Distributed Learning with Compressed Gradient Differences

Konstantin Mishchenko [1] **Eduard Gorbunov** [2] **Martin Takáč** [3] **Peter Richtárik** [1 4 2]

Abstract

Training very large machine learning models requires a distributed computing approach, with communication of the model updates often being the bottleneck. For this reason, several methods based on the compression (e.g., sparsification and/or quantization) of the updates were recently proposed, including QSGD (Alistarh et al., 2017), TernGrad (Wen et al., 2017), SignSGD (Bernstein et al., 2018), and DQGD (Khirirat et al., 2018). However, none of these methods are able to learn the gradients, which means that they necessarily suffer from several issues, such as the inability to converge to the true optimum in the batch mode, inability to work with a nonsmooth regularizer, and slow convergence rates. In this work we propose a new distributed learning method—DIANA—which resolves these issues via compression of *gradient differences*. We perform a theoretical analysis in the strongly convex and nonconvex settings and show that our rates are vastly superior to existing rates. Our analysis of block-quantization and differences between ℓ_2 and ℓ_∞ quantization closes the gaps in theory and practice. Finally, by applying our analysis technique to TernGrad, we establish the first convergence rate for this method.

1. Introduction

Big machine learning models are typically trained in a distributed fashion. In this paradigm, the training data is distributed across several workers (e.g., nodes of a cluster), all of which compute in parallel an update to the model based on their local data. For instance, they can all perform a single step of Gradient Descent (GD) or Stochastic Gradient Descent (SGD). These updates are then sent to a parameter server which performs aggregation (typically this means just averaging of the updates) and then broadcasts the aggregated updates back to the workers. The process is repeated until a good solution is found.

When doubling the amount of computational power, one usually expects to see the learning process finish in half time. If this is the case, the considered system is called to scale linearly. For various reasons, however, this does not happen, even to the extent that the system might become slower with more resources. At the same time, the surge of big data applications increased the demand for distributed optimization methods, often requiring new properties such as ability to find a sparse solution. It is, therefore, of great importance to design new methods that are versatile, efficient and scale linearly with the amount of available resources.

In fact, the applications vary a lot in their desiderata. There is a rising interest in federated learning (Konečný et al., 2016), where the main concerns include the communication cost and ability to use local data only in an attempt to provide a certain level of privacy. In high-dimensional machine learning problems, nonsmooth ℓ_1-penalty is often utilized, so one wants to have a support for proximable regularization. The efficiency of deep learning, in contrast, is dependent on heavy-ball momentum and nonconvex convergence to criticality, while sampling from the full dataset might not be an issue. In our work, we try to address all of these questions correspondingly.

1.1. Communication as the bottleneck

The the main aspects of distributed optimization efficiency are computational and communication complexity. In general, evaluating full gradients is intractable due to time and memory restrictions, so computation is made cheap by employing stochastic updates. On the other hand, in typical distributed computing architectures, communication is much slower (see Figure 1 for our experiments with communication cost of aggregating and broadcasting) than a stochastic update, and the design of a training algorithm needs to find a trade-off between them. There have been considered several ways of dealing with this issue.

One of the early approaches is to have each worker perform a block coordinate descent step (Richtárik and Takáč, 2016; Fercoq et al., 2014). By choosing the size of the

[1] King Abdullah University of Science and Technology, Kingdom of Saudi Arabia [2] Moscow Institute of Physics and Technology, Russian Federation [3] Lehigh University, USA [4] University of Edinburgh, United Kingdom. Correspondence to: Peter Richtárik <peter.richtarik@kaust.edu.sa>.

January 26, 2019

block, one directly chooses the amount of data that needs to be communicated. An alternative idea is for each worker to do more work between communication rounds (e.g., by employing a more powerful local solver, such as a second order method), so that computation roughly balances out with communication. The key methods in this sphere include CoCoA and its variants (Jaggi et al., 2014; Ma et al., 2015; 2017b;a; Smith et al., 2018), DANE (Shamir et al., 2014), DiSCO (Zhang and Xiao, 2015; Ma and Takáč, 2015), DANCE (Jahani et al., 2018) and AIDE (Reddi et al., 2016).

1.2. Update compression via randomized sparsification and/or quantization

Practitioners also suggested a number of heuristics to find a remedy for the communication botlleneck. Of special interest to this paper is the idea of compressing SGD updates, proposed by Seide et al. (2014). Building off of this work, Alistarh et al. (2017) designed a variant of SGD that guarantees convergence with compressed updates. Other works with SGD update structure include (Konečný and Richtárik, 2016; Bernstein et al., 2018; Khirirat et al., 2018). Despite proving a convergence rate, (Alistarh et al., 2017) also left many new questions open and introduced an additional, unexplained, heuristic of quantizing only vector blocks. Moreover, their analysis implicitly makes an assumption that all data should be available to each worker, which is hard and sometimes even impossible to satisfy.

In a concurrent with (Alistarh et al., 2017) work (Wen et al., 2017), the Terngrad method was analyzed for stochastic updates that in expectation have positive correlation with the vector pointing to the solution. While giving more intuition about convergence of quantized methods, this work used ℓ_∞ norm for quantization, unlike ℓ_2-quantization of (Alistarh et al., 2017). This introduces another question of which norm is better suited for applications, which we also aim to answer in this work.

1.3. The problem

In this paper we focus on the problem of training a machine learning model via regularized empirical risk minimization:

$$\min_{x\in\mathbb{R}^d} f(x) + R(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x). \qquad (1)$$

Above, $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed convex regularizer, and $f_i : \mathbb{R}^d \to \mathbb{R}$ is the loss of model x obtained on data points belonging to distribution $\mathcal{D}_i$:

$$f_i(x) \overset{\text{def}}{=} \mathbf{E}_{\zeta\sim\mathcal{D}_i}\phi(x,\zeta).$$

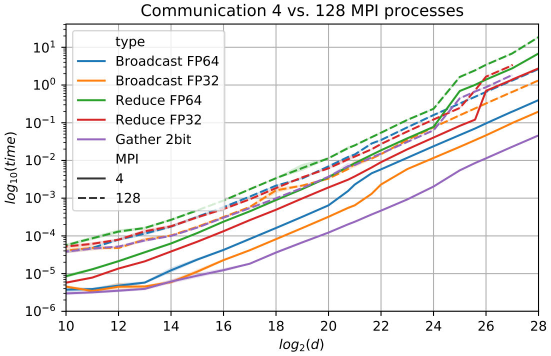

Figure 1. Typical communication cost using broadcast, reduce and gather for 64 and 32 FP using 4 (solid) resp 128 (dashed) MPI processes. See suppl. material for details about the network.

method	lin. rate	loc. data	non-smooth	momentum
DIANA (New!)	√	√	√	√
QSGD (Alistarh et al., 2017)	×	×	×	×
TernGrad (Wen et al., 2017)	×	×	×	×
DQGD (Khirirat et al., 2018)	√	√	×	×
QSVRG (Alistarh et al., 2017)	√	√	×	×

Table 1. Comparison of DIANA and related methods. Here “lin. rate” means that authors of corresponding paper prove linear convergence to either ball at the optimum or to the optimum itself, “loc. data” describes whether or not authors assume that f_i is available at node i only, “non-smooth” means support for a non-smooth regularizer, and “momentum” says whether or not authors consider momentum in their algorithm.

Note that we do not assume any kind of similarity between distributions $\mathcal{D}_1, \ldots, \mathcal{D}_n$. In contrast, we allow them to be arbitrary meaning that functions $f_1, \ldots, f_n$ may have completely different minimizers. Furthermore, we will assume that every f_i has its own smoothness parameter L_i, so it is permitted to use utterly non-uniform distributions. As a special case where $n = 1$, we recover the framework of one distribution, consideration of which leads to a purely stochastic method.

1.4. Notation

By $\text{sign}(t)$ we denote the sign of $t \in \mathbb{R}$ (-1 if $t < 0$, 0 if $t = 0$ and 1 if $t > 0$). The jth element of a vector $x \in \mathbb{R}^d$ is denoted as $x_{(j)}$. For $x = (x_{(1)}, \ldots, x_{(d)}) \in \mathbb{R}^d$ and $p \geq 1$, the ℓ_p norm of x is $\|x\|_p = (\sum_i |x_{(i)}|^p)^{1/p}$. Note that $\|x\|_1 \geq \|x\|_p \geq \|x\|_\infty$ for all x. By $\|x\|_0$ we denote the number of nonzero elements of x. For the detailed description of the notation see the Table 5 in the appendix.

2. Contributions

DIANA. We develop a distributed gradient-type method with compression of *gradient differences*, which we call DIANA (Algorithm 1).

Rate in the strongly convex case. We show that when applied to the smooth strongly convex minimization problem with arbitrary closed convex regularizer, DIANA has the iteration complexity $O\left(\max\left\{\sqrt{\frac{d}{m}}, \kappa\left(1+\frac{1}{n}\sqrt{\frac{d}{m}}\right)\right\}\ln\frac{1}{\varepsilon}\right)$, to a ball with center at the optimum (see Section 4.2, Theorem 2 and Corollary 1 for the details). In the case of decreasing stepsize we show $O\left(\frac{1}{\varepsilon}\right)$ iteration complexity (see Section 4.6, Theorem 5 and Corollary 2 for the details). Unlike in (Khirirat et al., 2018), in a noiseless regime our method converges to the exact optimum, and at a linear rate.

Rate in the non-convex case. We prove that DIANA also works for smooth non-convex problems with an indicator-like regularizer and get the iteration complexity $O\left(\frac{1}{\varepsilon^2}\max\left\{\frac{L^2(f(x^0)-f^*)^2}{n^2\alpha_p^2}, \frac{\sigma^4}{(1+n\alpha_p)^2}\right\}\right)$ (see Section 5, Theorem 4 and Corollary 3 for the details).

DIANA with momentum. We study momentum version of DIANA for the case of smooth non-convex objective with constant regularizer and $f_i = f$ (see Section J, Theorem 7 and Corollary 7 for the details).

First rate for Terngrad. We provide first convergence rate of TernGrad and provide new tight analysis of 1-bit QSGD under less restrictive assumptions for both smooth strongly convex objectives with arbitrary closed convex regularizer and non-convex objective with indicator-like regularizer (see Section 3.6 for the detailed comparison). Both of these methods are just special cases of our Algorithm 2 which is also a special case of Algorithm 1 with $\alpha = 0$ and $h_i^0 = 0$ for all i. We show that Algorithm 2 has $O\left(\frac{\kappa}{n\alpha_p}\right)$ iteration complexity of convergence to the ball with center at the optimum in the case of the smooth strongly convex minimization problem with arbitrary closed convex regularizer (see Theorem 10) and $O\left(\frac{1}{\varepsilon^2}\max\left\{\frac{L^2(f(x^0)-f(x^*))^2}{n^2\alpha_p^2}, \frac{\sigma^4}{(1+n\alpha_p)^2}\right\}\right)$ in the case of non-convex minimization problem with indicator-like regularizer (see Theorem 8 and Corollary 8).

QSGD and Terngrad with momentum. We study momentum version of DIANA for $\alpha = 0, h_i^0 = 0$ and, in particular, we propose momentum versions of (1-bit) QSGD and TernGrad the case of smooth non-convex objective with constant regularizer and $f_i = f$ (see Section K.3, Theorem 9 and Corollary 9 for the details).

Optimal norm power. We find the answer for the following question: *which ℓ_p norm to use for quantization in order to get the best iteration complexity of the algorithm?* It is easy to see that all the bounds that we propose depend on $\frac{1}{\alpha_p}$ where α_p is an increasing function of $1 \le p \le \infty$ (see Lemma 2 for the details). That is, *for both Algorithm 1 and 2 the iteration complexity reduces when p is growing and the best iteration complexity for our algorithms is achieved for $p = \infty$.* It implies that TernGrad has better iteration complexity than 1-bit QSGD.

Algorithm 1 DIANA (n nodes)

```
input learning rates $\alpha > 0$ and $\{\gamma^k\}_{k\ge 0}$, initial vectors $x^0, h_1^0, \dots, h_n^0 \in \mathbb{R}^d$ and $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$, quantization parameter $p \ge 1$, sizes of blocks $\{d_l\}_{l=1}^m$, momentum parameter $0 \le \beta < 1$
$v^0 = \nabla f(x^0)$
for $k = 0, 1, \dots$ do
    Broadcast $x^k$ to all workers
    for $i = 1, \dots, n$ in parallel do
        Sample $g_i^k$ such that $\mathbf{E}[g_i^k \mid x^k] = \nabla f_i(x^k)$
        $\Delta_i^k = g_i^k - h_i^k$
        Sample $\hat{\Delta}_i^k \sim \mathrm{Quant}_p(\Delta_i^k, \{d_l\}_{l=1}^m)$
        $h_i^{k+1} = h_i^k + \alpha\hat{\Delta}_i^k$
        $\hat{g}_i^k = h_i^k + \hat{\Delta}_i^k$
    end for
    $\hat{\Delta}^k = \frac{1}{n}\sum_{i=1}^n \hat{\Delta}_i^k$
    $\hat{g}^k = \frac{1}{n}\sum_{i=1}^n \hat{g}_i^k = h^k + \hat{\Delta}^k$
    $v^k = \beta v^{k-1} + \hat{g}^k$
    $x^{k+1} = \mathrm{prox}_{\gamma^k R}\left(x^k - \gamma^k v^k\right)$
    $h^{k+1} = \frac{1}{n}\sum_{i=1}^n h_i^{k+1} = h^k + \alpha\hat{\Delta}^k$
end for
```

First analysis for block-quantization. We give a first analysis of block-quantization (i.e. bucket-quantization). It was only mentioned in the paper (Alistarh et al., 2017) that it is possible to get better convergence via block-quantization, but the authors do not have rigorous analysis of block-quantization. So, we close this gap.

We summarize a few key features of our complexity results established in Table 1.

3. The Algorithm

In this section we describe our main method—DIANA. However, we first need to introduce several key concepts and ingredients that come together to make the algorithm.

3.1. Stochastic gradients

In each iteration k of DIANA, each node will sample an unbiased estimator of the local gradient. We assume that these gradients have bounded variance.

Assumption 1 (Stochastic gradients)**.** *For every $i = 1, 2, \dots, n$, $\mathbf{E}[g_i^k \mid x^k] = \nabla f_i(x^k)$. Moreover, the variance is bounded:*

$$\mathbf{E}\|g_i^k - \nabla f_i(x^k)\|_2^2 \le \sigma_i^2. \qquad (2)$$

Note that $g^k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n g_i^k$ is an unbiased estimator of $\nabla f(x^k)$:

$$\mathbf{E}[g^k \mid x^k] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^k) = \nabla f(x^k). \qquad (3)$$

Moreover, by independence of the random vectors $\{g_i^k - \nabla f_i(x^k)\}_{i=1}^n$, its variance is bounded above by

$$\mathbf{E}\left[\|g^k - \nabla f(x^k)\|_2^2 \mid x^k\right] \le \frac{\sigma^2}{n}, \qquad (4)$$

where $\sigma^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

3.2. Quantization

`DIANA` applies random compression (quantization) to gradient differences, which are then communicated to a parameter server. We now define the random quantization transformations used. Our first quantization operator transforms a vector $\Delta \in \mathbb{R}^d$ into a random vector $\hat{\Delta} \in \mathbb{R}^d$ whose entries belong to the set $\{-t, 0, t\}$ for some $t > 0$.

Definition 1 (p-quantization)**.** *Let $\Delta \in \mathbb{R}^d$ and let $p \ge 1$. If $\Delta = 0$, we define $\widetilde{\Delta} = \Delta$. If $\Delta \ne 0$, we define $\widetilde{\Delta}$ by setting*

$$\widetilde{\Delta}_{(j)} = \|\Delta\|_p \mathrm{sign}(\Delta_{(j)})\xi_{(j)}, \quad j = 1, 2, \dots, d, \qquad (5)$$

where $\xi_{(j)} \sim \mathrm{Be}\left(|\Delta_{(j)}|/\|\Delta\|_p\right)$ are Bernoulli random variables[1]*. Note that*

$$\widetilde{\Delta} = \|\Delta\|_p \,\mathrm{sign}(\Delta) \circ \xi, \qquad (6)$$

where sign *is applied elementwise, and $\circ$ denotes the Hadamard (i.e. elementwise) product. We say that $\widetilde{\Delta}$ is p-quantization of Δ. When sampling $\widetilde{\Delta}$, we shall write $\widetilde{\Delta} \sim \mathrm{Quant}_p(\Delta)$.*

In addition, we consider a block variant of p-quantization operators.

Definition 2 (block-p-quantization)**.** *Let $\Delta = (\Delta(1)^\top, \Delta(2)^\top, \dots, \Delta(m)^\top)^\top \in \mathbb{R}^d$, where $\Delta(1) \in \mathbb{R}^{d_1}, \dots, \Delta(m) \in \mathbb{R}^{d_m}$, $d_1 + \ldots + d_m = d$ and $d_l > 1$ for all $l = 1, \dots, m$. We say that $\hat{\Delta}$ is p-quantization of Δ with sizes of blocks $\{d_l\}_{l=1}^m$ and write $\hat{\Delta} \sim \mathrm{Quant}_p(\Delta, \{d_l\}_{l=1}^m)$ if $\hat{\Delta}(l) \sim \mathrm{Quant}_p(\Delta)$ for all $l = 1, \dots, m$.*

In other words, we quantize subvectors called *blocks* of the initial vector. Note that in the special case when $m = 1$ we get full quantization: $\mathrm{Quant}_p(\Delta, \{d_l\}_{l=1}^m) = \mathrm{Quant}_p(\Delta)$. Note that we do not assume independence of the quantization of blocks or independence of $\xi_{(j)}$.

The next result states that $\hat{\Delta}$ is an unbiased estimator of Δ, and gives a formula for its variance.

Lemma 1. *Let $\Delta \in \mathbb{R}^d$ and $\hat{\Delta} \sim \mathrm{Quant}_p(\Delta)$. Then for $l = 1, \dots, m$*

$$\mathbf{E}\hat{\Delta}(l) = \Delta(l), \qquad \mathbf{E}\|\hat{\Delta}(l) - \Delta(l)\|_2^2 = \Psi_l(\Delta), \qquad (7)$$

$$\mathbf{E}\hat{\Delta} = \Delta, \qquad \mathbf{E}\|\hat{\Delta} - \Delta\|_2^2 = \Psi(\Delta), \qquad (8)$$

where

$$x = (x(1)^\top, x(2)^\top, \dots, x(m)^\top)^\top,$$

$$\Psi_l(x) \overset{def}{=} \|x(l)\|_1\|x(l)\|_p - \|x(l)\|_2^2 \ge 0, \qquad (9)$$

$$\Psi(x) \overset{def}{=} \sum_{l=1}^m \Psi_l(x) \ge 0. \qquad (10)$$

Thus, $\hat{\Delta}$ is an unbiased estimator of Δ. Moreover, the variance of $\hat{\Delta}$ is a decreasing function of p, and is minimized for $p = \infty$.

3.3. Communication cost

If b bits are used to encode a float number, then at most $C(\hat{\Delta}) \overset{\text{def}}{=} \|\hat{\Delta}\|_0^{1/2}(\log\|\hat{\Delta}\|_0 + \log 2 + 1) + b$ bits are needed to communicate $\hat{\Delta}$ with Elias coding (see Theorem 3.3 in (Alistarh et al., 2017)). In our next result, we given an upper bound on the expected communication cost.

Theorem 1 (Expected sparsity)**.** *Let $0 \ne \Delta \in \mathbb{R}^{\tilde{d}}$ and $\widetilde{\Delta} \sim \mathrm{Quant}_p(\Delta)$ be its p-quantization. Then*

$$\mathbf{E}\|\widetilde{\Delta}\|_0 = \frac{\|\Delta\|_1}{\|\Delta\|_p} \le \|\Delta\|_0^{1-1/p} \le \tilde{d}^{1-1/p}, \qquad (11)$$

and

$$C_p \overset{def}{=} \mathbf{E}C(\widetilde{\Delta}) \le \frac{\|\Delta\|_1^{1/2}}{\|\Delta\|_p^{1/2}}(\log\tilde{d} + \log 2 + 1) + b. \qquad (12)$$

All expressions in (11) *and* (12) *are increasing functions of p.*

3.4. Proximal step

Given $\gamma > 0$, the proximal operator for the regularizer R is defined as

$$\mathrm{prox}_{\gamma R}(u) \overset{\text{def}}{=} \arg\min_v \left\{\gamma R(v) + \frac{1}{2}\|v - u\|_2^2\right\}.$$

The proximal operator of a closed convex function is nonexpansive. That is, for any $\gamma > 0$ and $u, v \in \mathbb{R}^d$,

$$\left\|\mathrm{prox}_{\gamma R}(u) - \mathrm{prox}_{\gamma R}(v)\right\|_2 \le \|u - v\|_2. \qquad (13)$$

[1] That is, $\xi_{(j)} = 1$ with probability $|\Delta_{(j)}|/\|\Delta\|_p$ (observe that this quantity is always upper bounded by 1) and $\xi_{(j)} = 0$ with probability $1 - |\Delta_{(j)}|/\|\Delta\|_p$.

3.5. DIANA

In DIANA, each machine $i \in \{1, 2, \ldots, n\}$ first computes a stochastic gradient g_i^k at current iterate x^k. We *do not* quantize this information and send it off to the parameter server as that approach would not converge for $R \neq 0$. Instead, we maintain memory h_i^k at each node i (initialized to arbitrary values), and *quantize the difference* $\delta_i^k \overset{def}{=} g_i^k - h_i^k$ *instead.* Both the node and the parameter server update h_i^k in an appropriate manner, and a proximal gradient descent step is taken with respect to direction $v^k = \beta v^{k-1} + \hat{g}^k$, where $0 \leq \beta \leq 1$ is a *momentum parameter*, whereas $\hat{g}^k$ is an unbiased estimator of the full gradient, assembled from the memory h_i^k and the transmitted quantized vectors. Note that we allows for block quantization for more flexibility. In practice, we want the transmitted quantized vectors to be much easier to communicate than the full dimensional vector in $\mathbb{R}^d$, which can be tuned by the choice of p defining the quantization norm, and the choice of blocks.

3.6. Relation to QSGD and TernGrad

If the initialization is done with $h^0 = 0$ and $\alpha = 0$, our method reduces to either 1-bit QSGD or TernGrad with $p = 2$ and $p = \infty$ respectively. We unify them in the Algorithm 2. We analyse this algorithm (i.e. DIANA with $\alpha = 0$ and $h_i^0 = 0$) in three cases: 1) smooth strongly convex objective with arbitrary closed convex regularizer; 2) smooth non-convex objective with constant regularizer; 3) smooth non-convex objective with constant regularizer for the momentum version of the algorithm. We notice, that in the original paper (Wen et al., 2017) authors do not provide the rate of convergence for TernGrad and we get the convergence rate for the three aforementioned situations as a special case of our results. Moreover, we emphasize that our analysis is new even for 1-bit QSGD, since in the original paper (Alistarh et al., 2017) authors consider only the case of bounded gradients ($\mathbf{E}\|g^k\|_2^2 \leq B^2$), which is very restrictive assumption, and they do not provide rigorous analysis of block-quantization as we do. In contrast, we consider more general case of block-quantization and assume only that the variance of the stochastic gradients is bounded, which is less restrictive assumption since the inequality $\mathbf{E}\|g^k\|_2^2 \leq B^2$ implies $\mathbf{E}\|g^k - \nabla f(x^k)\|_2^2 \leq \mathbf{E}\|g^k\|_2^2 \leq B^2$.

We obtain the convergence rate for arbitrary $p \geq 1$ for the three aforementioned cases (see Theorems 8, 9, 10, 11 and Corollaries 8, 9, 10 for the details) and all obtained bounds becomes better when p is growing, which means that TernGrad has better iteration complexity than QSGD and, more generally, the best iteration complexity attains for ℓ_∞ norm quantization.

4. Theory: Strongly Convex Case

4.1. Assumptions

Let us introduce two key assumptions of this section.

Assumption 2 (L–smoothness)**.** *We say that a function f is L-smooth if*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|_2^2 \qquad (14)$$

for any x and y.

