B=1$. Nevertheless, the approach of zeroing out scalar products does not apply to a quantization technique for the general set of vectors $\{a_i\}_{i=1}^n$, as quantization does not inherently enforce sparsity in vectors. Instead, we demonstrate below that a thoughtfully designed dependence between the unbiased quantizers can yield an even more substantial enhancement of the constant A, while the constant B remains comparatively smaller and equals zero. Further, we introduce a regime in which we attain theoretical improvements through correlation, explain why this regime is more encompassing than the one involving clients that send homogeneous data (homogeneous clients regime), and clarify why MARINA particularly excels within it.

3.3 ZERO-HESSIAN-VARIANCE REGIME

The concept of Hessian variance was introduced by Szlendak et al. [2021] and allowed the authors to refine the communication complexity analysis of MARINA. First, let us provide a formal definition of it.

Definition 4 (Hessian Variance)**.** *Let $L_{\pm}\geq 0$ be the smallest constant such that*

$$\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x)-\nabla f_i(y)\|^2-\|\nabla f(x)-\nabla f(y)\|^2 \leq L_{\pm}^2\|x-y\|^2, \quad x,y\in\mathbb{R}^d.$$

The quantity $L_{\pm}^2$ is called Hessian variance.

Our theoretical results cover the setting when $L_{\pm}=0$. It extends the case when clients are either homogeneous or nearly homogeneous with linear perturbations.

Proposition 2. *In the homogeneous clients regime Hessian variance is equal to 0. Moreover, if loss functions on all clients differ only by a linear term, then Hessian variance is equal to 0.*

The scenario in which $L_{\pm}=0$ holds is particularly advantageous for the MARINA algorithm. Owing to the structure of the local gradient updates (see line 8 of Algorithm 1), the vectors $\nabla f_i(x^{t+1})-\nabla f_i(x^t)$ that need to be compressed and transmitted from clients to the server during the communication round exhibit homogeneity in the zero-Hessian-variance regime. While achieving the zero-Hessian-variance regime in practice can be challenging, practical problems can indeed have $L_{\pm}$ values very close to zero. We delve into the theoretical properties of the correlated quantizers by Suresh et al. [2022] in the context of homogeneous data and illustrate their advantages over previously proposed quantizers on homogeneous data.

3.4 SUPERIOR QUANTIZERS FOR MARINA

We start with an introduction to baseline independent quantizers. For simplicity, we initially define them in a one-dimensional case and outline their properties. We focus on the homogeneous case where $a_i=a\in\mathbb{R}^d$ for all $i\in[n]$.

Definition 5. *Suppose that, for all $i\in[n]$, $a_i=a\in[l,r]$, $l,r\in\mathbb{R}$. Define independent randomized quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ such that $\mathcal{Q}_i(a_i)=r$ with probability $\frac{a_i-l}{r-l}$ and $\mathcal{Q}_i(a_i)=l$ otherwise, $i\in[n]$.*

Proposition 3. *Quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ from Definition 5 are individually unbiased. The MSE of the quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ can be bounded from above in the following way:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(a_i-\mathcal{Q}_i(a_i)\right)\right\|^2\right]\leq\frac{(r-l)^2}{4n}.$$

Let us generalize the quantizers from Definition 5 to multiple dimensions.

Definition 6. *Assume that each $a_i=a$ is a d-dimensional vector and that $\mathcal{Q}_i$ quantizes each coordiante independently as in Definition 5 with $l=-\|a\|$, $r=\|a\|$. We'll refer to them as Independent Quantization (IQ).*

Corollary 1. *Suppose each $a_i\in\mathbb{R}^d$, $i\in[n]$. Then the MSE of quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ of the set of vectors $\{a_i\}_{i=1}^n$ can be bounded from above in the following way:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(a_i-\mathcal{Q}_i(a_i)\right)\right\|^2\right]\leq\frac{d\|a\|^2}{n}.$$

Algorithm 2 provides a definition of one-dimensional correlated quantizers, that generalize their independent counterparts defined above. We aim to establish individual unbiasedness and bound the MSE of $\{\mathcal{Q}_i\}_{i=1}^n$, defined in Algorithm 2, in relation to the set of numbers $\{a_i\}_{i=1}^n$, in the homogeneous case when $a_i = a \in [l, r]$, for all $i \in [n]$.

Algorithm 2 CQ (ONE-DIMENSIONAL VARIANT) [Suresh et al., 2022]

Input: $a_1, a_2, \ldots, a_n, l, r \in \mathbb{R}$; $\forall i \in [n]$, $a_i \in [l, r]$
Generate π, a random permutation of $\{0, 1, \ldots, n-1\}$
for $i = 1$ to n **do**
  $y_i = \frac{a_i - l}{r - l}$.
  $U_i = \frac{\pi_i}{n} + \gamma_i$, where γ_i has a continuous uniform distribution $\mathrm{U}[0, 1/n)$.
  $\mathcal{Q}_i(a_i) = (r - l)1_{U_i < y_i}$.
end for
Output: $\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i(a_i)$.

Theorem 1. *Suppose all the inputs $a_i = a$, $i \in [n]$, lie in the range $[l, r]$. Then $\{\mathcal{Q}_i\}_{i=1}^n$ from Algorithm 2 are individually unbiased and the following upper bound on the MSE of the set of quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ holds true:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (a_i - \mathcal{Q}_i(a_i))\right\|^2\right] \leq \frac{(r-l)^2}{4n^2}.$$

The generalization to multiple dimensions is performed in the same way as in the independent case.

Definition 7. *Assume that each $a_i = a$ is a d-dimensional vector and that $\mathcal{Q}_i$ quantizes each coordiante independently as in Algorithm 2 with $l = -\|a\|$, $r = \|a\|$. We'll refer to them as Correlated Quantization (CQ).*

Corollary 2. *Suppose each $a_i = a \in \mathbb{R}^d$, $i \in [n]$. Then $\{\mathcal{Q}_i\}_{i=1}^n$ are individually unbiased and the MSE of quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ associated with the set of vectors $\{a_i\}_{i=1}^n$ can be bounded from above in the following way:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (a_i - \mathcal{Q}_i(a_i))\right\|^2\right] \leq \frac{d\|a\|^2}{n^2}.$$

Notice that in both Corollaries 1 and 2, the term $\|a\|^2$ in the numerator can be replaced with $\frac{1}{n}\sum_{i=1}^n \|a_i\|^2$. Consequently, IQ belongs to $\mathbb{U}\left(\frac{d}{n}, 0\right)$. Importantly, any client-wise independent quantization satisfying Assumption 5 will do so with $A = \frac{\widehat{\omega}}{n}$ and $B = 0$, where $\widehat{\omega}$ is independent of n. Conversely, CQ adheres to Assumption 5 with $A = \frac{d}{n^2}$. As a result, with a fixed value of d, CQ has an A constant that is smaller by a factor of $\mathcal{O}\left(\frac{1}{n}\right)$.

Further, we analyze MARINA in the zero-Hessian-variance regime with independent and correlated quantizers.

Proposition 4. *Let $L_\pm = 0$. Denote by $\mathcal{C}_{cor}$ the communication complexity per client in* MARINA *with CQ (Definition 7). Similarly, denote by $\mathcal{C}_{ind}$ the communication complexity per client in* MARINA *with IQ (Definition 6). Then*

$$\frac{\mathcal{C}_{ind}}{\mathcal{C}_{cor}} = \frac{1 + \sqrt{\frac{(1-p)}{p}\frac{d}{4n}}}{1 + \sqrt{\frac{(1-p)}{p}\frac{d}{4n^2}}}.$$

That is, $\forall p \in [0, 1]$, $\mathcal{C}_{cor} \leq \mathcal{C}_{ind}$. In particular, we show that $\mathcal{C}_{cor} = \mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1 + \frac{d}{n}\right\}\right)$ and $\mathcal{C}_{ind} = \mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1 + \frac{d}{\sqrt{n}}\right\}\right)$.

Experiments suggest that when $d = n \gg 1$, the complexity ratio is approximately 7.29 (see Section B.2). For a more detailed discussion on the complexities and the conditions on the relation between d and n under which the ratio can reach up to 32, please refer to the experimental Section 4.5.

3.5 COMBINATION WITH SPARSIFICATION

Combining different compression techniques often yields better results than using any single technique on its own [Safaryan et al., 2021, Wang et al., 2023]. Motivated by this observation, in [Szlendak et al., 2021, Section 2.4], the authors obtain general results for the composition of independent unbiased compressors and PermK sparsifiers [Szlendak et al., 2021]. We design a new compression Algorithm 3, incorporating correlated sparsification in the form of PermK and CQ.

Algorithm 3 PERMK+CQ (NEW)

Input: $a_1, a_2, \ldots, a_n \in \mathbb{R}^d$, $\tau \in \mathbb{N}$
Consider the $n \times d$ block diagonal matrix with τ blocks of size $\frac{n}{\tau} \times \frac{d}{\tau}$ filled with ones.
  1. Randomly permute the matrix rows, then the columns to obtain a matrix $M_{i,j}$ which indicates if the client i should send its j-th component.
  2. Zero out $a_{i,j}$ entries where $M_{i,j} = 0$. Scale the remaining entries by τ.
  3. Independently perform CQ (7) within τ groups of entries that each block was mapped into.
Output: sparse, quantized vectors $\{P\mathcal{Q}_i(a_i)\}_{i=1}^n$.

Notice that when $\tau = n$, we obtain PermK sparsifier. On the other hand, if $\tau = 1$, then the compressor behaves as CQ (when there is only one block, we do not perform any permutations). That is, when $1 < \tau < n$, we indeed have a compressor which combines CQ and PermK. As experiments suggest, for some values of $L_\pm$, PermK+CQ is better than PermK. It means, that our new compressor is more robust to the introduced noise than PermK.

