# Linearly Converging Error Compensated SGD

**Eduard Gorbunov**
MIPT, Yandex and Sirius, Russia
KAUST, Saudi Arabia

**Dmitry Kovalev**
KAUST, Saudi Arabia

**Dmitry Makarenko**
MIPT, Russia

**Peter Richtárik**
KAUST, Saudi Arabia

## Abstract

In this paper, we propose a unified analysis of variants of distributed `SGD` with arbitrary compressions and delayed updates. Our framework is general enough to cover different variants of quantized `SGD`, Error-Compensated `SGD` (`EC-SGD`) and `SGD` with delayed updates (`D-SGD`). Via a single theorem, we derive the complexity results for all the methods that fit our framework. For the existing methods, this theorem gives the best-known complexity results. Moreover, using our general scheme, we develop new variants of `SGD` that combine variance reduction or arbitrary sampling with error feedback and quantization and derive the convergence rates for these methods beating the state-of-the-art results. In order to illustrate the strength of our framework, we develop 16 new methods that fit this. In particular, we propose the first method called `EC-SGD-DIANA` that is based on error-feedback for biased compression operator and quantization of gradient differences and prove the convergence guarantees showing that `EC-SGD-DIANA` converges to the exact optimum asymptotically in expectation with constant learning rate for both convex and strongly convex objectives when workers compute full gradients of their loss functions. Moreover, for the case when the loss function of the worker has the form of finite sum, we modified the method and got a new one called `EC-LSVRG-DIANA` which is the first distributed stochastic method with error feedback and variance reduction that converges to the exact optimum asymptotically in expectation with a constant learning rate.

## 1 Introduction

We consider distributed optimization problems of the form

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

where $n$ is the number of workers/devices/clients/nodes. The information about function $f_i$ is stored on the $i$-th worker only. Problems of this form appear in the distributed or federated training of supervised machine learning models [42, 30]. In such applications, $x \in \mathbb{R}^d$ describes the parameters identifying a statistical model we wish to train, and $f_i$ is the (generalization or empirical) loss of model $x$ on the data accessible by worker $i$. If worker $i$ has access to data with distribution $\mathcal{D}_i$, then $f_i$ is assumed to have the structure

$$f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right]. \tag{2}$$

Dataset $\mathcal{D}_i$ may or may not be available to worker $i$ in its entirety. Typically, we assume that worker $i$ has only access to samples from $\mathcal{D}_i$. If the dataset is fully available, it is typically

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

finite, in which case we can assume that $f_i$ has the finite-sum form[1]:

$$f_i(x) = \tfrac{1}{m}\sum_{j=1}^{m} f_{ij}(x). \tag{3}$$

**Communication bottleneck.** The key bottleneck in practical distributed [14] and federated [30, 21] systems comes from the high cost of communication of messages among the clients needed to find a solution of sufficient qualities. Several approaches to addressing this communication bottleneck have been proposed in the literature.

In the very rare situation when it is possible to adjust the network architecture connecting the clients, one may consider a fully decentralized setup [6], and allow each client in each iteration to communicate to their neighbors only. One can argue that in some circumstances and in a certain sense, decentralized architecture may be preferable to centralized architectures [34]. Another natural way to address the communication bottleneck is to do more meaningful (which typically means more expensive) work on each client before each communication round. This is done in the hope that such extra work will produce more valuable messages to be communicated, which hopefully results in the need for fewer communication rounds. A popular technique of this type which is particularly relevant to Federated Learning is based in applying multiple *local updates* instead of a single update only. This is the main idea behind `Local-SGD` [43]; see also [4, 15, 22, 24, 29, 46, 50]. However, in this paper, we contribute to the line work which aims to resolve the communication bottleneck issue via *communication compression*. That is, the information that is normally exchanged—be it iterates, gradients or some more sophisticated vectors/tensors—is compressed in a lossy manner before communication. By applying compression, fewer bits are transmitted in each communication round, and one hopes that the increase in the number of communication rounds necessary to solve the problem, if any, is compensated by the savings, leading to a more efficient method overall.

**Error-feedback framework.** In order to address these issues, in this paper we study a broad class of distributed stochastic first order methods for solving problem (1) described by the iterative framework

$$x^{k+1} = x^k - \tfrac{1}{n}\sum_{i=1}^{n} v_i^k, \tag{4}$$

$$e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k, \qquad i = 1, 2, \dots, n. \tag{5}$$

In this scheme, $x^k$ represents the key iterate, $v_i^k$ is the contribution of worker $i$ towards the update in iteration $k$, $g_i^k$ is an unbiased estimator of $\nabla f_i(x^k)$ computed by worker $i$, $\gamma > 0$ is a fixed stepsize and $e_i^k$ is the error accumulated at node $i$ prior to iteration $k$ (we set to $e_i^0 = 0$ for all $i$). In order to better understand the role of the vectors $v_i^k$ and $e_i^k$, first consider the simple special case with $v_i^k \equiv \gamma g_i^k$. In this case, $e_i^k = 0$ for all $i$ and $k$, and method (4)–(5) reduces to distributed `SGD`:

$$x^{k+1} = x^k - \tfrac{\gamma}{n}\sum_{i=1}^{n} g_i^k. \tag{6}$$

However, by allowing to chose the vectors $v_i^k$ in a different manner, we obtain a more general update rule than what the `SGD` update (6) can offer. Stich and Karimireddy [46], who studied (4)–(5) in the $n = 1$ regime, show that this flexibility allows to capture several types of methods, including those employing i) compressed communication, ii) delayed gradients, and iii) minibatch gradient updates. While our general results apply to all these special cases and more, in order to not dilute the focus of the paper, in the main body of this paper we concentrate on the first use case—compressed communication—which we now describe.

**Error-compensated compressed gradient methods.** Note that in distributed `SGD` (6), each worker needs to know the aggregate gradient $g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$ to form $x^{k+1}$, which is needed before the next iteration can start. This can be achieved, for example, by each

[1]The implicit assumption that each worker contains exactly $m$ data points is for simplicity only; all our results have direct analogues in the general setting with $m_i$ data points on worker $i$.

worker $i$ communicating their gradient $g_i^k$ to all other workers. Alternatively, in a parameter server setup, a dedicated master node collects the gradients from all workers, and broadcasts their average $g^k$ to all workers. Instead of communicating the gradient vectors $g_i^k$, which is expensive in distributed learning in general and in federated learning in particular, and especially if $d$ is large, we wish to communicate other but closely related vectors which can be represented with fewer bits. To this effect, each worker $i$ sends the vector

$$v_i^k = \mathcal{C}(e_i^k + \gamma g_i^k), \qquad \forall i \in [n] \tag{7}$$

instead, where $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a (possibly randomized, and in such a case, drawn independently of all else in iteration $k$) compression operator used to reduce communication. We assume throughout that there exists $\delta \in (0, 1]$ such that the following inequality holds for all $x \in \mathbb{R}^d$

$$\mathbf{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \delta)\|x\|^2. \tag{8}$$

For any $k \geq 0$, the vector $e_i^{k+1} = \sum_{t=0}^{k} \gamma g_i^t - v_i^t$ captures the *error* accumulated by worker $i$ up to iteration $k$. This is the difference between the ideal `SGD` update $\sum_{t=0}^{k} \gamma g_i^t$ and the applied update $\sum_{t=0}^{k} v_i^t$. As we see in (7), at iteration $k$ the current error $e_i^k$ is added to the gradient update $\gamma g_i^k$—this is referred to as *error feedback*—and subsequently compressed, which defines the update vector $v_i^k$. Compression introduces additional error, which is added to $e_i^k$, and the process is repeated.

**Compression operators.** For a rich collection of specific operators satisfying (8), we refer the reader to Stich and Karimireddy [46] and Beznosikov et al. [7]. These include various unbiased or contractive sparsification operators such as RandK and TopK, respectively, and quantization operators such as natural compression and natural dithering [18]. Several additional comments related to compression operators are included in Section B.

## 2 Summary of Contributions

We now summarize the key contributions of this paper.

◇ **General theoretical framework.** In this work we propose a *general theoretical framework* for analyzing a wide class of methods that can be written in the the error-feedback form (4)-(5). We perform *complexity analysis* under $\mu$-strong quasi convexity (Assumption 3.1) and $L$-smoothness (Assumption 3.2) assumptions on the functions $f$ and $\{f_i\}$, respectively. Our analysis is based on an additional *parametric assumption* (using parameters $A$, $A'$, $B_1$, $B_1'$, $B_2$, $B_2'$, $C_1$, $C_2$, $D_1$, $D_1'$, $D_2$, $D_3$, $\eta$, $\rho_1$, $\rho_2$, $F_1$, $F_2$, $G$) on the relationship between the iterates $x^k$, stochastic gradients $g^k$, errors $e^k$ and a few other quantities (see Assumption 3.4, and the stronger Assumption 3.3). We prove a single theorem (Theorem 3.1) from which all our complexity results follow as special cases. That is, for each existing or new specific method, we *prove* that one (or both) of our parametric assumptions holds, and specify the parameters for which it holds. These parameters have direct impact on the theoretical rate of the method. A summary of the values of the parameters for all methods developed in this paper is provided in Table 5 in the appendix. We remark that the values of the parameters $A, A', B_1, B_1', B_2, B_2', C_1, C_2$ and $\rho_1, \rho_2$ influence the theoretical stepsize.

◇ **Sharp rates.** For existing methods covered by our general framework, our main convergence result (Theorem 3.1) recovers the best known rates for these methods up to constant factors.

◇ **Eight new error-compensated (EC) methods.** We study several specific EC methods for solving problem (1). First, we recover the `EC-SGD` method first analyzed in the $n = 1$ case by Stich and Karimireddy [46] and later in the general $n \geq 1$ case by Beznosikov et al. [7]. More importantly, we develop *eight new methods*: `EC-SGDsr`, `EC-GDstar`, `EC-SGD-DIANA`[2],

[2] Inspired by personal communication with D. Kovalev in November 2019 who shared a key algorithm and preliminary results of our paper, Stich [45] studied almost the same algorithm and also other related methods and independently derived convergence rates. Our work was finalized and submitted to NeurIPS 2020 in June 2020, while the results in [45] were obtained in Summer 2020 and appeared on arXiv in September 2020. Moreover, in our work, we obtain tighter rates (see Table 1 for the details).

Table 1: Complexity of Error-Compensated SGD methods established in this paper. Symbols: $\varepsilon$ = error tolerance; $\delta$ = contraction factor of compressor $\mathcal{C}$; $\omega$ = variance parameter of compressor $\mathcal{Q}$; $\kappa = L/\mu$; $\mathcal{L}$ = expected smoothness constant; $\sigma_*^2$ = variance of the stochastic gradients in the solution; $\zeta_*^2$ = average of $\|\nabla f_i(x^*)\|^2$; $\sigma^2$ = average of the uniform bounds for the variances of stochastic gradients of workers. `EC-GDstar`, `EC-LSVRGstar` and `EC-LSVRG-DIANA` are the first EC methods with a linear convergence rate without assuming that $\nabla f_i(x^*) = 0$ for all $i$. `EC-LSVRGstar` and `EC-LSVRG-DIANA` are the first EC methods with a linear convergence rate which do not require the computation of the full gradient $\nabla f_i(x^k)$ by all workers in each iteration. Out of these three methods, only `EC-LSVRG-DIANA` is practical. †`EC-GD-DIANA` is a special case of `EC-SGD-DIANA` where each worker $i$ computes the full gradient $\nabla f_i(x^k)$.

| Problem | Method | Alg # | Citation | Sec # | Rate (constants ignored) |
|---|---|---|---|---|---|
| (1)+(3) | `EC-SGDsr` | Alg 3 | **new** | J.1 | $\widetilde{\mathcal{O}}\left(\frac{\mathcal{L}}{\mu} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L(\sigma_*^2+\zeta_*^2/\delta)}}{\mu\sqrt{\delta\varepsilon}}\right)$ |
| (1)+(2) | `EC-SGD` | Alg 4 | [46] | J.2 | $\widetilde{\mathcal{O}}\left(\frac{\kappa}{\delta} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L(\sigma_*^2+\zeta_*^2/\delta)}}{\delta\mu\sqrt{\varepsilon}}\right)$ |
| (1)+(3) | `EC-GDstar` | Alg 5 | **new** | J.3 | $\mathcal{O}\left(\frac{\kappa}{\delta}\log\frac{1}{\varepsilon}\right)$ |
| (1)+(2) | `EC-SGD-DIANA` | Alg 6 | **new** | J.4 | Opt. I: $\widetilde{\mathcal{O}}\left(\omega + \frac{\kappa}{\delta} + \frac{\sigma^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma^2}}{\delta\mu\sqrt{\varepsilon}}\right)$ <br> Opt. II: $\widetilde{\mathcal{O}}\left(\frac{1+\omega}{\delta} + \frac{\kappa}{\delta} + \frac{\sigma^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma^2}}{\mu\sqrt{\delta\varepsilon}}\right)$ |
| (1)+(3) | `EC-SGDsr-DIANA` | Alg 7 | **new** | J.5 | Opt. I: $\widetilde{\mathcal{O}}\left(\omega + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma_*^2}}{\delta\mu\sqrt{\varepsilon}}\right)$ <br> Opt. II: $\widetilde{\mathcal{O}}\left(\frac{1+\omega}{\delta} + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\sigma_*^2}}{\mu\sqrt{\delta\varepsilon}}\right)$ |
| (1)+(2) | `EC-GD-DIANA`† | Alg 6 | **new** | J.4 | $\mathcal{O}\left(\left(\omega + \frac{\kappa}{\delta}\right)\log\frac{1}{\varepsilon}\right)$ |
| (1)+(3) | `EC-LSVRG` | Alg 8 | **new** | J.6 | $\widetilde{\mathcal{O}}\left(m + \frac{\kappa}{\delta} + \frac{\sqrt{L\zeta_*^2}}{\delta\mu\sqrt{\varepsilon}}\right)$ |
| (1)+(3) | `EC-LSVRGstar` | Alg 9 | **new** | J.7 | $\mathcal{O}\left(\left(m + \frac{\kappa}{\delta}\right)\log\frac{1}{\varepsilon}\right)$ |
| (1)+(3) | `EC-LSVRG-DIANA` | Alg 10 | **new** | J.8 | $\mathcal{O}\left(\left(\omega + m + \frac{\kappa}{\delta}\right)\log\frac{1}{\varepsilon}\right)$ |

`EC-SGDsr-DIANA`, `EC-GD-DIANA`, `EC-LSVRG`, `EC-LSVRGstar` and `EC-LSVRG-DIANA`. Some of these methods are designed to work with the expectation structure of the local functions $f_i$ given in (2), and others are specifically designed to exploit the finite-sum structure (3). All these methods follow the error-feedback framework (4)–(5), with $v_i^k$ chosen as in (7). They differ in how the gradient estimator $g_i^k$ is *constructed* (see Table 2 for a compact description of all these methods; formal descriptions can be found in the appendix). As we shall see, the existing `EC-SGD` method uses a rather naive gradient estimator, which renders it less efficient in theory and practice when compared to the best of our new methods. A key property of our parametric assumption described above is that it allows for the construction and modeling of more elaborate gradient estimators, which leads to new EC methods with superior theoretical and practical properties when compared to prior state of the art.