Assumption 3 (μ-strong convexity)**.** *A function f is called μ-strongly convex if for all x and y*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2. \qquad (15)$$

4.2. Iteration complexity

For $1 \leq p \leq +\infty$, define

$$\alpha_p(d) \overset{\text{def}}{=} \inf_{x \neq 0, x \in \mathbb{R}^d} \frac{\|x\|_2^2}{\|x\|_1 \|x\|_p}. \qquad (16)$$

Lemma 2. *α_p is increasing as a function of p and decreasing as a function of d. In particular, $\alpha_1 \leq \alpha_2 \leq \alpha_\infty$, and moreover,*

$$\alpha_1(d) = \frac{1}{d}, \quad \alpha_2(d) = \frac{1}{\sqrt{d}}, \quad \alpha_\infty(d) = \frac{2}{1 + \sqrt{d}}$$

and, as a consequence, for all positive integers $\widetilde{d}$ and d the following relations holds

$$\alpha_1(\widetilde{d}) = \alpha_1(d)\frac{d}{\widetilde{d}}, \quad \alpha_2(\widetilde{d}) = \alpha_2(d)\sqrt{\frac{d}{\widetilde{d}}},$$

$$\alpha_\infty(\widetilde{d}) = \alpha_\infty(d)\frac{1 + \sqrt{d}}{1 + \sqrt{\widetilde{d}}}.$$

Theorem 2. *Assume the functions $f_1, \ldots, f_n$ are L–smooth and μ–strongly convex. Choose stepsizes $\alpha > 0$ and $\gamma^k = \gamma > 0$, block sizes $\{d_l\}_{l=1}^m$, where $\widetilde{d} = \max_{l=1,\ldots,m} d_l$, and parameter $c > 0$ satisfying the following relations:*

$$\frac{1 + nc\alpha^2}{1 + nc\alpha} \leq \alpha_p \overset{def}{=} \alpha_p(\widetilde{d}), \qquad (17)$$

$$\gamma \leq \min\left\{\frac{\alpha}{\mu}, \frac{2}{(\mu + L)(1 + c\alpha)}\right\}. \qquad (18)$$

Define the Lyapunov function

$$V^k \overset{def}{=} \|x^k - x^*\|_2^2 + \frac{c\gamma^2}{n}\sum_{i=1}^n \|h_i^k - h_i^*\|_2^2, \qquad (19)$$

p	iteration complexity	$\kappa = \Theta(n)$	$\kappa = \Theta(n^2)$
1	$\max\left\{\frac{2d}{m}, (\kappa+1)A\right\}$ $A = \left(\frac{1}{2} - \frac{1}{n} + \frac{d}{nm}\right)$	$O\left(n + \frac{d}{m}\right)$	$O\left(n^2 + \frac{nd}{m}\right)$
2	$\max\left\{\frac{2\sqrt{d}}{\sqrt{m}}, (\kappa+1)B\right\}$ $B = \left(\frac{1}{2} - \frac{1}{n} + \frac{\sqrt{d}}{n\sqrt{m}}\right)$	$O\left(n + \sqrt{\frac{d}{m}}\right)$	$O\left(n^2 + \frac{n\sqrt{d}}{\sqrt{m}}\right)$
∞	$\max\left\{1 + \sqrt{\frac{d}{m}}, (\kappa+1)C\right\}$ $C = \left(\frac{1}{2} - \frac{1}{n} + \frac{1+\sqrt{\frac{d}{m}}}{2n}\right)$	$O\left(n + \sqrt{\frac{d}{m}}\right)$	$O\left(n^2 + \frac{n\sqrt{d}}{\sqrt{m}}\right)$

Table 2. The leading term of the iteration complexity of DIANA in the strongly convex case, according to Theorem 2 (see Corollary 1 and Lemma 2). Logarithmic dependence on $1/\epsilon$ is suppressed. Condition number: $\kappa \overset{\text{def}}{=} \frac{L}{\mu}$.

where x^* is the solution of (1) *and $h^* \overset{def}{=} \nabla f(x^*)$. Then for all $k \geq 0$,*

$$\mathbf{E}V^k \leq (1-\gamma\mu)^k V^0 + \frac{\gamma}{\mu}(1+nc\alpha)\frac{\sigma^2}{n}. \qquad (20)$$

This implies that as long as $k \geq \frac{1}{\gamma\mu}\log\frac{V^0}{\epsilon}$, we have $\mathbf{E}V^k \leq \epsilon + \frac{\gamma}{\mu}(1+nc\alpha)\frac{\sigma^2}{n}$.

In particular, if we set γ to be equal to the minimum in (18), then the leading term in the iteration complexity bound is

$$\frac{1}{\gamma\mu} = \max\left\{\frac{1}{\alpha}, \frac{(\mu+L)(1+c\alpha)}{2\mu}\right\}. \qquad (21)$$

Corollary 1. *Let $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\alpha_p}{2}$, $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$, and $\gamma = \min\left\{\frac{\alpha}{\mu}, \frac{2}{(L+\mu)(1+c\alpha)}\right\}$. Then the conditions* (17) *and* (18) *are satisfied, and the leading term in the iteration complexity bound is equal to*

$$\frac{1}{\gamma\mu} = \max\left\{\frac{2}{\alpha_p}, (\kappa+1)\left(\frac{1}{2} - \frac{1}{n} + \frac{1}{n\alpha_p}\right)\right\}. \qquad (22)$$

This is a decreasing function of p. Hence, from iteration complexity perspective, $p = +\infty$ is the optimal choice.

In Table 2 we calculate the leading term (43) in the complexity of DIANA for $p \in \{1, 2, +\infty\}$, each for two condition number regimes: $n = \kappa$ (standard) and $n = \kappa^2$ (large).

4.3. Matching the rate of gradient descent for quadratic size models

Note that as long as the model size is not too big; in particular, when $d = O(\min\{\kappa^2, n^2\})$, the linear rate of DIANA with $p \geq 2$ is $O(\kappa \log(1/\epsilon))$, which matches the rate of gradient descent.

4.4. Optimal number of nodes

In practice one has access to a finite data set, consisting of N data points, where N is very large, and wishes to solve an empirical risk minimization ("finite-sum") of the form

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{N}\sum_{i=1}^{N}\phi_i(x) + R(x), \qquad (23)$$

where each ϕ_i is L–smooth and μ-strongly convex. If $n \leq N$ compute nodes of a distributed system are available, one may partition the N functions into n groups, $G_1, \ldots, G_n$, each of size $|G_i| = N/n$, and define $f_i(x) = \frac{n}{N}\sum_{i\in G_i}\phi_i(x)$. Note that $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x)$. Note that each f_i is also L–smooth and μ–strongly convex.

This way, we have fit the original (and large) problem (23) into our framework. One may now ask the question: *How many many nodes n should we use (other things equal)?* If what we care about is iteration complexity, then insights can be gained by investigating Eq. (43). For instance, if $p = 2$, then the complexity is

$$W(n) \overset{\text{def}}{=} \max\left\{\frac{2\sqrt{d}}{\sqrt{m}}, (\kappa+1)\left(\frac{1}{2} - \frac{1}{n} + \frac{\sqrt{d}}{n\sqrt{m}}\right)\right\}.$$

The optimal choice is to choose n so that the term $-\frac{1}{n} + \frac{\sqrt{d}}{n\sqrt{m}}$ becomes (roughly) equal to $\frac{1}{2}$: $-\frac{1}{n} + \frac{\sqrt{d}}{n\sqrt{m}} = \frac{1}{2}$. This gives the formula for the optimal number of nodes

$$n^* = n(d) \overset{\text{def}}{=} 2\left(\sqrt{\frac{d}{m}} - 1\right),$$

and the resulting iteration complexity is $W(n^*) = \max\left\{\frac{2\sqrt{d}}{\sqrt{m}}, \kappa + 1\right\}$. Note that $n(d)$ is increasing in d. Hence, it makes sense to use more nodes for larger models (big d).

4.5. Optimal block quantization

If the dimension of the problem is large, it becomes reasonable to quantize vector's blocks, also called blocks. For example, if we had a vector which consists of 2 smaller subvectors each of which is proportional to the vector of all ones, we can transmit just the subvectors without any loss of information. In the real world, we have a similar situation when different parts of the parameter vector have different scale. A straightforward example is deep neural networks, layers of which have pairwise different scales. If we quantized the whole vector at once, we would zero most of the update for the layer with the smallest scale.

Moreover, our theory says that if we have n workers, then the iteration complexity increase of quantization is about $\frac{\sqrt{d}}{n}$. However, if quantization is applied to a block of size n^2, then this number becomes 1, implying that the complexity remains the same. Therefore, if one uses about 100 workers and splits the parameter vector into parts of size about 10,000, the algorithm will work as fast as SGD, while communicating bits instead of floats!

4.6. Decreasing stepsizes

We now provide a convergence result for DIANA with decreasing step sizes, obtaining a $\mathcal{O}(1/k)$ rate.

Theorem 3. *Assume that f is L-smooth, μ-strongly convex and we have access to its gradients with bounded noise. Set $\gamma^k = \frac{2}{\mu k+\theta}$ with some $\theta \geq 2\max\left\{\frac{\mu}{\alpha}, \frac{(\mu+L)(1+c\alpha)}{2}\right\}$ for some numbers $\alpha > 0$ and $c > 0$ satisfying $\frac{1+nc\alpha^2}{1+nc\alpha} \leq \alpha_p$. After k iterations of DIANA we have*

$$\mathbf{E}V^k \leq \frac{1}{\eta k+1}\max\left\{V^0, 4\frac{(1+nc\alpha)\sigma^2}{n\theta\mu}\right\},$$

where $\eta \overset{def}{=} \frac{\mu}{\theta}$, $V^k = \|x^k - x^*\|_2^2 + \frac{c\gamma^k}{n}\sum_{i=1}^n \|h_i^0 - h_i^*\|_2^2$ and σ is the standard deviation of the gradient noise.

Corollary 2. *If we choose $\alpha = \frac{\alpha_p}{2}$, $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$, $\theta = 2\max\left\{\frac{\mu}{\alpha}, \frac{(\mu+L)(1+c\alpha)}{2}\right\} = \frac{\mu}{\alpha_p}\max\left\{4, \frac{2(\kappa+1)}{n} + \frac{(\kappa+1)(n-2)}{n}\alpha_p\right\}$, then there are three regimes:*

1) *if $1 = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, then $\theta = \Theta\left(\frac{\mu}{\alpha_p}\right)$ and to achieve $\mathbf{E}V^k \leq \varepsilon$ we need at most $O\left(\frac{1}{\alpha_p}\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{n\mu^2}\right\}\frac{1}{\varepsilon}\right)$ iterations;*

2) *if $\frac{\kappa}{n} = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, then $\theta = \Theta\left(\frac{L}{n\alpha_p}\right)$ and to achieve $\mathbf{E}V^k \leq \varepsilon$ we need at most $O\left(\frac{\kappa}{n\alpha_p}\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{\mu L}\right\}\frac{1}{\varepsilon}\right)$ iterations;*

3) *if $\kappa\alpha_p = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, then $\theta = \Theta(L)$ and to achieve $\mathbf{E}V^k \leq \varepsilon$ we need at most $O\left(\kappa\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{\mu L n\alpha_p}\right\}\frac{1}{\varepsilon}\right)$ iterations.*

5. Theory: Nonconvex Case

In this section we consider the non-convex case.

Theorem 4. *Assume R is such that exists a closed convex set $\mathcal{X}$ satisfying 1) $\forall z \in \mathbb{R}^n$ $\text{prox}_{\gamma R}(z) \in \mathcal{X}$ and 2) $\forall z \in \mathcal{X}$ $z = \text{prox}_{\gamma R}(z)$ (e.g. indicator function of $\mathcal{X}$). Also assume that $h^* = 0$, f is L-smooth, stepsizes $\alpha > 0$ and $\gamma^k = \gamma > 0$ and parameter $c > 0$ satisfying $\frac{1+nc\alpha^2}{1+nc\alpha} \leq \alpha_p$, $\gamma \leq \frac{2}{L(1+c\alpha)}$ and $\overline{x}^k$ is chosen randomly from $\{x^1, \ldots, x^k\}$. If, further, every worker samples from the full dataset, then*

$$\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2 \leq \frac{2}{k}\frac{f(x^0) - f^* + c\gamma^2\|h^0\|_2^2}{\gamma(2 - L\gamma - c\alpha L\gamma)} + (1+cn\alpha)\frac{L\gamma}{2 - L\gamma - c\alpha L\gamma}\frac{\sigma^2}{n}.$$

Algorithm 2 DIANA with $\alpha = 0$ and $h_i^0 = 0$; QSGD for $p = 2$ (1-bit)/ TernGrad for $p = \infty$ (SGD), (Alistarh et al., 2017; Wen et al., 2017)

```
input learning rates $\{\gamma^k\}_{k\geq 0}$, initial vector $x^0$, quantization parameter $p \geq 1$, sizes of blocks $\{d_l\}_{l=1}^m$, momentum parameter $0 \leq \beta < 1$
$v^0 = \nabla f(x^0)$
for $k = 1, 2, \ldots$ do
    Broadcast $x^k$ to all workers
    for $i = 1, \ldots, n$ do in parallel do
        Sample $g_i^k$ such that $\mathbf{E}[g_i^k \mid x^k] = \nabla f_i(x^k)$
        Sample $\hat{g}_i^k \sim \text{Quant}_p(g_i^k, \{d_l\}_{l=1}^m)$
    end for
    $\hat{g}^k = \overline{g}^k = \frac{1}{n}\sum_{i=1}^n \hat{g}_i^k$
    $v^k = \beta v^{k-1} + \hat{g}^k$
    $x^{k+1} = \text{prox}_{\gamma^k R}\left(x^k - \gamma^k v^k\right)$
end for
```

Corollary 3. *Set $\alpha = \frac{\alpha_p}{2}$, $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$, $\gamma = \frac{n\alpha_p}{L(2+(n-2)\alpha_p)\sqrt{K}}$, $h^0 = 0$ and run the algorithm for K iterations. Then, the final accuracy is at most $\frac{2}{\sqrt{K}}\frac{L(2+(n-2)\alpha_p)}{n\alpha_p}(f(x^0) - f^*) + \frac{1}{\sqrt{K}}\frac{(2-\alpha_p)\sigma^2}{2+(n-2)\alpha_p}$.*

Moreover, if the first term in Corollary 3 is leading and $\frac{1}{n} = \Omega(\alpha_p)$, the resulting complexity is $O(\frac{1}{\sqrt{K}})$, i.e. the same as that of SGD. For instance, if sufficiently large mini-batches are used, the former condition holds, while for the latter it is enough to quantize vectors in blocks of size $O(n^2)$.

6. Convergence Rate of TernGrad

Here we give the convergence guarantees for TernGrad and provide upper bounds for this method. The method coincides with Algorithm 2 for the case when $p = \infty$. In the original paper (Wen et al., 2017) no convergence rate was given and *we close this gap.*

To maintain consistent notation we rewrite the TernGrad in notation which is close to the notation we used for DIANA. Using our notation it is easy to see that TernGrad is DIANA with $h_1^0 = h_2^0 = \ldots = h_n^0 = 0, \alpha = 0$ and $p = \infty$. Firstly, it means that $h_i^k = 0$ for all $i = 1, 2, \ldots, n$ and $k \geq 1$. What is more, this observation tells us that Lemma 3 holds for the iterates of TernGrad too. What is more, in the original paper (Wen et al., 2017) the quantization parameter p was chosen as ∞. We generalize the method and we don't restrict our analysis only on the case of ℓ_∞ sampling.

As it was in the analysis of DIANA our proofs for TernGrad work under Assumption 1.

7. Implementation and Experiments

Following advice from Alistarh et al. (2017), we encourage the use of *blocks* when quantizing large vectors. To this effect, a vector can decomposed into a number of blocks, each of which should then be quantized separately. If coordinates have different scales, as is the case in deep learning, it will prevent undersampling of those with typically smaller values. Moreover, our theoretical results predict that applying quantization to blocks or layers will result in superlinear acceleration.

In our convex experiments, the optimal values of α were usually around $\min_i \frac{1}{\sqrt{d_i}}$, where the minimum is taken with respect to blocks and d_i are their sizes.

Finally, higher mini-batch sizes make the sampled gradients less noisy, which in turn is favorable to more uniform differences $g_i^k - h_i^k$ and faster convergence.

7.1. DIANA with momentum works best

We implement DIANA, QSGD, TernGrad and DQGD in Python[2] using MPI4PY for processes communication. This is then tested on a machine with 24 cores, each is Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz. The problem considered is binary classification with logistic loss and ℓ_2 penalty, chosen to be of order $1/N$, where N is the total number of data points. We experiment with choices of α, choice of norm type p, different number of workers and search for optimal block sizes. h_i^0 is always set to be zero vector for all i. We observe that for ℓ_∞-norm the optimal block size is significantly bigger than for ℓ_2-norm. Detailed description of the experiments can be found in Section L. Here, however, we provide Figure 2 to show how vast the difference is with other methods.

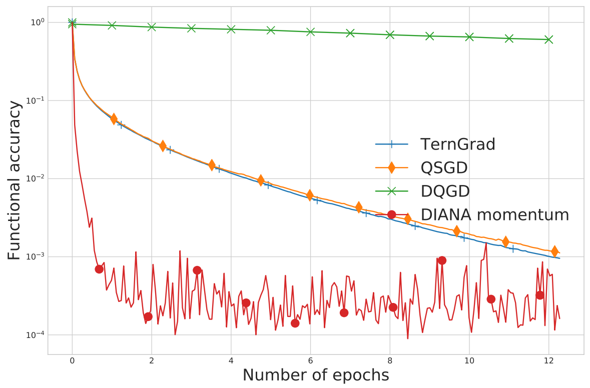

Figure 2. Comparison of the DIANA ($\beta = 0.95$) with QSGD, TernGrad and DQGD on the logistic regression problem for the "mushrooms" dataset.

7.2. DIANA vs MPI

In Figure 3 we compare the performance of DIANA vs. doing a MPI reduce operation with 32bit floats. The computing cluster had Cray Aries High Speed Network. However, for DIANA we used 2bit per dimension and have experienced a strange scaling behaviour, which was documented also in (Chunduri et al., 2017). In our case, this affected speed for alexnet and vgg_a beyond 64 or 32 MPI processes respectively. For more detailed experiments, see Section L.

7.3. Train and test accuracy on Mnist and Cifar10

In the next experiments, we run QSGD (Alistarh et al., 2017), TernGrad (Wen et al., 2017), SGD with momentum and DIANA on Mnist dataset and Cifar10 dataset for 3 epochs. We have selected 8 workers and run each method for learning rate from $\{0.1, 0.2, 0.05\}$. For QSGD, DIANA and TernGrad, we also tried various quantization bucket sizes in $\{32, 128, 512\}$. For QSGD we have chosen $2, 4, 8$ quantization levels. For DIANA we have chosen $\alpha \in \{0, 1.0/\sqrt{\text{quantization bucket sizes}}\}$ and have selected initial $h = 0$. For DIANA and SGD we also run a momentum version, with a momentum parameter in $\{0, 0.95, 0.99\}$. For DIANA we also run with two choices of norm ℓ_2 and ℓ_∞. For each experiment we have selected softmax cross entropy loss. Mnist-Convex is a simple DNN with no hidden layer, Mnist-DNN is a convolutional NN described here `https://github.com/floydhub/mnist/blob/master/ConvNet.py` and Cifar10-DNN is a convolutional DNN described here `https://github.com/kuangliu/pytorch-cifar/blob/master/models/lenet.py`.

In Figure 4 we show the best runs over all the parameters for all the methods. For Mnist-Convex SGD and DIANA makes use of the momentum and dominate all other algorithms. For Mnist-DNN situation is very similar. For Cifar10-DNN both DIANA and SGD significantly outperform other methods.

Acknowledgements

The work of Peter Richtárik was supported by the KAUST baseline funding scheme. The work of Martin Takáč was partially supported by the U.S. National Science Foundation, under award numbers NSF:CCF:1618717, NSF:CMMI:1663256 and NSF:CCF:1740796.

[2] The code will be made available online upon acceptance of this work.

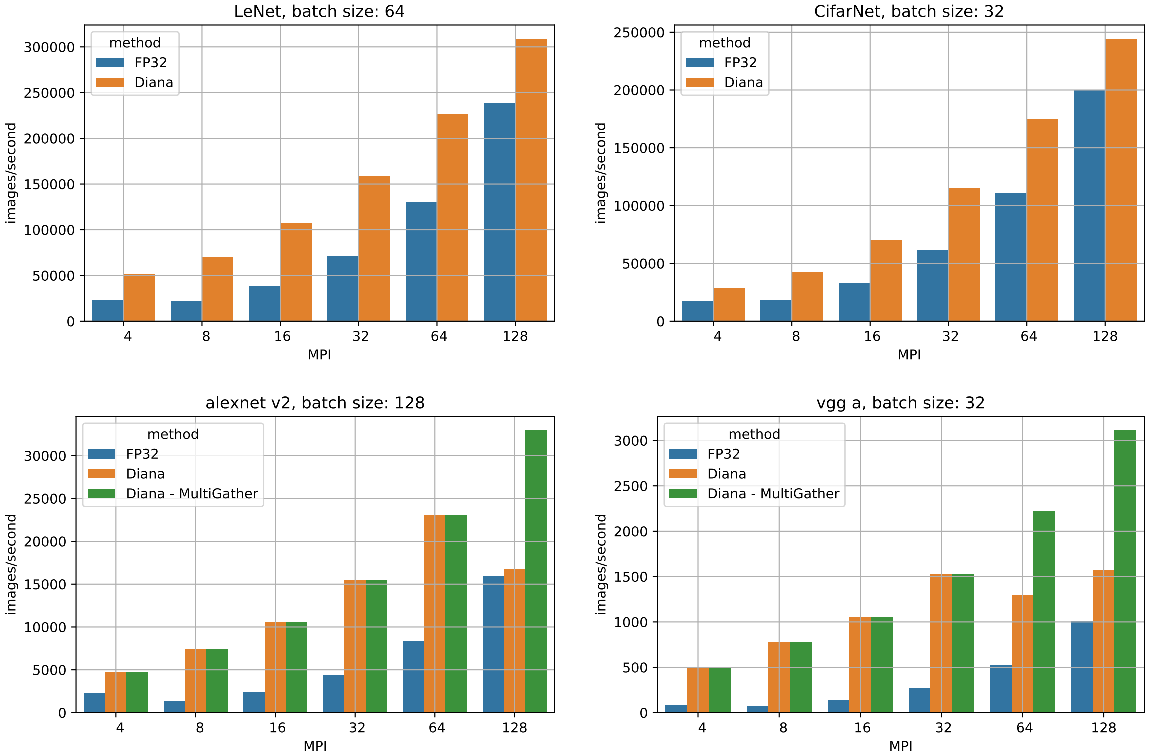

Figure 3. Comparison of performance (images/second) for various number of GPUs/MPI processes and sparse communication `DIANA` (2bit) vs. Reduce with 32bit float (FP32).