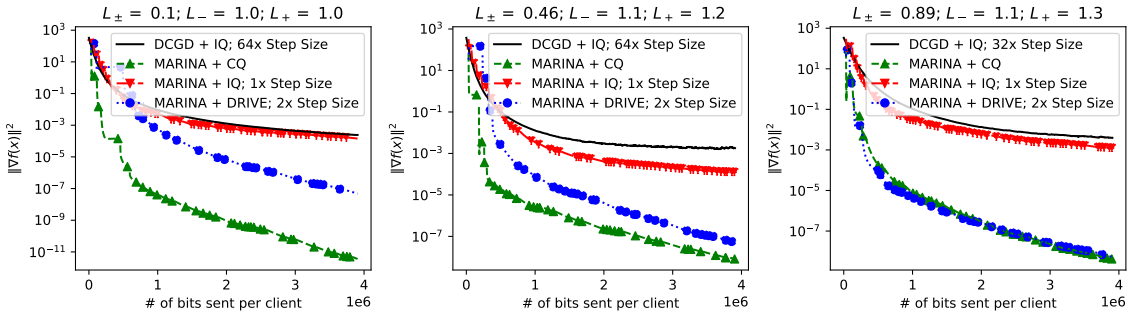

Figure 1: Comparison of CQ, IQ, and DRIVE with MARINA on quadratic optimization tasks with diverse $L_\pm$ values

Corollary 3. *The compressors described in Algorithm 3 use $32 + d/\tau$ bits per client and are individually unbiased. Moreover, if we assume that for each $i \in [n] : a_i = a$, then the mean square error of the quantizers $\{\mathcal{Q}_i\}_{i=1}^n$ can be bounded from above as:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(a_i - \mathcal{Q}_i(a_i))\right\|^2\right] \leq \frac{d\tau^2\|a\|^2}{n^2}.$$

Comparing this algorithm with CQ in the zero-Hessian-variance regime, we find that it communicates approximately τ times less data, albeit with its variance increased by a factor of τ^2.

4 EXPERIMENTS

We compare the performance of MARINA when combined with Correlated Quantization (CQ, see Definition 7), Independent Quantization (IQ, see Definition 6) and DRIVE [Vargaftik et al., 2021]. The latter serves as a robust non-correlated quantization baseline. Our primary objective is to ascertain whether our findings can be practically extended beyond the zero-Hessian-variance regime. In the plots, we depict the relationship between the total gradient norm and the volume of information communicated from clients to the server. To ensure a fair comparison of the various methods, we optimized the value of p and fine-tuned the stepsizes individually for each method and task. Where applicable, the selected stepsize is shown as a multiplier of the theoretical stepsize. For details, see Appendix D.3.

In Proposition 4 we prove that in zero-Hessian-variance regime MARINA+CQ has a lower communication complexity than MARINA+IQ. Below in Section 4.5 we also perform a numerical analysis to determine the ratio of communication complexities for MARINA+CQ and MARINA+IQ for different d and n. We explore whether it is possible to achieve a maximal speedup of 32.

4.1 QUADRATIC TASKS WITH VARIOUS $L_\pm$

We produced a range of quadratic optimization tasks with varying smoothness constants (see Figure 1). The procedures used to generate these tasks provide us with control over it (see Appendix for details) as $L_\pm$ values can be regulated. We opted for $d = 1024$, $n = 128$, regularization $\lambda = 0.001$, and noise scale $s \in \{0, 0.5, 1.0\}$. We can see that CQ outperforms IQ and is on par with, if not superior to, DRIVE even in tasks where $L_\pm$ substantially deviates from 0. We also included DCGD as a baseline.

We established our theory in the zero-Hessian-variance regime, but it becomes more challenging when $L_\pm \neq 0$. There is no theoretical stepsize for MARINA+CQ, but a fair comparison is imperative. In the absence of theory, a common approach is to choose optimal stepsizes for each method considered. We adjust the stepsizes by selecting the optimal ones as multiples of theoretical stepsizes by powers of 2 (see Appendix D.2 for details).

4.2 NON-CONVEX LOGISTIC REGRESSION

We examine MARINA combined with CQ in a non-convex scenario using a logistic regression problem formulated with a non-convex regularizer:

$$f(x) = \frac{1}{m}\sum_{k=1}^{m}\log\left(1 + \exp\left(-y_k a_k^T x\right)\right) + \lambda\sum_{j=1}^{d}\frac{x_j^2}{1 + x_j^2},$$

where $a_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ denote the training data, with $\lambda > 0$ as the regularization parameter. All our experiments utilized $\lambda = 0.1$. We obtained datasets from LibSVM [Chang and Lin, 2011] and partitioned their N entries into $n = d$ uniform segments. Table 3 provides a summary of these datasets. Additionally, for reference, we included DGD (Gradient Descent) in the comparison, which can be seen as MARINA with no compression. We specifically choose such setting to test our approach. Notice that, mainly, it is infeasible to calculate $L_\pm$ for this practical problem. In general, $L_\pm$ should be different from zero, and we do not

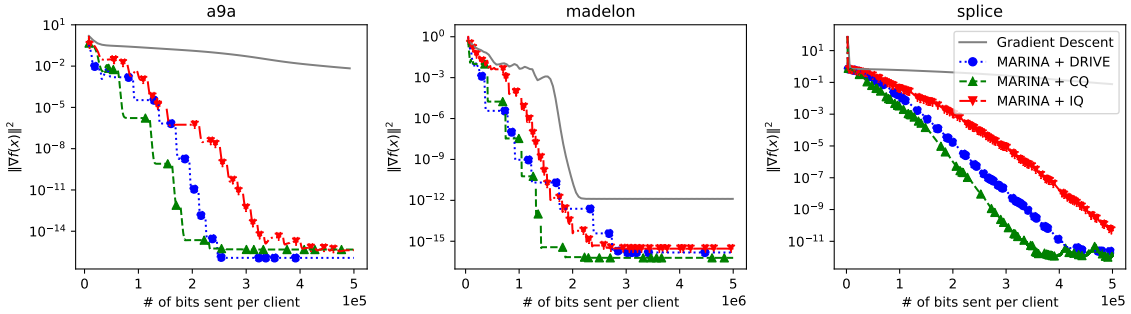

Figure 2: Comparison of CQ, IQ and DRIVE with MARINA on LibSVM datasets. The points represent the uncompressed rounds of the algorithm

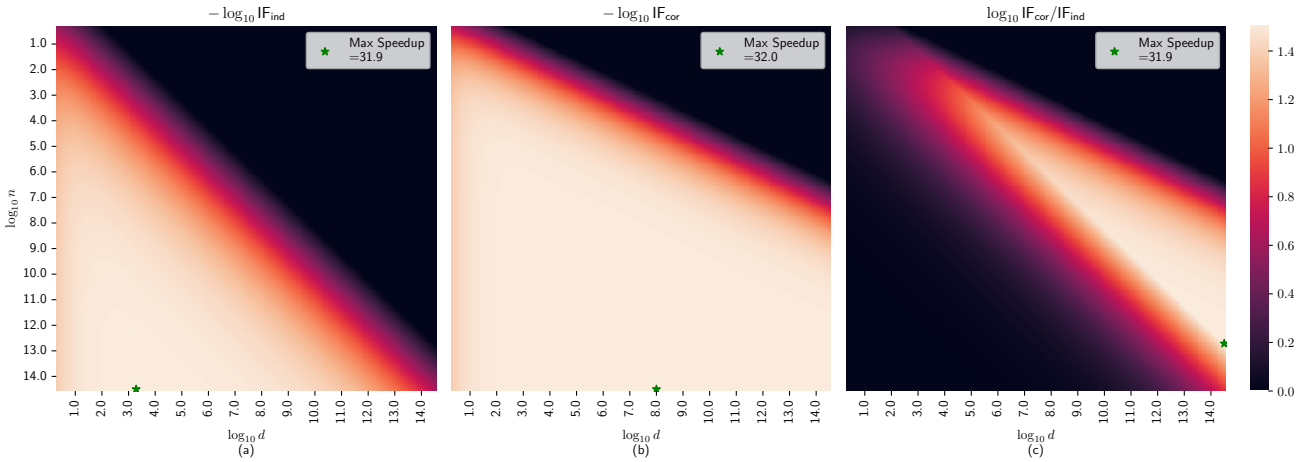

Figure 3: (a)/(b): Logarithmic speedup of MARINA with Correlated/Uncorrelated Quantization over Gradient Descent. (c): Logarithmic speedup of MARINA+CQ compared to MARINA+IQ

Table 3: Datasets and splitting of the data among clients

Dataset	$n = d$	N	$\lfloor N/d \rfloor$
a9a	123	32,561	264
madelon	500	2,000	4
splice	60	1,000	16

have a theory for $L_{\pm} \neq 0$. The results are in Figure 2. Our approach is mostly dominant even in $L_{\pm} \neq 0$ case against a strong baseline MARINA+DRIVE. We choose optimal stepsizes as in Section 4.1.

4.3 MULTI-LAYER PERCEPTRON

Experiments with an MLP classifier on the a9a dataset with 131 clients (see Figure 4), demonstrate that MARINA+CQ exhibits reduced complexity compared to MARINA+DRIVE, DCGD+IQ, MARINA+IQ. MARINA+CQ accommodates larger step sizes due to lower compression errors compared to MARINA+IQ, resulting in faster convergence in terms of loss (also, see Appendix D.2).