◇ **First linearly converging EC methods.** The key theoretical consequence of our general framework is the development of the *first linearly converging* error-compensated `SGD`-type methods for distributed training with biased communication compression. In particular, we design four such methods: two simple but impractical methods, `EC-GDstar` and `EC-LSVRGstar`, with rates $\mathcal{O}\left(\frac{\kappa}{\delta}\ln\frac{1}{\varepsilon}\right)$ and $\mathcal{O}\left(\left(m + \frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$, respectively, and two practical but more elaborate methods, `EC-GD-DIANA`, with rate $\mathcal{O}\left(\left(\omega + \frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$, and `EC-LSVRG-DIANA`, with rate $\mathcal{O}\left(\left(\omega + m + \frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$. In these rates, $\kappa = L/\mu$ is the condition number, $0 < \delta \leq 1$ is the contraction parameter associated with the compressor $\mathcal{C}$ used in (7), and $\omega$ is the variance parameter associated with a *secondary unbiased compressor*[3] $\mathcal{Q}$ which plays a key role in the construction of the gradient estimator $g_i^k$. The complexity of the first and third

[3] We assume that $\mathbf{E}\mathcal{Q}(x) = x$ and $\mathbf{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega\|x\|^2$ for all $x \in \mathbb{R}^d$.

Table 2: Error compensated methods developed in this paper. In all cases, $v_i^k = \mathcal{C}(e_i^k + \gamma g_i^k)$. The full descriptions of the algorithms are included in the appendix.

| Problem | Method | $g_i^k$ | Comment |
|---|---|---|---|
| (1) + (3) | `EC-SGDsr` | $\frac{1}{m}\sum\limits_{j=1}^{m} \xi_{ij} \nabla f_{ij}(x^k)$ | $\mathbf{E}[\xi_{ij}] = 1$<br>$\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right]$<br>$\leq 2\mathcal{L} D_{f_i}(x, x^*)$ |
| (1) + (2) | `EC-SGD` | $\nabla f_{\xi_i}(x^k)$ | |
| (1) | `EC-GDstar` | $\nabla f_i(x^k) - \nabla f_i(x^*)$ | known $\nabla f_i(x^*)\ \forall i$ |
| (1) + (2) | `EC-SGD-DIANA` | $\hat{g}_i^k - h_i^k + h^k$ | $\mathbf{E}\left[\hat{g}_i^k\right] = \nabla f_i(x^k)$<br>$\mathbf{E}_k\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2\right] \leq D_{1,i}$<br>$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}(\hat{g}_i^k - h_i^k)$<br>$h^k = \frac{1}{n}\sum\limits_{i=1}^{n} h_i^k$ |
| (1) + (3) | `EC-SGDsr-DIANA` | $\nabla f_{\xi_i^k}(x^k) - h_i^k + h^k$ | $\mathbf{E}\left[\nabla f_{\xi_i^k}(x^k)\right] = \nabla f_i(x^k)$<br>$\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right]$<br>$\leq 2\mathcal{L} D_{f_i}(x, x^*)$<br>$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}(\nabla f_{\xi_i^k}(x^k) - h_i^k)$<br>$h^k = \frac{1}{n}\sum\limits_{i=1}^{n} h_i^k$ |
| (1) + (3) | `EC-LSVRG` | $\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k)$<br>$+ \nabla f_i(w_i^k)$ | $l$ chosen uniformly from $[m]$<br>$w_i^{k+1} = \begin{cases} x^k, & \text{with prob. } p, \\ w_i^k, & \text{with prob. } 1-p \end{cases}$ |
| (1) + (3) | `EC-LSVRGstar` | $\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k)$<br>$+ \nabla f_i(w_i^k) - \nabla f_i(x^*)$ | $l$ chosen uniformly from $[m]$<br>$w_i^{k+1} = \begin{cases} x^k, & \text{with prob. } p, \\ w_i^k, & \text{with prob. } 1-p \end{cases}$ |
| (1) + (3) | `EC-LSVRG-DIANA` | $\hat{g}_i^k - h_i^k + h^k$<br>where<br>$\hat{g}_i^k = \nabla f_{il}(x^k)$<br>$- \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k)$ | $h_i^{k+1} = h_i^k + \alpha \mathcal{Q}(\hat{g}_i^k - h_i^k)$<br>$h^k = \frac{1}{n}\sum\limits_{i=1}^{n} h_i^k$<br>$l$ chosen uniformly from $[m]$<br>$w_i^{k+1} = \begin{cases} x^k, & \text{with prob. } p, \\ w_i^k, & \text{with prob. } 1-p \end{cases}$ |

methods does not depend on $m$ as they require the computation of the full gradient $\nabla f_i(x^k)$ for each $i$. The remaining two methods only need to compute $\mathcal{O}(1)$ stochastic gradients $\nabla f_{ij}(x^k)$ on each worker $i$.

The first two methods, while impractical, provided us with the intuition which enabled us to develop the practical variant. We include them in this paper due to their simplicity, because of the added insights they offer, and to showcase the flexibility of our general theoretical framework, which is able to describe them. `EC-GDstar` and `EC-LSVRGstar` are impractical since they require the knowledge of the gradients $\{\nabla f_i(x^*)\}$, where $x^*$ is an optimal solution of (1), which are obviously not known since $x^*$ is not known.

The only known linear convergence result for an error compensated `SGD` method is due to Beznosikov et al. [7], who require the computation of the full gradient of $f_i$ by each machine $i$ (i.e., $m$ stochastic gradients), and the additional assumption that $\nabla f_i(x^*) = 0$ for all $i$. We do not need such assumptions, thereby resolving a major theoretical issue with EC methods.

⋄ **Results in the convex case.** Our theoretical analysis goes beyond distributed optimization and recovers the results from Gorbunov et al. [11], Khaled et al. [25] (without regularization) in the special case when $v_i^k \equiv \gamma g_i^k$. As we have seen, in this case $e_i^k \equiv 0$ for all $i$ and $k$, and the error-feedback framework (4)–(5) reduces to distributed `SGD` (6). In this regime, the relation (19) in Assumption 3.4 becomes void, while relations (15) and (16) with $\sigma_{2,k}^2 \equiv 0$ are precisely those used by Gorbunov et al. [11] to analyze a wide array of `SGD` methods, including vanilla `SGD` [41], `SGD` with arbitrary sampling [13], as well as variance reduced methods such as SAGA [9], SVRG [20], LSVRG [17, 31], JacSketch [12], SEGA [16] and DIANA [37, 19]. Our theorem recovers the rates of all the methods just listed in both the convex case $\mu = 0$ Khaled et al. [25] and the strongly-convex case $\mu > 0$ Gorbunov et al. [11] under the more general Assumption 3.4.

$\diamond$ **DIANA with bi-directional quantization.** To illustrate how our framework can be used even in the case when $v_i^k \equiv \gamma g_i^k$, $e_i^k \equiv 0$, we develop analyze a new version of DIANA called DIANAsr-DQ that uses arbitrary sampling on every node and double quantization[4], i.e., unbiased compression not only on the workers' side but also on the master's one.

$\diamond$ **Methods with delayed updates.** Following Stich [44], we also show that our approach covers SGD with delayed updates [1, 3, 10] (D-SGD), and our analysis shows the best-known rate for this method. Due to the flexibility of our framework, we are able develop several new variants of D-SGD with and without quantization, variance reduction, and arbitrary sampling. Again, due to space limitations, we put these methods together with their convergence analyses in the appendix.

## 3 Main Result

In this section we present the main theoretical result of our paper. First, we introduce our assumption on $f$, which is a relaxation of $\mu$-strong convexity.

**Assumption 3.1** ($\mu$-strong quasi-convexity)**.** *Assume that function $f$ has a unique minimizer $x^*$. We say that function $f$ is strongly quasi-convex with parameter $\mu \geq 0$ if for all $x \in \mathbb{R}^d$*

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \tfrac{\mu}{2}\|x - x^*\|^2. \tag{9}$$

We allow $\mu$ to be zero, in which case $f$ is sometimes called *weakly quasi-convex* (see [44] and references therein). Second, we introduce the classical $L$-smoothness assumption.

**Assumption 3.2.** *$L$-smoothness We say that function $f$ is $L$-smooth if it is differentiable and its gradient is $L$-Lipschitz continuous, i.e., for all $x, y \in \mathbb{R}^d$*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{10}$$

It is a well-known fact [38] that $L$-smoothness of convex function $f$ implies that

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle) \stackrel{\text{def}}{=} 2LD_f(x, y). \tag{11}$$

We now introduce our key parametric assumption on the stochastic gradient $g^k$. This is a generalization of the assumption introduced by Gorbunov et al. [11] for the particular class of methods described covered by the EF framework (4)–(5).

**Assumption 3.3.** *For all $k \geq 0$, the stochastic gradient $g^k$ is an average of stochastic gradients $g_i^k$ such that*

$$g^k = \tfrac{1}{n}\sum_{i=1}^n g_i^k, \qquad \mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k). \tag{12}$$

*Moreover, there exist constants $A, \widetilde{A}, A', B_1, B_2, \widetilde{B}_1, \widetilde{B}_2, B_1', B_2', C_1, C_2, G, D_1, \widetilde{D}_1, D_1', D_2, D_3 \geq 0$, and $\rho_1, \rho_2 \in [0, 1]$ and two sequences of (probably random) variables $\{\sigma_{1,k}\}_{k\geq 0}$ and $\{\sigma_{2,k}\}_{k\geq 0}$, such that the following recursions hold:*

$$\tfrac{1}{n}\sum_{i=1}^n \left\|\bar{g}_i^k\right\|^2 \leq 2A(f(x^k) - f(x^*)) + B_1\sigma_{1,k}^2 + B_2\sigma_{2,k}^2 + D_1, \tag{13}$$

$$\tfrac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|g_i^k - \bar{g}_i^k\right\|^2 \mid x^k\right] \leq 2\widetilde{A}(f(x^k) - f(x^*)) + \widetilde{B}_1\sigma_{1,k}^2 + \widetilde{B}_2\sigma_{2,k}^2 + \widetilde{D}_1, \tag{14}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 2A'(f(x^k) - f(x^*)) + B_1'\sigma_{1,k}^2 + B_2'\sigma_{2,k}^2 + D_1', \tag{15}$$

$$\mathbf{E}\left[\sigma_{1,k+1}^2 \mid \sigma_{1,k}^2, \sigma_{2,k}^2\right] \leq (1 - \rho_1)\sigma_{1,k}^2 + 2C_1\left(f(x^k) - f(x^*)\right) + G\rho_1\sigma_{2,k}^2 + D_2, \tag{16}$$

$$\mathbf{E}\left[\sigma_{2,k+1}^2 \mid \sigma_{2,k}^2\right] \leq (1 - \rho_2)\sigma_{2,k}^2 + 2C_2\left(f(x^k) - f(x^*)\right), \tag{17}$$

[4] In the concurrent work (which appeared on arXiv after we have submitted our paper to NeurIPS) a similar method was independently proposed under the name of Artemis [40]. However, our analysis is more general, see all the details on this method in the appendix. This footnote was added to the paper during the preparation of the camera-ready version of our paper.

*where* $\bar{g}_i^k = \mathbf{E}\left[g_i^k \mid x^k\right]$.

Let us briefly explain the intuition behind the assumption and the meaning of the introduced parameters. First of all, we assume that the stochastic gradient at iteration $k$ is conditionally unbiased estimator of $\nabla f(x^k)$, which is a natural and commonly used assumption on the stochastic gradient in the literature. However, we explicitly do *not* require unbiasedness of $g_i^k$, which is very useful in some special cases. Secondly, let us consider the simplest special case when $g^k \equiv \nabla f(x^k)$ and $f_1 = \ldots = f_n = f$, i.e., there is no stochasticity/randomness in the method and the workers have the same functions. Then due to $\nabla f(x^*) = 0$, we have that

$$\|\nabla f(x^k)\|^2 \overset{(11)}{\leq} 2L(f(x^k) - f(x^*)),$$

which implies that Assumption 3.3 holds in this case with $A = A' = L$, $\widetilde{A} = 0$ and $B_1 = B_2 = \widetilde{B}_1 = \widetilde{B}_2 = B_1' = B_2' = C_1 = C_2 = D_1 = \widetilde{D}_1 = D_1' = D_2 = 0$, $\rho = 1$, $\sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0$.

In general, if $g^k$ satisfies Assumption 3.4, then parameters $A$, $\widetilde{A}$ and $A'$ are usually connected with the smoothness properties of $f$ and typically they are just multiples of $L$, whereas terms $B_1\sigma_{1,k}^2$, $B_2\sigma_{2,k}^2$, $\widetilde{B}_1\sigma_{1,k}^2$, $\widetilde{B}_2\sigma_{2,k}^2$, $B_1'\sigma_{1,k}^2$, $B_2'\sigma_{2,k}^2$ and $D_1$, $\widetilde{D}_1$, $D_1'$ appear due to the stochastic nature of $g_i^k$. Moreover, $\{\sigma_{1,k}^2\}_{k\geq 0}$ and $\{\sigma_{2,k}^2\}_{k\geq 0}$ are sequences connected with variance reduction processes and for the methods; without any kind of variance reduction these sequences contains only zeros. Parameters $B_1$ and $B_2$ are often 0 or small positive constants, e.g., $B_1 = B_2 = 2$, and $D_1$ characterizes the remaining variance in the estimator $g^k$ that is not included in the first two terms.

Inequalities (16) and (17) describe the variance reduction processes: one can interpret $\rho_1$ and $\rho_2$ as the *rates* of the variance reduction processes, $2C_1(f(x^k) - f(x^*))$ and $2C_2(f(x^k) - f(x^*))$ are "optimization" terms and, similarly to $D_1$, $D_2$ represents the remaining variance that is not included in the first two terms. Typically, $\sigma_{1,k}^2$ controls the variance coming from compression and $\sigma_{2,k}^2$ controls the variance taking its origin in finite-sum type randomization (i.e., subsampling) by each worker. In the case $\rho_1 = 1$ we assume that $B_1 = B_1' = C_1 = G = 0, D_2 = 0$ (for $\rho_2 = 1$ analogously), since inequality (16) becomes superfluous.

However, in our main result we need a slightly different assumption.