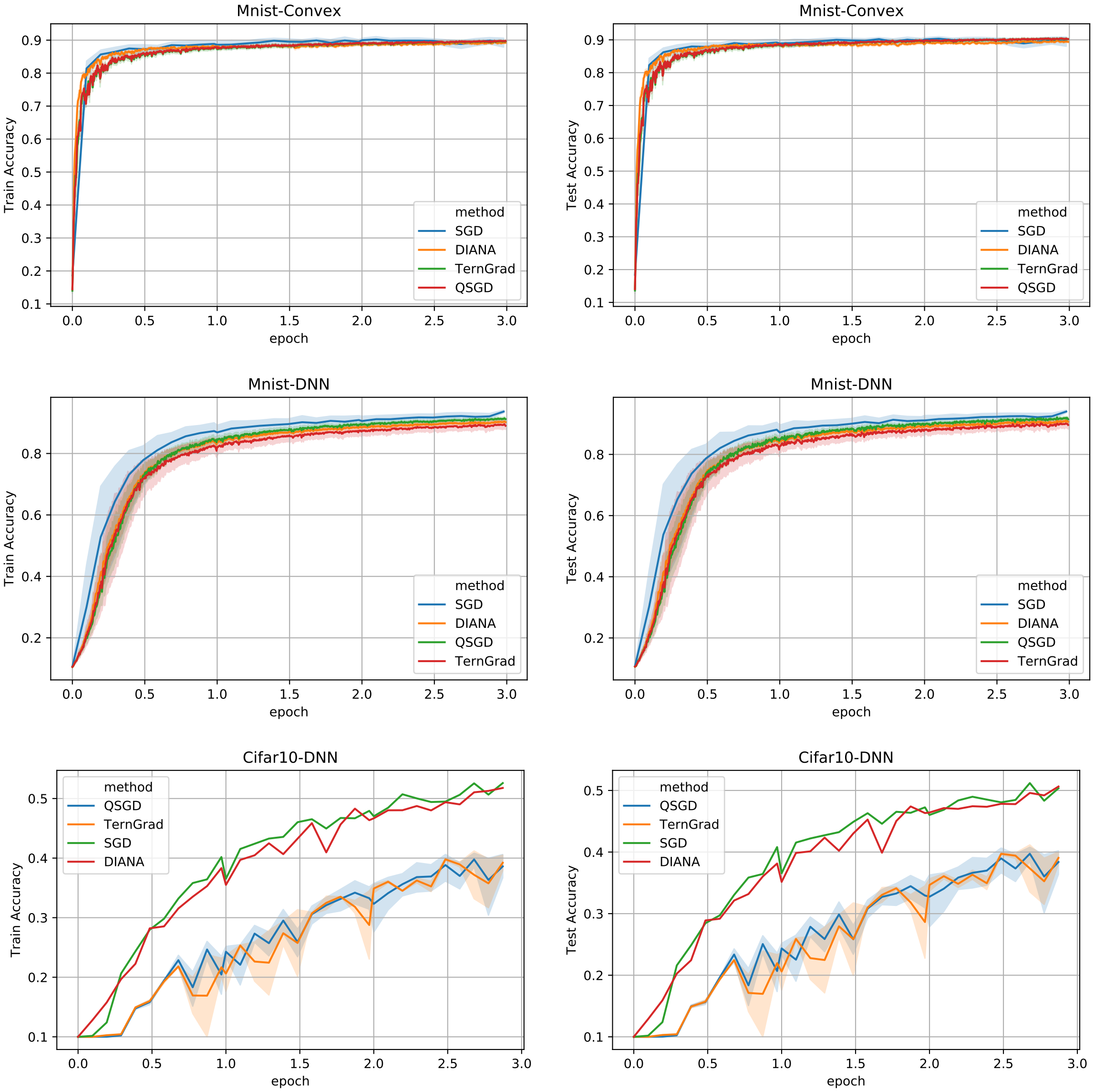

Figure 4. Evolution of training (left) and testing (right) accuracy for 3 different problems, using 4 algorithms: `DIANA`, `SGD`, `QSGD` and `TernGrad`. We have chosen the best runs over all tested hyper-parameters.

References

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimisation for non-convex problems. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

S. Chunduri, P. Coffman, S. Parker, and K. Kumaran. Performance analysis of mpi on cray xc40 xeon phi system. 2017.

O. Fercoq, Z. Qu, P. Richtárik, and M. Takáč. Fast distributed coordinate descent for minimizing non-strongly convex losses. *IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

M. Jaggi, V. Smith, M. Takáč, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems 27*, 2014.

M. Jahani, X. He, C. Ma, A. Mokhtari, D. Mudigere, A. Ribeiro, and M. Takáč. Efficient distributed hessian free algorithm for large-scale empirical risk minimization via accumulating sample strategy. *arXiv:1810.11507*, 2018.

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

J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

J. Konečný and P. Richtárik. Randomized distributed mean estimation: Accuracy vs communication. *arXiv preprint arXiv:1611.07555*, 2016.

C. Ma and M. Takáč. Partitioning data on features or samples in communication-efficient distributed optimization? *OptML@NIPS 2015, arXiv:1510.06688*, 2015.

C. Ma, V. Smith, M. Jaggi, M. I. Jordan, P. Richtárik, and M. Takáč. Adding vs. averaging in distributed primal-dual optimization. In *The 32nd International Conference on Machine Learning*, pages 1973–1982, 2015.

C. Ma, M. Jaggi, F. E. Curtis, N. Srebro, and M. Takáč. An accelerated communication-efficient primal-dual optimization framework for structured machine learning. *arXiv:1711.05305*, 2017a.

C. Ma, J. Konečný, M. Jaggi, V. Smith, M. I. Jordan, P. Richtárik, and M. Takáč. Distributed optimization with arbitrary local solvers. *Optimization Methods and Software*, 32(4):813–848, 2017b.

S. J. Reddi, J. Konečný, P. Richtárik, B. Póczós, and A. Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25, 2016.

F. Seide, H. Fu, J. Droppo, G. Li, and D. 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.

O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning, PMLR*, volume 32, pages 1000–1008, 2014.

V. Smith, S. Forte, C. Ma, M. Takáč, M. I. Jordan, and M. Jaggi. CoCoA: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18:1–49, 2018.

W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

Y. Zhang and L. Xiao. DiSCO: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, volume 37, pages 362–370, 2015.

A. Basic Identities and Inequalities

Smoothness and strong convexity. If f is L-smooth and μ-strongly convex, then for any vectors $x, y \in \mathbb{R}^d$ we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2. \tag{24}$$

Norm of a convex combination. For any $0 \leq \alpha \leq 1$ and $x, y \in \mathbb{R}^d$, we have

$$\|\alpha x + (1-\alpha)y\|_2^2 = \alpha\|x\|_2^2 + (1-\alpha)\|y\|_2^2 - \alpha(1-\alpha)\|x - y\|_2^2. \tag{25}$$

Variance decomposition. The (total) variance of a random vector $g \in \mathbb{R}^d$ is defined as the trace of its covariance matrix:

$$\mathbf{V}[g] \overset{\text{def}}{=} \mathrm{Tr}\left[\mathbf{E}\left[(g - \mathbf{E}g)(g - \mathbf{E}g)^\top\right]\right] = \mathbf{E}\left[\mathrm{Tr}\left[(g - \mathbf{E}g)(g - \mathbf{E}g)^\top\right]\right] = \mathbf{E}\|g - \mathbf{E}g\|_2^2 = \sum_{j=1}^d \mathbf{E}(g_{(j)} - \mathbf{E}g_{(j)})^2.$$

For any vector $h \in \mathbb{R}^d$, the variance of g can be decomposed as follows:

$$\mathbf{E}\|g - \mathbf{E}g\|_2^2 = \mathbf{E}\|g - h\|_2^2 - \|\mathbf{E}g - h\|_2^2. \tag{26}$$

B. Proof of Lemma 1

Note that the first part of (8) follows from the first part of (7) and the second part of (8) follows from the second part of (7) and

$$\|\hat{\Delta} - \Delta\|_2^2 = \sum_{l=1}^m \|\hat{\Delta}(l) - \Delta(l)\|_2^2.$$

Therefore, it is sufficient to prove (7). If $\Delta(l) = 0$, the statements follow trivially. Assume $\Delta(l) \neq 0$. In view of (5), we have

$$\mathbf{E}\hat{\Delta}_{(j)}(l) = \|\Delta(l)\|_p \mathrm{sign}(\Delta_{(j)}(l))\mathbf{E}\xi_{(j)} = \|\Delta(l)\|_p \mathrm{sign}(\Delta_{(j)}(l))|\Delta_{(j)}(l)|/\|\Delta(l)\|_p = \Delta_{(j)}(l),$$

which establishes the first claim. We can write

$$\begin{aligned}
\mathbf{E}\|\hat{\Delta}(l) - \Delta(l)\|_2^2 &= \mathbf{E}\sum_j (\hat{\Delta}_{(j)}(l) - \Delta_{(j)}(l))^2 \\
&= \mathbf{E}\sum_j (\hat{\Delta}_{(j)}(l) - \mathbf{E}\hat{\Delta}_{(j)}(l))^2 \\
&\overset{(5)}{=} \|\Delta(l)\|_p^2 \sum_j \mathrm{sign}^2(\Delta_{(j)}(l))\mathbf{E}(\xi_{(j)} - \mathbf{E}\xi_{(j)})^2 \\
&= \|\Delta(l)\|_p^2 \sum_j \mathrm{sign}^2(\Delta_{(j)}(l))\tfrac{|\Delta_{(j)}(l)|}{\|\Delta(l)\|_p}(1 - \tfrac{|\Delta_{(j)}(l)|}{\|\Delta(l)\|_p}) \\
&= \sum_j |\Delta_{(j)}(l)|(\|\Delta(l)\|_p - |\Delta_{(j)}(l)|) \\
&= \|\Delta(l)\|_1\|\Delta(l)\|_p - \|\Delta(l)\|_2.
\end{aligned}$$

C. Proof of Theorem 1

Let $1_{[\cdot]}$ denote the indicator random variable of an event. In view of (5), $\hat{\Delta}_{(j)} = \|\Delta\|_p \mathrm{sign}(\Delta_{(j)})\xi_{(j)}$, where $\xi_{(j)} \sim \mathrm{Be}(|\Delta_{(j)}|/\|\Delta\|_p)$. Therefore,

$$\|\hat{\Delta}\|_0 = \sum_{j=1}^d 1_{[\hat{\Delta}_{(j)} \neq 0]} = \sum_{j\,:\,\Delta_{(j)} \neq 0} 1_{[\xi_{(j)} = 1]},$$

which implies that

$$\mathbf{E}\|\hat{\Delta}\|_0 = \mathbf{E}\sum_{j\,:\,\Delta_{(j)}\neq 0}^{d} 1_{[\xi_{(j)}=1]} = \sum_{j\,:\,\Delta_{(j)}\neq 0}^{d} \mathbf{E}1_{[\xi_{(j)}=1]} = \sum_{j\,:\,\Delta_{(j)}\neq 0}^{d} \frac{|\Delta_{(j)}|}{\|\Delta\|_p} = \frac{\|\Delta\|_1}{\|\Delta\|_p}.$$

To establish the first clam, it remains to recall that for all $x \in \mathbb{R}^d$ and $1 \leq q \leq p \leq +\infty$, one has the bound

$$\|x\|_p \leq \|x\|_q \leq \|x\|_0^{1/q-1/p}\|x\|_p,$$

and apply it with $q = 1$.

The proof of the second claim follows the same pattern, but uses the concavity of $t \mapsto \sqrt{t}$ and Jensen's inequality in one step.

D. Proof of Lemma 2

$\alpha_p(d)$ is increasing as a function of p because $\|\cdot\|_p$ is decreasing as a function of p. Moreover, $\alpha_p(d)$ is decreasing as a function of d since if we have $d < b$ then

$$\alpha_p(b) = \inf_{x\neq 0, x\in\mathbb{R}^b} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p} \leqslant \inf_{x\neq 0, x\in\mathbb{R}^b_d} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p} = \inf_{x\neq 0, x\in\mathbb{R}^d} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p},$$

where $R_d^b \overset{\text{def}}{=} \{x \in \mathbb{R}^b : x_{(d+1)} = \ldots = x_{(b)} = 0\}$. It is known that $\frac{\|x\|_2}{\|x\|_1} \geq \frac{1}{\sqrt{d}}$, and that this bound is tight. Therefore,

$$\alpha_1(d) = \inf_{x\neq 0, x\in\mathbb{R}^d} \frac{\|x\|_2^2}{\|x\|_1^2} = \frac{1}{d}$$

and

$$\alpha_2(d) = \inf_{x\neq 0, x\in\mathbb{R}^d} \frac{\|x\|_2}{\|x\|_1} = \frac{1}{\sqrt{d}}.$$

Let us now establish that $\alpha_\infty(d) = \frac{2}{1+\sqrt{d}}$.Note that

$$\frac{\|x\|_2^2}{\|x\|_1\|x\|_\infty} = \frac{\left\|\frac{x}{\|x\|_\infty}\right\|_2^2}{\left\|\frac{x}{\|x\|_\infty}\right\|_1\left\|\frac{x}{\|x\|_\infty}\right\|_\infty} = \frac{\left\|\frac{x}{\|x\|_\infty}\right\|_2^2}{\left\|\frac{x}{\|x\|_\infty}\right\|_1}.$$

Therefore, w.l.o.g. one can assume that $\|x\|_\infty = 1$. Moreover, signs of coordinates of vector x does not influence aforementioned quantity. It means that w.l.o.g. one can consider only $x \in \mathbb{R}^d_+$ and, taking into account our non-restrictive assumption that $\|x\|_\infty = 1$, one can assume that $x_{(1)} = 1$. Thus, our goal now is to show that the minimal value of the function

$$f(x) = \frac{1 + x_{(2)}^2 + \ldots + x_{(d)}^2}{1 + x_{(2)} + \ldots + x_{(d)}}$$

on the set $M = \{x \in \mathbb{R}^d \mid x_{(1)} = 1, 0 \leq x_{(j)} \leq 1, j = 2,\ldots,d\}$ is equal to $\frac{2}{1+\sqrt{d}}$. By Cauchy-Schwartz inequality: $x_{(2)}^2 + \ldots + x_{(d)}^2 \geq \frac{(x_{(2)}+\ldots+x_{(d)})^2}{d-1}$ and it becomes equality if and only if all $x_{(j)}, j = 2,\ldots,d$ are equal. It means that if we fix $x_{(j)} = a$ for $j = 2,\ldots,d$ and some $0 \leq a \leq 1$ than the minimal value of the function

$$g(a) = \frac{1 + \frac{((d-1)a)^2}{d-1}}{1+(d-1)a} = \frac{1+(d-1)a^2}{1+(d-1)a}$$

on $[0,1]$ coincides with minimal value of f on M. The derivative

$$g'(a) = \frac{2(d-1)a}{1+(d-1)a} - \frac{(d-1)(1+(d-1)a^2)}{(1+(d-1)a)^2}$$

has the same sign on $[0,1]$ as the difference $a - \frac{1+(d-1)a^2}{2(1+(d-1)a)}$, which implies that g attains its minimal value on $[0,1]$ at such a that $a = \frac{1+(d-1)a^2}{2(1+(d-1)a)}$. It remains to find $a \in [0,1]$ which satisfies

$$a = \frac{1+(d-1)a^2}{2(1+(d-1)a)}, \quad a \in [0,1] \Longleftrightarrow (d-1)a^2 + 2a - 1 = 0, \quad a \in [0,1].$$

This quadratic equation has unique positive solution $a^* = \frac{-1+\sqrt{d}}{d-1} = \frac{1}{1+\sqrt{d}} < 1$. Direct calculations show that $g(a^*) = \frac{2}{1+\sqrt{d}}$. It implies that $\alpha(d) = \frac{2}{1+\sqrt{d}}$.

E. Quantization Lemmas

Consider iteration k of the DIANA method (Algorithm 1). Let $\mathbf{E}_{Q^k}$ be the expectation with respect to the randomness inherent in the quantization steps $\hat{\Delta}_i^k \sim \text{Quant}_p(\Delta_i^k, \{d_l\}_{l=1}^m)$ for $i = 1, 2, \ldots, n$ (i.e. we condition on everything else).

Lemma 3. *For all iterations $k \geq 0$ of DIANA and $i = 1, 2, \ldots, n$ we have the identities*

$$\mathbf{E}_{Q^k}\hat{g}_i^k = g_i^k, \qquad \mathbf{E}_{Q^k}\|\hat{g}_i^k - g_i^k\|_2^2 = \Psi(\Delta_i^k) \tag{27}$$

and

$$\mathbf{E}_{Q^k}\hat{g}^k = g^k \overset{def}{=} \frac{1}{n}\sum_{i=1}^n g_i^k, \qquad \mathbf{E}_{Q^k}\|\hat{g}^k - g^k\|_2^2 = \frac{1}{n^2}\sum_{i=1}^n \Psi(\Delta_i^k). \tag{28}$$

Furthermore, letting $h^* = \nabla f(x^*)$, and invoking Assumption 1, we have

$$\mathbf{E}\hat{g}^k = \nabla f(x^k), \qquad \mathbf{E}\|\hat{g}^k - h^*\|_2^2 \leq \mathbf{E}\|\nabla f(x^k) - h^*\|_2^2 + \left(\frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\Psi(\Delta_i^k)\right) + \frac{\sigma^2}{n}. \tag{29}$$

Proof. (i) Since $\hat{g}_i^k = h_i^k + \hat{\Delta}_i^k$ and $\Delta_i^k = g_i^k - h_i^k$, we can apply Lemma 1 and obtain

$$\mathbf{E}_{Q^k}\hat{g}_i^k = h_i^k + \mathbf{E}_{Q^k}\hat{\Delta}_i^k \overset{(8)}{=} h_i^k + \Delta_i^k = g_i^k.$$

Since $\hat{g}_i^k - g_i^k = \hat{\Delta}_i^k - \Delta_i^k$, applying the second part of Lemma 1 gives the second identity in (27).

(ii) The first part of (28) follows directly from the first part of (27):

$$\mathbf{E}_{Q^k}\hat{g}^k = \mathbf{E}_{Q^k}\left[\frac{1}{n}\sum_{i=1}^n \hat{g}_i^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{Q^k}\hat{g}_i^k \overset{(27)}{=} \frac{1}{n}\sum_{i=1}^n g_i^k \overset{(28)}{=} g^k.$$

The second part in (28) follows from the second part of (27) and independence of $\hat{g}_1^k, \ldots, \hat{g}_n^k$.

(iii) The first part of (29) follows directly from the first part of (28) and the assumption that g_i^k is and unbiased estimate of $\nabla f_i(x^k)$. It remains to establish the second part of (29). First, we shall decompose

$$\begin{aligned}\mathbf{E}_{Q^k}\|\hat{g}^k - h^*\|_2^2 &\overset{(26)}{=} \mathbf{E}_{Q^k}\|\hat{g}^k - \mathbf{E}_{Q^k}\hat{g}^k\|_2^2 + \|\mathbf{E}_{Q^k}\hat{g}^k - h^*\|_2^2 \\ &\overset{(28)}{=} \mathbf{E}_{Q^k}\|\hat{g}^k - g^k\|_2^2 + \|g^k - h^*\|_2^2 \\ &\overset{(28)}{=} \frac{1}{n^2}\sum_{i=1}^n \Psi(\Delta_i^k) + \|g^k - h^*\|_2^2.\end{aligned}$$

Further, applying variance decomposition (26), we get

$$\begin{aligned}\mathbf{E}\left[\|g^k - h^*\|_2^2 \mid x^k\right] &\overset{(26)}{=} \mathbf{E}\left[\|g^k - \mathbf{E}[g^k \mid x^k]\|_2^2 \mid x^k\right] + \|\mathbf{E}[g^k \mid x^k] - h^*\|_2^2 \\ &\overset{(3)}{=} \mathbf{E}\left[\|g^k - \nabla f(x^k)\|_2^2 \mid x^k\right] + \|\nabla f(x^k) - h^*\|_2^2 \\ &\overset{(4)}{\leq} \frac{\sigma^2}{n} + \|\nabla f(x^k) - h^*\|_2^2.\end{aligned}$$

Combining the two results, we get

$$\mathbf{E}[\mathbf{E}_{Q^k}\|\hat{g}^k - h^*\|_2^2 \mid x^k] \leq \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\Psi(\Delta_i^k) \mid x^k\right] + \frac{\sigma^2}{n} + \|\nabla f(x^k) - h^*\|_2^2.$$

After applying full expectation, and using tower property, we get the result.

□

Lemma 4. *Let x^* be a solution of* (1) *and let $h_i^* = \nabla f_i(x^*)$ for $i = 1, 2, \ldots, d$. For every i, we can estimate the first two moments of h_i^{k+1} as*

$$\begin{aligned}\mathbf{E}_{Q^k} h_i^{k+1} &= (1-\alpha)h_i^k + \alpha g_i^k,\\ \mathbf{E}_{Q^k}\|h_i^{k+1} - h_i^*\|_2^2 &= (1-\alpha)\|h_i^k - h_i^*\|_2^2 + \alpha\|g_i^k - h_i^*\|_2^2 - \alpha\left(\|\Delta_i^k\|_2^2 - \alpha\sum_{l=1}^m \|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p\right).\end{aligned} \tag{30}$$

Proof. Since

$$h_i^{k+1} = h_i^k + \alpha\hat{\Delta}_i^k \tag{31}$$

and $\Delta_i^k = g_i^k - h_i^k$, in view of Lemma 1 we have

$$\mathbf{E}_{Q^k} h_i^{k+1} \overset{(31)}{=} h_i^k + \alpha\mathbf{E}_{Q^k}\hat{\Delta}_i^k \overset{(8)}{=} h_i^k + \alpha\Delta_i^k = (1-\alpha)h_i^k + \alpha g_i^k, \tag{32}$$

which establishes the first claim. Further, using $\|\Delta_i^k\|_2^2 = \sum\limits_{l=1}^m \|\Delta_i^k(l)\|_2^2$ we obtain

$$\begin{aligned}\mathbf{E}_{Q^k}\|h_i^{k+1} - h_i^*\|_2^2 &\overset{(26)}{=} \|\mathbf{E}_{Q^k} h_i^{k+1} - h_i^*\|_2^2 + \mathbf{E}_{Q^k}\|h_i^{k+1} - \mathbf{E}_{Q^k} h_i^{k+1}\|_2^2\\ &\overset{(32)+(31)}{=} \|(1-\alpha)h_i^k + \alpha g_i^k - h_i^*\|_2^2 + \alpha^2\mathbf{E}_{Q^k}\|\hat{\Delta}_i^k - \mathbf{E}_{Q^k}\hat{\Delta}_i^k\|_2^2\\ &\overset{(8)}{=} \|(1-\alpha)(h_i^k - h_i^*) + \alpha(g_i^k - h_i^*)\|_2^2 + \alpha^2\sum_{l=1}^m(\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p - \|\Delta_i^k(l)\|_2^2)\\ &\overset{(25)}{=} (1-\alpha)\|h_i^k - h_i^*\|_2^2 + \alpha\|g_i^k - h_i^*\|_2^2 - \alpha(1-\alpha)\|\Delta_i^k\|_2^2\\ &\qquad +\alpha^2\sum_{l=1}^m(\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p) - \alpha^2\|\Delta_i^k\|_2^2\\ &= (1-\alpha)\|h_i^k - h_i^*\|_2^2 + \alpha\|g_i^k - h_i^*\|_2^2 + \alpha^2\sum_{l=1}^m(\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p) - \alpha\|\Delta_i^k\|_2^2.\end{aligned}$$

□

Lemma 5. *We have*

$$\mathbf{E}\left[\|\hat{g}^k - h^*\|_2^2 \mid x^k\right] \leq \|\nabla f(x^k) - h^*\|_2^2 + \left(\frac{1}{\alpha_p} - 1\right)\frac{1}{n^2}\sum_{i=1}^n \|\nabla f_i(x^k) - h_i^k\|_2^2 + \frac{\sigma^2}{\alpha_p n}. \tag{33}$$

Proof. Since $\alpha_p = \alpha_p(\max\limits_{l=1,\ldots,m} d_l)$ and $\alpha_p(d_l) = \inf\limits_{x\neq 0, x\in\mathbb{R}^{d_l}} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p}$, we have for a particular choice of $x = \Delta_i^k(l)$ that $\alpha_p \leq \alpha_p(d_l) \leq \frac{\|\Delta_i^k(l)\|_2^2}{\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p}$. Therefore,

$$\Psi(\Delta_i^k) = \sum_{l=1}^m \Psi_l(\Delta_i^k) = \sum_{l=1}^m(\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_\infty - \|\Delta_i^k(l)\|_2) \leq \sum_{l=1}^m\left(\frac{1}{\alpha_p} - 1\right)\|\Delta_i^k(l)\|_2^2 = \left(\frac{1}{\alpha_p} - 1\right)\|\Delta_i^k\|_2^2.$$

This can be applied to (29) in order to obtain

$$\begin{aligned} \mathbf{E}\left[\|\hat{g}^k - h^*\|_2^2 \mid x^k\right] &\leq \|\nabla f(x^k) - h^*\|_2^2 + \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\Psi(\Delta_i^k) \mid x^k\right] + \frac{\sigma^2}{n} \\ &\leq \|\nabla f(x^k) - h^*\|_2^2 + \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{\alpha_p} - 1\right)\mathbf{E}\left[\|\Delta_i^k\|_2^2 \mid x^k\right] + \frac{\sigma^2}{n}. \end{aligned}$$

Note that for every i we have $\mathbf{E}\left[\Delta_i^k \mid x^k\right] = \mathbf{E}\left[g_i^k - h_i^k \mid x\right] = \nabla f_i(x^k) - h_i^k$, so

$$\mathbf{E}\left[\|\Delta_i^k\|_2^2 \mid x^k\right] \overset{(26)}{=} \|\nabla f_i(x^k) - h_i^k\|_2^2 + \mathbf{E}\left[\|g_i^k - \nabla f_i(x^k)\|_2^2 \mid x^k\right] \leq \|\nabla f_i(x^k) - h_i^k\|_2^2 + \sigma_i^2.$$

Summing the produced bounds, we get the claim. □

F. Proof of Theorem 2

Proof. Note that x^* is a solution of (1) if and only if $x^* = \operatorname{prox}_{\gamma R}(x^* - \gamma h^*)$ (this holds for any $\gamma > 0$). Using this identity together with the nonexpansiveness of the proximaloperator, we shall bound the first term of the Lyapunov function:

$$\begin{aligned} \mathbf{E}_{Q^k}\|x^{k+1} - x^*\|_2^2 &= \mathbf{E}_{Q^k}\|\operatorname{prox}_{\gamma R}(x^k - \gamma\hat{g}^k) - \operatorname{prox}_{\gamma R}(x^* - \gamma h^*)\|_2^2 \\ &\overset{(13)}{\leq} \mathbf{E}_{Q^k}\|x^k - \gamma\hat{g}^k - (x^* - \gamma h^*)\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma\mathbf{E}_{Q^k}\left\langle \hat{g}^k - h^*, x^k - x^*\right\rangle + \gamma^2\mathbf{E}_{Q^k}\|\hat{g}^k - h^*\|_2^2 \\ &\overset{(28)}{=} \|x^k - x^*\|_2^2 - 2\gamma\left\langle g^k - h^*, x^k - x^*\right\rangle + \gamma^2\mathbf{E}_{Q^k}\|\hat{g}^k - h^*\|_2^2. \end{aligned}$$

Next, taking conditional expectation on both sides of the above inequality, and using (3), we get

$$\mathbf{E}\left[\mathbf{E}_{Q^k}\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq \|x^k - x^*\|_2^2 - 2\gamma\left\langle \nabla f(x^k) - h^*, x^k - x^*\right\rangle + \gamma^2\mathbf{E}\left[\mathbf{E}_{Q^k}\|\hat{g}^k - h^*\|_2^2 \mid x^k\right].$$

Taking full expectation on both sides of the above inequality, and applying the tower property and Lemma 3 leads to

$$\begin{aligned} \mathbf{E}\|x^{k+1} - x^*\|_2^2 &\leq \mathbf{E}\|x^k - x^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k) - h^*, x^k - x^*\right\rangle + \gamma^2\mathbf{E}\|\hat{g}^k - h^*\|_2^2 \\ &\overset{(29)}{\leq} \mathbf{E}\|x^k - x^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k) - h^*, x^k - x^*\right\rangle \\ &\qquad + \gamma^2\mathbf{E}\|\nabla f(x^k) - h^*\|_2^2 + \frac{\gamma^2}{n^2}\sum_{i=1}^n \left(\mathbf{E}\Psi(\Delta_i^k)\right) + \frac{\gamma^2\sigma^2}{n} \\ &\leq \mathbf{E}\|x^k - x^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k) - h^*, x^k - x^*\right\rangle \\ &\qquad + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2 + \frac{\gamma^2}{n^2}\sum_{i=1}^n \left(\mathbf{E}\Psi(\Delta_i^k)\right) + \frac{\gamma^2\sigma^2}{n}, \end{aligned} \tag{34}$$

where the last inequality follows from the identities $\nabla f(x^k) = \frac{1}{n}\sum_{i=1}^n f_i(x^k)$, $h^* = \frac{1}{n}\sum_{i=1}^n h_i^*$ and an application of Jensen's inequality.