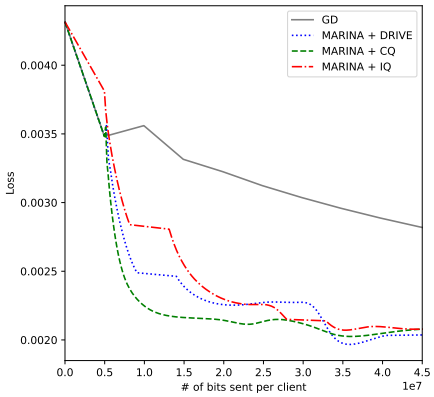

Figure 4: Comparison of CQ, IQ and DRIVE with MARINA on an MLP training task on the a9a dataset.

4.4 COMBINATION WITH PERMK

As it was mentioned in Section 3.5, combining different compression techniques, we may obtain better compressors.

In Algorithm 3, we proposed a compressor that combines PermK with CQ. We empirically measure the performance of PermK+CQ on the same synthetic quadratic optimization tasks as in Section 4.1 (see Figure 5). We set $d = 1024$, $n = 3072$, $\tau = \sqrt{d}$, the regularization $\lambda = 0.001$ and $s = 0.0$. We choose optimal stepsizes the same way as in Section 4.1.

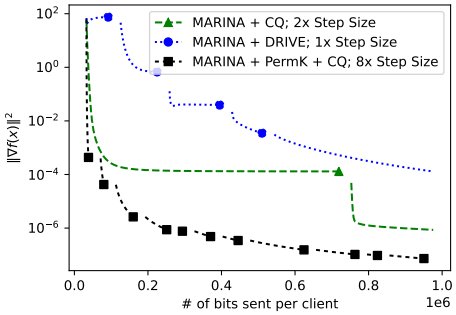

Figure 5: Comparison of PermK+CQ, CQ and DRIVE with MARINA on quadratic optimization task with $L_{\pm} = 0$

4.5 NUMERICAL COMPLEXITY ANALYSIS IN THE D-N PLANE

An Improvement Factor (IF) is a ratio of complexities of MARINA and GD (see Appendix B). To identify the region where CQ significantly outperforms IQ, we analyze the IFs of MARINA as functions of d and n, presuming we optimally choose the parameter p of the algorithm (see Appendix D.2.1). From plot (a) of Figure 3, we see that MARINA+IQ defaults to GD when $n \ll d$ and achieves the best possible speedup of x32 (owing to the compressor's 1-bit per coordinate behavior) when $n \gg d$. Conversely, CQ is distinguished by $d = n^2$, as shown in plot (b) of Figure 3. Consequently, plot (c) of Figure 3 reveals that CQ surpasses IQ by up to a factor of x32 when $\sqrt{d} < n < d$.

Acknowledgements

The work was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Anna Choromanska, MIkael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 192–204, San Diego, California, USA, 09–12 May 2015. PMLR.

Laurent Condat, Ivan Agarský, Grigory Malinovsky, and Peter Richtárik. Tamuna: Doubly accelerated distributed optimization with local training, compression, and partial participation, 2024.

Yury Demidovich, Grigory Malinovsky, Egor Shulgin, and Peter Richtárik. Mast: Model-agnostic sparsified training. *arXiv preprint arXiv:2311.16086*, 2023a.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the zoo of biased sgd. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23158–23171. Curran Associates, Inc., 2023b.

Yury Demidovich, Grigory Malinovsky, and Peter Richtárik. Streamlining in the riemannian realm: Efficient riemannian optimization with loopless variance reduction. *arXiv preprint arXiv:2403.06677*, 2024a.

Yury Demidovich, Petr Ostroukhov, Grigory Malinovsky, Samuel Horváth, Martin Takáč, Peter Richtárik, and Eduard Gorbunov. Methods with local steps and random reshuffling for generally smooth non-convex federated optimization. *arXiv preprint arXiv:2412.02781*, 2024b.

Eduard Gorbunov, Konstantin Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression, 2022.

Michal Grudzień, Grigory Malinovsky, and Peter Richtárik. Improving accelerated federated learning with compression and importance sampling. *ArXiv*, abs/2306.03240, 2023.

Samuel Horváth, Chen-Yu Ho, L'udovít Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Proceedings of MSML'22*, Aug 2022.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Peter Richtárik, and Sebastian Stich. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, 38(1):91–106, 2023.

Peter Kairouz and et. al. Advances and open problems in federated learning, 2019.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

D. 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 *International Conference on Algorithmic Learning Theory*, 2019.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Edward A. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10462–10472, 2019.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37 (3):50–60, 2020a.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020b.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtarik. Acceleration for compressed gradient descent in distributed and federated optimization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5895–5904. PMLR, 13–18 Jul 2020c.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6286–6295. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/li21a.html`.

Hyeontaek Lim, David G. Andersen, and Michael Kaminsky. 3lc: Lightweight and effective traffic compression for distributed machine learning. *ArXiv*, abs/1802.07389, 2018.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *ArXiv*, abs/1712.01887, 2017.

Prathamesh Mayekar and Himanshu Tyagi. Ratq: A universal fixed-length quantizer for stochastic optimization, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

Konstantin Mishchenko, Eduard A. Gorbunov, Martin Takác, and Peter Richtárik. Distributed learning with compressed gradient differences. *ArXiv*, abs/1901.09269, 2019.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtarik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine*

Learning Research, pages 15750–15769. PMLR, 17–23 Jul 2022.

Philipp Moritz, Robert Nishihara, Ion Stoica, and Michael I. Jordan. SparkNet: Training deep networks in spark. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *In Doklady AN USSR*, 269:543–547, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course (Applied Optimization)*. Kluwer Academic Publishers, 2004.

Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. Parallel training of dnns with natural gradient and parameter averaging. *arXiv: Neural and Evolutionary Computing*, 2014.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *arXiv preprint arXiv:2106.05203*, 2021.

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*, 11(2):557–580, apr 2021. doi: 10.1093/imaiai/iaab006. URL https://doi.org/10.1093%2Fimaiai%2Fiaab006.

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 *Interspeech*, 2014.

Igor Sokolov and Peter Richtárik. Marina-p: Superior performance in non-smooth federated optimization with adaptive stepsizes. *arXiv preprint arXiv:2412.17082*, 2024.

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3329–3337. JMLR.org, 2017.

Ananda Theertha Suresh, Ziteng Sun, Jae Hun Ro, and Felix Yu. Correlated quantization for distributed mean estimation and optimization, 2022.

RafałSzlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster distributed nonconvex optimization, 2021.

Alexander Tyurin, Lukang Sun, Konstantin Burlachenko, and Peter Richtárik. Sharper rates and flexible framework for nonconvex SGD with client and data sampling. *arXiv preprint arXiv:2206.02275*, 2022.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. DRIVE: One-bit distributed mean estimation, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jue Wang, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Re, and Ce Zhang. CocktailSGD: Fine-tuning foundation models over 500Mbps networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36058–36076. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wang23t.html.

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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019. ISSN 1367-5788.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4035–4043. PMLR, 06–11 Aug 2017.

URL https://proceedings.mlr.press/v70/zhang17e.html.

Zeyuan Allen Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Symposium on the Theory of Computing*, 2017.

Correlated Quantization for Faster Nonconvex Distributed Optimization (Supplementary Material)

Andrei Panferov[1] **Yury Demidovich**[1] **Ahmad Rammal**[1] **Peter Richtárik**[1]

[1]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

A QUANTIZERS IN HOMOGENEOUS DATA REGIME

A.1 ANALYSIS OF INDEPENDENT QUANTIZATION

A.1.1 Proof of Proposition 3

Proof. Let us calculate the first moment of $\mathcal{Q}_i(x_i)$: for every $i \in [n]$, we have that

$$\mathbb{E}\left[\mathcal{Q}_i(x_i)\right] = \frac{r(x_i - l)}{r - l} + \frac{l(r - x_i)}{r - l} = x_i.$$

Therefore, $\{\mathcal{Q}_i\}_{i=1}^n$ are individually unbiased. Further, let us calculate the variance of $\mathcal{Q}_i(x_i)$: for every i we obtain that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Q}_i^2(x_i)\right] - \left(\mathbb{E}\left[\mathcal{Q}_i(x_i)\right]\right)^2 &= \frac{r^2(x_i - l)}{r - l} + \frac{l^2(r - x_i)}{r - l} - x_i^2 \\
&= \frac{(r-l)^2(x_i - l)(r - x_i)}{(r-l)^2} \\
&= (r-l)^2 \frac{x_i - l}{r - l}\left(1 - \frac{x_i - l}{r - l}\right) \\
&\leq \frac{(r-l)^2}{4}.
\end{aligned}$$

Since the quantizers are independent and identically distributed, we obtain the following bound on the mean square error of the set $\{\mathcal{Q}_i\}_{i=1}^n$:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (x_i - \mathcal{Q}_i(x_i))\right\|^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathcal{Q}_i(x_i) - x_i\right\|^2\right] \leq \frac{(r-l)^2}{4n}.$$

□

A.1.2 Proof of Corollary 1

Proof. We apply the result of Theorem 1 coordinate-wise and sum the variances. □

A.2 ANALYSIS OF CORRELATED QUANTIZATION

A.2.1 Proof of Theorem 1

Proof. As shown in the proof of Theorem 2 in Suresh et al. [2022], $\mathbb{E}\left[\mathcal{Q}_i(a)\right] = a$ for all $a \in \mathbb{R}^d$, meaning that they are individually unbiased. We further analyze the variance in the homogeneous data regime.

We first show the result when $l = 0$ and $r = 1$, one can obtain the final result by rescaling the quantizer operation $(r-l)\cdot\mathcal{Q}_i\left(\frac{a_i}{r-l}\right)$.