**Assumption 3.4.** *For all $k \geq 0$, the stochastic gradient $g^k$ is an unbiased estimator of $\nabla f(x^k)$:*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k). \tag{18}$$

*Moreover, there exist non-negative constants $A', B_1', B_2', C_1, C_2, F_1, F_2, G, D_1', D_2, D_3 \geq 0, \rho_1, \rho_2 \in [0,1]$ and two sequences of (probably random) variables $\{\sigma_{1,k}\}_{k\geq 0}$ and $\{\sigma_{2,k}\}_{k\geq 0}$ such that inequalities* (15)*,* (16) *and* (17) *hold and*

$$3L\sum_{k=0}^{K} w_k \mathbf{E}\|e^k\|^2 \quad \leq \quad \frac{1}{4}\sum_{k=0}^{K} w_k \mathbf{E}\left[f(x^k) - f(x^*)\right] + F_1\sigma_{1,0}^2 + F_2\sigma_{2,0}^2 + \gamma D_3 W_K \tag{19}$$

*for all $k, K \geq 0$, where $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$ and $\{W_K\}_{K\geq 0}$ and $\{w_k\}_{k\geq 0}$ are defined as*

$$W_K = \sum_{k=0}^{K} w_k, \quad w_k = (1-\eta)^{-(k+1)}, \quad \eta = \min\left\{\frac{\gamma\mu}{2}, \frac{\rho_1}{4}, \frac{\rho_2}{4}\right\}. \tag{20}$$

This assumption is more flexible than Assumption 3.3 and helps us to obtain a unified analysis of all methods falling in the error-feedback framework. We emphasize that in this assumption we do not assume that (13) and (14) hold *explicitly*. Instead of this, we introduce inequality (19), which is the key tool that helps us to analyze the effect of error-feedback and comes from the analysis from [46] with needed adaptations connected with the first three inequalities. As we show in the appendix, this inequality can be derived for `SGD` with error compensation and delayed updates under Assumption 3.3 and, in particular, using (13) and (14). As before, $D_3$ hides a variance that is not handled by variance reduction processes and $F_1$ and $F_2$ are some constants that typically depend on $L, B_1, B_2, \rho_1, \rho_2$ and $\gamma$.

We now proceed to stating our main theorem.

**Theorem 3.1.** *Let Assumptions 3.1, 3.2 and 3.4 be satisfied and* $\gamma \leq 1/4(A'+C_1M_1+C_2M_2)$. *Then for all* $K \geq 0$ *we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq (1-\eta)^K \frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma} + 4\gamma\left(D_1' + M_1D_2 + D_3\right) \tag{21}$$

*when* $\mu > 0$ *and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma K} + 4\gamma\left(D_1' + M_1D_2 + D_3\right) \tag{22}$$

*when* $\mu = 0$, *where* $\eta = \min\left\{\gamma\mu/2, \rho_1/4, \rho_2/4\right\}$, $T^k \stackrel{def}{=} \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$ *and* $M_1 = \frac{4B_1'}{3\rho_1}$, $M_2 = \frac{4\left(B_2'+\frac{4}{3}G\right)}{3\rho_2}$.

All the complexity results summarized in Table 1 follow from this theorem; the detailed proofs are included in the appendix. Furthermore, in the appendix we include similar results but for methods employing *delayed* updates. The methods, and all associated theory is included there, too.

## 4 Numerical Experiments

To justify our theory, we conduct several numerical experimentson logistic regression problem with $\ell_2$-regularization:

$$\min_{x\in\mathbb{R}^d}\left\{f(x) = \frac{1}{N}\sum_{i=1}^N \log\left(1+\exp\left(-y_i\cdot(Ax)_i\right)\right) + \frac{\mu}{2}\|x\|^2\right\}, \tag{23}$$

where $N$ is a number of features, $x \in \mathbb{R}^d$ represents the weights of the model, $A \in \mathbb{R}^{N\times d}$ is a feature matrix, vector $y \in \{-1,1\}^N$ is a vector of labels and $(Ax)_i$ denotes the $i$-th component of vector $Ax$. Clearly, this problem is $L$-smooth and $\mu$-strongly convex with $L = \mu + \lambda_{\max}(A^\top A)/4N$, where $\lambda_{\max}(A^\top A)$ is a largest eigenvalue of $A^\top A$. The datasets were taken from LIBSVM library [8], and the code was written in Python 3.7 using standard libraries. Our code is available at `https://github.com/eduardgorbunov/ef_sigma_k`.

We simulate parameter-server architecture using one machine with Intel(R) Core(TM) i7-9750 CPU 2.60 GHz in the following way. First of all, we always use such $N$ that $N = n\cdot m$ and consider $n = 20$ and $n = 100$ workers. The choice of $N$ for each dataset that we consider is stated in Table 3. Next, we shuffle the data and split in $n$ groups of size $m$. To emulate

Table 3: Summary of datasets: $N$ = total # of data samples; $d$ = # of features.

| | `a9a` | `w8a` | `gisette` | `mushrooms` | `madelon` | `phishing` |
|---|---|---|---|---|---|---|
| $N$ | 32,000 | 49,700 | 6,000 | 8,000 | 2,000 | 11,000 |
| $d$ | 123 | 300 | 5,000 | 112 | 500 | 68 |

the work of workers, we use a single machine and run the methods with the parallel loop in series. Since in these experiments we study sample complexity and number of bits used for communication, this setup is identical to the real parameter-server setup in this sense.

In all experiments we use the stepsize $\gamma = 1/L$ and $\ell_2$-regularization parameter $\mu = 10^{-4}\lambda_{\max}(A^\top A)/4N$. The starting point $x^0$ for each dataset was chosen so that $f(x^0) - f(x^*) \sim 10$. In experiments with stochastic methods we used batches of size 1 and uniform sampling for simplicity. For `LSVRG`-type methods we choose $p = 1/m$.

**Compressing stochastic gradients.** The results for `a9a`, `madelon` and `phishing` can be found in Figure 1 (included here) and for `w8a`, `mushrooms` and `gisette` in Figure 3 (in the Appendix). We choose number of components for TopK operator of the order $\max\{1, d/100\}$. Clearly, in these experiments we see two levels of noise. For some datasets, like `a9a`, `phishing` or `mushrooms`, the noise that comes from the stochasticity of the gradients dominates the noise coming from compression. Therefore, methods such as `EC-SGD` and `EC-SGD-DIANA` start to oscillate around a larger value of the loss function than other methods we consider. `EC-LSVRG` reduces the largest source of noise and, as a result, finds a better approximation of the solution. However, at some point, it reaches another level of the loss function and starts to oscillate there due to the noise coming from compression. Finally, `EC-LSVRG-DIANA` reduces the variance of both types, and as a result, finds an even better approximation of the solution. In contrast, for the `madelon` dataset, both noises are of the same order, and therefore, `EC-LSVRG` and `EC-SGD-DIANA` behave similarly to `EC-SGD`. However, `EC-LSVRG-DIANA` again reduces both types of noise effectively and finds a better approximation of the solution after a given number of epochs. In the experiments with `w8a` and `gisette` datasets, the noise produced by compression is dominated by the noise coming from the stochastic gradients. As a result, we see that the `DIANA`-trick is not needed here.

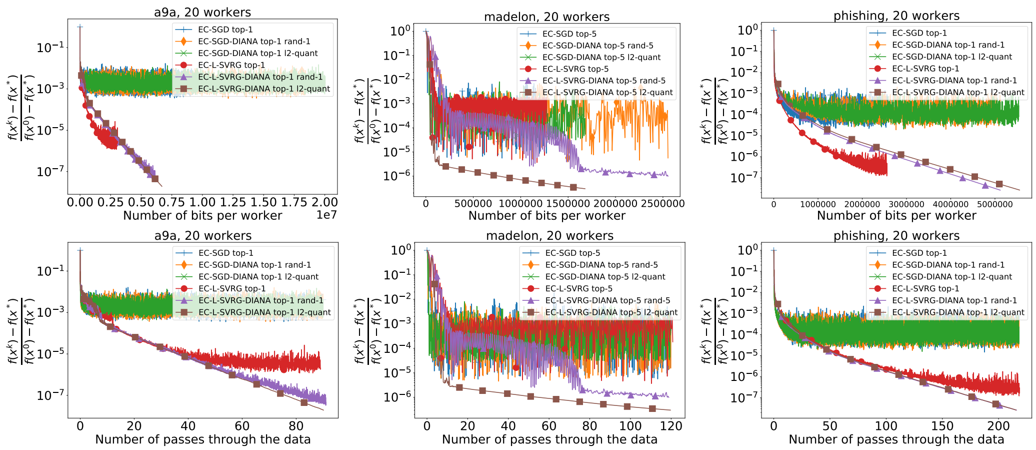


Figure 1: Trajectories of `EC-SGD`, `EC-SGD-DIANA`, `EC-LSVRG` and `EC-LSVRG-DIANA` applied to solve logistic regression problem with 20 workers.

**Compressing full gradients.** In order to show the effect of `DIANA`-type variance reduction itself, we consider the case when all workers compute the full gradients of their functions, see Figure 2 (included here) and Figures 4–7 (in the Appendix). Clearly, for all datasets except `mushrooms`, `EC-GD` with constant stepsize converges to a neighborhood of the solution only, while `EC-GDstar` and `EC-GD-DIANA` converge with linear rate asymptotically to the exact solution. `EC-GDstar` always show the best performance, however, it is impractical: we used a very good approximation of the solution to apply this method. In contrast, `EC-DIANA` converges slightly slower and requires more bits for communication; but it is practical and shows better performance than `EC-GD`. On the `mushrooms` datasets, `EC-GD` does not reach the oscillation region after the given number of epochs, therefore, it is preferable there.

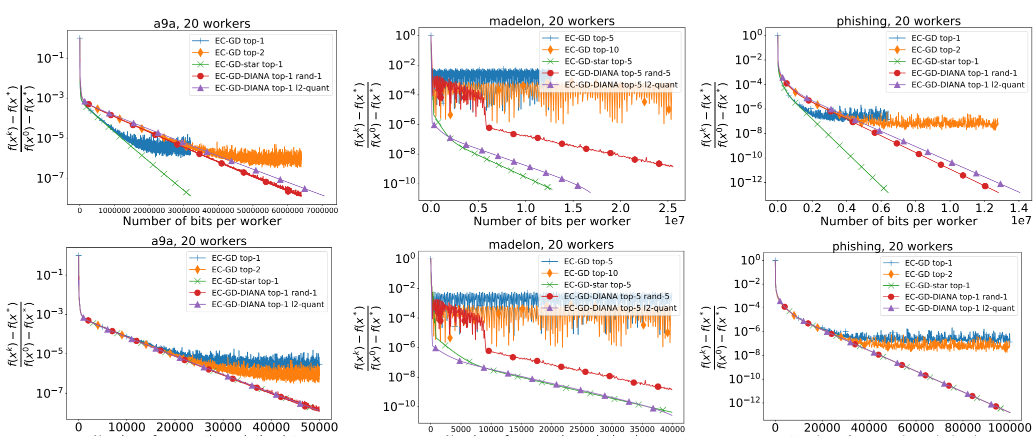


Figure 2: Trajectories of `EC-GD`, `EC-GD-star` and `EC-DIANA` applied to solve logistic regression problem with 20 workers.

## Broader Impact

Our contribution is primarily theoretical. Therefore, a broader impact discussion is not applicable.

## Acknowledgments and Disclosure of Funding

The work of Peter Richtárik, Eduard Gorbunov and Dmitry Kovalev was supported by KAUST Baseline Research Fund. Part of this work was done while E. Gorbunov was a research intern at KAUST. The work of E. Gorbunov in Sections 4, A–D, H, and K was also partially supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project No. 0714-2020-0005, and in Sections 3, E–G, I, J – by RFBR, project number 19-31-51001.

### References

[1] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

[2] 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.

[3] Y. Arjevani, O. Shamir, and N. Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. *arXiv preprint arXiv:1806.10188*, 2018.

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

[5] J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar. SignSGD with majority vote is communication efficient and fault tolerant. In *ICLR*, 2019.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

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

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[9] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[10] H. R. Feyzmahdavian, A. Aytekin, and M. Johansson. An asynchronous mini-batch algorithm for regularized stochastic optimization. *IEEE Transactions on Automatic Control*, 61(12):3740–3754, 2016.

[11] E. Gorbunov, F. Hanzely, and P. Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

[12] R. M. Gower, P. Richtárik, and F. Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arXiv preprint arXiv:1805.02632*, 2018.

[13] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209, 2019.

[14] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[15] F. Haddadpour and M. Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

[16] F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pages 2082–2093, 2018.

[17] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

[18] S. Horváth, C.-Y. Ho, Ľudovít Horváth, A. N. Sahu, M. Canini, and P. Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.

[19] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

[20] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

[21] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[22] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

[23] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error feedback fixes signSGD and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261, 2019.

[24] A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

[25] A. Khaled, O. Sebbouh, N. Loizou, R. M. Gower, and P. Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *arXiv preprint arXiv:2006.11573*, 2020.

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

[27] A. Koloskova, S. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487, 2019.

[28] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi. Decentralized deep learning with arbitrary communication compression. *ICLR*, page arXiv:1907.09356, 2020. URL `https://arxiv.org/abs/1907.09356`.

[29] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich. A unified theory of decentralized SGD with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020.

[30] 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.

[31] D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

[32] R. Leblond, F. Pedregosa, and S. Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *The Journal of Machine Learning Research*, 19(1):3140–3207, 2018.

[33] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

[34] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

[35] X. Liu, Y. Li, J. Tang, and M. Yan. A double residual compression algorithm for efficient distributed learning. *arXiv preprint arXiv:1910.07561*, 2019.

[36] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

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

[38] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[39] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takáč. SGD and Hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758, 2018.

[40] C. Philippenko and A. Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in federated learning. *arXiv preprint arXiv:2006.14591*, 2020.

[41] H. Robbins and S. Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985.

[42] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *International conference on machine learning*, pages 1000–1008, 2014.

[43] S. U. Stich. Local SGD converges fast and communicates little. *International Conference on Learning Representations (ICLR)*, page arXiv:1805.09767, 2019. URL `https://arxiv.org/abs/1805.09767`.

[44] S. U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

[45] S. U. Stich. On communication compression for distributed optimization on heterogeneous data. *arXiv preprint arXiv:2009.02388*, 2020.

[46] S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

[47] S. U. Stich, J.-B. Cordonnier, and M. Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[48] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pages 6155–6165, 2019.

[49] 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.

[50] B. Woodworth, K. K. Patel, S. U. Stich, Z. Dai, B. Bullins, H. B. McMahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? *arXiv preprint arXiv:2002.07839*, 2020.