Averaging over the identities (30) for $i = 1, 2, \ldots, n$ in Lemma 4, we get

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{Q^k}\|h_i^{k+1} - h_i^*\|_2^2 = \frac{1-\alpha}{n}\sum_{i=1}^n \|h_i^k - h_i^*\|_2^2 + \frac{\alpha}{n}\sum_{i=1}^n \|g_i^k - h_i^*\|_2^2 - \frac{\alpha}{n}\sum_{i=1}^n \left(\|\Delta_i^k\|_2^2 - \alpha\sum_{l=1}^m \|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p\right).$$

Applying expectation to both sides, and using the tower property, we get

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|h_i^{k+1} - h_i^*\|_2^2 = \frac{1-\alpha}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k - h_i^*\|_2^2 + \frac{\alpha}{n}\sum_{i=1}^n \mathbf{E}\|g_i^k - h_i^*\|_2^2 - \frac{\alpha}{n}\sum_{i=1}^n \mathbf{E}\left[\|\Delta_i^k\|_2^2 - \alpha\sum_{l=1}^m \|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p\right]. \tag{35}$$

Since

$$\mathbf{E}[\|g_i^k - h_i^*\|_2^2 \mid x^k] \overset{(26)}{=} \|\nabla f_i(x^k) - h_i^*\|_2^2 + \mathbf{E}[\|g_i^k - \nabla f_i(x^k)\|_2^2 \mid x^k] \overset{(2)}{\leq} \|\nabla f_i(x^k) - h_i^*\|_2^2 + \sigma_i^2,$$

the second term on the right hand side of (35) can be bounded above as

$$\mathbf{E}\|g_i^k - h_i^*\|_2^2 \leq \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2 + \sigma_i^2. \tag{36}$$

Plugging (36) into (35) leads to the estimate

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\|h_i^{k+1} - h_i^*\|_2^2 \quad &\leq \quad \frac{1-\alpha}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k - h_i^*\|_2^2 + \frac{\alpha}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2 + \alpha\sigma^2 \\ &\quad -\frac{\alpha}{n}\sum_{i=1}^n \mathbf{E}\left[\|\Delta_i^k\|_2^2 - \alpha\sum_{l=1}^m \|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p\right]. \end{aligned} \tag{37}$$

Adding (34) with the $c\gamma^2$ multiple of (37), we get an upper bound one the Lyapunov function:

$$\begin{aligned} \mathbf{E}V^{k+1} \quad \leq \quad & \mathbf{E}\|x^k - x^*\|_2^2 + \frac{(1-\alpha)c\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k - h_i^*\|_2^2 \\ & +\frac{\gamma^2(1+\alpha c)}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k) - h^*, x^k - x^*\right\rangle \\ & +\frac{\gamma^2}{n^2}\sum_{i=1}^n\sum_{l=1}^m \mathbf{E}\underbrace{\left[\left(\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p - \|\Delta_i^k(l)\|_2^2\right) - n\alpha c\left(\|\Delta_i^k(l)\|_2^2 - \alpha\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p\right)\right]}_{\overset{\text{def}}{=} T_i^k(l)} \\ & +(nc\alpha + 1)\frac{\gamma^2\sigma^2}{n}. \end{aligned} \tag{38}$$

We now claim that due to our choice of α and c, we have $T_i^k(l) \leq 0$ for all $\Delta_i^k(l) \in \mathbb{R}^{d_l}$, which means that we can bound this term away by zero. Indeed, note that $T_k^i(l) = 0$ for $\Delta_i^k(l) = 0$. If $\Delta_i^k(l) \neq 0$, then $T_k^i(l) \leq 0$ can be equivalently written as

$$\frac{1 + nc\alpha^2}{1 + nc\alpha} \leq \frac{\|\Delta_i^k(l)\|_2^2}{\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p}.$$

However, this inequality holds since in view of the first inequality in (17) and the definitions of α_p and $\alpha_p(d_l)$, we have

$$\frac{1 + nc\alpha^2}{1 + nc\alpha} \overset{(17)}{\leq} \alpha_p \leq \alpha_p(d_l) \overset{(16)}{=} \inf_{x \neq 0, x \in \mathbb{R}^{d_l}} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p} \leq \frac{\|\Delta_i^k(l)\|_2^2}{\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p}.$$

Therefore, from (38) we get

$$\begin{aligned} \mathbf{E}V^{k+1} \quad \leq \quad & \mathbf{E}\|x^k - x^*\|_2^2 + \frac{(1-\alpha)c\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k - h_i^*\|_2^2 \\ & +\frac{\gamma^2(1+\alpha c)}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k) - h^*, x^k - x^*\right\rangle \\ & +(nc\alpha + 1)\frac{\gamma^2\sigma^2}{n}. \end{aligned} \tag{39}$$

The next trick is to split $\nabla f(x^k)$ into the average of $\nabla f_i(x^k)$ in order to apply strong convexity of each term:

$$\begin{aligned}\mathbf{E}\left\langle \nabla f(x^k) - h^*, x^k - x^* \right\rangle &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\langle \nabla f_i(x^k) - h_i^*, x^k - x^* \right\rangle \\ &\overset{(24)}{\geq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\frac{\mu L}{\mu+L}\|x^k - x^*\|_2^2 + \frac{1}{\mu+L}\|\nabla f_i(x^k) - h_i^*\|_2^2\right) \\ &= \frac{\mu L}{\mu+L}\mathbf{E}\|x^k - x^*\|_2^2 + \frac{1}{\mu+L}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2. \end{aligned} \quad (40)$$

Plugging these estimates into (39), we obtain

$$\begin{aligned}\mathbf{E}V^{k+1} \quad &\leq \quad \left(1 - \frac{2\gamma\mu L}{\mu+L}\right)\mathbf{E}\|x^k - x^*\|_2^2 + \frac{(1-\alpha)c\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k - h_i^*\|_2^2 \\ &\quad + \left(\gamma^2(1+\alpha c) - \frac{2\gamma}{\mu+L}\right)\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k) - h_i^*\|_2^2 \\ &\quad + (nc\alpha + 1)\frac{\gamma^2\sigma^2}{n}. \end{aligned} \quad (41)$$

Notice that in view of the second inequality in (18), we have $\gamma^2(1+\alpha c) - \frac{2\gamma}{\mu+L} \leq 0$. Moreover, since f_i is μ–strongly convex, we have $\mu\|x^k - x^*\|_2^2 \leq \langle \nabla f_i(x^k) - h_i^*, x^k - x^* \rangle$. Applying the Cauchy-Schwarz inequality to further bound the right hand side, we get the inequality $\mu\|x^k - x^*\|_2 \leq \|\nabla f_i(x^k) - h_i^*\|_2$. Using these observations, we can get rid of the term on the second line of (41) and absorb it with the first term, obtaining

$$\mathbf{E}V^{k+1} \leq \left(1 - 2\gamma\mu + \gamma^2\mu^2 + c\alpha\gamma^2\mu^2\right)\mathbf{E}\|x^k - x^*\|_2^2 + \frac{(1-\alpha)c\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k - h_i^*\|_2^2 + (nc\alpha+1)\frac{\gamma^2\sigma^2}{n}. \quad (42)$$

It follows from the second inequality in (18) that $1 - 2\gamma\mu + \gamma^2\mu^2 + c\alpha\gamma^2\mu^2 \leq 1 - \gamma\mu$. Moreover, the first inequality in (18) implies that $1 - \alpha \leq 1 - \gamma\mu$. Consequently, from (42) we obtain the recursion

$$\mathbf{E}V^{k+1} \leq (1-\gamma\mu)\mathbf{E}V^k + (nc\alpha+1)\frac{\gamma^2\sigma^2}{n}.$$

Finally, unrolling the recurrence leads to

$$\begin{aligned}\mathbf{E}V^k &\leq (1-\gamma\mu)^k V^0 + \sum_{l=0}^{k-1}(1-\gamma\mu)^l\gamma^2(1+nc\alpha)\frac{\sigma^2}{n} \\ &\leq (1-\gamma\mu)^k V^0 + \sum_{l=0}^{\infty}(1-\gamma\mu)^l\gamma^2(1+nc\alpha)\frac{\sigma^2}{n} \\ &= (1-\gamma\mu)^k V^0 + \frac{\gamma}{\mu}(1+nc\alpha)\frac{\sigma^2}{n}.\end{aligned}$$

□

G. Proof of Corollary 1

Corollary 4. *Let* $\kappa = \frac{L}{\mu}$, $\alpha = \frac{\alpha_p}{2}$, $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$, *and* $\gamma = \min\left\{\frac{\alpha}{\mu}, \frac{2}{(L+\mu)(1+c\alpha)}\right\}$. *Then the conditions* (17) *and* (18) *are satisfied, and the leading term in the iteration complexity bound is equal to*

$$\frac{1}{\gamma\mu} = \max\left\{\frac{2}{\alpha_p}, (\kappa+1)\left(\frac{1}{2} - \frac{1}{n} + \frac{1}{n\alpha_p}\right)\right\}. \quad (43)$$

This is a decreasing function of p. *Hence, from iteration complexity perspective,* $p = +\infty$ *is the optimal choice.*

Proof. Condition (18) is satisfied since $\gamma = \min\left\{\frac{\alpha}{\mu}, \frac{2}{(L+\mu)(1+c\alpha)}\right\}$. Now we check that (17) is also satisfied:

$$
\begin{aligned}
\frac{1+nc\alpha^2}{1+nc\alpha}\frac{1}{\alpha_p} &= \frac{1+n\cdot\frac{4(1-\alpha_p)}{n\alpha_p^2}\cdot\frac{\alpha_p^2}{4}}{1+n\cdot\frac{4(1-\alpha_p)}{n\alpha_p^2}\cdot\frac{\alpha_p}{2}}\cdot\frac{1}{\alpha_p}\\
&= \frac{2-\alpha_p}{\alpha_p+2(1-\alpha_p)}\\
&= 1.
\end{aligned}
$$

Since $\alpha = \frac{\alpha_p}{2}$ and $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$ we have

$$
1+\alpha c = 1 + \frac{2(1-\alpha_p)}{n\alpha_p} = 1 - \frac{2}{n} + \frac{2}{n\alpha_p}
$$

and, therefore,

$$
\frac{1}{\gamma\mu} = \max\left\{\frac{1}{\alpha}, \frac{L+\mu}{2\mu}(1+c\alpha)\right\} = \max\left\{\frac{2}{\alpha_p}, (\kappa+1)\left(\frac{1}{2}-\frac{1}{n}+\frac{1}{n\alpha_p}\right)\right\},
$$

which is a decreasing function of p, because α_p increases when p increases. □

H. Strongly convex case: decreasing stepsize

Lemma 6. *Let a sequence $\{a^k\}_k$ satisfy inequality $a^{k+1} \le (1-\gamma^k\mu)a^k + (\gamma^k)^2 N$ for any positive $\gamma^k \le \gamma_0$ with some constants $\mu > 0, N > 0, \gamma_0 > 0$. Further, let $\theta \ge \frac{2}{\gamma_0}$ and take C such that $N \le \frac{\mu\theta}{4}C$ and $a_0 \le C$. Then, it holds*

$$
a^k \le \frac{C}{\frac{\mu}{\theta}k+1}
$$

if we set $\gamma^k = \frac{2}{\mu k+\theta}$.

Proof. We will show the inequality for a^k by induction. Since inequality $a_0 \le C$ is one of our assumptions, we have the initial step of the induction. To prove the inductive step, consider

$$
a^{k+1} \le (1-\gamma^k\mu)a^k + (\gamma^k)^2 N \le \left(1-\frac{2\mu}{\mu k+\theta}\right)\frac{\theta C}{\mu k+\theta} + \theta\mu\frac{C}{(\mu k+\theta)^2}.
$$

To show that the right-hand side is upper bounded by $\frac{\theta C}{\mu(k+1)+\theta}$, one needs to have, after multiplying both sides by $(\mu k+\theta)(\mu k+\mu+\theta)(\theta C)^{-1}$,

$$
\left(1-\frac{2\mu}{\mu k+\theta}\right)(\mu k+\mu+\theta) + \mu\frac{\mu k+\mu+\theta}{\mu k+\theta} \le \mu k+\theta,
$$

which is equivalent to

$$
\mu - \mu\frac{\mu k+\mu+\theta}{\mu k+\theta} \le 0.
$$

The last inequality is trivially satisfied for all $k \ge 0$. □

Theorem 5. *Assume that f is L-smooth, μ-strongly convex and we have access to its gradients with bounded noise. Set $\gamma^k = \frac{2}{\mu k+\theta}$ with some $\theta \ge 2\max\left\{\frac{\mu}{\alpha}, \frac{(\mu+L)(1+c\alpha)}{2}\right\}$ for some numbers $\alpha > 0$ and $c > 0$ satisfying $\frac{1+nc\alpha^2}{1+nc\alpha} \le \alpha_p$. After k iterations of* `DIANA` *we have*

$$
\mathbf{E}V^k \le \frac{1}{\eta k+1}\max\left\{V^0, 4\frac{(1+nc\alpha)\sigma^2}{n\theta\mu}\right\},
$$

where $\eta \overset{def}{=} \frac{\mu}{\theta}$, $V^k = \|x^k - x^*\|_2^2 + \frac{c\gamma^k}{n}\sum_{i=1}^n \|h_i^0 - h_i^*\|_2^2$ and σ is the standard deviation of the gradient noise.

Proof. To get a recurrence, let us recall an upper bound we have proved before:

$$\mathbf{E}V^{k+1} \le (1-\gamma^k\mu)\mathbf{E}V^k + (\gamma^k)^2(1+nc\alpha)\frac{\sigma^2}{n}.$$

Having that, we can apply Lemma 6 to the sequence $\mathbf{E}V^k$. The constants for the lemma are: $N = (1+nc\alpha)\frac{\sigma^2}{n}$, $C = \max\left\{V^0, 4\frac{(1+nc\alpha)\sigma^2}{n\theta\mu}\right\}$, and μ is the strong convexity constant. □

Corollary 5. *If we choose* $\alpha = \frac{\alpha_p}{2}$, $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$, $\theta = 2\max\left\{\frac{\mu}{\alpha}, \frac{(\mu+L)(1+c\alpha)}{2}\right\} = \frac{\mu}{\alpha_p}\max\left\{4, \frac{2(\kappa+1)}{n} + \frac{(\kappa+1)(n-2)}{n}\alpha_p\right\}$, *then there are three regimes:*

1) *if* $1 = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, *then* $\theta = \Theta\left(\frac{\mu}{\alpha_p}\right)$ *and to achieve* $\mathbf{E}V^k \le \varepsilon$ *we need at most* $O\left(\frac{1}{\alpha_p}\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{n\mu^2}\right\}\frac{1}{\varepsilon}\right)$ *iterations;*

2) *if* $\frac{\kappa}{n} = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, *then* $\theta = \Theta\left(\frac{L}{n\alpha_p}\right)$ *and to achieve* $\mathbf{E}V^k \le \varepsilon$ *we need at most* $O\left(\frac{\kappa}{n\alpha_p}\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{\mu L}\right\}\frac{1}{\varepsilon}\right)$ *iterations;*

3) *if* $\kappa\alpha_p = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, *then* $\theta = \Theta(L)$ *and to achieve* $\mathbf{E}V^k \le \varepsilon$ *we need at most* $O\left(\kappa\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{\mu L n\alpha_p}\right\}\frac{1}{\varepsilon}\right)$ *iterations.*

Proof. First of all, let us show that $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$ and α satisfy inequality $\frac{1+nc\alpha^2}{1+nc\alpha} \le \alpha_p$:

$$\begin{aligned}\frac{1+nc\alpha^2}{1+nc\alpha}\frac{1}{\alpha_p} &= \frac{1+n\cdot\frac{4(1-\alpha_p)}{n\alpha_p^2}\cdot\frac{\alpha_p^2}{4}}{1+n\cdot\frac{4(1-\alpha_p)}{n\alpha_p^2}\cdot\frac{\alpha_p}{2}}\cdot\frac{1}{\alpha_p}\\ &= \frac{2-\alpha_p}{\alpha_p+2(1-\alpha_p)}\\ &= 1.\end{aligned}$$

Moreover, since

$$1+c\alpha = 1+\frac{2(1-\alpha_p)}{n\alpha_p} = \frac{2+(n-2)\alpha_P}{n\alpha_p}$$

we can simplify the definition of θ:

$$\begin{aligned}\theta &= 2\max\left\{\frac{\mu}{\alpha}, \frac{(\mu+L)(1+c\alpha)}{2}\right\}\\ &= \frac{\mu}{\alpha_p}\max\left\{4, \frac{2(\kappa+1)}{n} + \frac{(\kappa+1)(n-2)}{n}\alpha_p\right\}\\ &= \Theta\left(\frac{\mu}{\alpha_p}\max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}\right).\end{aligned}$$

Using Theorem 5, we get in the case:

1) if $1 = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, then $\theta = \Theta\left(\frac{\mu}{\alpha_p}\right)$, $\eta = \Theta(\alpha_p)$, $\frac{4(1+nc\alpha)\sigma^2}{n\theta\mu} = \Theta\left(\frac{(1-\alpha_p)\sigma^2}{n\mu^2}\right)$ and to achieve $\mathbf{E}V^k \le \varepsilon$ we need at most $O\left(\frac{1}{\alpha_p}\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{n\mu^2}\right\}\frac{1}{\varepsilon}\right)$ iterations;

2) if $\frac{\kappa}{n} = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, then $\theta = \Theta\left(\frac{L}{n\alpha_p}\right)$, $\eta = \Theta\left(\frac{\alpha_p n}{\kappa}\right)$, $\frac{4(1+nc\alpha)\sigma^2}{n\theta\mu} = \Theta\left(\frac{(1-\alpha_p)\sigma^2}{\mu L}\right)$ and to achieve $\mathbf{E}V^k \le \varepsilon$ we need at most $O\left(\frac{\kappa}{n\alpha_p}\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{\mu L}\right\}\frac{1}{\varepsilon}\right)$ iterations;

3) if $\kappa\alpha_p = \max\left\{1, \frac{\kappa}{n}, \kappa\alpha_p\right\}$, then $\theta = \Theta\left(L\right)$, $\eta = \Theta\left(\frac{1}{\kappa}\right)$, $\frac{4(1+nc\alpha)\sigma^2}{n\theta\mu} = \Theta\left(\frac{(1-\alpha_p)\sigma^2}{\mu L n\alpha_p}\right)$ and to achieve $\mathbf{E}V^k \le \varepsilon$ we need at most $O\left(\kappa\max\left\{V^0, \frac{(1-\alpha_p)\sigma^2}{\mu L n \alpha_p}\right\}\frac{1}{\varepsilon}\right)$ iterations.