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^n a - \sum_{i=1}^n \mathcal{Q}_i(a)\right)^2\right] &= \sum_{i=1}^n \mathbb{E}\left[\left(a-\mathcal{Q}_i(a)\right)^2\right] + \sum_{i=1}^n\sum_{j\neq i}\mathbb{E}\left[\left(a-\mathcal{Q}_i(a)\right)\left(a-\mathcal{Q}_j(a)\right)\right] \\
&= \sum_{i=1}^n a(1-a) + \sum_{i=1}^n\sum_{j\neq i}\mathbb{E}\left[\left(a-\mathcal{Q}_i(a)\right)\left(a-\mathcal{Q}_j(a)\right)\right] \\
&= na(1-a) + \sum_{i=1}^n\sum_{j\neq i}\left(\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right] - a^2\right)
\end{aligned}$$

where the second equality uses the fact that $\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}$ is a Bernoulli random variable with parameter a. We now calculate $\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right]$ for $i \neq j$:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right] &= \mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\frac{\pi_j}{n}+\gamma_j<a}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\frac{\pi_j}{n}+\gamma_j<a}\left(\mathbb{1}_{\pi_i>\pi_j} + \mathbb{1}_{\pi_j>\pi_i}\right)\right]
\end{aligned}$$

Notice that since $\delta_j < \frac{1}{n}$ then $\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\frac{\pi_j}{n}+\gamma_j<a}\mathbb{1}_{\pi_i>\pi_j} = \mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}$. Therefore,

$$\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right] = \mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right] + \mathbb{E}\left[\mathbb{1}_{\frac{\pi_j}{n}+\gamma_j<a}\mathbb{1}_{\pi_j>\pi_i}\right].$$

Let us calculate $\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right]$:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right] &= \sum_{k=0}^{n-1}\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\mathbb{1}_{\pi_i=k}\right] \\
&= \sum_{k=0}^{n-1}\mathbb{E}\left[\mathbb{1}_{n\gamma_i<na-k}\mathbb{1}_{k>\pi_j}\mathbb{1}_{\pi_i=k}\right] \\
&\overset{\text{indep}}{=} \sum_{k=0}^{n-1}\mathbb{E}\left[\mathbb{1}_{n\gamma_i<nx-k}\right]\mathbb{E}\left[\mathbb{1}_{k>\pi_j}\mathbb{1}_{\pi_i=k}\right] \\
&= \sum_{k=0}^{n-1}\mathbb{P}\left(\mathbb{1}_{n\gamma_i<na-k}\right)\mathbb{P}\left(\{k>\pi_j\}\cap\{\pi_i=k\}\right).
\end{aligned}$$

We have that

$$\mathbb{P}\left(\{k>\pi_j\}\cap\{\pi_i=k\}\right) = \mathbb{P}\left(\{k>\pi_j\}\,|\pi_i=k\right)\mathbb{P}\left(\pi_i=k\right) = \frac{k}{n\left(n-1\right)}.$$

Therefore,

$$\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right] = \frac{1}{n\left(n-1\right)}\sum_{k=0}^{n-1}k\mathbb{P}\left(\mathbb{1}_{n\gamma_i<na-k}\right).$$

We have that

$$\mathbb{P}\left(\mathbb{1}_{n\gamma_i<na-k}\right)=\begin{cases}1, & \text{if } k<\lfloor na\rfloor\\ na-\lfloor na\rfloor, & \text{if } k=\lfloor na\rfloor\\ 0, & \text{if } k>\lfloor na\rfloor.\end{cases}$$

With that we get

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right]&=\frac{1}{n\left(n-1\right)}\left(\sum_{k=0}^{\lfloor na\rfloor-1}k+\lfloor na\rfloor\left(na-\lfloor na\rfloor\right)\right)\\&=\frac{1}{n\left(n-1\right)}\left(\frac{\lfloor na\rfloor\left(\lfloor na\rfloor-1\right)}{2}+\lfloor na\rfloor\left(na-\lfloor na\rfloor\right)\right)\\&=\frac{\lfloor na\rfloor}{n\left(n-1\right)}\left(\frac{\lfloor na\rfloor-1}{2}+na-\lfloor na\rfloor\right)\\&=\frac{\lfloor na\rfloor}{n\left(n-1\right)}\left(na-\frac{\lfloor na\rfloor+1}{2}\right)\\&=\frac{na+\left(\lfloor na\rfloor-na\right)}{n\left(n-1\right)}\left(na-\frac{\lfloor na\rfloor+1}{2}\right)\\&=\frac{1}{2}\left(\frac{na^2}{(n-1)}-\frac{a}{n-1}+\frac{na-\lfloor na\rfloor}{n\left(n-1\right)}\left(\lfloor na\rfloor+1-na\right)\right).\end{aligned}$$

Let $c_a=\left(na-\lfloor na\rfloor\right)\left(\lfloor na\rfloor+1-na\right)$. We have that:

$$\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right]=\frac{1}{2}\frac{1}{n\left(n-1\right)}\left(n^2a^2-na+c_a\right)$$

Using the symmetry between i and j, we have that: $\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right]=\mathbb{E}\left[\mathbb{1}_{\frac{\pi_j}{n}+\gamma_j<a}\mathbb{1}_{\pi_j>\pi_i}\right]$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right]&=\mathbb{E}\left[\mathbb{1}_{\frac{\pi_i}{n}+\gamma_i<a}\mathbb{1}_{\pi_i>\pi_j}\right]+\mathbb{E}\left[\mathbb{1}_{\frac{\pi_j}{n}+\gamma_j<a}\mathbb{1}_{\pi_j>\pi_i}\right]\\&=\frac{1}{n\left(n-1\right)}\left(n^2a^2-na+c_a\right).\end{aligned}$$

Further,

$$\begin{aligned}\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right]-a^2&=\frac{1}{n\left(n-1\right)}\left(na^2-na+c_a\right)\\&=\frac{1}{n\left(n-1\right)}\left(-na\left(1-a\right)+c_a\right).\end{aligned}$$

Therefore, we have

$$\sum_{i=1}^{n}\sum_{j\neq i}\left(\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right]-a^2\right)=-na\left(1-a\right)+c_a.$$

With that we get,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n}a-\sum_{i=1}^{n}\mathcal{Q}_i(a)\right)^2\right]&=na(1-a)+\sum_{i=1}^{n}\sum_{j\neq i}\left(\mathbb{E}\left[\mathcal{Q}_i(a)\mathcal{Q}_j(a)\right]-a^2\right)\\&=na(1-a)-na\left(1-a\right)+c_a\\&=c_a\\&=\left(na-\lfloor na\rfloor\right)\left(\lfloor na\rfloor+1-na\right)\\&=\left(na-\lfloor na\rfloor\right)\left(1-\left(na-\lfloor na\rfloor\right)\right)\\&\leq\frac{1}{4}.\end{aligned}$$

We get the variance by dividing this equality by n^2. Therefore the variance of our quantizers is upper bounded by $\frac{1}{4n^2}$. $\square$

A.2.2 Proof of Corollary 2

Proof. We apply the result of Theorem 1 coordinate-wise and sum the variances. □

A.3 PROOF OF COROLLARY 3

Proof. Let us first show the amount of bits the compressor uses. First we notice that the image of the each block contains n/τ clients processing the same d/τ coordinates. Thus, quantized vectors require $32 + d/\tau$ bits per client, and the permutations require no extra communications, since they can be seeded.

We denote Correlated Quantization by $\mathcal{Q}$, PermK by $\mathcal{P}$.

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i\left(\mathcal{P}_i\left(a_i\right)\right)-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{P}_i\left(a_i\right)-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] \\
&\quad+\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i\left(\mathcal{P}_i\left(a_i\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\mathcal{P}_i\left(a_i\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i\left(\mathcal{P}_i\left(a_i\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\mathcal{P}_i\left(a_i\right)\right\|^2\right] \\
&\leq \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i\left(\mathcal{P}_i\left(a\right)\right)-a\right\|^2\right].
\end{aligned}$$

Notice that in the homogeneous case all the clients are equivalent, so the Correlated Quantization withing the image of each block can be percieved to use the same set of clients of size n/τ. Let us denote the set of coordinates attributed to the image of the k-th block as $a_{k,j} = a_k : j \in [d/\tau]$. Since Correlated Quantization is independent coordinate-wise, the square error is additive coordinate-wise and the images of the blocks do not intersect coordinate-wise, we can freely move the sum over the blocks in and out of the norm.