# Appendix: Linearly Converging Error Compensated SGD

## Contents

# A Missing Plots

## A.1 Compressing stochastic gradients

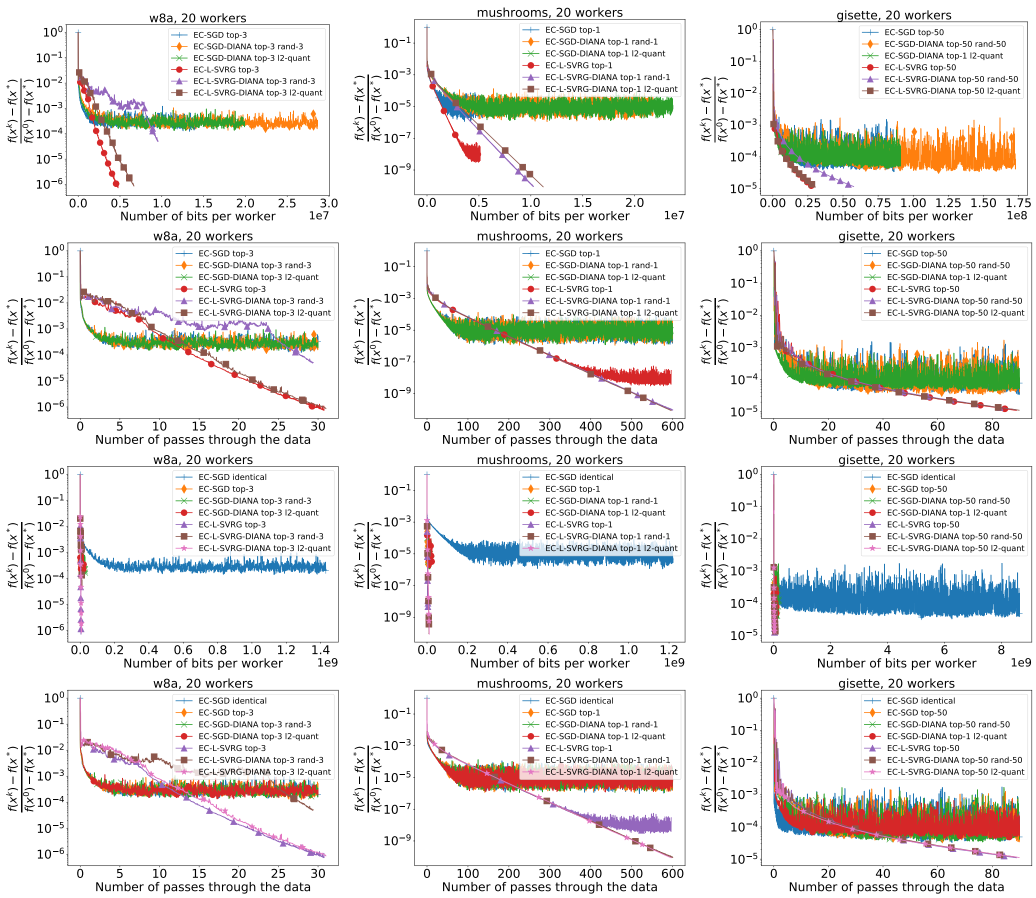


Figure 3: Trajectories of EC-SGD, EC-SGD-DIANA, EC-LSVRG and EC-LSVRG-DIANA applied to solve logistic regression problem with 20 workers. EC-SGD identical corresponds to SGD with error compensation with trivial compression operator $\mathcal{C}(x) = x$, i.e., it is just parallel SGD.

## A.2 Compressing full gradients

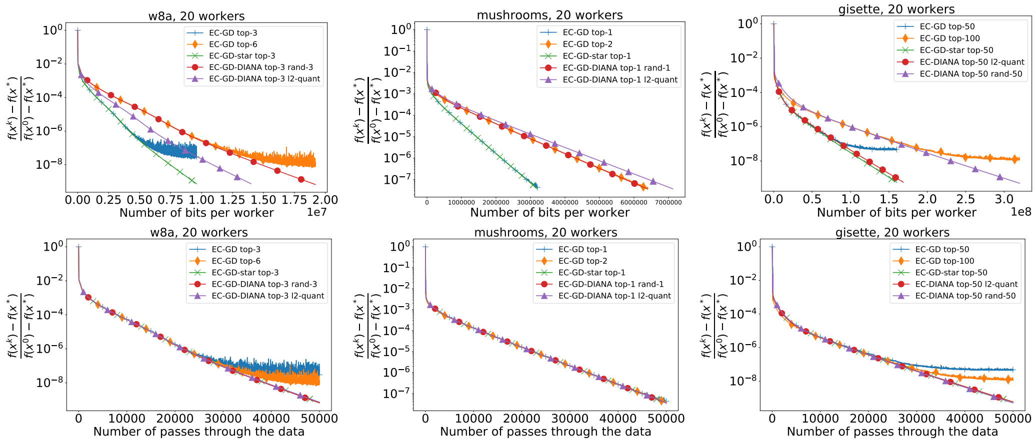


Figure 4: Trajectories of `EC-GD`, `EC-GD-star` and `EC-DIANA` applied to solve logistic regression problem with 20 workers.

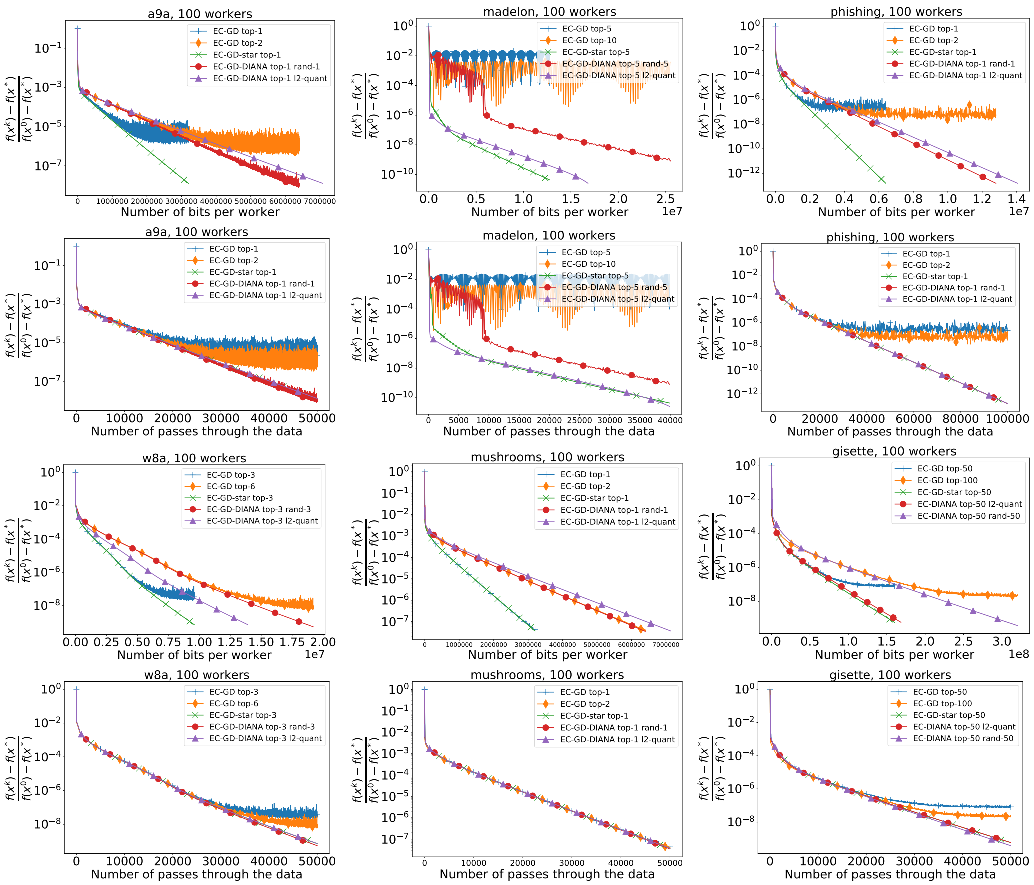


Figure 5: Trajectories of EC-GD, EC-GD-star and EC-DIANA applied to solve logistic regression problem with 100 workers.

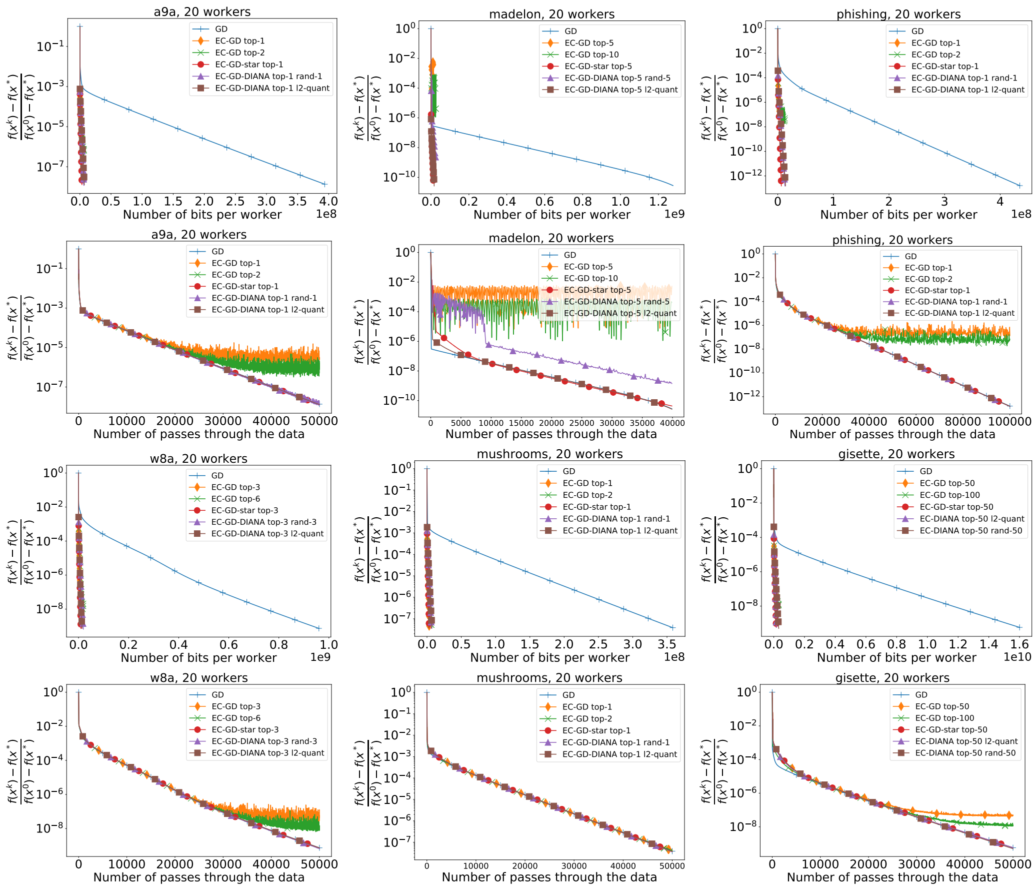


Figure 6: Trajectories of `EC-GD`, `EC-GD-star`, `EC-DIANA` and `GD` applied to solve logistic regression problem with 20 workers.

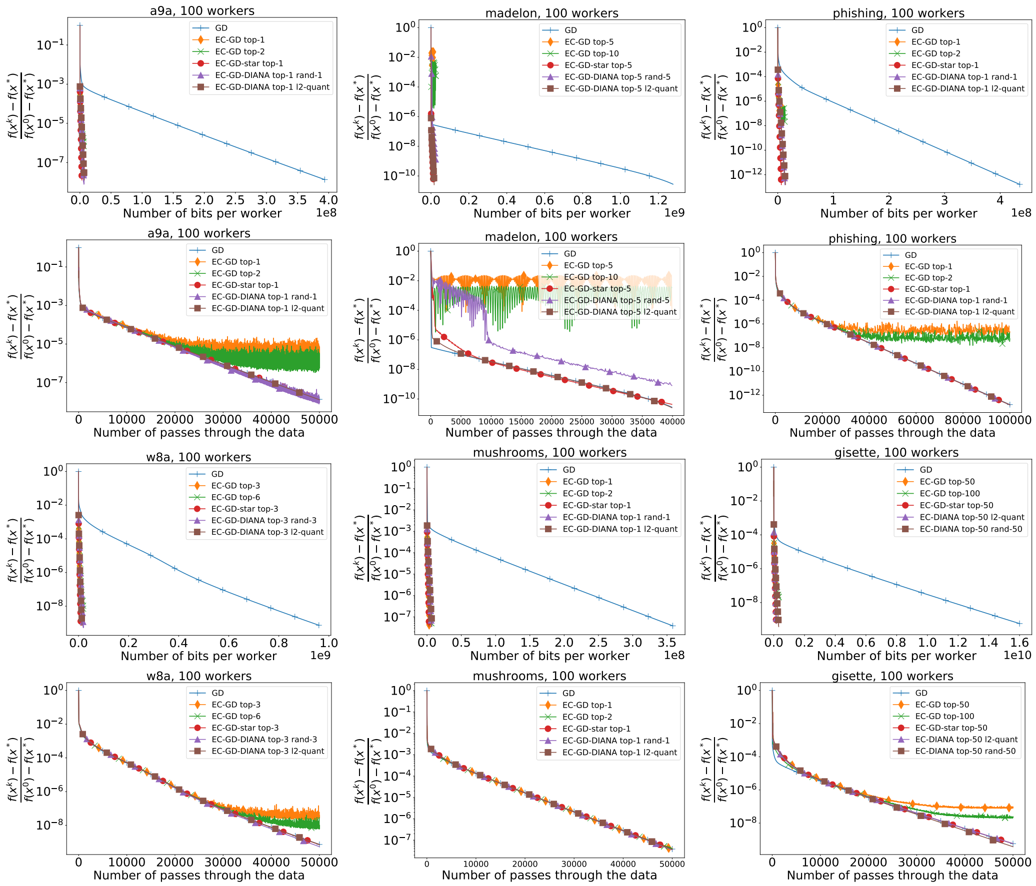


Figure 7: Trajectories of EC-GD, EC-GD-star, EC-DIANA and GD applied to solve logistic regression problem with 100 workers.

# B Compression Operators: Extra Commentary

Communication efficient distributed `SGD` methods based on the idea of communication compression exists in two distinct varieties: i) methods based on unbiased compression operators, and ii) methods based on biased compression operators. The first class of methods is much mire developed than the latter since it is easier to theoretically analyze unbiased operators. The subject of this paper is the study of the latter and dramatically less developed and understood class.

## B.1 Unbiased compressors

By unbiased compression operators we mean randomized mappings $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}$ satisfying the relations

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

for some $\omega \geq 0$. While operators satisfying the above relations are often in the literature called *quantization operators*, this class includes compressors which perform sparsification as well.

Among the first methods using unbiased compressors developed in this field are `QSGD` [2], `TernGrad` [49] and `DQGD` [26]. The first analysis of `QSGD` and `TernGrad` without bounded gradients assumptions was proposed in [37], which contains the best known results for `QSGD` and `TernGrad`. However, existing guarantees in the strongly convex case for `QGSD`, `TernGrad`, and `DQGD` establish linear convergence to some neighborhood of the solution only, even if the workers quantize the full gradients of their functions. This problem was resolved by Mishchenko et al. [37], who proposed the first method, called `DIANA`, which uses quantization for communication and enjoys the linear rate of convergence to the exact optimum asymptotically in the strongly convex case when workers compute the full gradients of their functions in each iteration. Unlike all previous approaches, `DIANA` is based on the quantization of gradient differences rather than iterates or gradients. In essence, DIANA is a technique for reducing the variance introduced by quantization. Horváth et al. [19] generalized the `DIANA` method to the case of more general quantization operators. Moreover, the same authors developed a new method called `VR-DIANA` specially designed to solve problems (1) with the individual functions having the finite sum structure (3).

## B.2 Biased compressors

By biased compressors we mean (possibly) randomized mappings $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}$ satisfying the average contraction relation

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

for some $\delta > 0$.