□

I. Non-convex analysis

Theorem 6. *Assume R is such that exists a closed convex set $\mathcal{X}$ satisfying 1) $\forall z \in \mathbb{R}^n$ $\mathrm{prox}_{\gamma R}(z) \in \mathcal{X}$ and 2) $\forall z \in \mathcal{X}$ $z = \mathrm{prox}_{\gamma R}(z)$ (e.g. indicator function of $\mathcal{X}$). Also assume that $h^* = 0$, f is L-smooth, stepsizes $\alpha > 0$ and $\gamma^k = \gamma > 0$ and parameter $c > 0$ satisfying $\frac{1+nc\alpha^2}{1+nc\alpha} \le \alpha_p$, $\gamma \le \frac{2}{L(1+c\alpha)}$ and $\overline{x}^k$ is chosen randomly from $\{x^1, \ldots, x^k\}$. If, further, every worker samples from the full dataset, then*

$$\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2 \le \frac{2}{k}\frac{f(x^0) - f^* + c\gamma^2\|h^0\|_2^2}{\gamma\left(2 - L\gamma - c\alpha L\gamma\right)} + (1 + cn\alpha)\frac{L\gamma}{2 - L\gamma - c\alpha L\gamma}\frac{\sigma^2}{n}.$$

Proof. Since $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma\overline{g}^k)$and $x^k = \mathrm{prox}_{\gamma R}(x^k)$, due to non-expansiveness we have $\|x^{k+1} - x^k\|_2 \le \|x^k - \gamma\overline{g}^k - x^k\|_2 = \gamma\|\overline{g}^k\|_2$. Moreover, by smoothness of f

$$\begin{aligned}
\mathbf{E}f(x^{k+1}) &\overset{(14)}{\le} \mathbf{E}f(x^k) + \mathbf{E}\left\langle \nabla f(x^k), x^{k+1} - x^k\right\rangle + \frac{L}{2}\mathbf{E}\|x^{k+1} - x^k\|_2^2 \\
&\overset{(28)}{\le} \mathbf{E}f(x^k) - \gamma\mathbf{E}\|\nabla f(x^k)\|_2^2 + \frac{L\gamma^2}{2}\mathbf{E}\|\hat{g}^k\|_2^2 \\
&\overset{(29)}{\le} \mathbf{E}f(x^k) - \left(\gamma - \frac{L\gamma^2}{2}\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 + \frac{L\gamma^2}{2}\frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\Psi(\Delta_i^k)\right] + \frac{L\gamma^2}{2n^2}\sum_{i=1}^n \sigma_i^2.
\end{aligned}$$

Denote $\Lambda^k \overset{\text{def}}{=} f(x^k) - f^* + c\frac{L\gamma^2}{2}\frac{1}{n}\sum_{i=1}^n \|h_i^k\|_2^2$. Due to the assumption about sampling from the full dataset, we have $h_i^* = h^* = 0$ and we can plug into equation (30) that $\mathbf{E}g_i^k = \mathbf{E}\nabla f(x^k)$ and $\mathbf{E}\|g_i^k\|_2^2 \le \mathbf{E}\|\nabla f(x^k)\|_2^2 + \sigma_i^2$. If we add it the bound above, we get

$$\begin{aligned}
\mathbf{E}\Lambda^{k+1} &= \mathbf{E}f(x^{k+1}) - f^* + c\frac{L\gamma^2}{2}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|h_i^{k+1}\|_2^2 \\
&\le \mathbf{E}f(x^k) - f^* - \gamma\left(1 - \frac{L\gamma}{2} - \frac{c\alpha L\gamma}{2}\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 + (1-\alpha)c\frac{L\gamma^2}{2}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k\|_2^2 \\
&\quad + \frac{L\gamma^2}{2}\frac{1}{n^2}\sum_{i=1}^n\sum_{l=1}^m \mathbf{E}\underbrace{\left[(\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p - \|\Delta_i^k(l)\|_2^2) - nc\alpha(\|\Delta_i^k(l)\|_2^2 - \alpha\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p)\right]}_{\overset{\text{def}}{=}T_i^k(l)} \\
&\quad + (1 + cn\alpha)\frac{L\gamma^2}{2}\frac{\sigma^2}{n}.
\end{aligned}$$

We now claim that due to our choice of α and c, we have $T_i^k(l) \le 0$ for all $\Delta_i^k(l) \in \mathbb{R}^{d_l}$, which means that we can bound this term away by zero. Indeed, note that $T_k^i(l) = 0$ for $\Delta_i^k(l) = 0$. If $\Delta_i^k(l) \ne 0$, then $T_k^i(l) \le 0$ can be equivalently written as

$$\frac{1 + nc\alpha^2}{1 + nc\alpha} \le \frac{\|\Delta_i^k(l)\|_2^2}{\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p}.$$

However, this inequality holds since in view of the first inequality in (17) and the definitions of α_p and $\alpha_p(d_l)$, we have

$$\frac{1 + nc\alpha^2}{1 + nc\alpha} \overset{(17)}{\le} \alpha_p \le \alpha_p(d_l) \overset{(16)}{=} \inf_{x \ne 0, x \in \mathbb{R}^{d_l}} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p} \le \frac{\|\Delta_i^k(l)\|_2^2}{\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p}.$$

Putting all together we have

$$\mathbf{E}\Lambda^{k+1} \le \mathbf{E}f(x^k) - f^* + c\frac{L\gamma^2}{2}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|h_i^k\|_2^2 - \gamma\left(1 - \frac{L\gamma}{2} - \frac{c\alpha L\gamma}{2}\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 + (1 + cn\alpha)\frac{L\gamma^2}{2}\frac{\sigma^2}{n}.$$

Due to $\gamma \le \frac{2}{L(1+c\alpha)}$ the coefficient before $\|\nabla f(x^k)\|_2^2$ is positive. Therefore, we can rearrange the terms and rewrite the last bound as

$$\mathbf{E}[\|\nabla f(x^k)\|_2^2] \le 2\frac{\mathbf{E}\Lambda^k - \mathbf{E}\Lambda^{k+1}}{\gamma(2 - L\gamma - c\alpha L\gamma)} + (1 + cn\alpha)\frac{L\gamma}{2 - L\gamma - c\alpha L\gamma}\frac{\sigma^2}{n}.$$

Summing from 0 to $k-1$ results in telescoping of the right-hand side, giving

$$\sum_{l=1}^{k} \mathbf{E}[\|\nabla f(x^l)\|_2^2] \le 2\frac{\Lambda^0 - \mathbf{E}\Lambda^k}{\gamma(2 - L\gamma - c\alpha L\gamma)} + k(1 + cn\alpha)\frac{L\gamma}{2 - L\gamma - c\alpha L\gamma}\frac{\sigma^2}{n}.$$

Note that $\mathbf{E}\Lambda^k$ is non-negative and, thus, can be dropped. After that, it suffices to divide both sides by k and rewrite the left-hand side as $\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2$ where expectation is taken w.r.t. all randomness. □

Corollary 6. *Set $\alpha = \frac{\alpha_p}{2}$, $c = \frac{4(1-\alpha_p)}{n\alpha_p^2}$, $\gamma = \frac{n\alpha_p}{L(2+(n-2)\alpha_p)\sqrt{K}}$, $h^0 = 0$ and run the algorithm for K iterations. Then, the final accuracy is at most $\frac{2}{\sqrt{K}}\frac{L(2+(n-2)\alpha_p)}{n\alpha_p}(f(x^0) - f^*) + \frac{1}{\sqrt{K}}\frac{(2-\alpha_p)\sigma^2}{2+(n-2)\alpha_p}$.*

Proof. Our choice of α and c implies

$$1 + c\alpha = \frac{2 + (n-2)\alpha_p}{n\alpha_p}, \quad 1 + cn\alpha = \frac{2 - \alpha_p}{\alpha_p}.$$

Using this and the inequality $\gamma = \frac{n\alpha_p}{L(2+(n-2)\alpha_p)\sqrt{K}} \le \frac{n\alpha_p}{L(2+(n-2)\alpha_p)}$ we get $2 - L\gamma - c\alpha L\gamma = 2 - (1 + c\alpha)L\gamma \ge 1$. Putting all together we obtain $\frac{2}{K}\frac{f(x^0)-f^*+c\gamma^2\|h^0\|_2^2}{\gamma(2-L\gamma-c\alpha L\gamma)} + (1 + cn\alpha)\frac{L\gamma}{2-L\gamma-c\alpha L\gamma}\frac{\sigma^2}{n} = \frac{2}{\sqrt{K}}\frac{L(2+(n-2)\alpha_p)}{n\alpha_p}(f(x^0) - f^*) + \frac{1}{\sqrt{K}}\frac{(2-\alpha_p)\sigma^2}{2+(n-2)\alpha_p}$. □

J. Momentum version of DIANA

Theorem 7. *Assume that f is L-smooth, $R \equiv const$, $h_i^0 = 0$ and $f_i = f$ for all i. Choose $0 \le \alpha < \alpha_p$, $\beta \ne 1 - \alpha$ and $\gamma < \frac{1-\beta^2}{2L\left(\omega+\frac{\alpha(\omega-1)}{1-\xi}\right)}$, where $\xi \overset{def}{=} \max\{1-\alpha, \beta\}$, such that $\frac{\beta^2}{(1-\beta)^2(1-\xi)} \le \frac{1-\beta^2-2L\gamma\left(\omega+\frac{\alpha(\omega-1)}{1-\xi}\right)}{\gamma^2L^2\delta}$, where $\delta \overset{def}{=} 1 + \frac{2}{n}\left(\frac{1}{\alpha_p} - 1\right)\left(1 + \frac{\alpha}{|1-\alpha-\beta|}\right)$ and $\omega \overset{def}{=} \frac{n-1}{n} + \frac{1}{n\alpha_p}$, and sample $\overline{x}^k$ uniformly from $\{x^0, \dots, x^{k-1}\}$. Then*

$$\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2 \le \frac{4(f(z^0) - f^*)}{\gamma k} + 2\gamma\frac{L\sigma^2}{(1-\beta)^2 n}\left(\frac{3}{\alpha_p} - 2\right) + 2\gamma^2\frac{L^2\beta^2\sigma^2}{(1-\beta)^5 n}\left(\frac{3}{\alpha_p} - 2\right).$$

Proof. The main idea of the proof is to find virtual iterates z^k whose recursion would satisfy $z^{k+1} = z^k - \frac{\gamma}{1-\beta}\hat{g}^k$. Having found it, we can prove convergence by writing a recursion on $f(z^k)$. One possible choice is defined below:

$$z^k \overset{\text{def}}{=} x^k - \frac{\gamma\beta}{1-\beta}v^{k-1}, \tag{44}$$

where for the edge case $k = 0$ we simply set $v^{-1} = 0$ and $z^0 = x^0$. Although z^k is just a slight perturbation of x^k, applying smoothness inequality (14) to it produces a more convenient bound than the one we would have if used x^k. But first of all, let us check that we have the desired recursion for z^{k+1}:

$$\begin{aligned} z^{k+1} &\overset{(51)}{=} x^{k+1} - \frac{\gamma\beta}{1-\beta}v^k \\ &= x^k - \frac{\gamma}{1-\beta}v^k \\ &= x^k - \frac{\gamma\beta}{1-\beta}v^{k-1} - \frac{\gamma}{1-\beta}\hat{g}^k \\ &\overset{(51)}{=} z^k - \frac{\gamma}{1-\beta}\hat{g}^k. \end{aligned}$$

Now, it is time to apply smoothness of f:

$$\begin{aligned}\mathbf{E}f(z^{k+1}) &\leq \mathbf{E}\left[f(z^k) + \left\langle\nabla f(z^k), z^{k+1}-z^k\right\rangle + \frac{L}{2}\|z^{k+1}-z^k\|_2^2\right]\\ &\overset{(51)}{=} \mathbf{E}\left[f(z^k) - \frac{\gamma}{1-\beta}\left\langle\nabla f(z^k), \hat{g}^k\right\rangle + \frac{L\gamma^2}{2(1-\beta)^2}\|\hat{g}^k\|_2^2\right]. \qquad (45)\end{aligned}$$

The scalar product in (52) can be bounded using the fact that for any vectors a and b one has $-\langle a,b\rangle = \frac{1}{2}(\|a-b\|_2^2 - \|a\|_2^2 - \|b\|_2^2)$. In particular,

$$\begin{aligned}-\frac{\gamma}{1-\beta}\left\langle\nabla f(z^k),\nabla f(x^k)\right\rangle &= \frac{\gamma}{2(1-\beta)}\left(\|\nabla f(x^k)-\nabla f(z^k)\|_2^2 - \|\nabla f(x^k)\|_2^2 - \|\nabla f(z^k)\|_2^2\right)\\ &\leq \frac{\gamma}{2(1-\beta)}\left(L^2\|x^k-z^k\|_2^2 - \|\nabla f(x^k)\|_2^2\right)\\ &= \frac{\gamma^3L^2\beta^2}{2(1-\beta)^3}\|v^{k-1}\|_2^2 - \frac{\gamma}{2(1-\beta)}\|\nabla f(x^k)\|_2^2.\end{aligned}$$

The next step is to come up with an inequality for $\mathbf{E}\|v^k\|_2^2$. Since we initialize $v^{-1}=0$, one can show by induction that

$$v^k = \sum_{l=0}^k \beta^l\hat{g}^{k-l}.$$

Define $B \overset{\text{def}}{=} \sum_{l=0}^k\beta^l = \frac{1-\beta^{k+1}}{1-\beta}$. Then, by Jensen's inequality

$$\mathbf{E}\|v^k\|_2^2 = B^2\mathbf{E}\left\|\sum_{l=0}^k\frac{\beta^l}{B}\hat{g}^{k-l}\right\|_2^2 \leq B^2\sum_{l=0}^k\frac{\beta^l}{B}\mathbf{E}\|\hat{g}^{k-l}\|_2^2.$$

Since $\alpha < \alpha_p \leq \alpha_p(d_l) \leq \frac{\|\Delta_i^k(l)\|_2^2}{\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p}$ for all i, k and l, we have

$$\|\Delta_i^k(l)\|_2^2 - \alpha\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p \geq (\alpha_p-\alpha)\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p \geq 0$$

for the case when $\Delta_i^k(l)\neq 0$. When $\Delta_i^k(l)=0$ we simply have $\|\Delta_i^k(l)\|_2^2 - \alpha\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p = 0$. Taking into account this and the following equality

$$\|\Delta_i^k\|_2^2 - \alpha\sum_{l=1}^m\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p = \sum_{l=1}^m\left(\|\Delta_i^k(l)\|_2^2 - \alpha\|\Delta_i^k(l)\|_1\|\Delta_i^k(l)\|_p\right),$$

we get from (30)

$$\begin{aligned}\mathbf{E}\|h_i^k\|_2^2 &\leq (1-\alpha)\mathbf{E}\left[\|h_i^{k-1}\|_2^2\right] + \alpha\mathbf{E}\left[\|g_i^{k-1}\|_2^2\right]\\ &\leq (1-\alpha)^2\mathbf{E}\left[\|h_i^{k-2}\|_2^2\right] + \alpha(1-\alpha)\mathbf{E}\left[\|g_i^{k-2}\|_2^2\right] + \alpha\mathbf{E}\left[\|g_i^{k-1}\|_2^2\right]\\ &\leq \ldots \leq (1-\alpha)^k\underbrace{\|h_i^0\|_2^2}_{0} + \alpha\sum_{j=0}^{k-1}(1-\alpha)^j\mathbf{E}\left[\|g_i^{k-1-j}\|_2^2\right]\\ &= \alpha\sum_{j=0}^{k-1}(1-\alpha)^j\mathbf{E}\|\nabla f(x^{k-1-j})\|_2^2 + \alpha\sum_{j=0}^{k-1}(1-\alpha)^j\sigma_i^2\\ &\leq \alpha\sum_{j=0}^{k-1}(1-\alpha)^j\mathbf{E}\|\nabla f(x^{k-1-j})\|_2^2 + \alpha\cdot\frac{\sigma_i^2}{1-(1-\alpha)}\\ &= \alpha\sum_{j=0}^{k-1}(1-\alpha)^j\mathbf{E}\|\nabla f(x^{k-1-j})\|_2^2 + \sigma_i^2 \leq \alpha\sum_{j=0}^{k-1}\max\left\{(1-\alpha)^j,\beta^j\right\}\mathbf{E}\|\nabla f(x^{k-1-j})\|_2^2 + \sigma_i^2.\end{aligned}$$

Under our special assumption inequality (33) gives us

$$\begin{aligned}
\mathbf{E}\left[\|\hat{g}^k\|_2^2\right] &\le \mathbf{E}\|\nabla f(x^k)\|_2^2 + \left(\frac{1}{\alpha_p}-1\right)\frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\underbrace{\left[\|\nabla f(x^k)-h_i^k\|_2^2\right]}_{\le 2\|\nabla f(x^k)\|_2^2+2\|h_i^k\|_2^2} + \frac{\sigma^2}{\alpha_p n}\\
&\le \left(1+\frac{2}{n}\left(\frac{1}{\alpha_p}-1\right)\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 + \frac{2}{n^2}\left(\frac{1}{\alpha_p}-1\right)\sum_{i=1}^n \mathbf{E}\|h_i^k\|_2^2 + \frac{\sigma^2}{\alpha_p n}\\
&\le \left(1+\frac{2}{n}\left(\frac{1}{\alpha_p}-1\right)\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 + \frac{2\alpha}{n}\left(\frac{1}{\alpha_p}-1\right)\sum_{j=0}^{k-1}\max\left\{(1-\alpha)^j,\beta^j\right\}\mathbf{E}\|\nabla f(x^{k-1-j})\|_2^2\\
&\quad + \left(\frac{1}{\alpha_p}-1\right)\frac{2\sigma^2}{n} + \frac{\sigma^2}{\alpha_p n}.
\end{aligned}$$

Using this, we continue our evaluation of $\mathbf{E}\|v^k\|_2^2$:

$$\begin{aligned}
\mathbf{E}\|v^k\|_2^2 \quad\le\quad & B\sum_{l=0}^k \beta^l\left(1+\frac{2}{n}\left(\frac{1}{\alpha_p}-1\right)\right)\mathbf{E}\|\nabla f(x^{k-l})\|_2^2 + B\left(\frac{1}{\alpha_p}-1\right)\frac{2\alpha}{n}\sum_{l=0}^k\sum_{j=0}^{k-l-1}\beta^l(1-\alpha)^j\mathbf{E}\|\nabla f(x^{k-l-1-j})\|_2^2\\
& +B\sum_{l=0}^k\beta^l\left(\left(\frac{1}{\alpha_p}-1\right)\frac{2\sigma^2}{n}+\frac{\sigma^2}{\alpha_p n}\right).
\end{aligned}$$

Now we are going to simplify the double summation:

$$\begin{aligned}
\sum_{l=0}^k\sum_{j=0}^{k-l-1}\beta^l(1-\alpha)^j\mathbf{E}\|\nabla f(x^{k-l-1-j})\|_2^2 \quad &= \quad \sum_{l=0}^k\sum_{j=0}^{k-l-1}\beta^l(1-\alpha)^{k-l-1-j}\mathbf{E}\|\nabla f(x^j)\|_2^2\\
&= \quad \sum_{j=0}^{k-1}\mathbf{E}\|\nabla f(x^j)\|_2^2\sum_{l=0}^{k-j-1}\beta^l(1-\alpha)^{k-l-1-j}\\
&= \quad \sum_{j=0}^{k-1}\mathbf{E}\|\nabla f(x^j)\|_2^2\cdot\frac{(1-\alpha)^{k-j}-\beta^{k-j}}{1-\alpha-\beta}\\
&\le \quad \sum_{j=0}^{k}\mathbf{E}\|\nabla f(x^j)\|_2^2\cdot\frac{\max\left\{(1-\alpha)^{k-j},\beta^{k-j}\right\}}{|1-\alpha-\beta|}\\
&= \quad \frac{1}{|1-\alpha-\beta|}\sum_{j=0}^k\max\left\{(1-\alpha)^j,\beta^j\right\}\mathbf{E}\|\nabla f(x^{k-j})\|_2^2.
\end{aligned}$$

Note that $B\overset{\text{def}}{=}\sum\limits_{l=0}^k\beta^l\le\frac{1}{1-\beta}$. Putting all together we get

$$\mathbf{E}\|v^k\|_2^2 \quad\le\quad \frac{\delta}{1-\beta}\sum_{l=0}^k\max\left\{(1-\alpha)^l,\beta^l\right\}\mathbf{E}\|\nabla f(x^{k-l})\|_2^2 + \frac{\sigma^2}{n(1-\beta)^2}\left(\frac{3}{\alpha_p}-2\right),$$

where $\delta\overset{\text{def}}{=}1+\frac{2}{n}\left(\frac{1}{\alpha_p}-1\right)\left(1+\frac{\alpha}{|1-\alpha-\beta|}\right)$, and as a result

$$\frac{\gamma^3L^2\beta^2}{2(1-\beta)^3}\mathbf{E}\|v^{k-1}\|_2^2 \quad\le\quad \frac{\gamma^3L^2\beta^2\delta}{2(1-\beta)^4}\sum_{l=0}^{k-1}\max\left\{(1-\alpha)^{k-1-l},\beta^{k-1-l}\right\}\mathbf{E}\|\nabla f(x^l)\|_2^2 + \frac{\gamma^3L^2\beta^2\sigma^2}{2n(1-\beta)^5}\left(\frac{3}{\alpha_p}-2\right).$$

Putting all together and using designation $\xi \overset{\text{def}}{=} \max\{1-\alpha, \beta\}$ we have

$$\begin{aligned}
\mathbf{E}\left[f(z^{k+1})\right] &\leq \mathbf{E}\left[f(z^k)\right] - \frac{\gamma}{2(1-\beta)}\left(1 - \frac{L\gamma\omega}{1-\beta}\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 \\
&\quad + \left(\frac{L\gamma^2\alpha(\omega-1)}{2(1-\beta)^2} + \frac{\gamma^3L^2\beta^2\delta}{2(1-\beta)^4}\right)\sum_{l=0}^{k-1}\xi^{k-1-l}\mathbf{E}\|\nabla f(x^l)\|_2^2 \\
&\quad + \frac{\sigma^2}{n}\left(\frac{3}{\alpha_p} - 2\right)\left(\frac{L\gamma^2}{2(1-\beta)^2} + \frac{\gamma^3L^2\beta^2}{2(1-\beta)^5}\right).
\end{aligned}$$

Telescoping this inequality from 0 to $k-1$, we get

$$\begin{aligned}
\mathbf{E}f(z^k) - f(z^0) &\leq k\frac{\sigma^2}{n}\left(\frac{3}{\alpha_p} - 2\right)\left(\frac{L\gamma^2}{2(1-\beta)^2} + \frac{\gamma^3L^2\beta^2}{2(1-\beta)^5}\right) \\
&\quad + \frac{\gamma}{2}\sum_{l=0}^{k-2}\left(\left(\frac{L\gamma\alpha(\omega-1)}{(1-\beta)^2} + \frac{\gamma^2L^2\beta^2\delta}{(1-\beta)^4}\right)\sum_{k'=l+1}^{k-1}\xi^{k'-1-l} + \frac{L\gamma\omega}{(1-\beta)^2} - \frac{1}{1-\beta}\right)\mathbf{E}\|\nabla f(x^l)\|_2^2 \\
&\quad + \frac{\gamma}{2}\left(\frac{L\gamma\omega}{(1-\beta)^2} - \frac{1}{1-\beta}\right)\mathbf{E}\|\nabla f(x^{k-1})\|_2^2 \\
&\leq k\frac{\sigma^2}{n}\left(\frac{3}{\alpha_p} - 2\right)\left(\frac{L\gamma^2}{2(1-\beta)^2} + \frac{\gamma^3L^2\beta^2}{2(1-\beta)^5}\right) \\
&\quad + \frac{\gamma}{2}\sum_{l=0}^{k-1}\left(\frac{\gamma^2L^2\beta^2\delta}{(1-\beta)^4(1-\xi)} + \frac{L\gamma}{(1-\beta)^2}\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right) - \frac{1}{1-\beta}\right)\mathbf{E}\|\nabla f(x^l)\|_2^2
\end{aligned}$$

It holds $f^* \leq f(z^k)$ and our assumption on β implies that $\frac{\gamma^2L^2\beta^2\delta}{(1-\beta)^4(1-\xi)} + \frac{L\gamma}{(1-\beta)^2}\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right) - \frac{1}{1-\beta} \leq -\frac{1}{2}$, so it all results in

$$\frac{1}{k}\sum_{l=0}^{k-1}\|\nabla f(x^l)\|_2^2 \leq \frac{4(f(z^0) - f^*)}{\gamma k} + 2\gamma\frac{L\sigma^2}{(1-\beta)^2n}\left(\frac{3}{\alpha_p} - 2\right) + 2\gamma^2\frac{L^2\beta^2\sigma^2}{(1-\beta)^5n}\left(\frac{3}{\alpha_p} - 2\right).$$

Since $\overline{x}^k$ is sampled uniformly from $\{x^0, \ldots, x^{k-1}\}$, the left-hand side is equal to $\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2$. Also note that $z^0 = x^0$. □

Corollary 7. *If we set* $\gamma = \frac{1-\beta^2}{2\sqrt{k}L\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right)}$ *and* β *such that* $\frac{\beta^2}{(1-\beta)^2(1-\xi)} \leq \frac{4k\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right)}{\delta}$ *with* $k > 1$*, then the accuracy after* k *iterations is at most*

$$\frac{1}{\sqrt{k}}\left(\frac{8L\zeta(f(x^0) - f^*)}{1-\beta^2} + \frac{(1+\beta)\sigma^2}{\zeta\alpha_p n(1-\beta)}\left(\frac{3}{\alpha_p} - 2\right)\right) + \frac{1}{k}\frac{(1+\beta)^4\beta^2\sigma^2}{2(1-\beta)\zeta\alpha_p n}\left(\frac{3}{\alpha_p} - 2\right),$$

where $\zeta = \omega + \frac{\alpha(\omega-1)}{1-\xi}$.