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Q}_i(\mathcal{P}_i(a))-a\right\|^2\right] &= \sum_{k=1}^{\tau}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n/\tau}\tau\mathcal{Q}_i\left(a_k\right)-a_k\right\|\right] \\
&= \mathbb{E}\left[\left\|\frac{\tau}{n}\sum_{i=1}^{n/\tau}\mathcal{Q}_i\left(a\right)-a\right\|\right] \\
&= \frac{d\left\|a\right\|^2}{(n/\tau)^2} \\
&= \frac{d\tau^2\left\|a\right\|^2}{n^2}.
\end{aligned}$$

□

B COMPLEXITY ANALYSIS

B.1 PROOF OF PROPOSITION 4

Proof. Szlendak et al. [2021] demonstrated that by integrating MARINA with a compressor that satisfies the AB-inequality and by choosing the stepsize

$$\gamma \leq \left(L_- + \sqrt{\frac{(1-p)}{p}\left((A-B)L_+^2 + BL_\pm^2\right)}\right)^{-1},$$

MARINA can identify a point $\widehat{x}^T$, for which $\mathbb{E}\left[\|f(\widehat{x}^T)\|^2\right] \leq \frac{2\Delta^0}{\gamma T}$. Notice, that in the homogeneous scenario, $L_- = L_+ = L$ and $L_\pm = 0$. Thus, we can rewrite the upper bound on the stepsize

$$\gamma \leq \frac{1}{L}\left(1+\sqrt{\frac{(1-p)}{p}(A-B)}\right)^{-1}.$$

Without quantization, each client will send d coordinates, each composed of 32 bits, which is $32d$ bits in total. With Correlated and Independent Quantizations, each client will send d bits, plus the gradient's norm (32 bits). So $32+d$ in total. In MARINA, the expected number of bits sent per client in each step is

$$p\,(32d) + (1-p)\,(32+d)\,.$$

To achieve an approximately stationary point $\widehat{x}$ such that $\mathbb{E}\left[\|f(\widehat{x})\|^2\right] \leq \varepsilon^2$, we require

$$T = \frac{2\Delta_0}{\varepsilon^2} L\left(1+\sqrt{\frac{1-p}{p}(A-B)}\right)$$

algorithm steps. Consequently, the overall communication complexity per client is:

$$\begin{aligned}
\mathcal{C}(p) &= \left(p\,(32d) + (1-p)\,(32+d)\right) T \\
&= \frac{2\Delta_0}{\varepsilon^2} L\left(p\,(32d) + (1-p)\,(32+d)\right)\left(1+\sqrt{\frac{(1-p)}{p}(A-B)}\right) \\
&= \underbrace{\frac{2\Delta_0}{\varepsilon^2} L\,(32d)}_{\text{GD Rate}} \underbrace{\left(p+(1-p)\,\frac{32+d}{32d}\right)\left(1+\sqrt{\frac{(1-p)}{p}(A-B)}\right)}_{\text{Improvement Factor}}.
\end{aligned}$$

Correlated Quantizers
$\{\mathcal{Q}_i\}_{i\in[n]} \in \mathbb{U}\left(\frac{d}{n^2}, 0\right)$, therefore:

$$\mathcal{C}_{\text{cor}}\,(p) = \frac{2\Delta_0}{\varepsilon^2} L\,(32d)\left(p+(1-p)\,\frac{32+d}{32d}\right)\left(1+\sqrt{\frac{(1-p)}{p}\frac{d}{4n^2}}\right).$$

Applying Lemma 12 from Szlendak et al. [2021], we obtain that $\mathcal{C}_{\text{cor}} = \mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1+\frac{d}{n}\right\}\right)$.

Independent Quantizers
$\{\mathcal{Q}_i\}_{i\in[n]} \in \mathbb{U}\left(\frac{d}{4n}, 0\right)$, therefore:

$$\mathcal{C}_{\text{ind}}\,(p) = \frac{2\Delta_0}{\varepsilon^2} L\,(32d)\left(p+(1-p)\,\frac{32+d}{32d}\right)\left(1+\sqrt{\frac{(1-p)}{p}\frac{d}{4n}}\right).$$

Applying Lemma 12 from Szlendak et al. [2021], we obtain that $\mathcal{C}_{\text{ind}} = \mathcal{O}\left(\frac{\Delta^0 L}{\varepsilon^2}\min\left\{d, 1+\frac{d}{\sqrt{n}}\right\}\right)$.

Clearly, $\forall p \in [0,1]$, $\mathcal{C}_{\text{cor}}\,(p) \leq \mathcal{C}_{\text{ind}}\,(p)$.

□

B.2 EXTENDED COMPLEXITY ANALYSIS: THE CASE OF N=D

Correlated Quantizers
Given $d = n \gg 1$, denoting $\frac{d}{4n^2} = \frac{1}{4n} = b$ and $\frac{32+d}{32d} = \frac{1}{32} + \frac{1}{n} = a$, the complexity can be simplified using the fact that

$b \to 0$ implies $p \to 0$:

$$\frac{-a\sqrt{b}}{2p^{3/2}} + (1-a)\left(1 + \sqrt{\frac{b}{p}}\right) = 0,$$
$$\frac{a\sqrt{b}}{2} = (1-a)\left(p^{3/2} + p\sqrt{b}\right),$$
$$p \approx \left(\frac{a}{2(1-a)}\right)^{2/3} b^{1/3}.$$

Substituting this into the Improvement Factor over GD, given by $\text{IF} = \frac{\mathcal{C}_{\text{cor}}}{\mathcal{C}_{\text{GD}}}$, we get:

$$\text{IF}_{\text{cor}} = a + b^{1/3}\left(\left(a(1-a)^2/2\right)^{2/3} + \left(2a^2(1-a)\right)^{2/3}\right) + o\left(b^{1/3}\right) =$$
$$= \frac{32+d}{32d} + o(1) = \frac{1}{32} + o(1) \approx 0.03125.$$

Thus, MARINA with Correlated Quantization demands approximately 0.03 times fewer bits communicated than Gradient Descent to find an ε-solution.

Uncorrelated Quantizers
By setting $d = n$, we denote $\frac{d}{4n} = \frac{1}{4} = b$ and $\frac{32+d}{32d} = \frac{1}{32} + \frac{1}{d} = a$. Therefore, the problem of finding the optimal p can be reduced to minimizing the function

$$\mathcal{C}(p) = \left(p + (1-p)\left(\frac{1}{32} + o(1)\right)\right)\left(1 + \sqrt{\frac{1}{4p} - \frac{1}{4}}\right).$$

Solving it numerically we get

$$\lim_{n=d\to\infty} p \approx 0.02105,$$

leading to

$$\text{IF}_{\text{ind}} \approx 0.2277.$$

Hence, MARINA with Independent Quantization requires approximately 0.23 times fewer bits communicated than Gradient Descent to find an ε-solution.

The speedup due to correlation is then

$$\frac{\text{IF}_{\text{ind}}}{\text{IF}_{\text{cor}}} \approx \frac{0.2277}{0.03125} \approx 7.29.$$

C IMPROVED ANALYSIS OF MARINA

Szlendak et al. [2021] analyzed the MARINA algorithm under the assumption of individual unbiasedness (4). This algorithm employs compressed vectors to compute their average. While the assumption of individual unbiasedness guarantees the unbiasedness of the average when using independent compressors, allowing for correlated compressors at times offers a guarantee of the average's unbiasedness even without the need for individual unbiasedness. Moreover, similar to Tyurin et al. [2022], we can further refine the assumption with weights, allowing for even more sophisticated compressors.

Assumption 6 (Weighted AB-Inequality [Tyurin et al., 2022])**.** *Consider a random mapping $\mathcal{S} : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}^d$ to which we refer as "combinatorial compressor", such that, for all $a_i \in \mathbb{R}^d,\ i \in [n]$, $\mathbb{E}\left[\mathcal{S}\left(a_1, \ldots, a_n\right)\right] = \frac{1}{n}\sum_{i=1}^n a_i$. Assume that there exist $A, B \geq 0$ and weights $w_1, \ldots, w_n \in \mathbb{R}_+ : \sum_{i=1}^n w_i = 1$, such that, for all $a_i \in \mathbb{R}^d,\ i \in [n]$,*

$$\mathbb{E}\left[\left\|\mathcal{S}\left(a_1, \ldots, a_n\right) - \frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right] \leq \frac{A}{n}\sum_{i=1}^n \frac{1}{nw_i}\left\|a_i\right\|^2 - B\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2.$$

The set of combinatorial compressors that satisfy this assumption is denoted by $\mathbb{S}\left(A, B, \{w_i\}_{i=1}^n\right)$.

Assumption 7. *Given a set of weights $w_1, \ldots, w_n \in \mathbb{R}_+ : \sum_{i=1}^n w_i = 1$, let $L_{+,w} \geq 0$ be the smallest constant such that $\frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_{+,w}^2 \|x-y\|^2$, for all $x, y \in \mathbb{R}^d$.*

Assumption 8. *Given a set of weights $w_1, \ldots, w_n \in \mathbb{R}_+ : \sum_{i=1}^n w_i = 1$, let $L_{\pm,w}^2$ be the smallest constant such that*

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 - \|\nabla f(x) - \nabla f(y)\|^2 \leq L_{\pm,w}^2 \|x-y\|^2, \quad x, y \in \mathbb{R}^d.$$

We refer to the quantity $L_{\pm,w}^2$ by the name of weighted Hessian variance.

We refine the analysis of MARINA under Assumption 6.

Theorem 2. *Suppose that $\mathcal{S}^t \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$, for all $t \in \mathbb{N}$, and that Assumptions 1, 6, 7 and 8 hold. Then, for all $T > 0$ and for the stepsize $0 < \gamma \leq \left(L_- + \sqrt{\frac{1-p}{p}\left((A-B)L_{+,w}^2 + BL_{\pm,w}^2\right)}\right)^{-1}$, the iterates produced by* MARINA *satisfy $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{2\Delta_0}{\gamma T}$ where $\Delta_0 = f(x^0) - f^*$ and $\widehat{x}^T$ is chosen uniformly at random from $x^0, x^1, \ldots, x^{T-1}$.*

This contribution fundamentally shares the same goal as the contribution with the analysis of MARINA with correlated quantizers: we replace the prevalent framework of individual standalone compressors found in existing literature with a framework of dependent compressors. In the first contribution, quantizers are correlated, whereas in the second contribution, the compressors are not necessarily individually unbiased, but their average is.

It was originally used for analyzing sampling schemes combined with the PAGE method [Li et al., 2021] in non-distributed optimization (see [Tyurin et al., 2022]). However, we employ it for compressors and improve the communication complexity of a different method, MARINA, which is used in distributed optimization. Our work demonstrates that parameters such as $L_{\pm,w}$ and $L_{+,w}$ play a pivotal role in influencing the convergence of this variance-reduced algorithm.