Perhaps the most popular biased compression operator is TopK, which takes vector $x$ as input and substitutes all coordinates of $x$ by zero except the $k$ components with the largest absolute values. However, such a greedy approach applied to simple distributed `SGD` and even distributed `GD` can break the convergence of the method even when applied to simple functions in small dimensions, and may even lead to exponential divergence [7]. The *error-feedback* framework described in [23, 46, 47] and studies in this paper can fix this problem, and it remains the only known mechanism that does so for all compressors described in (8). This is one of the main motivations for the study of the error-feedback mechanism. For instance, error feedback can fix convergence issues with methods like `sign-SGD` [5]. The analysis of error feedback by Karimireddy et al. [23], Stich and Karimireddy [46], Stich et al. [47] works either under the assumption that the second moment of the stochastic gradient is uniformly bounded or only for the single-worker case. Recently Beznosikov et al. [7] proposed the first analysis of `SGD` with error feedback for the general case of multiple workers without bounded second moment assumption. There is another line of works [27, 28] where authors apply arbitrary compressions in the decentralized setup. This approach has better potential than a centralized one in terms of reducing the communication cost. However, in this paper, we study only centralized architecture.

## C Further Notation and Definitions

In what follows it will be useful to denote

$$v^k \stackrel{\text{def}}{=} \tfrac{1}{n}\sum_i v_i^k, \quad g^k \stackrel{\text{def}}{=} \tfrac{1}{n}\sum_i g_i^k, \quad e^k \stackrel{\text{def}}{=} \tfrac{1}{n}\sum_i e_i^k.$$

By aggregating identities (5) across all $i$, we get $e^{k+1} = e^k + \gamma g^k - v^k$. In our proofs we also use the perturbed iterates technique [32, 36] based on the analysis of the following sequence

$$\tilde{x}^k = x^k - e^k. \tag{24}$$

This sequence satisfies very useful for the analysis relation:

$$\tilde{x}^{k+1} \stackrel{(24)}{=} x^{k+1} - e^{k+1} \stackrel{(4),(5)}{=} x^k - v^k - (e^k + \gamma g^k - v^k) = x^k - e^k - \gamma g^k \stackrel{(24)}{=} \tilde{x}^k - \gamma g^k. \tag{25}$$

### C.1 Quantization operators

**Definition C.1.** *We say that stochastic mapping $Q(x) : \mathbb{R}^d \to \mathbb{R}^d$ is a quantization operator if there exists such $\omega > 0$ that for any $x \in \mathbb{R}^d$*

$$\mathbf{E}\left[\mathcal{Q}(x)\right] = x, \quad \mathbf{E}\left[\|\mathcal{Q}(x) - x\|^2\right] \le \omega\|x\|^2. \tag{26}$$

Below we enumerate some classical compression and quantization operators (see more in [7]).

1. **TopK sparsification.** This compression operator is defined as follows:

$$\mathcal{C}(x) = \sum_{i=1}^{K} x_{(i)} e_{(i)}$$

   where $|x_{(1)}| \ge |x_{(2)}| \ge \ldots \ge |x_{(d)}|$ are components of $x$ sorted in the decreasing order of their absolute values, $e_1, \ldots, e_d$ is the standard basis in $\mathbb{R}^d$ and $K$ is some number from $[d]$. Clearly, TopK is a biased compression operator. One can show that TopK satisfies (8) with $\delta = \frac{K}{d}$ [7].

2. **RandK sparsification** operator is defined as

$$\mathcal{Q}(x) = \frac{d}{K}\sum_{i\in S} x_i e_i$$

   where $S$ is a random subset of $[d]$ sampled from the uniform distribution on the all subset of $[d]$ with cardinality $K$. RandK is an unbiased compression operator satisfying (26) with $\omega = \frac{d}{K}$.

3. **$\ell_p$-quantization.** By $\ell_2$-quantization we mean the following random operator:

$$\mathcal{Q}(x) = \|x\|_p \mathrm{sign}(x) \circ \xi$$

   where $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ is an $\ell_p$-norm of vector $x$, $\mathrm{sign}(x)$ is a component-wise sign of vector $x$, $a \circ b$ defines a component-wise product of vectors $a$ and $b$ and $\xi = (\xi_1, \ldots, \xi_d)^\top$ is a random vector such that

$$\xi_i = \begin{cases} 1, & \text{with probability } \frac{|x_i|}{\|x\|_p}, \\ 0, & \text{with probability } 1 - \frac{|x_i|}{\|x\|_p}. \end{cases}$$

   One can show that this operator satisfies (26). In particular, if $p = 2$ it satisfies (26) with $\omega = \sqrt{d} - 1$ and if $p = \infty$, then $\omega = \frac{1+\sqrt{d}}{2} - 1$ (see [37]).

We assume that $\mathcal{C}$ is any operator which enjoys the following contractive property: there exists a constant $0 < \delta \le 1$ such that

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

# D Basic Inequalities, Identities and Technical Lemmas

## D.1 Basic inequalities

For all $a, b, x_1, \ldots, x_n \in \mathbb{R}^d$, $\beta > 0$ and $p \in (0,1]$ the following inequalities hold

$$\langle a, b\rangle \leq \frac{\|a\|^2}{2\beta} + \frac{\beta\|b\|^2}{2}, \tag{27}$$

$$\langle a - b, a + b\rangle = \|a\|^2 - \|b\|^2, \tag{28}$$

$$\frac{1}{2}\|a\|^2 - \|b\|^2 \leq \|a + b\|^2, \tag{29}$$

$$\|a + b\|^2 \leq (1+\beta)\|a\|^2 + (1 + {}^1\!/\!_\beta)\|b\|^2, \tag{30}$$

$$\left\|\sum_{i=1}^n x_n\right\|^2 \leq n\sum_{i=1}^n \|x_i\|^2, \tag{31}$$

$$\left(1 - \frac{p}{2}\right)^{-1} \leq 1 + p, \tag{32}$$

$$\left(1 + \frac{p}{2}\right)(1-p) \leq 1 - \frac{p}{2}. \tag{33}$$

## D.2 Identities and inequalities involving random variables

**Variance decomposition.** For a random vector $\xi \in \mathbb{R}^d$ and any deterministic vector $x \in \mathbb{R}^d$ the variance can be decomposed as

$$\mathbf{E}\left[\|\xi - \mathbf{E}\xi\|^2\right] = \mathbf{E}\left[\|\xi - x\|^2\right] - \|\mathbf{E}\xi - x\|^2 \tag{34}$$

**Tower property of mathematical expectation.** For random variables $\xi, \eta \in \mathbb{R}^d$ we have

$$\mathbf{E}[\xi] = \mathbf{E}[\mathbf{E}[\xi \mid \eta]] \tag{35}$$

under assumption that all expectations in the expression above are well-defined.

**Lemma D.1** (Lemma 14 from [46]). *For any $\tau$ vectors $a_1, \ldots, a_\tau \in \mathbb{R}^d$ and $\xi_1, \ldots, \xi_\tau$ zero-mean random vectors in $\mathbb{R}^d$, each $\xi_t$ conditionally independent of $\{\xi_i\}_{i=1}^{t-1}$ for all $1 \leq t \leq \tau$ the following inequality holds*

$$\mathbf{E}\left[\left\|\sum_{t=1}^\tau (a_t + \xi_t)\right\|^2\right] \leq \tau\sum_{t=1}^\tau \|a_t\|^2 + \sum_{t=1}^\tau \mathbf{E}\|\xi_t\|^2. \tag{36}$$

## D.3 Technical lemmas

**Lemma D.2** (see also Lemma 2 from [44]). *Let $\{r_k\}_{k\geq 0}$ satisfy*

$$r_K \leq \frac{a}{\gamma W_K} + c_1\gamma + c_2\gamma^2 \tag{37}$$

*for all $K \geq 0$ with some constants $a, c_2 > 0$, $c_1 \geq 0$ where $\{w_k\}_{k\geq 0}$ and $\{W_K\}_{K\geq 0}$ are defined in (20), $\gamma \leq \frac{1}{d}$. Then for all $K$ such that $\frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{K} \leq \min\{\rho_1, \rho_2\}$ and*

$$\gamma = \min\left\{\frac{1}{d}, \frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}\right\} \tag{38}$$

*we have that*

$$r_K = \widetilde{\mathcal{O}}\left(da\exp\left(-\min\left\{\frac{\mu}{d}, \rho_1, \rho_2\right\}K\right) + \frac{c_1}{\mu K} + \frac{c_2}{\mu^2 K^2}\right). \tag{39}$$

*Proof.* Since $W_K \geq w_K = (1-\eta)^{-(K+1)}$ we have

$$r_K \leq (1-\eta)^{K+1}\frac{a}{\gamma} + c_1\gamma + c_2\gamma^2 \leq \frac{a}{\gamma}\exp\left(-\eta(K+1)\right) + c_1\gamma + c_2\gamma^2. \tag{40}$$

Next we consider two possible situations.

1. If $\frac{1}{d} \geq \frac{\ln\left(\max\{2,\min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}$ then we choose $\gamma = \frac{\ln\left(\max\{2,\min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}$ and get that

$$\begin{aligned} r_K &\overset{(40)}{\leq} \frac{a}{\gamma}\exp\left(-\eta(K+1)\right) + c_1\gamma + c_2\gamma^2 \\ &= \widetilde{\mathcal{O}}\left(a\mu K\exp\left(-\min\left\{\rho_1,\rho_2,\frac{\ln\left(\max\{2,\min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{K}\right\}K\right)\right) \\ &\quad +\widetilde{\mathcal{O}}\left(\frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \end{aligned}$$

   Since $\frac{\ln\left(\max\{2,\min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{K} \leq \min\{\rho_1,\rho_2\}$ we have

$$\begin{aligned} r_K &= \widetilde{\mathcal{O}}\left(a\mu K\exp\left(-\ln\left(\max\left\{2,\min\left\{\frac{a\mu^2K^2}{c_1},\frac{a\mu^3K^3}{c_2}\right\}\right\}\right)\right)\right) \\ &\quad +\widetilde{\mathcal{O}}\left(\frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \end{aligned}$$

2. If $\frac{1}{d} \leq \frac{\ln\left(\max\{2,\min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}$ then we choose $\gamma = \frac{1}{d}$ which implies that

$$\begin{aligned} r_K &\overset{(40)}{\leq} da\exp\left(-\min\left\{\frac{\mu}{d},\frac{\rho_1}{4},\frac{\rho_2}{4}\right\}(K+1)\right) + \frac{c_1}{d} + \frac{c_2}{d^2} \\ &= \widetilde{\mathcal{O}}\left(da\exp\left(-\min\left\{\frac{\mu}{d},\rho_1,\rho_2\right\}K\right) + \frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \end{aligned}$$

Combining the obtained bounds we get the result. □

**Lemma D.3.** *Let $\{r_k\}_{k\geq 0}$ satisfy*

$$r_K \leq \frac{a}{\gamma K} + \frac{b_1\gamma}{K} + \frac{b_2\gamma^2}{K} + c_1\gamma + c_2\gamma^2 \tag{41}$$

*for all $K \geq 0$ with some constants $a > 0$, $b_1, b_2, c_1, c_2 \geq 0$ where $\gamma \leq \gamma_0$. Then for all $K$ and*

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{a}{b_1}}, \sqrt[3]{\frac{a}{b_2}}, \sqrt{\frac{a}{c_1K}}, \sqrt[3]{\frac{a}{c_2K}}\right\}$$

*we have that*

$$r_K = \mathcal{O}\left(\frac{a}{\gamma_0K} + \frac{\sqrt{ab_1}}{K} + \frac{\sqrt[3]{a^2b_2}}{K} + \sqrt{\frac{ac_1}{K}} + \frac{\sqrt[3]{a^2c_2}}{K^{2/3}}\right). \tag{42}$$

*Proof.* We have

$$\begin{aligned} r_K &\leq \frac{a}{\gamma K} + \frac{b_1\gamma}{K} + \frac{b_2\gamma^2}{K} + c_1\gamma + c_2\gamma^2 \\ &\leq \frac{a}{\min\left\{\gamma_0, \sqrt{\frac{a}{b_1}}, \sqrt[3]{\frac{a}{b_2}}, \sqrt{\frac{a}{c_1K}}, \sqrt[3]{\frac{a}{c_2K}}\right\}K} + \frac{b_1}{K}\cdot\sqrt{\frac{a}{b_1}} + \frac{b_2}{K}\cdot\sqrt[3]{\frac{a}{b_2}} \\ &\quad +c_1\cdot\sqrt{\frac{a}{c_1K}} + c_2\left(\sqrt[3]{\frac{a}{c_2K}}\right)^2 \\ &= \mathcal{O}\left(\frac{a}{\gamma_0K} + \frac{\sqrt{ab_1}}{K} + \frac{\sqrt[3]{a^2b_2}}{K} + \sqrt{\frac{ac_1}{K}} + \frac{\sqrt[3]{a^2c_2}}{K^{2/3}}\right). \end{aligned}$$

□

# E Proofs for Section 3

## E.1 A lemma

**Lemma E.1** (See also Lemma 8 from [46])**.** *Let Assumptions 3.1, 3.4 and 3.2 be satisfied and* $\gamma \leq 1/4(A'+C_1M_1+C_2M_2)$*. Then for all* $k \geq 0$ *we have*

$$\frac{\gamma}{2}\mathbf{E}\left[f(x^k) - f(x^*)\right] \leq (1-\eta)\mathbf{E}T^k - \mathbf{E}T^{k+1} + \gamma^2(D_1' + M_1D_2) + 3L\gamma\mathbf{E}\|e^k\|^2, \tag{43}$$

*where* $T^k \overset{def}{=} \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$ *and* $M_1 = \frac{4B_1'}{3\rho_1}$*,* $M_2 = \frac{4\left(B_2'+\frac{4}{3}G\right)}{3\rho_2}$*.*

*Proof.* We start with the upper bound for $\mathbf{E}\|\tilde{x}^{k+1} - x^*\|^2$. First of all, by definition of $\tilde{x}^k$ we have

$$\begin{aligned}
\|\tilde{x}^{k+1} - x^*\|^2 &\overset{(25)}{=} \|\tilde{x}^k - x^* - \gamma g^k\|^2 \\
&= \|\tilde{x}^k - x^*\|^2 - 2\gamma\langle \tilde{x}^k - x^*, g^k\rangle + \gamma^2\|g^k\|^2 \\
&= \|\tilde{x}^k - x^*\|^2 - 2\gamma\langle x^k - x^*, g^k\rangle + \gamma^2\|g^k\|^2 + 2\gamma\langle x^k - \tilde{x}^k, g^k\rangle.
\end{aligned}$$

Taking conditional expectation $\mathbf{E}\left[\cdot \mid x^k\right]$ from the both sides of the previous inequality we get

$$\begin{aligned}
\mathbf{E}\left[\|\tilde{x}^{k+1} - x^*\|^2 \mid x^k\right] &\overset{(18),(15)}{\leq} \|\tilde{x}^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k)\rangle \\
&\quad + \gamma^2\left(2A'(f(x^k) - f(x^*)) + B_1'\sigma_{1,k}^2 + B_2'\sigma_{2,k}^2 + D_1'\right) \\
&\quad + 2\gamma\langle x^k - \tilde{x}^k, \nabla f(x^k)\rangle \\
&\overset{(9)}{\leq} \|\tilde{x}^k - x^*\|^2 - \gamma\mu\|x^k - x^*\|^2 - \gamma(2 - 2A'\gamma)(f(x^k) - f(x^*)) \\
&\quad + \gamma^2B_1'\sigma_{1,k}^2 + \gamma^2B_2'\sigma_{2,k}^2 + \gamma^2D_1' \\
&\quad + 2\gamma\langle x^k - \tilde{x}^k, \nabla f(x^k)\rangle.
\end{aligned} \tag{44}$$