Proof. Our choice of $\gamma = \frac{1-\beta^2}{2\sqrt{k}L\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right)}$ implies that

$$\frac{\beta^2}{(1-\beta)^2(1-\xi)} \leq \frac{4k\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right)}{\delta} \iff \frac{\beta^2}{(1-\beta)^2(1-\xi)} \leq \frac{1 - \beta^2 - 2L\gamma\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right)}{\gamma^2L^2\delta}.$$

After that it remains to put $\gamma = \frac{1-\beta^2}{2\sqrt{k}L\left(\omega + \frac{\alpha(\omega-1)}{1-\xi}\right)}$ in $\frac{4(f(z^0)-f^*)}{\gamma k} + 2\gamma\frac{L\sigma^2}{(1-\beta)^2n}\left(\frac{3}{\alpha_p} - 2\right) + 2\gamma^2\frac{L^2\beta^2\sigma^2}{(1-\beta)^5n}\left(\frac{3}{\alpha_p} - 2\right)$ to get the desired result. □

K. Analysis of DIANA with $\alpha = 0$ and $h_i^0 = 0$

K.1. Technical lemmas

First of all, we notice that since TernGrad coincides with Diana, having $h_i^k = 0, i, k \geq 1$, $\alpha = 0$ and $p = \infty$, all inequalities from Lemma 3 holds for the iterates of TernGrad as well because $\Delta_i^k = g_i^k$ and $\hat{\Delta}_i^k = \hat{g}_i^k$.

Lemma 7. *Assume $\gamma \leq \frac{n\alpha_p}{L((n-1)\alpha_p+1)}$. Then*

$$2\gamma\mu\left(1 - \frac{\gamma L((n-1)\alpha_p+1)}{2n\alpha_p}\right) \geq \gamma\mu. \tag{46}$$

Proof. Since $\gamma \leq \frac{n\alpha_p}{L((n-1)\alpha_p+1)}$ we have

$$2\gamma\mu\left(1 - \frac{\gamma L((n-1)\alpha_p+1)}{2n\alpha_p}\right) \geq 2\gamma\mu\left(1 - \frac{1}{2}\right) = \gamma\mu.$$

□

Lemma 8. *Assume $\gamma \leq \frac{1}{L(1+\kappa(1-\alpha_p)/(n\alpha_p))}$, where $\kappa \overset{def}{=} \frac{L}{\mu}$ is the condition number of f. Then*

$$r \geq \gamma\mu, \tag{47}$$

where $r = 2\mu\gamma - \gamma^2\left(\mu L + \frac{L^2(1-\alpha_p)}{n\alpha_p}\right)$.

Proof. Since $\gamma \leq \frac{1}{L(1+\kappa(1-\alpha_p)/(n\alpha_p))} = \frac{\mu n\alpha_p}{\mu n\alpha_p L + L^2(1-\alpha_p)}$ we have

$$n\alpha_p r = \gamma\left(2\mu n\alpha_p - \gamma\left(\mu n\alpha_p L + L^2(1-\alpha_p)\right)\right) \geq \gamma\mu n\alpha_p,$$

whence $r \geq \gamma\mu$. □

Lemma 9. *Assume $\gamma \leq \frac{2n\alpha_p}{(\mu+L)(2+(n-2)\alpha_p)}$. Then*

$$2\gamma\mu - \gamma^2\mu^2\left(1 + \frac{2(1-\alpha_p)}{n\alpha_p}\right) \quad \geq \quad \gamma\mu. \tag{48}$$

Proof. Since $\gamma \leq \frac{2n\alpha_p}{(\mu+L)(2+(n-2)\alpha_p)}$ we have

$$\gamma\mu \leq \frac{2\mu n\alpha_p}{(\mu+L)(2+(n-2)\alpha_p)} \leq \frac{(\mu+L)n\alpha_p}{(\mu+L)(2+(n-2)\alpha_p)} = \frac{n\alpha_p}{2+(n-2)\alpha_p},$$

whence

$$2\gamma\mu - \gamma^2\mu^2\left(1 + \frac{2(1-\alpha_p)}{n\alpha_p}\right) \geq 2\gamma\mu - \gamma\mu\frac{n\alpha_p}{2+(n-2)\alpha_p}\left(1 + \frac{2(1-\alpha_p)}{n\alpha_p}\right) = 2\gamma\mu - \gamma\mu = \gamma\mu.$$

□

Lemma 10. *Assume that each function f_i is L-smooth and R is such that exists a closed convex set $\mathcal{X}$ satisfying 1) $\forall z \in \mathbb{R}^n$ $\mathrm{prox}_{\gamma R}(z) \in \mathcal{X}$ and 2) $\forall z \in \mathcal{X}$ $z = \mathrm{prox}_{\gamma R}(z)$ (e.g. indicator function of $\mathcal{X}$). Then for the iterates of Algorithm 2 with $\gamma^k = \gamma$ we have*

$$\mathbf{E}\Theta^{k+1} \quad \leq \quad \mathbf{E}\Theta^k + \left(\frac{\gamma^2 L}{2} - \gamma\right)\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{\gamma^2 L}{2n^2}\left(\frac{1}{\alpha_p} - 1\right)\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|_2^2\right] + \frac{\gamma^2 L\sigma^2}{2n}, \tag{49}$$

where $\Theta^k = f(x^k) - f(x^*)$ and $\sigma^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

Proof. Since $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma\overline{g}^k)$and $x^k = \mathrm{prox}_{\gamma R}(x^k)$, due to non-expansiveness we have $\|x^{k+1} - x^k\|_2 \leq \|x^k - \gamma\overline{g}^k - x^k\|_2 = \gamma\|\overline{g}^k\|_2$. Moreover, from the L-smoothness of f we have

$$\begin{aligned}
\mathbf{E}\Theta^{k+1} &\leq \mathbf{E}\Theta^k + \mathbf{E}\left[\langle\nabla f(x^k), x^{k+1} - x^k\rangle\right] + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\
&= \mathbf{E}\Theta^k - \gamma\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{\gamma^2 L}{2}\mathbf{E}\left[\left\|\hat{g}^k\right\|_2^2\right] \\
&\overset{(29)}{\leq} \mathbf{E}\Theta^k + \left(\frac{\gamma^2 L}{2} - \gamma\right)\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{\gamma^2 L}{2n^2}\sum_{i=1}^n\sum_{l=1}^m \mathbf{E}\left[\|g_i^k(l)\|_1\|g_i^k(l)\|_p - \|g_i^k(l)\|_2^2\right] + \frac{\gamma^2 L}{2n^2}\sum_{i=1}^n \sigma_i^2,
\end{aligned}$$

where the first equality follows from $x^{k+1} - x^k = \hat{g}^k$, $\mathbf{E}\left[\hat{g}^k \mid x^k\right] = \nabla f(x^k)$ and the tower property of mathematical expectation. By definition $\alpha_p(d_l) = \inf\limits_{x\neq 0, x\in\mathbb{R}^{d_l}} \frac{\|x\|_2^2}{\|x\|_1\|x\|_p} = \left(\sup\limits_{x\neq 0, x\in\mathbb{R}^{d_l}} \frac{\|x\|_1\|x\|_p}{\|x\|_2^2}\right)^{-1}$ and $\alpha_p = \alpha_p(\max\limits_{l=1,\ldots,m} d_l)$ which implies

$$\begin{aligned}
\mathbf{E}\left[\|g_i^k(l)\|_1\|g_i^k(l)\|_p - \|g_i^k(l)\|_2^2\right] &= \mathbf{E}\left[\|g_i^k(l)\|_2^2\left(\frac{\|g_i^k(l)\|_1\|g_i^k(l)\|_p}{\|g_i^k(l)\|_2^2} - 1\right)\right] \leq \left(\frac{1}{\alpha_p(d_l)} - 1\right)\mathbf{E}\|g_i^k(l)\|_2^2 \\
&\leq \left(\frac{1}{\alpha_p} - 1\right)\mathbf{E}\|g_i^k(l)\|_2^2.
\end{aligned}$$

Since $\|g_i^k\|_2^2 = \sum\limits_{l=1}^m \|g_i^k(l)\|_2^2$ we have

$$\mathbf{E}\Theta^{k+1} \leq \mathbf{E}\Theta^k + \left(\frac{\gamma^2 L}{2} - \gamma\right)\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{\gamma^2 L}{2n^2}\left(\frac{1}{\alpha_p} - 1\right)\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|_2^2\right] + \frac{\gamma^2 L\sigma^2}{2n},$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. □

K.2. Non-convex analysis for indicator-like regularizer

Theorem 8. *Assume R is such that exists a closed convex set $\mathcal{X}$ satisfying 1) $\forall z \in \mathbb{R}^n$ $\mathrm{prox}_{\gamma R}(z) \in \mathcal{X}$ and 2) $\forall z \in \mathcal{X}$ $z = \mathrm{prox}_{\gamma R}(z)$ (e.g. indicator function of $\mathcal{X}$), f is L-smooth, $\gamma \leq \frac{n\alpha_p}{L((n-1)\alpha_p+1)}$ and $\overline{x}^k$ is chosen randomly from $\{x^0, \ldots, x^{k-1}\}$. If, further, every worker samples from the full dataset, then*

$$\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2 \leq \frac{2}{k}\frac{f(x^0) - f(x^*)}{\gamma\left(2 - \gamma\frac{L((n-1)\alpha_p+1)}{n\alpha_p}\right)} + \frac{\gamma L\sigma^2}{n\alpha_p}.$$

Proof. Recall that we defined Θ^k as $f(x^k) - f(x^*)$ in Lemma 10. From (49) we have

$$\mathbf{E}\Theta^{k+1} \leq \mathbf{E}\Theta^k + \left(\frac{\gamma^2 L}{2} - \gamma\right)\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{\gamma^2 L}{2n^2}\left(\frac{1}{\alpha_p} - 1\right)\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|_2^2\right] + \frac{\gamma^2 L\sigma^2}{2n}.$$

Using the standard decomposition

$$\mathbf{E}\left[\|g_i^k\|_2^2\right] = \underbrace{\mathbf{E}\left[\|\nabla f_i(x^k)\|_2^2\right]}_{\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right]} + \sigma_i^2$$

we get

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|_2^2\right] \leq \mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \sigma^2.$$

Putting all together we obtain

$$\mathbf{E}\Theta^{k+1} \quad \leq \quad \mathbf{E}\Theta^k + \left(\frac{\gamma^2 L}{2}\cdot\frac{(n-1)\alpha_p+1}{n\alpha_p} - \gamma\right)\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{\gamma^2 L\sigma^2}{2n\alpha_p}. \tag{50}$$

Since $\gamma \leq \frac{n\alpha_p}{L((n-1)\alpha_p+1)}$ the factor $\left(\frac{\gamma^2 L}{2}\cdot\frac{(n-1)\alpha_p+1}{n\alpha_p} - \gamma\right)$ is negative and therefore

$$\mathbf{E}\left[\|\nabla f(x^k)\|_2^2\right] \quad \leq \quad \frac{\mathbf{E}\Theta^k - \mathbf{E}\Theta^{k+1}}{\gamma\left(1-\gamma\frac{L((n-1)\alpha_p+1)}{2n\alpha_p}\right)} + \frac{\gamma L\sigma^2}{2n\alpha_p - \gamma L((n-1)\alpha_p+1)}.$$

Telescoping the previous inequality from 0 to $k-1$ and using $\gamma \leq \frac{n\alpha_p}{L((n-1)\alpha_p+1)}$ we obtain

$$\frac{1}{k}\sum_{l=0}^{k-1}\mathbf{E}\left[\|\nabla f(x^l)\|_2^2\right] \quad \leq \quad \frac{2}{k}\frac{\mathbf{E}\Theta^0 - \mathbf{E}\Theta^k}{\gamma\left(2-\gamma\frac{L((n-1)\alpha_p+1)}{n\alpha_p}\right)} + \frac{\gamma L\sigma^2}{n\alpha_p}.$$

It remains to notice that left-hand side is just $\mathbf{E}\left[\|\nabla f(\overline{x}^k)\|_2^2\right]$, $\Theta^k \geq 0$ and $\Theta^0 = f(x^0) - f(x^*)$. □

Corollary 8. *If we choose* $\gamma = \frac{n\alpha_p}{L((n-1)\alpha_p+1)\sqrt{K}}$ *then the rate we get is* $\frac{2}{\sqrt{K}}L\frac{(n-1)\alpha_p+1}{n\alpha_p}\left(f(x^0)-f(x^*)\right) + \frac{1}{\sqrt{K}}\frac{\sigma^2}{(1+(n-1)\alpha_p)}$.

Proof. Our choice of $\gamma = \frac{n\alpha_p}{L((n-1)\alpha_p+1)\sqrt{K}} \leq \frac{n\alpha_p}{L((n-1)\alpha_p+1)}$ implies that $2 - \gamma\frac{L((n-1)\alpha_p+1)}{n\alpha_p} \geq 1$. After that it remains to notice that for our choice of $\gamma = \frac{n\alpha_p}{L((n-1)\alpha_p+1)\sqrt{K}}$ we have $\frac{2}{K}\frac{f(x^0)-f(x^*)}{\gamma\left(2-\gamma\frac{L((n-1)\alpha_p+1)}{n\alpha_p}\right)} + \frac{\gamma L\sigma^2}{n\alpha_p} = \frac{2}{\sqrt{K}}L\frac{(n-1)\alpha_p+1}{n\alpha_p}\left(f(x^0)-f(x^*)\right) + \frac{1}{\sqrt{K}}\frac{\sigma^2}{(1+(n-1)\alpha_p)}$. □

K.3. Momentum version

Theorem 9. *Assume that f is L-smooth, $R \equiv const$, $\alpha = 0$, $h_i = 0$ and $f_i = f$ for all i. Choose $\beta < 1$ and $\gamma < \frac{1-\beta^2}{2L\omega}$ such that $\frac{\beta^2}{(1-\beta)^3} \leq \frac{1-\beta^2-2L\gamma\omega}{\gamma^2L^2\omega}$, where $\omega \overset{def}{=} \frac{n-1}{n} + \frac{1}{n\alpha_p}$ and sample $\overline{x}^k$ uniformly from $\{x^0,\ldots,x^{k-1}\}$. Then*

$$\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2 \leq \frac{4(f(z^0)-f^*)}{\gamma k} + 2\gamma\frac{L\sigma^2}{\alpha_p n(1-\beta)^2} + 2\gamma^2\frac{L^2\beta^2\sigma^2}{(1-\beta)^5\alpha_p n}.$$

Proof. The main idea of the proof is to find virtual iterates z^k whose recursion would satisfy $z^{k+1} = z^k - \frac{\gamma}{1-\beta}\hat{g}^k$. Having found it, we can prove convergence by writing a recursion on $f(z^k)$. One possible choice is defined below:

$$z^k \overset{\text{def}}{=} x^k - \frac{\gamma\beta}{1-\beta}v^{k-1}, \tag{51}$$

where for the edge case $k=0$ we simply set $v^{-1}=0$ and $z^0 = x^0$. Although z^k is just a slight perturbation of x^k, applying smoothness inequality (14) to it produces a more convenient bound than the one we would have if used x^k. But first of all, let us check that we have the desired recursion for z^{k+1}:

$$\begin{aligned} z^{k+1} &\overset{(51)}{=} x^{k+1} - \frac{\gamma\beta}{1-\beta}v^k \\ &= x^k - \frac{\gamma}{1-\beta}v^k \\ &= x^k - \frac{\gamma\beta}{1-\beta}v^{k-1} - \frac{\gamma}{1-\beta}\hat{g}^k \\ &\overset{(51)}{=} z^k - \frac{\gamma}{1-\beta}\hat{g}^k. \end{aligned}$$

Now, it is time to apply smoothness of f:

$$\begin{aligned}
\mathbf{E} f(z^{k+1}) &\le \mathbf{E}\left[f(z^k) + \left\langle \nabla f(z^k), z^{k+1} - z^k\right\rangle + \frac{L}{2}\|z^{k+1}-z^k\|_2^2\right]\\
&\overset{(51)}{=} \mathbf{E}\left[f(z^k) - \frac{\gamma}{1-\beta}\left\langle \nabla f(z^k), \hat{g}^k\right\rangle + \frac{L\gamma^2}{2(1-\beta)^2}\|\hat{g}^k\|_2^2\right].
\end{aligned} \tag{52}$$

Under our special assumption inequality (33) simplifies to

$$\mathbf{E}\left[\|\hat{g}^k\|_2^2 \mid x^k\right] \le \|\nabla f(x^k)\|_2^2 + \left(\frac{1}{\alpha_p} - 1\right)\frac{1}{n}\|\nabla f(x^k)\|_2^2 + \frac{\sigma^2}{\alpha_p n}.$$

The scalar product in (52) can be bounded using the fact that for any vectors a and b one has $-\langle a, b\rangle = \frac{1}{2}(\|a-b\|_2^2 - \|a\|_2^2 - \|b\|_2^2)$. In particular,

$$\begin{aligned}
-\frac{\gamma}{1-\beta}\left\langle \nabla f(z^k), \nabla f(x^k)\right\rangle &= \frac{\gamma}{2(1-\beta)}\left(\|\nabla f(x^k) - \nabla f(z^k)\|_2^2 - \|\nabla f(x^k)\|_2^2 - \|\nabla f(z^k)\|_2^2\right)\\
&\le \frac{\gamma}{2(1-\beta)}\left(L^2\|x^k - z^k\|_2^2 - \|\nabla f(x^k)\|_2^2\right)\\
&= \frac{\gamma^3 L^2\beta^2}{2(1-\beta)^3}\|v^{k-1}\|_2^2 - \frac{\gamma}{2(1-\beta)}\|\nabla f(x^k)\|_2^2.
\end{aligned}$$

The next step is to come up with an inequality for $\mathbf{E}\|v^k\|_2^2$. Since we initialize $v^{-1} = 0$, one can show by induction that

$$v^k = \sum_{l=0}^{k} \beta^l \hat{g}^{k-l}.$$

Define $B \overset{\text{def}}{=} \sum_{l=0}^k \beta^l = \frac{1-\beta^{k+1}}{1-\beta}$. Then, by Jensen's inequality

$$\mathbf{E}\|v^k\|_2^2 = B^2\mathbf{E}\left\|\sum_{l=0}^k \frac{\beta^l}{B}\hat{g}^{k-l}\right\|_2^2 \le B^2\sum_{l=0}^k \frac{\beta^l}{B}\mathbf{E}\|\hat{g}^{k-l}\|_2^2 \le B\sum_{l=0}^k \beta^l\left(\left(\frac{n-1}{n} + \frac{1}{n\alpha_p}\right)\mathbf{E}\|\nabla f(x^{k-l})\|_2^2 + \frac{\sigma^2}{\alpha_p n}\right).$$

Note that $B \le \frac{1}{1-\beta}$, so

$$\frac{\gamma^3 L^2\beta^2}{2(1-\beta)^3}\mathbf{E}\|v^{k-1}\|_2^2 \le \frac{\gamma^3 L^2\beta^2}{2(1-\beta)^5}\frac{\sigma^2}{\alpha_p n} + \omega\frac{\gamma^3 L^2\beta^2}{2(1-\beta)^4}\sum_{l=0}^{k-1}\beta^{k-1-l}\mathbf{E}\|\nabla f(x^l)\|_2^2$$

with $\omega \overset{\text{def}}{=} \frac{n-1}{n} + \frac{1}{n\alpha_p}$. We, thus, obtain

$$\begin{aligned}
\mathbf{E} f(z^{k+1}) &\le \mathbf{E} f(z^k) - \frac{\gamma}{2(1-\beta)}\left(1 - \frac{L\gamma\omega}{1-\beta}\right)\mathbf{E}\|\nabla f(x^k)\|_2^2 + \frac{L\gamma^2\sigma^2}{2n\alpha_p(1-\beta)^2} + \frac{\gamma^3 L^2\beta^2\sigma^2}{2(1-\beta)^5\alpha_p n}\\
&\quad + \omega\frac{\gamma^3 L^2\beta^2}{2(1-\beta)^4}\sum_{l=0}^{k-1}\beta^{k-1-l}\mathbf{E}\|\nabla f(x^l)\|_2^2.
\end{aligned}$$

Telescoping this inequality from 0 to $k-1$, we get

$$\begin{aligned}
\mathbf{E} f(z^k) - f(z^0) &\le k\left(\frac{L\gamma^2\sigma^2}{2\alpha_p n(1-\beta)^2} + \frac{\gamma^3 L^2\beta^2\sigma^2}{2(1-\beta)^5\alpha_p n}\right)\\
&\quad + \frac{\gamma}{2}\sum_{l=0}^{k-2}\left(\omega\frac{\gamma^2 L^2\beta^2}{(1-\beta)^4}\sum_{k'=l+1}^{k-1}\beta^{k'-1-l} + \frac{L\gamma\omega}{(1-\beta)^2} - \frac{1}{1-\beta}\right)\|\nabla f(x^l)\|_2^2\\
&\quad + \frac{\gamma}{2}\left(\frac{L\gamma\omega}{(1-\beta)^2} - \frac{1}{1-\beta}\right)\mathbf{E}\|\nabla f(x^{k-1})\|_2^2\\
&\le k\left(\frac{L\gamma^2\sigma^2}{2\alpha_p n(1-\beta)^2} + \frac{\gamma^3 L^2\beta^2\sigma^2}{2(1-\beta)^5\alpha_p n}\right) + \frac{\gamma}{2}\sum_{l=0}^{k-1}\left(\omega\frac{\gamma^2 L^2\beta^2}{(1-\beta)^5} + \frac{L\gamma\omega}{(1-\beta)^2} - \frac{1}{1-\beta}\right)\|\nabla f(x^l)\|_2^2.
\end{aligned}$$

It holds $f^* \leq f(z^k)$ and our assumption on β implies that $\omega\frac{\gamma^2 L^2 \beta^2}{(1-\beta)^5} + \frac{L\gamma\omega}{(1-\beta)^2} - \frac{1}{1-\beta} \leq -\frac{1}{2}$, so it all results in

$$\frac{1}{k}\sum_{l=0}^{k-1}\|\nabla f(x^l)\|_2^2 \leq \frac{4(f(z^0)-f^*)}{\gamma k} + 2\gamma\frac{L\sigma^2}{\alpha_p n(1-\beta)^2} + 2\gamma^2\frac{L^2\beta^2\sigma^2}{(1-\beta)^5\alpha_p n}.$$

Since $\overline{x}^k$ is sampled uniformly from $\{x^0,\ldots,x^{k-1}\}$, the left-hand side is equal to $\mathbf{E}\|\nabla f(\overline{x}^k)\|_2^2$. Also note that $z^0 = x^0$. □

Corollary 9. *If we set $\gamma = \frac{1-\beta^2}{2\sqrt{k}L\omega}$, where $\omega = \frac{n-1}{n} + \frac{1}{n\alpha_p}$, and β such that $\frac{\beta^2}{(1-\beta)^3} \leq 4k\omega$ with $k > 1$, then the accuracy after k iterations is at most*

$$\frac{1}{\sqrt{k}}\left(\frac{8L\omega(f(x^0)-f^*)}{1-\beta^2} + \frac{(1+\beta)\sigma^2}{\omega\alpha_p n(1-\beta)}\right) + \frac{1}{k}\frac{(1+\beta)^4\beta^2\sigma^2}{2(1-\beta)\omega\alpha_p n}.$$

Proof. Our choice of $\gamma = \frac{1-\beta^2}{2\sqrt{k}L\omega}$ implies that

$$\frac{\beta^2}{(1-\beta)^3} \leq \frac{1-\beta^2-2L\gamma\omega}{\gamma^2 L^2\omega} \Longleftrightarrow \frac{\beta^2}{(1-\beta)^3} \leq 4k\omega.$$

After that it remains to put $\gamma = \frac{1-\beta^2}{2\sqrt{k}L\omega}$ in $\frac{4(f(z^0)-f^*)}{\gamma k} + 2\gamma\frac{L\sigma^2}{\alpha_p n(1-\beta)^2} + 2\gamma^2\frac{L^2\beta^2\sigma^2}{(1-\beta)^5\alpha_p n}$ to get the desired result. □

K.4. Strongly convex analysis

Theorem 10. *Assume that each function f_i is μ-strongly convex and L-smooth. Choose stepsizes $\gamma^k = \gamma > 0$ satisfying*

$$\gamma \leq \frac{2n\alpha_p}{(\mu+L)(2+(n-2)\alpha_p)}. \tag{53}$$

If we run Algorithm 2 for k iterations with $\gamma^k = \gamma$, then

$$\mathbf{E}\left[\|x^k - x^*\|_2^2\right] \leq (1-\gamma\mu)^k\|x^0-x^*\|_2^2 + \frac{\gamma}{\mu}\left(\frac{\sigma^2}{n\alpha_p} + \frac{2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n\|h_i^*\|_2^2\right),$$

where $\sigma^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ and $h_i^* = \nabla f_i(x^*)$.