Furthermore, the AB-inequality, even when used independently, proves useful for simpler problems of MSE minimization. It decomposes the bound on the MSE in a natural way, allowing us to compare and analyze different sets of compressors, and it is generally tight.

Proof of Theorem 2. In the proof, we follow closely the analysis of [Gorbunov et al., 2022] and adapt it to utilize the power of weighted Hessian variance (Assumption 8) and weighted AB assumption (Assumption 6). We bound the term $\mathbb{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right]$ in a similar way to [Gorbunov et al., 2022], but make use of the weighted AB assumption. Other steps are essentially identical, but refine the existing analysis through weighted Hessian variance.

First, we recall the following lemmas.

Lemma 1 (Li et al. [2021])**.** *Suppose that L_- is finite and let $x^{t+1} = x^t - \gamma g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\gamma > 0$, we have*

$$f(x^{t+1}) \leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - x^t\right\|^2. \tag{2}$$

Lemma 2 (Richtárik et al. [2021])**.** *Let $a, b > 0$. If $0 \leq \gamma \leq \frac{1}{\sqrt{a}+b}$, then $a\gamma^2 + b\gamma \leq 1$. Moreover, the bound is tight up to the factor of 2 since $\frac{1}{\sqrt{a}+b} \leq \min\left\{\frac{1}{\sqrt{a}}, \frac{1}{b}\right\} \leq \frac{2}{\sqrt{a}+b}$.*

Next, we get an upper bound of $\mathbb{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2 \middle| x^{t+1}\right]$.

Lemma 3. *Let us consider g^{t+1} from Algorithm 1 and assume, that Assumptions 1, 4, 6, 7 and 8 hold, then*

$$\begin{aligned}\mathbb{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2 \middle| x^{t+1}\right] \quad &\leq \quad (1-p)\left((A-B)L_{+,w}^2 + BL_{\pm,w}^2\right)\left\|x^{t+1} - x^t\right\|^2 \\ &\quad + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2. \end{aligned} \tag{3}$$

Proof. In the view of definition of g^{t+1}, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|g^{t+1}-\nabla f(x^{t+1})\right\|^2\middle| x^{t+1}\right]\\
&=(1-p)\mathbb{E}\left[\left\|g^t+S^t\left(\{\nabla f_i(x^{t+1})-\nabla f_i(x^t)\}_{i=1}^n\right)-\nabla f(x^{t+1})\right\|^2\middle| x^{t+1}\right]\\
&=(1-p)\mathbb{E}\left[\left\|S^t\left(\{\nabla f_i(x^{t+1})-\nabla f_i(x^t)\}_{i=1}^n\right)-\nabla f(x^{t+1})+\nabla f(x^t)\right\|^2\middle| x^{t+1}\right]\\
&+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2.
\end{aligned}$$

In the last inequality we used unbiasedness of S^t. Next, from weighted AB inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|g^{t+1}-\nabla f(x^{t+1})\right\|^2\middle| x^{t+1}\right]\\
&\leq(1-p)\mathbb{E}\left[\left\|S^t\left(\{\nabla f_i(x^{t+1})-\nabla f_i(x^t)\}_{i=1}^n\right)-\nabla f(x^{t+1})+\nabla f(x^t)\right\|^2\middle| x^{t+1}\right]\\
&\quad+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2\\
&\leq(1-p)\left(\frac{A}{n}\left(\sum_{i=1}^n\frac{1}{nw_i}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2\right)-B\left\|\nabla f(x^{t+1})-\nabla f(x^t)\right\|^2\right)\\
&\quad+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2\\
&=(1-p)\Bigg((A-B)\left(\sum_{i=1}^n\frac{1}{n^2w_i}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2\right)\\
&\quad+B\left(\sum_{i=1}^n\frac{1}{n^2w_i}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2-\left\|\nabla f(x^{t+1})-\nabla f(x^t)\right\|^2\right)\Bigg)\\
&\quad+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2.
\end{aligned}$$

Using the definition of $L_{+,w}$ and $L_{\pm,w}$, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|g^{t+1}-\nabla f(x^{t+1})\right\|^2\right]&\leq(1-p)\left((A-B)L_{+,w}^2+BL_{\pm,w}^2\right)\left\|x^{t+1}-x^t\right\|^2\\
&\quad+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2.
\end{aligned}$$

□

We are ready to prove Theorem 2. Defining

$$\begin{aligned}
\Phi^t&\overset{\text{def}}{=}f(x^t)-f^{\inf}+\frac{\gamma}{2p}\left\|g^t-\nabla f(x^t)\right\|^2,\\
\widehat{L}^2&\overset{\text{def}}{=}(A-B)L_{+,w}^2+BL_{\pm,w}^2,
\end{aligned}$$

and using inequalities (2) and (3), we get

$$\begin{aligned}
\mathbb{E}\left[\Phi^{t+1}\right]&\leq\mathbb{E}\left[f(x^t)-f^{\inf}-\frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L_-}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\frac{\gamma}{2}\left\|g^t-\nabla f(x^t)\right\|^2\right]\\
&\quad+\frac{\gamma}{2p}\mathbb{E}\left[(1-p)\widehat{L}^2\left\|x^{t+1}-x^t\right\|^2+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2\right]\\
&=\mathbb{E}\left[\Phi^t\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\quad+\left(\frac{\gamma(1-p)\widehat{L}^2}{2p}-\frac{1}{2\gamma}+\frac{L_-}{2}\right)\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\\
&\leq\mathbb{E}\left[\Phi^t\right]-\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right],
\end{aligned}$$

where in the last inequality we use

$$\frac{\gamma(1-p)\widehat{L}^2}{2p} - \frac{1}{2\gamma} + \frac{L}{2} \leq 0,$$

following from the stepsize choice and Lemma 2.

Summing up inequalities $\mathbb{E}\left[\Phi^{t+1}\right] \leq \mathbb{E}\left[\Phi^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$ for $t = 0, 1, \ldots, T-1$ and rearranging the terms, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \quad \leq \quad \frac{2}{\gamma T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\Phi^t\right] - \mathbb{E}\left[\Phi^{t+1}\right]\right) = \frac{2\left(\mathbb{E}\left[\Phi^0\right] - \mathbb{E}\left[\Phi^T\right]\right)}{\gamma T} \leq \frac{2\Delta_0}{\gamma T},$$

since $g^0 = \nabla f(x^0)$ and $\Phi^T \geq 0$. Finally, using the tower property and the definition of $\hat{x}^T$ from Algorithm 4, we obtain the desired result.

Theorem 2 is proven.

C.1 EXAMPLE: DISTRIBUTED MEAN ESTIMATION

Distributed Mean Estimation algorithms are commonly assessed based on their MSE [Suresh et al., 2017, Mayekar and Tyagi, 2019, Vargaftik et al., 2021, Suresh et al., 2022], with it often being bounded by a factor of its input's average square norm. Naturally, such algorithms fit into Weighted AB-inequality with certain A, uniform weights $w_i = \frac{1}{n}$ and $B = 0$, allowing for their incorporation into MARINA, in accordance with Theorem 2.

C.2 EXAMPLE: IMPORTANCE SAMPLING

In this example, we consider a combinatorial compressor which is a composition of unbiased independent compressors verifying Definition 1, with importance sampling [Tyurin et al., 2022].

Let us recall the definition. Fix $\tau > 0$. For all $k \in [\tau]$, we define i.i.d. random variables

$$\chi_k = \begin{cases} 1 & \text{with probability } q_1 \\ 2 & \text{with probability } q_2 \\ & \vdots \\ n & \text{with probability } q_n, \end{cases}$$

where $(q_1, \ldots, q_n) \in \mathcal{S}^n$ (simple simplex). A sampling

$$S(a_1, \ldots, a_n) \overset{\text{def}}{=} \frac{1}{\tau}\sum_{k=1}^{\tau}\frac{a_{\chi_k}}{nq_{\chi_k}}$$

is called the Importance sampling. Using the result from Tyurin et al. [2022], we get:

$$\mathbb{E}\left[\left\|\frac{1}{\tau}\sum_{k=1}^{\tau}\frac{a_{\chi_k}}{nq_{\chi_k}} - \frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] = \frac{1}{\tau}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nq_i}\left\|a_i\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right).$$

In particular for $\tau = 1$ we get:

$$\mathbb{E}\left[\left\|S(a_1, \ldots, a_n) - \frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{nq_i}\left\|a_i\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2. \tag{4}$$

With this method, instead of all n clients participating in the compressed rounds, only one client (selected randomly based on its "importance") sends their compressed vector. Next we establish whether a composition of unbiased compressors with importance sampling satisfies Assumption 6.

Lemma 4. *Let us assume that an importance sampling function S satisfies* (4) *with probabilities* q_i*, and some random compressor* $\mathcal{Q}$ *satisfies Definition 1. Then*

$$\mathbb{E}\left[\left\|S\left(\mathcal{Q}\left(a_1\right),\ldots,\mathcal{Q}\left(a_n\right)\right)-\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]\leq\frac{(1+\omega)}{n^2}\sum_{i=1}^n\frac{1}{q_i}\left\|a_i\right\|^2-\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2.$$

Thus, a composition of unbiased independent compressors with importance sampling with $\tau=1$ *yields a combinatorial compressor* $S\in\mathbb{S}\left(\omega+1,1,\left\{\frac{L_i}{\sum_{i=1}^n L_i}\right\}_{i=1}^n\right)$.