Next,

$$-\|x^k - x^*\|^2 = -\|\tilde{x}^k - x^* + x^k - \tilde{x}^k\|^2 \overset{(29)}{\leq} -\frac{1}{2}\|\tilde{x}^k - x^*\|^2 + \|x^k - \tilde{x}^k\|^2. \tag{45}$$

Using Fenchel-Young inequality we derive an upper bound for the inner product from (44):

$$\langle x^k - \tilde{x}^k, \nabla f(x^k)\rangle \overset{(27)}{\leq} L\|x^k - \tilde{x}^k\|^2 + \frac{1}{4L}\|\nabla f(x^k)\|^2 \overset{(11)}{\leq} L\|x^k - \tilde{x}^k\|^2 + \frac{1}{2}(f(x^k) - f(x^*)). \tag{46}$$

Combining previous three inequalities we get

$$\begin{aligned}
\mathbf{E}\left[\|\tilde{x}^{k+1} - x^*\|^2 \mid x^k\right] &\overset{(44)-(46)}{\leq} \left(1 - \frac{\gamma\mu}{2}\right)\|\tilde{x}^k - x^*\|^2 - \gamma\left(1 - 2A'\gamma\right)\left(f(x^k) - f(x^*)\right) \\
&\quad + \gamma^2B_1'\sigma_{1,k}^2 + \gamma^2B_2'\sigma_{2,k}^2 + \gamma^2D_1' \\
&\quad + \gamma(2L + \mu)\|x^k - \tilde{x}^k\|^2.
\end{aligned} \tag{47}$$

Taking into account that $T^k = \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$ with $M_1 = \frac{4B_1'}{3\rho_1}$ and $M_2 = \frac{4\left(B_2'+\frac{4}{3}G\right)}{3\rho_2}$, using the tower property (35) of mathematical expectation together with

$\gamma \le \frac{1}{4(A'+C_1M_1+C_2M_2)}$, we conclude

$$\begin{aligned}
\mathbf{E}\left[T^{k+1}\right] &\overset{(47)}{\le} \left(1-\frac{\gamma\mu}{2}\right)\mathbf{E}\|\tilde{x}^k - x^*\|^2 - \gamma\left(1-2A'\gamma\right)\mathbf{E}\left[f(x^k)-f(x^*)\right] + M_1\gamma^2\mathbf{E}\left[\sigma_{1,k+1}^2\right] \\
&\quad M_2\gamma^2\mathbf{E}\left[\sigma_{2,k+1}^2\right] + \gamma^2 B_1'\sigma_{1,k}^2 + \gamma^2 B_2'\sigma_{2,k}^2 + \gamma^2 D_1' + \gamma(2L+\mu)\mathbf{E}\|x^k - \tilde{x}^k\|^2 \\
&\overset{(16),(17)}{\le} \left(1-\frac{\gamma\mu}{2}\right)\mathbf{E}\|\tilde{x}^k - x^*\|^2 + \left(1+\frac{B_1'}{M_1}-\rho_1\right)M_1\gamma^2\mathbf{E}\left[\sigma_{1,k}^2\right] \\
&\quad + \left(1+\frac{B_2'+M_1G\rho_1}{M_2}-\rho_2\right)M_2\gamma^2\mathbf{E}\left[\sigma_{2,k}^2\right] + \gamma^2(D_1'+M_1D_2) \\
&\quad -\gamma\left(1-2(A'+C_1M_1+C_2M_2)\gamma\right)\mathbf{E}\left[f(x^k)-f(x^*)\right] + \gamma(2L+\mu)\mathbf{E}\|x^k-\tilde{x}^k\|^2 \\
&\le \left(1-\frac{\gamma\mu}{2}\right)\mathbf{E}\|\tilde{x}^k - x^*\|^2 + \left(1-\frac{\rho_1}{4}\right)M_1\gamma^2\mathbf{E}\left[\sigma_{1,k}^2\right] + \left(1-\frac{\rho_2}{4}\right)M_2\gamma^2\mathbf{E}\left[\sigma_{2,k}^2\right] \\
&\quad -\frac{\gamma}{2}\mathbf{E}\left[f(x^k)-f(x^*)\right] + \gamma(2L+\mu)\mathbf{E}\|x^k-\tilde{x}^k\|^2 + \gamma^2(D_1'+M_1D_2).
\end{aligned}$$

Since $L \ge \mu$, $\tilde{x}^k = x^k - e^k$ and $\eta \overset{\text{def}}{=} \min\{\frac{\gamma\mu}{2}, \frac{\rho_1}{4}, \frac{\rho_2}{4}\}$ the last inequality implies

$$\frac{\gamma}{2}\mathbf{E}\left[f(x^k)-f(x^*)\right] \le (1-\eta)\mathbf{E}T^k - \mathbf{E}T^{k+1} + \gamma^2(D_1'+M_1D_2) + 3L\gamma\mathbf{E}\|e^k\|^2,$$

which concludes the proof. □

### E.2 Proof of Theorem 3.1

*Proof.* Form Lemma E.1 we have

$$\frac{\gamma}{2}\mathbf{E}\left[f(x^k)-f(x^*)\right] \le (1-\eta)\mathbf{E}T^k - \mathbf{E}T^{k+1} + \gamma^2(D_1'+M_1D_2) + 3L\gamma\mathbf{E}\|e^k\|^2.$$

Summing up these inequalities for $k = 0,\ldots,K$ with weights $w_k = (1-\eta)^{-(k+1)}$ we get

$$\begin{aligned}
\frac{1}{2}\sum_{k=0}^K w_k\mathbf{E}\left[f(x^k)-f(x^*)\right] &\le \sum_{k=0}^K\left(\frac{w_k(1-\eta)}{\gamma}\mathbf{E}T^k - \frac{w_k}{\gamma}\mathbf{E}T^{k+1}\right) + \gamma(D_1'+M_1D_2)\sum_{k=0}^K w_k \\
&\quad +3L\sum_{k=0}^K w_k\mathbf{E}\|e^k\|^2 \\
&\overset{(19),(20)}{\le} \sum_{k=0}^K\left(\frac{w_{k-1}}{\gamma}\mathbf{E}T^k - \frac{w_k}{\gamma}\mathbf{E}T^{k+1}\right) + F_1\sigma_{1,0}^2 + F_2\sigma_{2,0}^2 \\
&\quad +\gamma^2(D_1'+M_1D_2+D_3)W_K + \frac{1}{4}\sum_{k=0}^K w_k\mathbf{E}\left[f(x^k)-f(x^*)\right].
\end{aligned}$$

Rearranging the terms and using $\bar{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$ together with Jensen's inequality we obtain

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \le \frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma W_K} + 4\gamma\left(D_1'+M_1D_2+D_3\right).$$

Finally, using the definition of the sequences $\{W_K\}_{K\ge 0}$ and $\{w_k\}_{k\ge 0}$ we derive that if $\mu >$, then $W_K \ge w_K \ge (1-\eta)^{-K}$ and we get (21). In the case when $\mu = 0$ we have $w_k = 1$ and $W_K = K$ which implies (22). □

## F SGD as a Special Case

In this section we want to show that our approach is general enough to cover many existing methods of SGD type. Consider the following situation:

$$v^k = \gamma g^k, \quad e^0 = 0. \tag{48}$$

It implies that $e^k = 0$ for all $k \geq 0$ and the updates rules (4)-(5) gives us a simple SGD:

$$x^{k+1} = x^k - \gamma g^k. \tag{49}$$

The following lemma formally shows that SGD under general enough assumptions satisfies Assumption 3.4.

**Lemma F.1.** *Let Assumptions 3.1 and 3.2 be satisfies and inequalities* (18), (15), (16) *and* (17) *hold. Then for the method* (49) *inequality* (19) *holds with $F_1 = F_2 = 0$ and $D_3 = 0$ for all $k \geq 0$.*

*Proof.* Since $e^k = 0$ and $f(x^k) \geq f(x^*)$ for all $k \geq 0$ we get

$$3L\sum_{k=0}^K w_k \mathbf{E}\|e^k\|^2 = 0 \leq \frac{1}{4}\sum_{k=0}^K w_k \mathbf{E}\left[f(x^k) - f(x^*)\right]$$

which concludes the proof. □

It implies that all methods considered in [11] fit our framework. Moreover, using Theorem 3.1 we derive the following result.

**Theorem F.1.** *Let Assumptions 3.1 and 3.2 be satisfied, inequalities* (18), (15), (16), (17) *hold and $\gamma \leq 1/4(A' + C_1M_1 + C_2M_2)$. Then for the method* (49) *for all $K \geq 0$ we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\rho_1}{4}, \frac{\rho_2}{4}\right\}\right)^K \frac{4T^0}{\gamma} + 4\gamma\left(D_1' + M_1D_2\right),$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4T^0}{\gamma K} + 4\gamma\left(D_1' + M_1D_2\right)$$

*when $\mu = 0$, where $T^k \stackrel{def}{=} \|x^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$ and $M_1 = \frac{4B_1'}{3\rho_1}$, $M_2 = \frac{4\left(B_2' + \frac{4}{3}G\right)}{3\rho_2}$.*

In particular, if $\sigma_{2,k}^2 \equiv 0$, then our assumption coincides with the key assumption from [11] and our theorem recovers the same rates as in [11] when $\mu > 0$. The case when $\mu = 0$ was not considered in [11], while in our analysis we get it for free.

# G Distributed SGD with Compression and Error Compensation

In this section we consider the scenario when compression and error-feedback is applied in order to reduce the communication cost of the method, i.e., we consider SGD with error compensation and compression (EC-SGD) which has updates of the form (4)-(5) with

$$
\begin{aligned}
g^k &= \frac{1}{n}\sum_{i=1}^n g_i^k \\
v^k &= \frac{1}{n}\sum_{i=1}^n v_i^k, \quad v_i^k = C(e_i^k + \gamma g_i^k) && (50)\\
e^k &= \frac{1}{n}\sum_{i=1}^n e_i^k, \quad e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k = e_i^k + \gamma g_i^k - C(e_i^k + \gamma g_i^k). && (51)
\end{aligned}
$$

Moreover, we assume that $e_i^0 = 0$ for $i = 1, \ldots, n$.

**Lemma G.1.** *Let Assumptions 3.1 and 3.2 be satisfied, Assumption 3.3 holds and*[5]

$$
\gamma \le \min\left\{\frac{\delta}{4\mu}, \sqrt{\frac{\delta}{96L\left(\frac{2A}{\delta} + \widetilde{A} + \frac{2}{1-\rho_1}\left(\frac{C_1}{\rho_1} + \frac{2GC_2}{\rho_2(1-\rho_2)}\right)\left(\frac{2B_1}{\delta} + \widetilde{B}_1\right) + \frac{2C_2\left(\frac{2B_2}{\delta} + \widetilde{B}_2\right)}{\rho_2(1-\rho_2)}\right)}}\right\}, \tag{52}
$$

*where* $M_1 = \frac{4B_1'}{3\rho_1}$ *and* $M_2 = \frac{4\left(B_2' + \frac{4}{3}G\right)}{3\rho_2}$. *Then* EC-SGD *satisfies Assumption 3.3, i.e., inequality* (19) *holds with the following parameters:*

$$
F_1 = \frac{24L\gamma^2}{\delta\rho_1(1-\eta)}\left(\frac{2B_1}{\delta} + \widetilde{B}_1\right), \quad F_2 = \frac{24L\gamma^2}{\delta\rho_2(1-\eta)}\left(\frac{2G}{1-\rho_1}\left(\frac{2B_1}{\delta} + \widetilde{B}_1\right) + \frac{2B_2}{\delta} + \widetilde{B}_2\right), \tag{53}
$$

$$
D_3 = \frac{6L\gamma}{\delta}\left(\frac{D_2}{\rho_1}\left(\frac{2B_1}{\delta} + \widetilde{B}_1\right) + \frac{2D_1}{\delta} + \widetilde{D}_1\right). \tag{54}
$$

*Proof.* First of all, we derive an upper bound for the second moment of $e_i^{k+1}$:

$$
\begin{aligned}
\mathbf{E}\|e_i^{k+1}\|^2 &\overset{(51),(35)}{=} \mathbf{E}\left[\mathbf{E}\left[\|e_i^k + \gamma g_i^k - C(e_i^k + \gamma g_i^k)\|^2 \mid e_i^k, g_i^k\right]\right] \\
&\overset{(8)}{\le} (1-\delta)\mathbf{E}\|e_i^k + \gamma g_i^k\|^2 \\
&\overset{(35),(34)}{=} (1-\delta)\mathbf{E}\|e_i^k + \gamma \bar{g}_i^k\|^2 + (1-\delta)\gamma^2\mathbf{E}\|g_i^k - \bar{g}_i^k\|^2 \\
&\overset{(30)}{\le} (1-\delta)(1+\beta)\mathbf{E}\|e_i^k\|^2 + (1-\delta)\left(1 + \frac{1}{\beta}\right)\gamma^2\mathbf{E}\|\bar{g}_i^k\|^2 \\
&\quad + (1-\delta)\gamma^2\mathbf{E}\|g_i^k - \bar{g}_i^k\|^2.
\end{aligned}
$$

Summing up these inequalities for $i = 1, \ldots, n$ we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|e_i^{k+1}\|^2 &\le (1-\delta)(1+\beta)\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|e_i^k\|^2 \\
&\quad + (1-\delta)\left(1 + \frac{1}{\beta}\right)\gamma^2\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|\bar{g}_i^k\|^2 + (1-\delta)\gamma^2\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|g_i^k - \bar{g}_i^k\|^2 (55)
\end{aligned}
$$

[5]When $\rho_1 = 1$ and $\rho_2 = 1$ one can always set the parameters in such a way that $B_1 = \widetilde{B}_1 = B_2 = \widetilde{B}_2 = C_1 = C_2 = 0$, $D_2 = 0$. In this case we assume that $\frac{2}{1-\rho_1}\left(\frac{C_1}{\rho_1} + \frac{2GC_2}{\rho_2(1-\rho_2)}\right)\left(\frac{2B_1}{\delta} + \widetilde{B}_1\right) + \frac{2C_2\left(\frac{2B_2}{\delta} + \widetilde{B}_2\right)}{\rho_2(1-\rho_2)} = 0$.