Proof. In the similar way as we did in the proof of Theorem 2 one can derive inequality (34) for the iterates of `TernGrad`:

$$\begin{aligned}\mathbf{E}\|x^{k+1}-x^*\|_2^2 \quad &\leq \quad \mathbf{E}\|x^k-x^*\|_2^2 - 2\gamma\mathbf{E}\left\langle\nabla f(x^k)-h^*, x^k-x^*\right\rangle \\ &\quad + \frac{\gamma^2}{n}\sum_{i=1}^n\mathbf{E}\|\nabla f_i(x^k)-h_i^*\|_2^2 + \frac{\gamma^2}{n^2}\sum_{i=1}^n\left(\mathbf{E}\Psi(g_i^k)\right) + \frac{\gamma^2\sigma^2}{n}.\end{aligned}$$

By definition $\alpha_p(d_l) = \inf\limits_{x\neq 0, x\in\mathbb{R}^{d_l}}\frac{\|x\|_2^2}{\|x\|_1\|x\|_p} = \left(\sup\limits_{x\neq 0, x\in\mathbb{R}^{d_l}}\frac{\|x\|_1\|x\|_p}{\|x\|_2^2}\right)^{-1}$ and $\alpha_p = \alpha_p(\max\limits_{l=1,\ldots,m} d_l)$ which implies

$$\begin{aligned}\mathbf{E}\left[\Psi_l(g_i^k)\right] \quad &= \quad \mathbf{E}\left[\|g_i^k(l)\|_1\|g_i^k(l)\|_p - \|g_i^k(l)\|_2^2\right] = \mathbf{E}\left[\|g_i^k(l)\|_2^2\left(\frac{\|g_i^k(l)\|_1\|g_i^k(l)\|_p}{\|g_i^k(l)\|_2^2} - 1\right)\right] \\ &\leq \quad \left(\frac{1}{\alpha_p(d_l)} - 1\right)\mathbf{E}\|g_i^k(l)\|_2^2 \leq \left(\frac{1}{\alpha_p} - 1\right)\mathbf{E}\|g_i^k(l)\|_2^2.\end{aligned}$$

Moreover,

$$\|g_i^k\|_2^2 = \sum_{l=1}^m\|g_i^k(l)\|_2^2, \quad \Psi(g_i^k) = \sum_{l=1}^m\Psi_l(g_i^k).$$

This helps us to get the following inequality

$$
\begin{aligned}
\mathbf{E}\|x^{k+1}-x^*\|_2^2 \quad &\le \quad \mathbf{E}\|x^k-x^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k)-h^*, x^k-x^*\right\rangle \\
&\qquad + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k)-h_i^*\|_2^2 + \frac{\gamma^2}{n^2}\left(\frac{1}{\alpha_p}-1\right)\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|_2^2\right] + \frac{\gamma^2\sigma^2}{n}.
\end{aligned}
$$

Using tower property of mathematical expectation and $\mathbf{E}\left[\|g_i^k\|_2^2 \mid x^k\right] = \mathbf{E}\left[\|g_i^k - \nabla f_i(x^k)\|_2^2 \mid x^k\right] + \|\nabla f_i(x^k)\|_2^2 \le \sigma_i^2 + \|\nabla f_i(x^k)\|_2^2$ we obtain

$$
\mathbf{E}\|g_i^k\|_2^2 \le \mathbf{E}\|\nabla f_i(x^k)\|_2^2 + \sigma_i^2 \le 2\mathbf{E}\|\nabla f_i(x^k)-h_i^*\|_2^2 + 2\|h_i^*\|_2^2 + \sigma_i^2,
$$

where the last inequality follows from the fact that for all $x, y \in \mathbb{R}^n$ the inequality $\|x+y\|_2^2 \le 2\left(\|x\|_2^2+\|y\|_2^2\right)$ holds. Putting all together we have

$$
\begin{aligned}
\mathbf{E}\|x^{k+1}-x^*\|_2^2 \quad &\le \quad \mathbf{E}\|x^k-x^*\|_2^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x^k)-h^*, x^k-x^*\right\rangle \\
&\qquad + \frac{\gamma^2}{n}\left(1+\frac{2(1-\alpha_p)}{n\alpha_p}\right)\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k)-h_i^*\|_2^2 + \frac{2\gamma^2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2 + \frac{\gamma^2\sigma^2}{n\alpha_p}.
\end{aligned}
$$

Using the splitting trick (40) we get

$$
\begin{aligned}
\mathbf{E}\|x^{k+1}-x^*\|_2^2 \quad &\le \quad \left(1-\frac{2\gamma\mu L}{\mu+L}\right)\mathbf{E}\|x^k-x^*\|_2^2 + \frac{1}{n}\left(\gamma^2\left(1+\frac{2(1-\alpha_p)}{n\alpha_p}\right)-\frac{2\gamma}{\mu+L}\right)\sum_{i=1}^n \mathbf{E}\|\nabla f_i(x^k)-h_i^*\|_2^2 \\
&\qquad + \frac{2\gamma^2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2 + \frac{\gamma^2\sigma^2}{n\alpha_p}.
\end{aligned} \tag{54}
$$

Since $\gamma \le \frac{2n\alpha_p}{(\mu+L)(2+(n-2)\alpha_p)}$ the term $\left(\gamma^2\left(1+\frac{2(1-\alpha_p)}{n\alpha_p}\right)-\frac{2\gamma}{\mu+L}\right)$ is non-negative. Moreover, since f_i is μ–strongly convex, we have $\mu\|x^k-x^*\|_2^2 \le \langle \nabla f_i(x^k)-h_i^*, x^k-x^*\rangle$. Applying the Cauchy-Schwarz inequality to further bound the right hand side, we get the inequality $\mu\|x^k-x^*\|_2 \le \|\nabla f_i(x^k)-h_i^*\|_2$. Using these observations, we can get rid of the second term in the (54) and absorb it with the first term, obtaining

$$
\begin{aligned}
\mathbf{E}\|x^{k+1}-x^*\|_2^2 \quad &\le \quad \left(1-2\gamma\mu+\gamma^2\mu^2\left(1+\frac{2(1-\alpha_p)}{n\alpha_p}\right)\right)\mathbf{E}\|x^k-x^*\|_2^2 + \frac{2\gamma^2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2 + \frac{\gamma^2\sigma^2}{n\alpha_p} \\
&\overset{(48)}{\le} \quad (1-\gamma\mu)\mathbf{E}\|x^k-x^*\|_2^2 + \gamma^2\left(\frac{\sigma^2}{n\alpha_p}+\frac{2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2\right).
\end{aligned}
$$

Finally, unrolling the recurrence leads to

$$
\begin{aligned}
\mathbf{E}\|x^k-x^*\|_2^2 &\le (1-\gamma\mu)^k\|x^0-x^*\|_2^2 + \sum_{l=0}^{k-1}(1-\gamma\mu)^l\gamma^2\left(\frac{\sigma^2}{n\alpha_p}+\frac{2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2\right) \\
&\le (1-\gamma\mu)^k\|x^0-x^*\|_2^2 + \sum_{l=0}^{\infty}(1-\gamma\mu)^l\gamma^2\left(\frac{\sigma^2}{n\alpha_p}+\frac{2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2\right) \\
&= (1-\gamma\mu)^k\|x^0-x^*\|_2^2 + \frac{\gamma}{\mu}\left(\frac{\sigma^2}{n\alpha_p}+\frac{2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2\right).
\end{aligned}
$$

□

K.5. Decreasing stepsize

Theorem 11. *Assume that f is L-smooth, μ-strongly convex and we have access to its gradients with bounded noise. Set $\gamma^k = \frac{2}{\mu k+\theta}$ with some $\theta \ge \frac{(\mu+L)(2+(n-2)\alpha_p)}{2n\alpha_p}$. After k iterations of Algorithm 2 we have*

$$
\mathbf{E}\|x^k-x^*\|_2^2 \le \frac{1}{\eta k+1}\max\left\{\|x^0-x^*\|_2^2, \frac{4}{\mu\theta}\left(\frac{\sigma^2}{n\alpha_p}+\frac{2(1-\alpha_p)}{n^2\alpha_p}\sum_{i=1}^n \|h_i^*\|_2^2\right)\right\},
$$

where $\eta \overset{def}{=} \frac{\mu}{\theta}$, $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ *and* $h_i^* = \nabla f_i(x^*)$.

Proof. To get a recurrence, let us recall an upper bound we have proved before in Theorem 10:

$$\mathbf{E}\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma^k \mu)\mathbf{E}\|x^k - x^*\|_2^2 + (\gamma^k)^2 \left(\frac{\sigma^2}{n\alpha_p} + \frac{2(1-\alpha_p)}{n^2\alpha_p} \sum_{i=1}^n \|h_i^*\|_2^2 \right).$$

Having that, we can apply Lemma 6 to the sequence $\mathbf{E}\|x^k - x^*\|_2^2$. The constants for the Lemma are: $N = \left(\frac{\sigma^2}{n\alpha_p} + \frac{2(1-\alpha_p)}{n^2\alpha_p} \sum_{i=1}^n \|h_i^*\|_2^2 \right)$ and $C = \max\left\{ \|x^0 - x^*\|_2^2, \frac{4}{\mu\theta}\left(\frac{\sigma^2}{n\alpha_p} + \frac{2(1-\alpha_p)}{n^2\alpha_p} \sum_{i=1}^n \|h_i^*\|_2^2 \right) \right\}$. □

Corollary 10. *If we choose* $\theta = \frac{(\mu+L)(2+(n-2)\alpha_p)}{2n\alpha_p}$, *then to achieve* $\mathbf{E}\|x^k - x^*\|_2^2 \leq \varepsilon$ *we need at most* $O\left(\frac{\kappa(1+n\alpha_p)}{n\alpha_p} \max\left\{ \|x^0 - x^*\|_2^2, \frac{n\alpha_p}{(1+n\alpha_p)\mu L}\left(\frac{\sigma^2}{n\alpha_p} + \frac{1-\alpha_p}{n^2\alpha_p} \sum_{i=1}^n \|h_i^*\|_2^2 \right) \right\} \frac{1}{\varepsilon} \right)$ *iterations, where* $\kappa \overset{def}{=} \frac{L}{\mu}$ *is the condition number of* f.

Proof. If $\theta = \frac{(\mu+L)(2+(n-2)\alpha_p)}{2n\alpha_p} = \Theta\left(\frac{L(1+n\alpha_p)}{n\alpha_p} \right)$, then $\eta = \Theta\left(\frac{n\alpha_p}{\kappa(1+n\alpha_p)} \right)$ and $\frac{1}{\mu\theta} = \Theta\left(\frac{n\alpha_p}{\mu L(1+n\alpha_p)} \right)$. Putting all together and using the bound from Theorem 11 we get the desired result. □

L. Detailed Numerical Experiments

L.1. Performance of DIANA, QSGD and Terngrad on the Rosenbrock function

In Figure 5 we illustrate the workings of DIANA, QSGD and TernGrad with 2 workers on the 2-dimensional (nonconvex) Rosenbrock function:

$$f(x, y) = (x - 1)^2 + 10(y - x^2)^2,$$

decomposed into average of $f_1 = (x + 16)^2 + 10(y - x^2)^2 + 16y$ and $f_2 = (x - 18)^2 + 10(y - x^2)^2 - 16y + \text{const}$. Each worker has access to its own piece of the Rosenbrock function with parameter $a = 1$ and $b = 10$. The gradients used are not stochastic, and we use 1-bit version of QSGD, so it also coincides with QGD in that situation. For all methods, its parameters were carefully tuned except for momentum and α, which were simply set to 0.9 and 0.5 correspondingly. We see that DIANA vastly outperforms the competing methods.

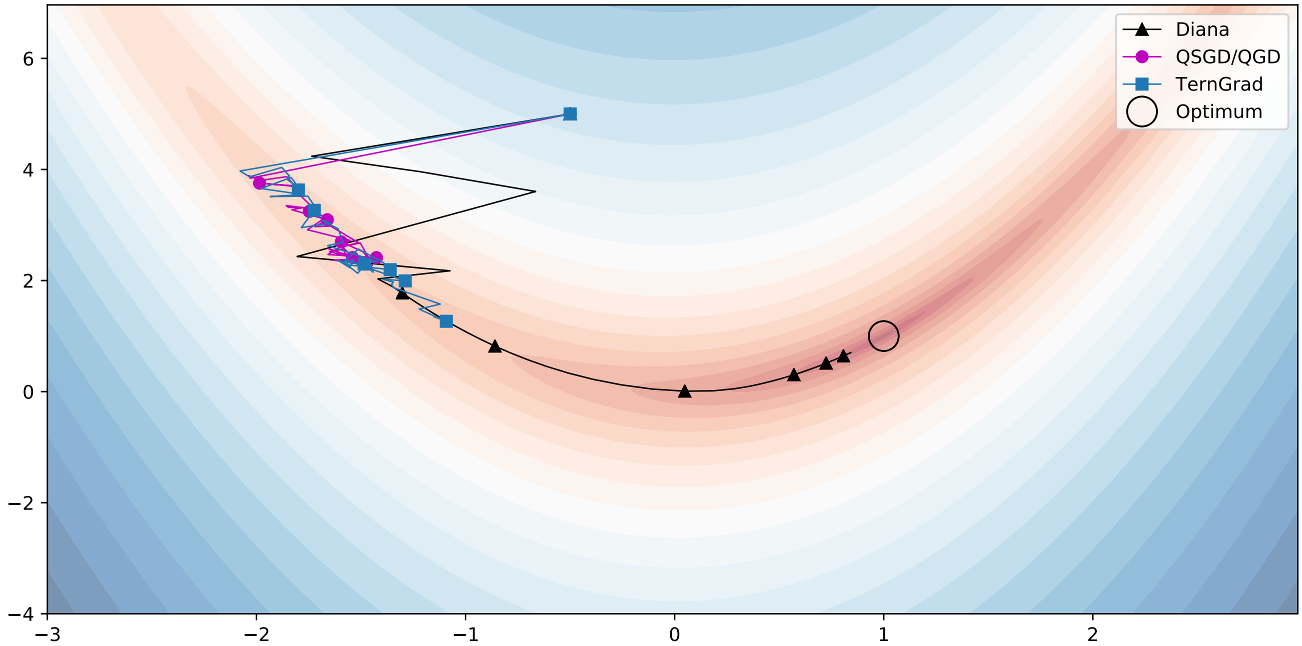

Figure 5. Illustration of the workings of DIANA, QSGD and TernGrad on the Rosenbrock function.

L.2. Logistic regression

We consider the logistic regression problem with ℓ_2 and ℓ_1 penalties for mushrooms dataset from LIBSVM. In our experiments we use ℓ_1-penalty coefficient $l_1 = 2 \cdot 10^{-3}$ and ℓ_2-penalty coefficient $l_2 = \frac{L}{n}$. The coefficient l_1 is adjusted in order to have sparse enough solution ($\approx 20\%$ non-zero values). The main goal of this series of experiment is to compare the optimal parameters for ℓ_2 and ℓ_∞ quantization.

L.2.1. What α is better to choose?

We run DIANA with zero momentum ($\beta = 0$) and obtain in our experiments that, actually, it is not important what α to choose for both ℓ_2 and ℓ_∞ quantization. The only thing that we need to control is that α is small enough.

L.2.2. What is the optimal block-size?

Since α is not so important, we run DIANA with $\alpha = 10^{-3}$ and zero momentum ($\beta = 0$) for different block sizes (see Figure 6). For the choice of ℓ_∞ quantization in our experiments it is always better to use full quantization. In the case of ℓ_2 quantization it depends on the regularization: if the regularization is big then optimal block-size ≈ 25 (dimension of the full vector of parameters is $d = 112$), but if the regularization is small it is better to use small block sizes.

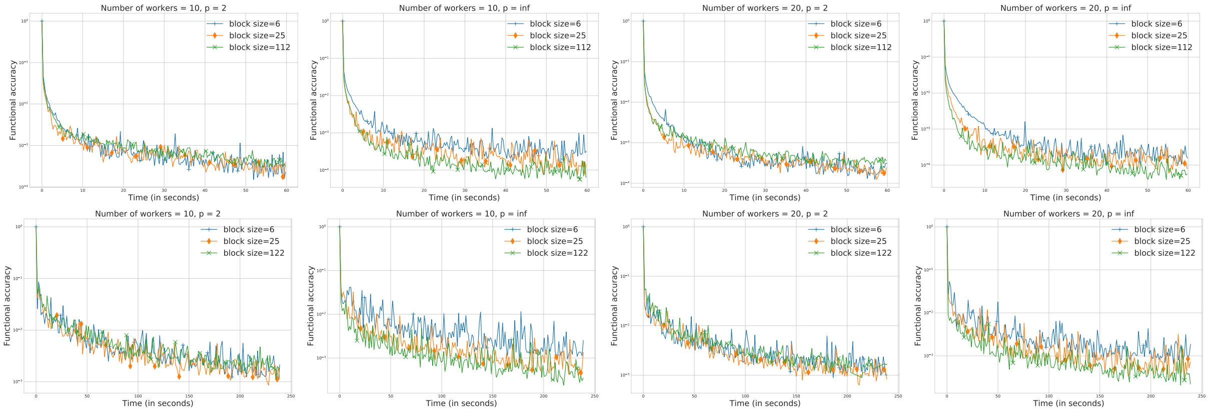

Figure 6. Comparison of the influence of the block sizes on convergence for ”mushrooms” (first row), ”a5a” (second row) datasets.

Dataset	n	d	Number of workers	Quantization	Optimal block size (approx.)
mushrooms	8124	112	10	ℓ_2	25
mushrooms	8124	112	10	ℓ_∞	112
mushrooms	8124	112	20	ℓ_2	25
mushrooms	8124	112	20	ℓ_∞	112
a5a	6414	122	10	ℓ_2	25
a5a	6414	122	10	ℓ_∞	112
a5a	6414	122	20	ℓ_2	25
a5a	6414	122	20	ℓ_∞	112

Table 3. Approximate optimal number of blocks for different dataset and configurations. Momentum equals zero for all experiments.

L.2.3. `DIANA` VS `QSGD` VS `TernGrad` VS `DQGD`

We compare `DIANA` (with momentum) with `QSGD`, `TernGrad` and `DQGD` on the ”mushrooms” dataset (See Figure)

L.3. MPI - broadcast, reduce and gather

In our experiments, we are running 4 MPI processes per physical node. Nodes are connected by Cray Aries High Speed Network.

We utilize 3 MPI collective operations, Broadcast, Reduce and Gather. When implementing `DIANA`, we could use P2P communication, but based on our experiments, we found that using Gather to collect data from workers significantly outperformed P2P communications.

In Figure 7 we show the duration of different communications for various MPI processes and message length. Note that Gather 2bit do not scale linearly (as would be expected). It turns out, we are not the only one who observed such a weird behavior when using cray MPI implementation (see (Chunduri et al., 2017) for a nice study obtained by a team from Argonne National Laboratory). To correct for the unexpected behavior, we have performed MPI Gather multiple times on shorter vectors, such that the master node obtained all data, but in much faster time (see *Multi-Gather 2bit*).

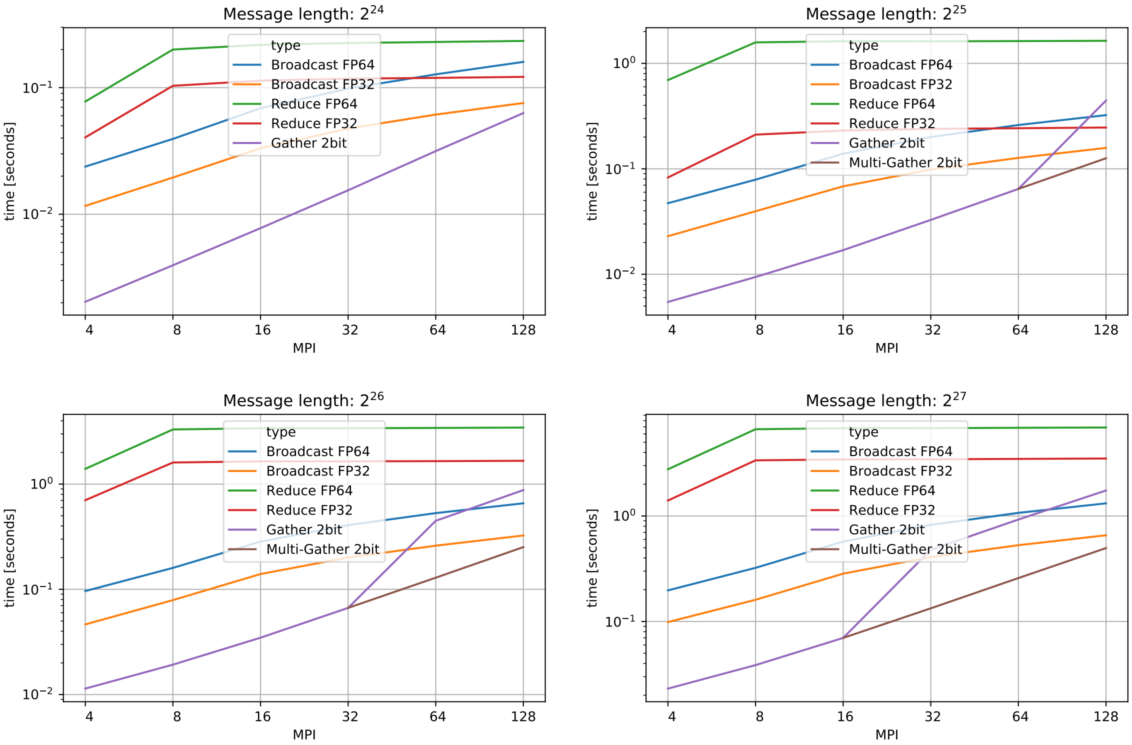

Figure 7. Time to communicate a vectors with different lengths for different methods as a function of # of MPI processes. One can observe that Gather 2bit is not having nice scaling. We also show that the proposed Multi-Gather communication still achieves a nice scaling when more MPI processes are used.

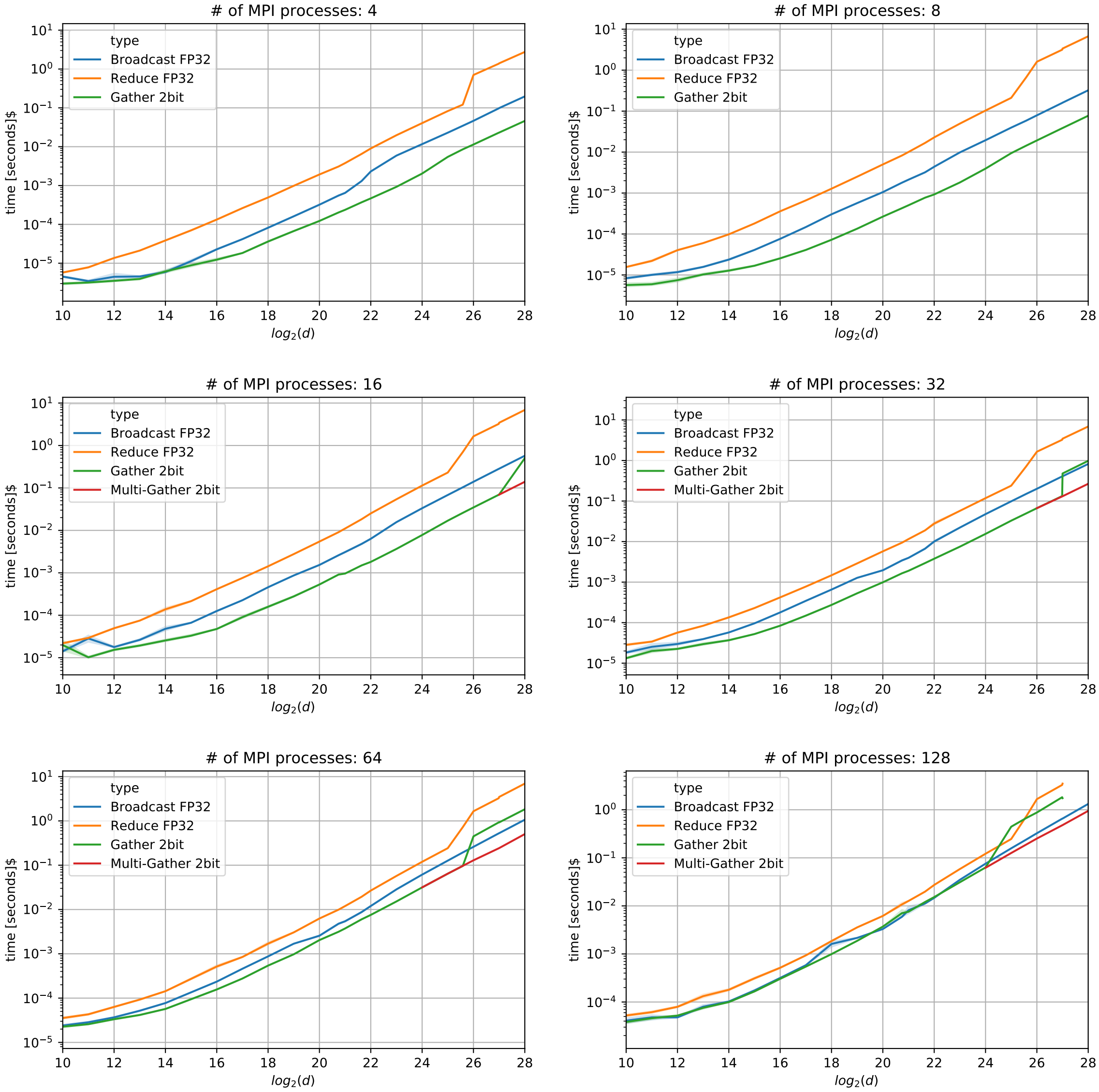

Figure 8. The duration of communication for MPI Broadcast, MPI Reduce and MPI Gather. We show how the communication time depends on the size of the vector in $\mathbb{R}^d$ (x-axis) for various # of MPI processes. In this experiment, we have run 4 MPI processes per computing node. For Broadcast and Reduce we have used a single precision floating point number. For Gather we used 2bits per dimension. For longer vectors and large number of MPI processes, one can observe that Gather has a very weird scaling issue. It turned out to be some weird behaviour of Cray-MPI implementation.