Proof. Using tower property we have

$$\begin{aligned}\mathbb{E}\left[\left\|S\left(\mathcal{Q}\left(a_1\right),\ldots,\mathcal{Q}\left(a_n\right)\right)-\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]&=\mathbb{E}\left[\left\|S\left(\mathcal{Q}\left(a_1\right),\ldots,\mathcal{Q}\left(a_n\right)\right)-\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)\right\|^2\right]\\&\quad+\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)-\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right].\end{aligned}$$

Let us bound the second term:

$$\begin{aligned}&\mathbb{E}\left[\left\|S\left(\mathcal{Q}\left(a_1\right),\ldots,\mathcal{Q}\left(a_n\right)\right)-\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)\right\|^2\right]\\&\leq\frac{1}{n}\sum_{i=1}^n\frac{1}{nq_i}\mathbb{E}\left[\left\|\mathcal{Q}\left(a_i\right)\right\|^2\right]-\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)\right\|^2\right]\\&=\frac{1}{n}\sum_{i=1}^n\frac{1}{nq_i}\mathbb{E}\left[\left\|\mathcal{Q}\left(a_i\right)-a_i\right\|^2\right]+\frac{1}{n}\sum_{i=1}^n\frac{1}{nq_i}\left\|a_i\right\|^2-\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)\right\|^2\right]\\&\leq(1+\omega)\frac{1}{n}\sum_{i=1}^n\frac{1}{nq_i}\left\|a_i\right\|^2-\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)\right\|^2\right]\\&=(1+\omega)\frac{1}{n}\sum_{i=1}^n\frac{1}{nq_i}\left\|a_i\right\|^2-\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)-a_i\right\|^2\right]-\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\left\|S\left(\mathcal{Q}\left(a_1\right),\ldots,\mathcal{Q}\left(a_n\right)\right)-\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]&=\mathbb{E}\left[\left\|S\left(\mathcal{Q}\left(a_1\right),\ldots,\mathcal{Q}\left(a_n\right)\right)-\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)\right\|^2\right]\\&\quad+\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\mathcal{Q}\left(a_i\right)-\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]\\&\leq\frac{(1+\omega)}{n^2}\sum_{i=1}^n\frac{1}{q_i}\left\|a_i\right\|^2-\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2.\end{aligned}$$

□

Theorem 3. *Let Assumptions 1, 7, 8 hold. Let Assumption 3 hold for all* f_i *with* L_i, $i\in[n]$*. Given combinatorial compressors based on importance sampling* $\mathcal{S}\in\mathbb{S}\left(\omega+1,1,\left\{\frac{L_i}{\sum_{i=1}^n L_i}\right\}_{i=1}^n\right)$*, assume that* $0<\gamma\leq\left(L_-+L_{avg}\sqrt{\frac{1-p}{p}\left(\omega+1\right)}\right)^{-1}$

where $L_{avg} = \frac{1}{n}\sum_{i=1}^{n} L_i$. *Then for all* $T \geq 0$ *the iterates produced by* MARINA *satisfy* $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{2\Delta_0}{\gamma T}$, *where* $\Delta_0 = f(x^0) - f^*$ *and* $\widehat{x}^T$ *is chosen uniformly at random from* $x^0, x^1, \ldots, x^{T-1}$.

Proof of Theorem 3. From [Tyurin et al., 2022, Section F], we know that by setting $q_i = \frac{L_i}{\sum_{i=1}^n L_i}$ for importance sampling, we obtain a sampling with $L_{+,w}^2 = L_{\pm,w}^2 = \left(\frac{1}{n}\sum_{i=1}^n L_i\right)^2$ satisfying Assumptions 7 and 8 with $w_i = \frac{L_i}{\sum_{i=1}^n L_i}$. The proof is then complete by applying Theorem 2 and Lemma 4.

Theorem 3 is proven.

Corollary 4. *Suppose assumptions of Theorem 3 hold. Then the communication complexity of a run of* MARINA *method with importance sampling combinatorial compressors in order to reach an approximately stationary point is upper bounded by* $\mathcal{O}\left(\frac{\Delta^0}{\varepsilon^2}\min\left\{dL_-, \frac{dL_-}{n} + \frac{d\sqrt{\omega+1}L_{avg}}{\sqrt{n}}\right\}\right)$.

Proof of Corollary 4. To get an ε-solution it's sufficient to have T iterations such that:

$$\frac{2\Delta_0}{\gamma T} < \varepsilon^2 \iff \frac{2\Delta_0}{\gamma\varepsilon^2} < T.$$

By taking $\gamma = \left(L_- + L_{avg}\sqrt{\frac{1-p}{p}(\omega+1)}\right)^{-1}$, we get $T > \frac{2\Delta_0}{\varepsilon^2}\left(L_- + L_{avg}\sqrt{\frac{1-p}{p}(\omega+1)}\right)$.

Since we're doing an importance sampling with $\tau = 1$, each round, with a probability $1-p$, only one client sends, on average, $\beta = \mathcal{O}(1)$ bits per coordinate. So the number of bits sent by round by a client is on average: $\left((1-p)\frac{\beta}{n} + 32p\right)d$. The total complexity over all iterations is:

$$\left((1-p)\frac{\beta}{n} + 32p\right)d \times T = \frac{2\Delta_0}{\varepsilon^2}d\left((1-p)\frac{\beta}{n} + 32p\right)\left(L_- + L_{avg}\sqrt{\frac{1-p}{p}(\omega+1)}\right).$$

In particular if we take $p = \frac{1}{32n}$ we get a complexity of $\mathcal{O}\left(\frac{\Delta^0}{\varepsilon^2}\left(\frac{dL_-}{n} + \frac{d\sqrt{\omega+1}L_{avg}}{\sqrt{n}}\right)\right)$ and if we take $p = 1$ we get a communication complexity $\mathcal{O}\left(\frac{\Delta^0}{\varepsilon^2}dL_-\right)$.

Therefore, the communication complexity is upper-bounded by $\mathcal{O}\left(\frac{\Delta^0}{\varepsilon^2}\min\left\{dL_-, \frac{dL_-}{n} + \frac{d\sqrt{\omega+1}L_{avg}}{\sqrt{n}}\right\}\right)$.

Corollary 4 is proven.

Since L_{avg} can be $\sqrt{n}$ times smaller than L_+, MARINA with importance sampling can converge up to $\sqrt{n}$ times faster than the original method.

C.3 EXPERIMENTS: WEIGHTED MARINA

We synthesized various quadratic optimization tasks with different smoothness constants L_i (see Figure 6). We choose $d = 1024$, $n = 128$, the regularization $\lambda = 0.001$, and the noise scale $s \in \{0.0, 10.0\}$. We generated tasks so that the difference between $\max_i L_i$ and $\min_i L_i$ increases. Our experiments show that in various regimes MARINA combined with ISCC based on DRIVE has lower communication complexity than MARINA simply combined with DRIVE.

D ADDITIONAL EXPERIMENTS DETAILS

D.1 DESCRIPTION OF COMPRESSORS

Table 4 provides a comparative analysis of the compressors used in all the experiments. Notably, DRIVE transmits an equivalent number of bits per coordinate as CQ and IQ, given $d = 2^k$ for some integer k. To ensure a balanced comparison, we aim to choose d as a power of 2 wherever possible.

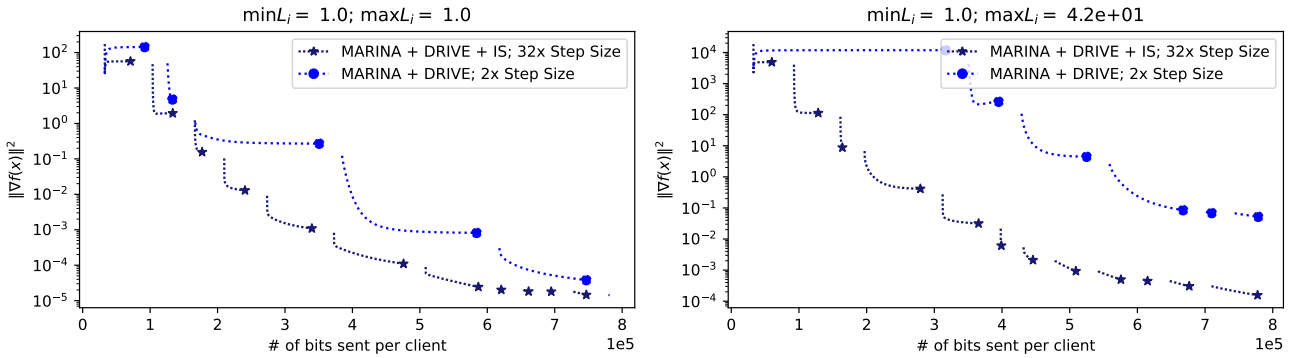

Figure 6: Comparison of DRIVE with or without Importance Sampling (IS) with MARINA on quadratic optimization tasks with diverse $L_{\pm}$ values

Algorithm 4 MARINA with combinatorial compression

Input: initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, probability $p \in (0, 1]$, number of iterations T
$g^0 = \nabla f(x^0)$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $c_t \sim \text{Bern}(p)$
  $x^{t+1} = x^t - \gamma g^t$
  Generate a combinatorial compressor $\mathbf{S}^t$
  $g^{t+1} = \nabla f(x^{t+1})$ if $c_t = 1$, and $g^{t+1} = g^t + \mathbf{S}^t\left(\{\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\}_{i=1}^n\right)$ otherwise
end for

D.2 OPTIMAL SELECTION OF PARAMETERS

D.2.1 Identifying the Optimal Probability p

Our objective is to determine the optimal probability p_{opt} that reduces the communication complexity to its minimum. This is described by the equation:

$$\mathcal{C}(p) = \frac{2\Delta_0}{\varepsilon^2}\left(32dp + \alpha\left(1-p\right)\right)\left(L_- + \sqrt{\frac{(1-p)}{p}\left((A-B)L_+^2 + BL_{\pm}^2\right)}\right),$$

where α represents the expected total number of bits communicated to the server during the compressed round of MARINA. When $B = 0$, which is the case for quantization, the expression simplifies to:

$$\mathcal{C}(p) = \frac{2\Delta_0}{\varepsilon^2}\left(32dp + \alpha\left(1-p\right)\right)\left(L_- + L_+\sqrt{\frac{(1-p)}{p}A}\right).$$

We solve this problem numerically. This takes into account each problem's L_+, L_-, and each compressor's α, A to obtain p_{opt}. Specifically for CQ and its variants where $L_{\pm} \neq 0$, A isn't explicitly defined and we extrapolate the equations from the zero-Hessian-variance regime.