Consider $\beta = \frac{\delta}{2(1-\delta)}$. For this choice of $\beta$ we have

$$
\begin{aligned}
(1-\delta)(1+\beta) &= (1-\delta)\left(1+\frac{\delta}{2(1-\delta)}\right) = 1-\frac{\delta}{2}\\
(1-\delta)\left(1+\frac{1}{\beta}\right) &= (1-\delta)\left(1+\frac{2(1-\delta)}{\delta}\right) = \frac{(1-\delta)(2-\delta)}{\delta} \leq \frac{2(1-\delta)}{\delta}.
\end{aligned}
$$

Using this we continue our derivations:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|e_i^{k+1}\|^2 &\leq \left(1-\frac{\delta}{2}\right)\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|e_i^k\|^2 + \frac{2\gamma^2(1-\delta)}{\delta}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|\bar{g}_i^k\|^2\\
&\quad +(1-\delta)\gamma^2\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|g_i^k-\bar{g}_i^k\|^2\\
&\overset{(13),(14)}{\leq} \left(1-\frac{\delta}{2}\right)\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|e_i^k\|^2 + 2\gamma^2(1-\delta)\left(\frac{2A}{\delta}+\widetilde{A}\right)\mathbf{E}\left[f(x^k)-f(x^*)\right]\\
&\quad +\gamma^2(1-\delta)\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)\mathbf{E}\sigma_{1,k}^2 + \gamma^2(1-\delta)\left(\frac{2B_2}{\delta}+\widetilde{B}_2\right)\mathbf{E}\sigma_{2,k}^2\\
&\quad +\gamma^2(1-\delta)\left(\frac{2D_1}{\delta}+\widetilde{D}_1\right). \qquad (56)
\end{aligned}
$$

Unrolling the recurrence above we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|e_i^{k+1}\|^2 &\overset{(56)}{\leq} 2\gamma^2(1-\delta)\left(\frac{2A}{\delta}+\widetilde{A}\right)\sum_{l=0}^k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\left[f(x^l)-f(x^*)\right]\\
&\quad +\gamma^2(1-\delta)\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)\sum_{l=0}^k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{1,l}^2\\
&\quad +\gamma^2(1-\delta)\left(\frac{2B_2}{\delta}+\widetilde{B}_2\right)\sum_{l=0}^k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{2,l}^2\\
&\quad +\gamma^2(1-\delta)\left(\frac{2D_1}{\delta}+\widetilde{D}_1\right)\sum_{l=0}^k\left(1-\frac{\delta}{2}\right)^{k-l} \qquad (57)
\end{aligned}
$$

which implies

$$
\begin{aligned}
3L\sum_{k=0}^K w_k\mathbf{E}\|e^k\|^2 &\overset{(51)}{=} 3L\sum_{k=0}^K w_k\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n e_i^k\right\|^2 \overset{(31)}{\leq} 3L\sum_{k=0}^K w_k\frac{1}{n}\sum_{i=1}^n\mathbf{E}\left\|e_i^k\right\|^2\\
&\overset{(57)}{\leq} \frac{6L\gamma^2(1-\delta)}{1-\frac{\delta}{2}}\left(\frac{2A}{\delta}+\widetilde{A}\right)\sum_{k=0}^K\sum_{l=0}^k w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\left[f(x^l)-f(x^*)\right]\\
&\quad +\frac{3L\gamma^2(1-\delta)}{1-\frac{\delta}{2}}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)\sum_{k=0}^K\sum_{l=0}^k w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{1,l}^2\\
&\quad +\frac{3L\gamma^2(1-\delta)}{1-\frac{\delta}{2}}\left(\frac{2B_2}{\delta}+\widetilde{B}_2\right)\sum_{k=0}^K\sum_{l=0}^k w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{2,l}^2\\
&\quad +\frac{3L\gamma^2(1-\delta)}{1-\frac{\delta}{2}}\left(\frac{2D_1}{\delta}+\widetilde{D}_1\right)\sum_{k=0}^K\sum_{l=0}^k w_k\left(1-\frac{\delta}{2}\right)^{k-l}. \qquad (58)
\end{aligned}
$$

In the remaining part of the proof we derive upper bounds for three terms in the right-hand side of the previous inequality. First of all, recall that $w_k = (1-\eta)^{-(k+1)}$ and

$\eta = \min\left\{\frac{\gamma\mu}{2}, \frac{\rho_1}{4}, \frac{\rho_2}{4}\right\}$. It implies that for all $0 \le i < k$ we have

$$
\begin{aligned}
w_k &= (1-\eta)^{-(k-j+1)}(1-\eta)^{-j} \overset{(32)}{\le} w_{k-j}(1+2\eta)^j \\
&\le w_{k-j}(1+\gamma\mu)^j \overset{(52)}{\le} w_{k-j}\left(1+\frac{\delta}{4}\right)^j, \qquad (59) \\
w_k &= (1-\eta)^{-(k-j+1)}(1-\eta)^{-j} \overset{(32)}{\le} w_{k-j}(1+2\eta)^j \\
&\le w_{k-j}\left(1+\frac{\min\{\rho_1,\rho_2\}}{2}\right)^j. \qquad (60)
\end{aligned}
$$

For simplicity, we introduce new notation: $r_k \overset{\text{def}}{=} \mathbf{E}\left[f(x^k) - f(x^*)\right]$. Using this we get

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l} r_l &\overset{(59)}{\le} \sum_{k=0}^{K}\sum_{l=0}^{k} w_l r_l \left(1+\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\delta}{2}\right)^{k-l} \\
&\overset{(33)}{\le} \sum_{k=0}^{K}\sum_{l=0}^{k} w_l r_l \left(1-\frac{\delta}{4}\right)^{k-l} \\
&\le \left(\sum_{k=0}^{K} w_k r_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{4}\right)^k\right) = \frac{4}{\delta}\sum_{k=0}^{K} w_k r_k. \qquad (61)
\end{aligned}
$$

Next, we apply our assumption on $\sigma_{2,k}^2$ and derive that

$$
\begin{aligned}
\mathbf{E}\sigma_{2,k+1}^2 &\overset{(17)}{\le} (1-\rho_2)\mathbf{E}\sigma_{2,k}^2 + 2C_2\underbrace{\mathbf{E}\left[f(x^k)-f(x^*)\right]}_{r_k} \\
&\le (1-\rho_2)^{k+1}\sigma_{2,0}^2 + 2C_2\sum_{l=0}^{k}(1-\rho_2)^{k-l} r_l, \qquad (62)
\end{aligned}
$$

hence

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{2,l}^2 &\le \sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_2)^l\sigma_{2,0}^2 \\
&\quad + \frac{2C_2}{1-\rho_2}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l} w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_2)^{l-t} r_t.
\end{aligned}
$$

Using this and

$$
\begin{aligned}
w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_2)^{l-t} &\overset{(59)}{\le} w_l\left(1+\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_2)^{l-t} \\
&\overset{(33),(60)}{\le} \left(1-\frac{\delta}{4}\right)^{k-l}\left(1+\frac{\rho_2}{2}\right)^{l-t}(1-\rho_2)^{l-t} w_t \\
&\overset{(33)}{\le} \left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_2}{2}\right)^{l-t} w_t
\end{aligned}
$$

we derive

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{2,l}^2 &\le \sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_2}{2}\right)^{l} w_0\sigma_{2,0}^2 \\
&\quad + \frac{2C_2}{1-\rho_2}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l}\left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_2}{2}\right)^{l-t} w_t r_t \\
&\le w_0\sigma_{2,0}^2\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{4}\right)^{k}\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\rho_2}{2}\right)^{k}\right) \\
&\quad \frac{2C_2}{1-\rho_2}\left(\sum_{k=0}^{K} w_k r_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{4}\right)^{k}\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\rho_2}{2}\right)^{k}\right) \\
&= \frac{8\sigma_{2,0}^2}{\delta\rho_2(1-\eta)} + \frac{16C_2}{\delta\rho_2(1-\rho_2)}\sum_{k=0}^{K} w_k r_k. \qquad (63)
\end{aligned}
$$

Similarly, we estimate $\sigma_{1,k}^2$:

$$
\begin{aligned}
\mathbf{E}\sigma_{1,k+1}^2 &\overset{(16)}{\le} (1-\rho_1)\mathbf{E}\sigma_{1,k}^2 + 2C_1\underbrace{\mathbf{E}\left[f(x^k)-f(x^*)\right]}_{r_k} + G\rho_1\mathbf{E}\sigma_{2,k}^2 + D_2 \\
&\le (1-\rho_1)^{k+1}\sigma_{1,0}^2 + 2C_1\sum_{l=0}^{k}(1-\rho_1)^{k-l}r_l + G\rho_1\sum_{l=0}^{k}(1-\rho_1)^{k-l}\mathbf{E}\sigma_{2,k}^2 \\
&\quad + D_2\sum_{l=0}^{k}(1-\rho_1)^{l} \\
&\le (1-\rho_1)^{k+1}\sigma_{1,0}^2 + 2C_1\sum_{l=0}^{k}(1-\rho_1)^{k-l}r_l + G\rho_1\sum_{l=0}^{k}(1-\rho_1)^{k-l}\mathbf{E}\sigma_{2,k}^2 \\
&\quad + D_2\sum_{l=0}^{\infty}(1-\rho_1)^{l} \\
&= (1-\rho_1)^{k+1}\sigma_{1,0}^2 + 2C_1\sum_{l=0}^{k}(1-\rho_1)^{k-l}r_l + G\rho_1\sum_{l=0}^{k}(1-\rho_1)^{k-l}\mathbf{E}\sigma_{2,k}^2 \\
&\quad + \frac{D_2}{\rho_1}. \qquad (64)
\end{aligned}
$$

Using this we get

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{1,l}^2 &\le \sigma_{1,0}^2\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_1)^{l} \\
&\quad + \frac{2C_1}{1-\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l} w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_1)^{l-t}r_t \\
&\quad + \frac{G\rho_1}{1-\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l} w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_1)^{l-t}\mathbf{E}\sigma_{2,t}^2 \\
&\quad + \frac{D_2}{\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l} w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_1)^{l-t}. \qquad (65)
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
w_k\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_1)^{l-t} &\overset{(59)}{\leq} w_l\left(1+\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\delta}{2}\right)^{k-l}(1-\rho_1)^{l-t}\\
&\overset{(33),(60)}{\leq} \left(1-\frac{\delta}{4}\right)^{k-l}\left(1+\frac{\rho_1}{2}\right)^{l-t}(1-\rho_1)^{l-t}w_t\\
&\overset{(33)}{\leq} \left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_1}{2}\right)^{l-t}w_t,
\end{aligned}
$$

hence

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{1,l}^2 &\overset{(65)}{\leq} w_0\sigma_{1,0}^2\sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_1}{2}\right)^{l}\\
&\quad+\frac{2C_1}{1-\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l}\left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_1}{2}\right)^{l-t}w_t r_t\\
&\quad+\frac{G\rho_1}{1-\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k}\sum_{t=0}^{l}\left(1-\frac{\delta}{4}\right)^{k-l}\left(1-\frac{\rho_1}{2}\right)^{l-t}w_t\mathbf{E}\sigma_{2,t}^2\\
&\quad+\frac{D_2}{\rho_1}\left(\sum_{k=0}^{K}w_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{2}\right)^{k}\right)\left(\sum_{k=0}^{\infty}(1-\rho_1)^k\right)\\
&\leq w_0\sigma_{1,0}^2\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{4}\right)^{k}\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\rho_1}{2}\right)^{k}\right)\\
&\quad+\frac{2C_1}{1-\rho_1}\left(\sum_{k=0}^{K}w_k r_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{4}\right)^{k}\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\rho_1}{2}\right)^{k}\right)\\
&\quad+\frac{G\rho_1}{1-\rho_1}\left(\sum_{k=0}^{K}w_k\mathbf{E}\sigma_{2,k}^2\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\delta}{4}\right)^{k}\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{\rho_1}{2}\right)^{k}\right)\\
&\quad+\frac{2D_2}{\delta\rho_1}W_K\\
&= \frac{8\sigma_{1,0}^2}{\delta\rho_1(1-\eta)}+\frac{16C_1}{\delta\rho_1(1-\rho_1)}\sum_{k=0}^{K}w_k r_k+\frac{8G}{\delta(1-\rho_1)}\sum_{k=0}^{K}w_k\mathbf{E}\sigma_{2,k}^2\\
&\quad+\frac{2D_2}{\delta\rho_1}W_K. \qquad (66)
\end{aligned}
$$

For the third term in the right-hand side of previous inequality we have

$$
\begin{aligned}
\frac{8G}{\delta(1-\rho_1)}\sum_{k=0}^K w_k \mathbf{E}\sigma_{2,k}^2 &\overset{(62)}{\leq} \frac{8G\sigma_{2,0}^2}{\delta(1-\rho_1)}\sum_{k=0}^K w_k(1-\rho_2)^k \\
&\quad + \frac{16GC_2}{\delta(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K\sum_{l=0}^k w_k(1-\rho_2)^{k-l}r_l \\
&\overset{(60)}{\leq} \frac{8G\sigma_{2,0}^2 w_0}{\delta(1-\rho_1)}\sum_{k=0}^K \left(1+\frac{\rho_2}{2}\right)^k(1-\rho_2)^k \\
&\quad + \frac{16GC_2}{\delta(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K\sum_{l=0}^k \left(1+\frac{\rho_2}{2}\right)^{k-l}(1-\rho_2)^{k-l}w_l r_l \\
&\overset{(33)}{\leq} \frac{8G\sigma_{2,0}^2 w_0}{\delta(1-\rho_1)}\sum_{k=0}^\infty \left(1-\frac{\rho_2}{2}\right)^k \\
&\quad + \frac{16GC_2}{\delta(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K\sum_{l=0}^k \left(1-\frac{\rho_2}{2}\right)^{k-l} w_l r_l \\
&\leq \frac{16G\sigma_{2,0}^2 w_0}{\delta\rho_2(1-\rho_1)} + \frac{16GC_2}{\delta(1-\rho_1)(1-\rho_2)}\left(\sum_{k=0}^K w_k r_k\right)\left(\sum_{k=0}^\infty\left(1-\frac{\rho_2}{2}\right)^k\right) \\
&= \frac{16G\sigma_{2,0}^2}{\delta\rho_2(1-\rho_1)(1-\eta)} + \frac{32GC_2}{\delta\rho_2(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K w_k r_k \qquad (67)
\end{aligned}
$$

Combining inequalities (66) and (67) we get

$$
\begin{aligned}
\sum_{k=0}^K\sum_{l=0}^k w_k\left(1-\frac{\delta}{2}\right)^{k-l}\mathbf{E}\sigma_{1,l}^2 &\leq \frac{8\sigma_{1,0}^2}{\delta\rho_1(1-\eta)} + \frac{16}{\delta(1-\rho_1)}\left(\frac{C_1}{\rho_1}+\frac{2GC_2}{\rho_2(1-\rho_2)}\right)\sum_{k=0}^K w_k r_k \\
&\quad + \frac{16G\sigma_{2,0}^2}{\delta\rho_2(1-\rho_1)(1-\eta)} + \frac{2D_2}{\delta\rho_1}W_K \qquad (68)
\end{aligned}
$$

Finally, we estimate the last term in the right-hand side of (58):

$$
\sum_{k=0}^K\sum_{l=0}^k w_k\left(1-\frac{\delta}{2}\right)^{k-l} \leq \left(\sum_{k=0}^K w_k\right)\left(\sum_{k=0}^\infty\left(1-\frac{\delta}{2}\right)^k\right) = \frac{2}{\delta}W_K. \qquad (69)
$$