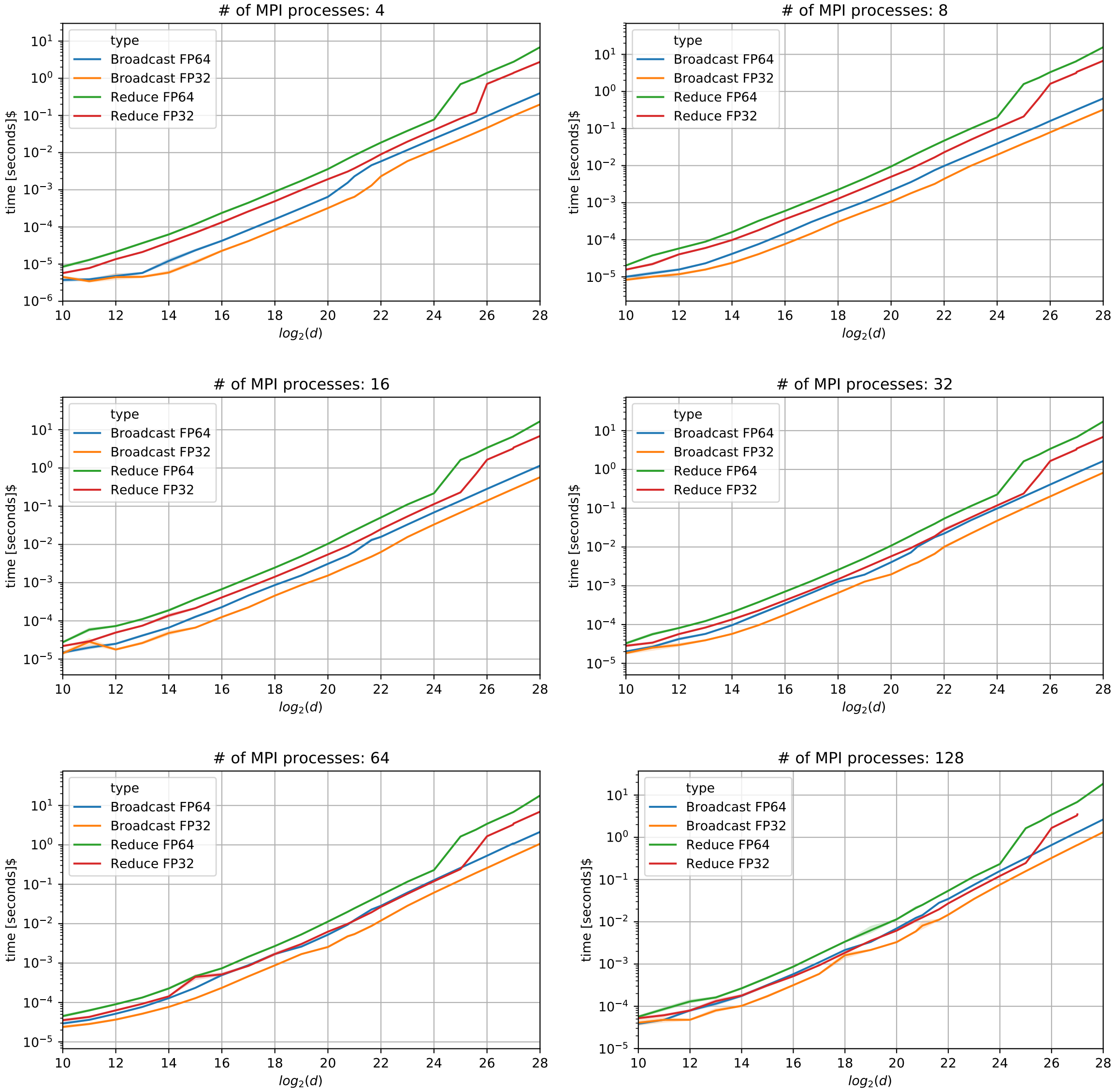

Figure 9. The duration of communication for MPI Broadcast, MPI Reduce for single precision (FP32) and double precision (FP64) floating numbers. We show how the communication time depends on the size of the vector in $\mathbb{R}^d$ (x-axis) for various # of MPI processes. In this experiment, we have run 4 MPI processes per computing node. We have used Cray implementation of MPI.

L.4. Performance of GPU

In Table 4 we list the DNN networks we have experimented in this paper.

model	d	# classes	input
LeNet	3.2M	10	$28 \times 28 \times 3$
CifarNet	1.7M	10	$32 \times 32 \times 3$
alexnet v2	50.3M	1,000	$224 \times 224 \times 3$
vgg a	132.8M	1,000	$224 \times 224 \times 3$

Table 4. Deep Neural Networks used in the experiments. The structure of the DNN is taken from `https://github.com/tensorflow/models/tree/master/research/slim`.

Figure 10 shows the performance of a single P100 GPU for different batch size, DNN network and operation.

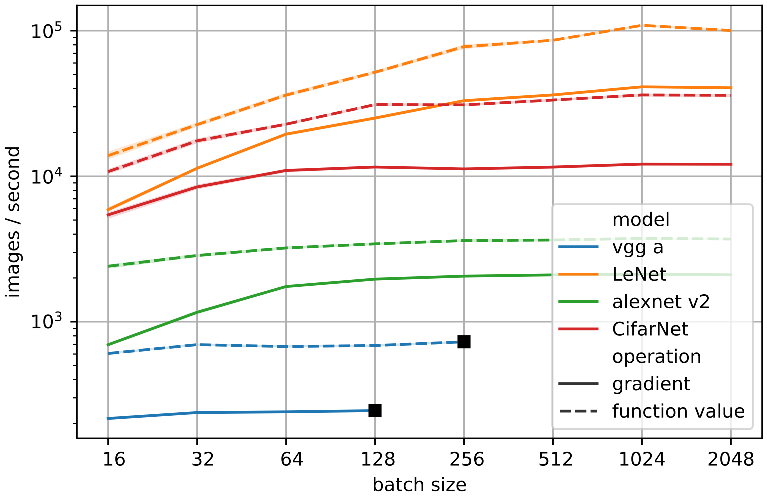

Figure 10. The performance (images/second) of NVIDIA Tesla P100 GPU on 4 different problems as a function of batch size. We show how different choice of batch size affects the speed of function evaluation and gradient evaluation. For vgg a, we have run out of memory on GPU for batch size larger than 128 (gradient evaluation) and 256 for function evaluation. Clearly, this graph suggest that choosing small batch size leads to small utilization of GPU. Note that using larger batch size do not necessary reduce the training process.

L.5. Diana vs. TenGrad, SGD and QSGD

In Figure 11 we compare the performance of DIANA vs. doing a MPI reduce operation with 32bit floats. The computing cluster had Cray Aries High Speed Network. However, for DIANA we used 2bit per dimension, we have experienced an weird scaling behaviour, which was documented also in(Chunduri et al., 2017). In our case, this affected speed for alexnet and vgg_a beyond 64 or 32 MPI processes respectively. For more detailed experiments, see Section L.3. In order to improve the speed of Gather, we impose a Multi-Gather strategy, when we call Gather multiple-times on shorter vectors. This significantly improved the communication cost of Gather (see Figure 8) and leads to much nicer scaling – see green bars – DIANA-MultiGather in Figure 11).

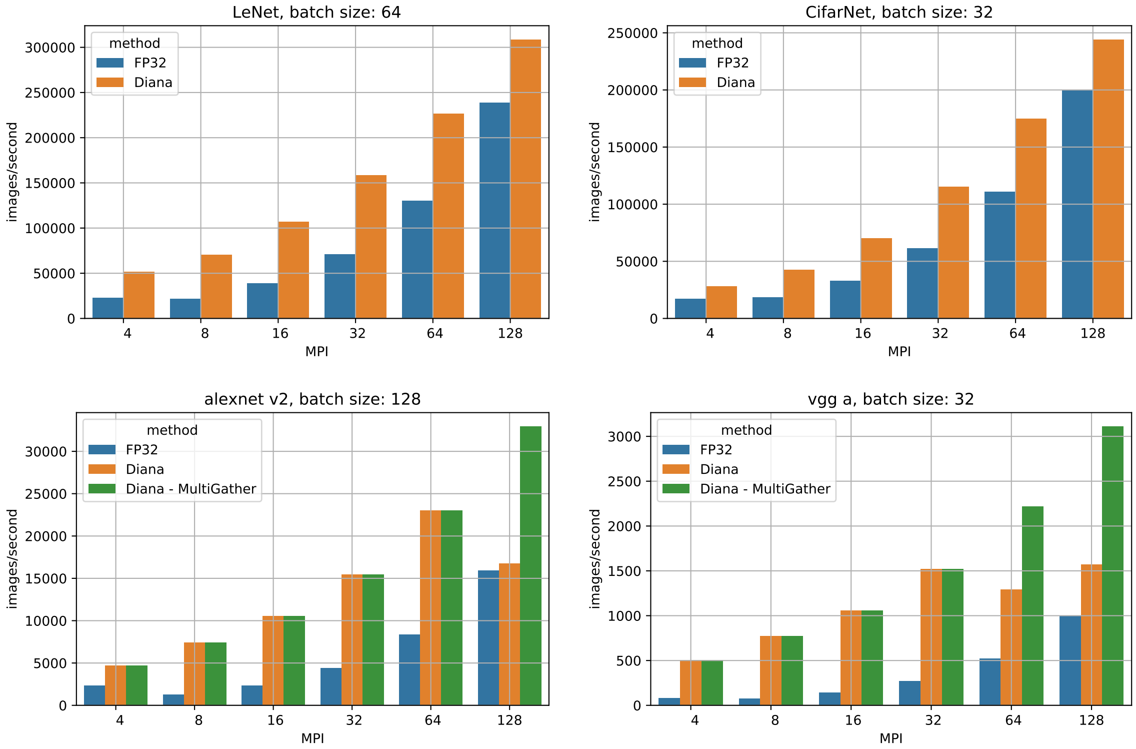

Figure 11. Comparison of performance (images/second) for various number of GPUs/MPI processes and sparse communication DIANA (2bit) vs. Reduce with 32bit float (FP32). We have run 4 MPI processes on each node. Each MPI process is using single P100 GPU. Note that increasing MPI from 4 to 8 will not bring any significant improvement for FP32, because with 8 MPI processes, communication will happen between computing nodes and will be significantly slower when compare to the single node communication with 4MPI processes.

In the next experiments, we run QSGD (Alistarh et al., 2017), TernGrad (Wen et al., 2017), SGD with momentum and DIANA on Mnist dataset and Cifar10 dataset for 3 epochs. We have selected 8 workers and run each method for learning rate from $\{0.1, 0.2, 0.05\}$. For QSGD, DIANA and TernGrad, we also tried various quantization bucket sizes in $\{32, 128, 512\}$. For QSGD we have chosen $2, 4, 8$ quantization levels. For DIANA we have chosen $\alpha \in \{0, 1.0/\sqrt{\text{quantization bucket sizes}}\}$ and have selected initial $h = 0$. For DIANA and SGD we also run a momentum version, with a momentum parameter in $\{0, 0.95, 0.99\}$. For DIANA we also run with two choices of norm ℓ_2 and ℓ_∞. For each experiment we have selected softmax cross entropy loss. Mnist-Convex is a simple DNN with no hidden layer, Mnist-DNN is a convolutional NN described here `https://github.com/floydhub/mnist/blob/master/ConvNet.py` and Cifar10-DNN is a convolutional DNN described here `https://github.com/kuangliu/pytorch-cifar/blob/master/models/lenet.py`. In Figure 12 we show the best runs over all the parameters for all the methods. For Mnist-Convex SGD and DIANA makes use of the momentum and dominate all other algorithms. For Mnist-DNN situation is very similar. For Cifar10-DNN both DIANA and SGD have significantly outperform other methods.

In Figure 13 show the evolution of sparsity of the quantized gradient for the 3 problems and DIANA, QSGD and TernGrad.

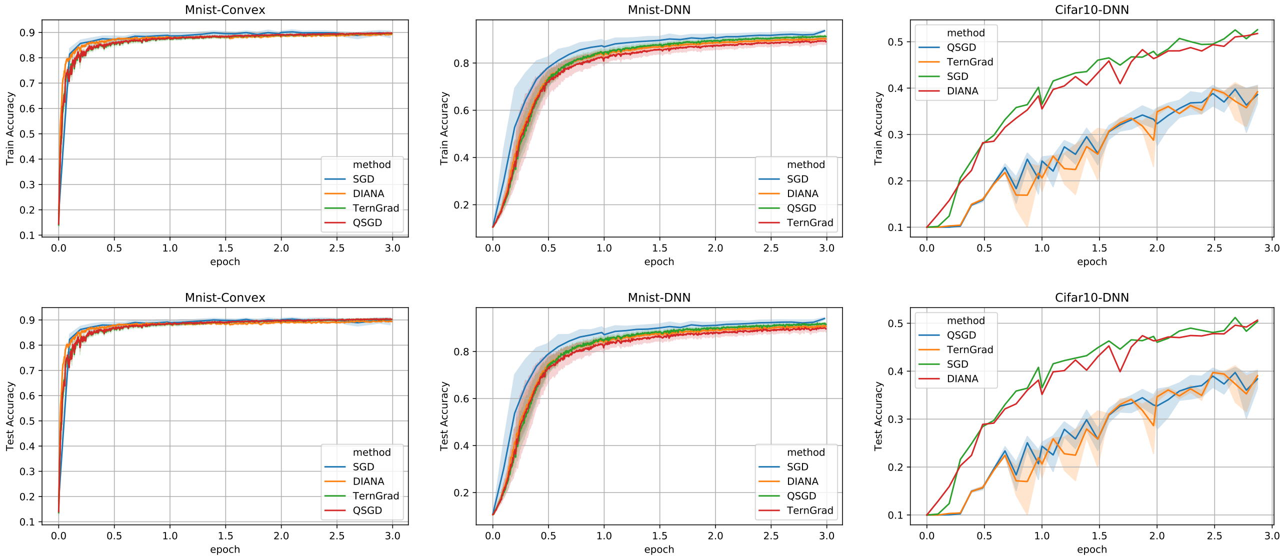

Figure 12. Evolution of training and testing accuracy for 3 different problems, using 4 algorithms: `DIANA`, `SGD`, `QSGD` and `TernGrad`. We have chosen the best runs over all tested hyper-parameters.

For Mnist-DNN, it seems that the quantized gradients are becoming sparser as the training progresses.

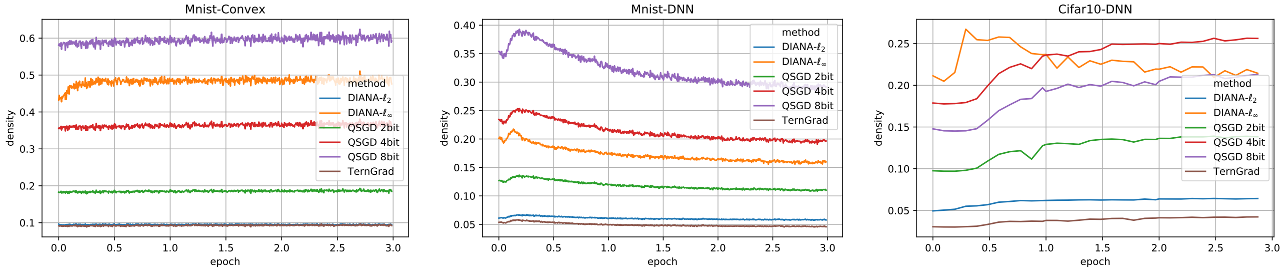

Figure 13. Evolution of sparsity of the quantized gradient for 3 different problems and 3 algorithms.

L.6. Computational Cost

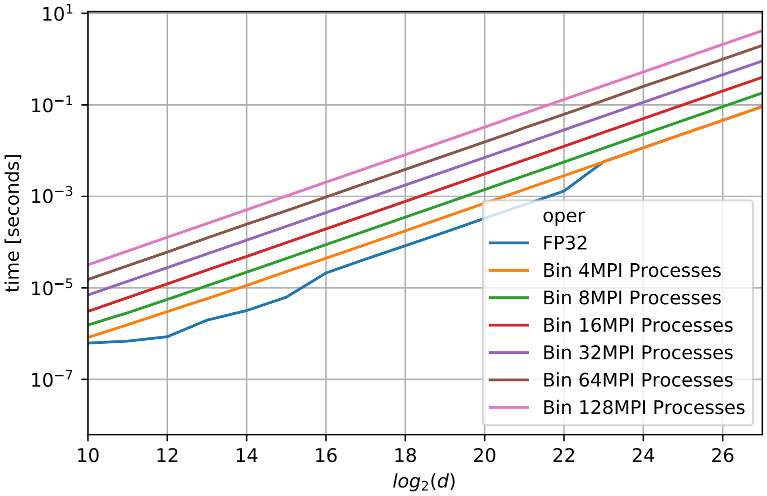

Figure 14. Comparison of a time needed to update weights after a reduce vs. the time needed to update the weights when using a sparse update from DIANA using 4-128 MPI processes and 10% sparsity.

Notation	Definition	First appearance
$f(x)$	Objective function, $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$	Eq.(1)
$R(x)$	Regularizer	Eq. (1)
n	Size of the dataset	Eq. (1)
d	Dimension of vector x	Eq. (11)
$\mathrm{sign}(t)$	The sign of t (-1 if $t < 0$, 0 if $t = 0$ and 1 if $t > 1$)	Eq. (6)
$x_{(j)}$	The j-th element of $x \in \mathbb{R}^d$	Eq. (6)
$x(l)$	The l-th subvector of $x = (x(1)^\top, x(2)^\top, \ldots, x(m)^\top)^\top$, $x(l) \in \mathbb{R}^{d_l}$, $\sum\limits_{l=1}^{m} d_l = d$	Def. 2
$\lVert x\rVert_p, p \geq 1$	ℓ_p norm of x: $\lVert x\rVert_p = \left(\sum\limits_{j=1}^{d} \lvert x_{(j)}\rvert^p\right)^{\frac{1}{p}}$ for $1 \leq p < \infty$ $\lVert x\rVert_\infty = \max\limits_{j=1,\ldots,d} \lVert x\rVert_\infty$	Eq. (14)
$\lVert x\rVert_0$	Number of nonzero elements of x	Eq. (11)
L	Lipschitz constant of the gradient of f w. r. t. ℓ_2 norm	Eq. (14)
μ	Strong convexity constant of f w. r. t. ℓ_2 norm	Eq. (15)
κ	Condition number of function f: $\kappa = \frac{L}{\mu}$	Cor. 1
g_i^k	Stochastic gradient of function f_i at the point $x = x^k$	Eq. (2)
g^k	Stochastic gradient of function f at the point $x = x^k$: $g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$	Eq. (3)
σ_i^2	Variance of the stochastic gradient g_i^k	Eq. (2)
σ^2	Variance of the stochastic gradient g^k: $\sigma^2 = \frac{1}{n}\sum\limits_{i=1}^{n} \sigma_i^2$	Eq. (3)
h_i^k	Stochastic approximation of the $\nabla f_i(x^*)$; $h_i^{k+1} = h_i^k + \alpha\hat{\Delta}_i^k$	Alg. 1
Δ_i^k	$\Delta_i^k = g_i^k - h_i^k$	Alg. 1
$\mathrm{Quant}_p(\Delta)$	Full p-quantization of vector Δ	Def. 1
$\mathrm{Quant}_p(\Delta, \{d_l\}_{l=1}^m)$	Block-p-quantization of vector Δ with block sizes $\{d_l\}_{l=1}^m$	Def. 2
d_l	Size of the l-th block for quantization	Def. 2
m	Number of blocks for quantization	Def. 2
α, γ^k	Learning rates	Alg. 1
β	Momentum parameter	Alg. 1
$\hat{\Delta}_i^k$	Block-p-quantization of $\Delta_i^k = g_i^k - h_i^k$	Alg. 1
$\hat{\Delta}$	$\hat{\Delta}^k = \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_i^k$	Alg. 1
$\hat{g}_i^k$	Stochastic approximation of $\nabla f_i(x^k)$; $\hat{g}_i^k = h_i^k + \hat{\Delta}_i^k$	Alg. 1
$\hat{g}^k$	$g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$	Alg. 1
v^k	Stochastic gradient with momentum: $v^k = \beta v^{k-1} + \hat{g}^k$	Alg. 1
h^{k+1}	$h^{k+1} = \frac{1}{n}\sum_{i=1}^{n} h_i^{k+1}$	Alg. 1
$\mathrm{prox}_{\gamma R}(u)$	$\arg\min\limits_{v}\left\{\gamma R(v) + \frac{1}{2}\lVert v - u\rVert_2^2\right\}$	Alg. 1
$\Psi_l(x)$	Variance of the l-th quantized block: $\Psi_l(x) = \lVert x(l)\rVert_1 \lVert x(l)\rVert_p - \lVert x(l)\rVert_2^2$	Eq. (7)
$\Psi(x)$	Variance of the block-p-quantized vector: $\Psi(x) = \sum\limits_{l=1}^{m} \Psi_l(x)$	Eq. (8)
$\alpha_p(d)$	$\alpha_p(d) = \inf\limits_{x\neq 0, x\in\mathbb{R}^d} \frac{\lVert x\rVert_2^2}{\lVert x\rVert_1 \lVert x\rVert_p}$	Eq. 16
$\widetilde{d}$	$\widetilde{d} = \max\limits_{l=1,\ldots,m} d_l$	Th. 2
α_p	$\alpha_p = \alpha_p(\widetilde{d})$	Th. 2
c	Such number that $\frac{1+nc\alpha^2}{1+nc\alpha} \leq \alpha_p$	Th. 2
x^*	Solution of the problem (1)	Eq. (19)
h_i^*	$h_i^* = \nabla f_i(x^*)$	Th. 2
V^k	Lyapunov function $V^k = \lVert x^k - x^*\rVert_2^2 + \frac{c\gamma^2}{n}\sum_{i=1}^{n} \lVert h_i^k - h_i^*\rVert$	Th. 2
$\delta, \zeta, \omega, \xi$	Parameters for the proof of momentum version of `DIANA`	Th. 7
η, θ, N, C	Parameters for the decreasing stepsizes results	Th. 5
$\mathbf{E}_{Q^k}$	Expectation w. r. t. the randomness coming from quantization	Lem. 3

Table 5. The table of all notations we use in this paper