D.2.2 Optimization of the Step Size

Having determined the value of p, we proceed to increase the step size. We increment the step size in multiples of 2 ($2, 4, 8$, etc.) of the theoretically optimal step size. Our aim is to identify the step size that ensures the algorithm's best performance at $4 \cdot 10^6$ bits communicated from each client to the server. That number was chosen as sufficiently large to demonstrate relative convergence between different algorithms. The convergence plots, as well as details about the selected optimal step sizes can be found in Figure 7.

Table 4: Comparison of the compressors used: A and B constants from AB-inequality 5 and the number of bits sent per client

Compressor	A	B	# of bits per client
CQ (Definition 7)	$d/(4n^2)$	0	$32+d$
IQ (Definition 6)	$d/(4n)$	0	$32+d$
DRIVE [Vargaftik et al., 2021]	$(\pi/2-1)/n$	0	$32+2^{\lceil \log_2 d \rceil}$
No compression	0	0	$32\cdot d$

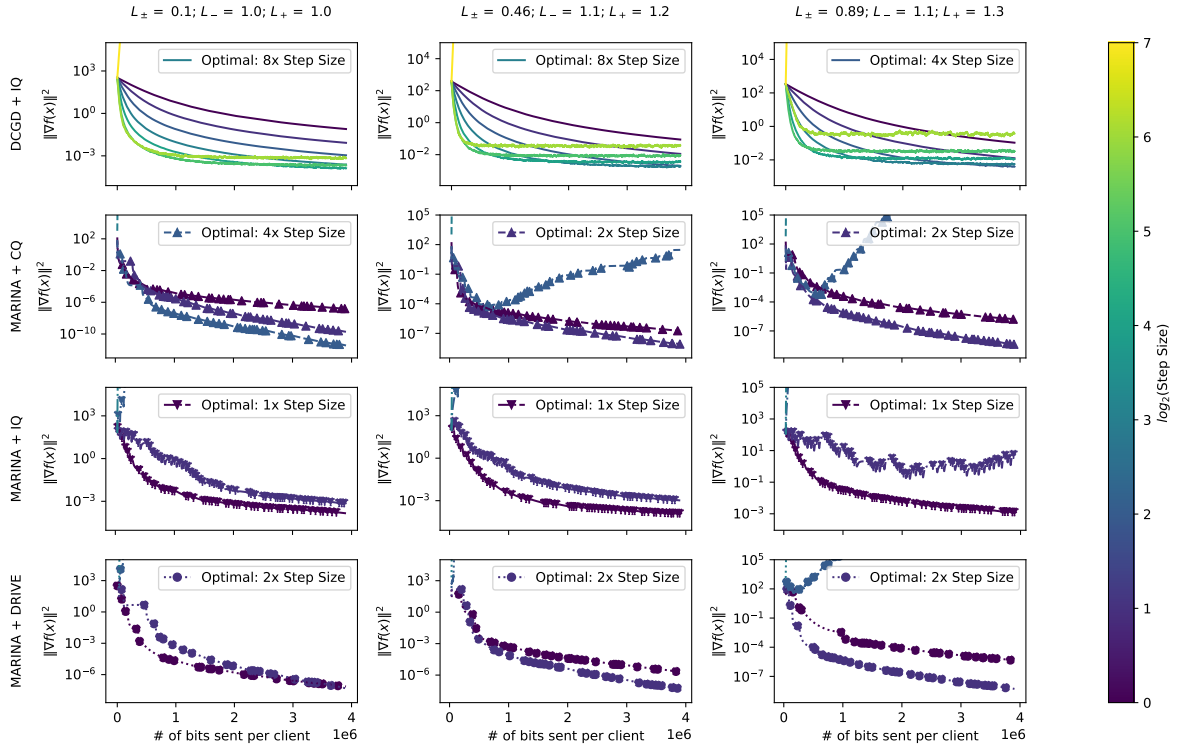

Figure 7: Convergence of CQ, IQ, and DRIVE with MARINA with different step sizes on quadratic optimization tasks with diverse $L_\pm$ values

In Figure 8 we provide the optimal stepsize selection procedure for the MLP classifier experiment on the a9a dataset, involving 131 clients. The largest step size was chosen such that the median of five optimization runs still converged.

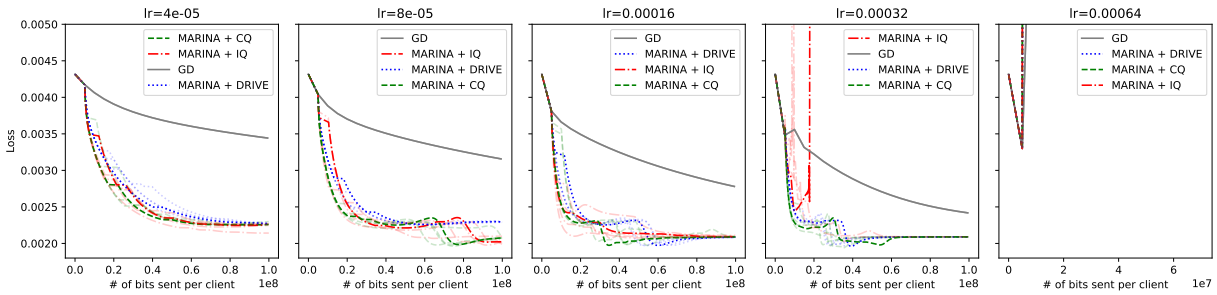

Figure 8: The optimal step size selection procedure for the MLP optimization problem. The largest step size was selected for which the median out of five optimization runs still converged.

Algorithm 5 GENERATE QUADRATIC OPTIMIZATION TASK WITH CONTROLLED $L_{\pm}$ [Tyurin et al., 2022]

Input: number nodes n, dimension d, regularizer λ, and noise scale s.

for $i = 1$ to n **do**
  Generate random noises $\nu_i^s = 1 + s\xi_i^s$ and $\nu_i^b = s\xi_i^b$, i.i.d $\xi_i^s, \xi_i^b \sim \text{NormalDistribution}(0, 1)$.
  Take vector $b_i = \frac{\nu_i^s}{4}\left(-1 + \nu_i^b, 0, \ldots, 0\right) \in \mathbb{R}^d$.
  Take the initial tridiagonal matrix

$$\mathbf{A}_i = \frac{\nu_i^s}{4}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{d\times d}.$$

end for
Take the mean of matrices $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i$.
Find the minimum eigenvalue $\lambda_{min}(\mathbf{A})$.
for $i = 1$ to n **do**
  Update matrix $\mathbf{A}_i = \mathbf{A}_i + (\lambda - \lambda_{min}(\mathbf{A}))\mathbf{I}$.
end for
Take starting point $x^0 = (\sqrt{d}, 0, \ldots, 0)$.
Output: matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$, vectors $b_1, \ldots, b_n$, starting point x^0.

Algorithm 6 GENERATE QUADRATIC OPTIMIZATION TASK WITH CONTROLLED L_i [Tyurin et al., 2022]

Input: number nodes n, dimension d and noise scale s.

for $i = 1$ to n **do**
  Generate random noises $\nu_i^s = 1 + s\xi_i^s$, i.i.d $\xi_i^s \sim \text{ExponentialDistribution}(1)$.
  Generate random noises $\nu_i^b = s\xi_i^b$, i.i.d $\xi_i^b \sim \text{NormalDistribution}(0, 1)$.
  Take vector $b_i = \left(-1 + \nu_i^b, 0, \ldots, 0\right) \in \mathbb{R}^d$.
  Take the initial tridiagonal matrix

$$\mathbf{A}_i = \frac{\nu_i^s}{4}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{d\times d}.$$

end for
Take starting point $x^0 = (\sqrt{d}, 0, \ldots, 0)$.
Output: matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$, vectors $b_1, \ldots, b_n$, starting point x^0.

D.3 QUADRATIC OPTIMIZATION TASKS GENERATION

Similar to Tyurin et al. [2022], we provide the algorithms used to generate artificial quadratic optimization tasks. Algorithm 5 and Algorithm 6 allow us to control the smoothness constants $L_{\pm}$ and L_i, respectively, via the noise scales.

E AUXILIARY FACTS

E.1 PROOF OF PROPOSITION 2

Proof. Since all $f_i(x), i \in [n]$, are equal to $f(x)$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(y)\|^2 = \|\nabla f(x) - \nabla f(y)\|^2, \quad x, y \in \mathbb{R}^d. \tag{5}$$

It immediately implies that the Hessian variance $L_{\pm}^2$ is equal to zero.

In case when functions are identical up to some random linear perturbation, assume that, for $i \in [n]$, $f_i(x) = \varphi(x) + b_i^\top x + c_i$, where $\varphi(x) : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function, $b_i \in \mathbb{R}^d$, $c_i \in \mathbb{R}$. Observe that in this case (5) also holds true, and, therefore, $L_{\pm}^2 = 0$. □