Plugging inequalities (61), (63), (68), (69) and $\frac{1-\delta}{1-\frac{\delta}{2}} \leq 1$ in (58) we obtain

$$
\begin{aligned}
3L\sum_{k=0}^K w_k\mathbf{E}\|e^k\|^2 &\leq \frac{24L\left(\frac{2A}{\delta}+\widetilde{A}+\frac{2}{1-\rho_1}\left(\frac{C_1}{\rho_1}+\frac{2GC_2}{\rho_2(1-\rho_2)}\right)\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2C_2\left(\frac{2B_2}{\delta}+\widetilde{B}_2\right)}{\rho_2(1-\rho_2)}\right)\gamma^2}{\delta}\sum_{k=0}^K w_k r_k \\
&\quad + \frac{24L\gamma^2}{\delta\rho_1(1-\eta)}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)\sigma_{1,0}^2 \\
&\quad + \frac{24L\gamma^2}{\delta\rho_2(1-\eta)}\left(\frac{2G}{1-\rho_1}\left(\frac{2B_1}{\delta}+\tilde{B}_1\right)+\frac{2B_2}{\delta}+\widetilde{B}_2\right)\sigma_{2,0}^2 \\
&\quad + \frac{6L\gamma^2}{\delta}\left(\frac{D_2}{\rho_1}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2D_1}{\delta}+\widetilde{D}_1\right)W_K.
\end{aligned}
$$

Taking into account that $\gamma \leq \sqrt{\frac{\delta}{96L\left(\frac{2A}{\delta}+\widetilde{A}+\frac{2}{1-\rho_1}\left(\frac{C_1}{\rho_1}+\frac{2GC_2}{\rho_2(1-\rho_2)}\right)\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2C_2\left(\frac{2B_2}{\delta}+\widetilde{B}_2\right)}{\rho_2(1-\rho_2)}\right)}}$, $F_1 = \frac{24L\gamma^2}{\delta\rho_1(1-\eta)}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)$, $F_2 = \frac{24L\gamma^2}{\delta\rho_2(1-\eta)}\left(\frac{2G}{1-\rho_1}\left(\frac{2B_1}{\delta}+\tilde{B}_1\right)+\frac{2B_2}{\delta}+\widetilde{B}_2\right)$ and $D_3 =$

$\frac{6L\gamma}{\delta}\left(\frac{D_2}{\rho_1}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2D_1}{\delta}+\widetilde{D}_1\right)$ we get

$$3L\sum_{k=0}^{K} w_k \mathbf{E}\|e^k\|^2 \quad\leq\quad \frac{1}{4}\sum_{k=0}^{K} w_k r_k + F_1\sigma_{1,0}^2 + F_2\sigma_{2,0}^2 + \gamma D_3.$$

□

As a direct application of Lemma G.1 and Theorem 3.1 we get the following result.

**Theorem G.1.** *Let Assumptions 3.1 and 3.2 be satisfied, Assumption 3.3 holds and*

$$\gamma \quad\leq\quad \frac{1}{4(A'+C_1M_1+C_2M_2)},$$

$$\gamma \quad\leq\quad \min\left\{\frac{\delta}{4\mu}, \sqrt{\frac{\delta}{96L\left(\frac{2A}{\delta}+\widetilde{A}+\frac{2}{1-\rho_1}\left(\frac{C_1}{\rho_1}+\frac{2GC_2}{\rho_2(1-\rho_2)}\right)\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2C_2\left(\frac{2B_2}{\delta}+\widetilde{B}_2\right)}{\rho_2(1-\rho_2)}\right)}}\right\},$$

*where* $M_1 = \frac{4B_1'}{3\rho_1}$ *and* $M_2 = \frac{4\left(B_2'+\frac{4}{3}G\right)}{3\rho_2}$. *Then for all* $K \geq 0$ *we have*

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \leq (1-\eta)^K\,\frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma} + 4\gamma\left(D_1'+M_1D_2+D_3\right),$$

*when* $\mu > 0$ *and*

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \leq \frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma K} + 4\gamma\left(D_1'+M_1D_2+D_3\right)$$

*when* $\mu = 0$, *where* $\eta = \min\left\{\gamma\mu/2, \rho_1/4, \rho_2/4\right\}$, $T^k \overset{def}{=} \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$ *and*

$$F_1 = \frac{24L\gamma^2}{\delta\rho_1(1-\eta)}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right), \quad F_2 = \frac{24L\gamma^2}{\delta\rho_2(1-\eta)}\left(\frac{2G}{1-\rho_1}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2B_2}{\delta}+\widetilde{B}_2\right),$$

$$D_3 = \frac{6L\gamma}{\delta}\left(\frac{D_2}{\rho_1}\left(\frac{2B_1}{\delta}+\widetilde{B}_1\right)+\frac{2D_1}{\delta}+\widetilde{D}_1\right).$$

# H SGD with Delayed Updates

In this section we consider the SGD with delayed updates (D-SGD) [1, 33, 10, 3, 46]. This method has updates of the form (4)-(5) with

$$
\begin{aligned}
g^k &= \frac{1}{n}\sum_{i=1}^n g_i^k \\
v^k &= \frac{1}{n}\sum_{i=1}^n v_i^k, \quad v_i^k = \begin{cases}\gamma g_i^{k-\tau}, & \text{if } t \geq \tau,\\ 0, & \text{if } t < \tau\end{cases} \qquad (70)\\
e^k &= \frac{1}{n}\sum_{i=1}^n e_i^k, \quad e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k = \gamma \sum_{t=1}^{\tau} g_i^{k+1-t}, \qquad (71)
\end{aligned}
$$

where the summation is performed only for non-negative indices. Moreover, we assume that $e_i^0 = 0$ for $i = 1, \ldots, n$.

For convenience we also introduce new constant:

$$\hat{A} = A' + L\tau. \tag{72}$$

**Lemma H.1.** *Let Assumptions 3.1 and 3.2 be satisfied, inequalities* (15)*,* (16) *and* (17) *hold and*[6]

$$\gamma \leq \min\left\{\frac{1}{2\tau\mu}, \frac{1}{8\sqrt{L\tau\left(\hat{A} + \frac{2B_1'C_1}{\rho_1(1-\rho_1)} + \frac{2B_2'C_2}{\rho_2(1-\rho_2)} + \frac{4B_1'GC_2}{\rho_2(1-\rho_1)(1-\rho_2)}\right)}}\right\}, \tag{73}$$

*where* $M_1 = \frac{4B_1'}{3\rho_1}$ *and* $M_2 = \frac{4\left(B_2' + \frac{4}{3}G\right)}{3\rho_2}$. *Then* D-SGD *satisfies Assumption 3.4, i.e., inequality* (19) *holds with the following parameters:*

$$F_1 = \frac{6\gamma^2 L B_1' \tau(2+\rho_1)}{\rho_1}, \quad F_2 = \frac{6\gamma^2\tau L(2+\rho_2)}{\rho_2}\left(\frac{2B_1'G}{1-\rho_1} + B_2'\right), \tag{74}$$

$$D_3 = 3\gamma\tau L\left(D_1' + \frac{2B_1'D_2}{\rho_1}\right). \tag{75}$$

*Proof.* First of all, we derive an upper bound for the second moment of $e_i^k$:

$$
\begin{aligned}
\mathbf{E}\|e^k\|^2 &\overset{(71)}{=} \gamma^2 \mathbf{E}\left[\left\|\sum_{t=1}^{\tau} g^{k-t}\right\|^2\right] \\
&\overset{(36)}{\leq} \gamma^2\tau\sum_{t=1}^{\tau}\mathbf{E}\left[\|\nabla f(x^{k-t})\|^2\right] + \gamma^2\sum_{t=1}^{\tau}\mathbf{E}\left[\|g^{k-t} - \nabla f(x^{k-t})\|^2\right] \\
&\overset{(34)}{\leq} \gamma^2\tau\sum_{t=1}^{\tau}\mathbf{E}\left[\|\nabla f(x^{k-t})\|^2\right] + \gamma^2\sum_{t=1}^{\tau}\mathbf{E}\left[\|g^{k-t}\|^2\right] \\
&\overset{(15),(11)}{\leq} 2\gamma^2\underbrace{(A' + L\tau)}_{\hat{A}}\sum_{t=1}^{\tau}\mathbf{E}\left[f(x^{k-t}) - f(x^*)\right] + \gamma^2 B_1'\sum_{t=1}^{\tau}\mathbf{E}\sigma_{1,k-t}^2 \\
&\quad + \gamma^2 B_2'\sum_{t=1}^{\tau}\mathbf{E}\sigma_{2,k-t}^2 + \gamma^2\tau D_1' \qquad (76)
\end{aligned}
$$

[6]When $\rho_1 = 1$ and $\rho_2 = 1$ one can always set the parameters in such a way that $B_1 = B_1' = B_2 = B_2' = C_1 = C_2 = 0$, $D_2 = 0$. In this case we assume that $\frac{2B_1'C_1}{\rho_1(1-\rho_1)} = \frac{2B_2'C_2}{\rho_2(1-\rho_2)} = 0$.

which implies

$$
\begin{aligned}
3L\sum_{k=0}^{K} w_k\mathbf{E}\|e^k\|^2 &\overset{(76)}{\leq} 6\gamma^2 L\hat{A}\sum_{k=0}^{K}\sum_{t=1}^{\tau} w_k\mathbf{E}\left[f(x^{k-t})-f(x^*)\right] \\
&\quad +3\gamma^2 LB_1'\sum_{k=0}^{K}\sum_{t=1}^{\tau} w_k\mathbf{E}\sigma_{1,k-t}^2 \\
&\quad +3\gamma^2 LB_2'\sum_{k=0}^{K}\sum_{t=1}^{\tau} w_k\mathbf{E}\sigma_{2,k-t}^2 + 3\gamma^2\tau LD_1'W_K \qquad (77)
\end{aligned}
$$

In the remaining part of the proof we derive upper bounds for four terms in the right-hand side of the previous inequality. First of all, recall that $w_k = (1-\eta)^{-(k+1)}$ and $\eta = \min\left\{\frac{\gamma\mu}{2}, \frac{\rho_1}{4}, \frac{\rho_2}{4}\right\}$. It implies that for all $0\leq i<k$ and $0\leq t\leq\tau$ we have

$$
\begin{aligned}
w_k &= (1-\eta)^{-(k-t+1)}(1-\eta)^{-t} \overset{(32)}{\leq} w_{k-t}(1+2\eta)^t \\
&\leq w_{k-t}(1+\gamma\mu)^t \overset{(73)}{\leq} w_{k-t}\left(1+\frac{1}{2\tau}\right)^t \leq w_{k-t}\exp\left(\frac{t}{2\tau}\right) \leq 2w_{k-t}, \qquad (78) \\
w_k &= (1-\eta)^{-(k-j+1)}(1-\eta)^{-j} \overset{(32)}{\leq} w_{k-j}(1+2\eta)^j \leq w_{k-j}\left(1+\frac{\min\{\rho_1,\rho_2\}}{2}\right)^j . (79)
\end{aligned}
$$

For simplicity, we introduce new notation: $r_k \overset{\text{def}}{=} \mathbf{E}\left[f(x^k)-f(x^*)\right]$. Using this we get

$$
\sum_{k=0}^{K}\sum_{t=1}^{\tau} w_k r_{k-t} \overset{(78)}{\leq} \sum_{k=0}^{K}\sum_{t=1}^{\tau} 2w_{k-t}r_{k-t} \leq 2\tau\sum_{k=0}^{K} w_k r_k \qquad (80)
$$

Similarly, we estimate the second term in the right-hand side of (79):

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{t=1}^{\tau} w_k\mathbf{E}\sigma_{1,k-t}^2 &\leq \sum_{k=0}^{K}\sum_{t=1}^{\tau} 2w_{k-t}\mathbf{E}\sigma_{1,k-t}^2 \leq 2\tau\sum_{k=0}^{K} w_k\mathbf{E}\sigma_{1,k}^2 \\
&\overset{(64)}{\leq} 2\tau\sigma_{1,0}^2\sum_{k=0}^{K} w_k(1-\rho_1)^k + \frac{4C_1\tau}{1-\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k} w_k(1-\rho_1)^{k-l}r_l \\
&\quad +\frac{2G\rho_1\tau}{1-\rho_1}\sum_{k=0}^{K}\sum_{l=0}^{k} w_k(1-\rho_1)^{k-l}\mathbf{E}\sigma_{2,l}^2 + \frac{2\tau D_2}{\rho}W_K. \qquad (81)
\end{aligned}
$$

For the first term in the right-hand side of previous inequality we have

$$
\begin{aligned}
2\tau\sigma_{1,0}^2\sum_{k=0}^{K} w_k(1-\rho_1)^k &\overset{(79)}{\leq} 2\tau\sigma_{1,0}^2\sum_{k=0}^{K}\left(1+\frac{\rho_1}{2}\right)^{k+1}(1-\rho_1)^k \\
&\overset{(33)}{\leq} 2\tau\left(1+\frac{\rho_1}{2}\right)\sigma_{1,0}^2\sum_{k=0}^{K}\left(1-\frac{\rho_1}{2}\right)^k \\
&\leq \tau(2+\rho_1)\sigma_{1,0}^2\sum_{k=0}^{\infty}\left(1-\frac{\rho_1}{2}\right)^k \leq \frac{2\tau(2+\rho_1)\sigma_{1,0}^2}{\rho_1}. \qquad (82)
\end{aligned}
$$

The second term in the right-hand side of (81) can be upper bounded in the following way:

$$
\begin{aligned}
\frac{4C_1\tau}{1-\rho_1}\sum_{k=0}^K\sum_{l=0}^k w_k(1-\rho_1)^{k-l}r_l &\overset{(79)}{\leq} \frac{4C_1\tau}{1-\rho_1}\sum_{k=0}^K\sum_{l=0}^k w_l r_l\left(1+\frac{\rho_1}{2}\right)^{k-l}(1-\rho_1)^{k-l}\\
&\overset{(33)}{\leq} \frac{4C_1\tau}{1-\rho_1}\sum_{k=0}^K\sum_{l=0}^k w_l r_l\left(1-\frac{\rho_1}{2}\right)^{k-l}\\
&\leq \frac{4C_1\tau}{1-\rho_1}\left(\sum_{k=0}^K w_k r_k\right)\left(\sum_{k=0}^\infty\left(1-\frac{\rho_1}{2}\right)^k\right)\\
&\leq \frac{8C_1\tau}{\rho_1(1-\rho_1)}\sum_{k=0}^K w_k r_k. \qquad (83)
\end{aligned}
$$

Repeating similar steps we estimate the third term in the right-hand side of (81):

$$
\begin{aligned}
\frac{2G\rho_1\tau}{1-\rho_1}\sum_{k=0}^K\sum_{l=0}^k w_k(1-\rho_1)^{k-l}\mathbf{E}\sigma_{2,l}^2 &\leq \frac{4G\tau}{1-\rho_1}\sum_{k=0}^K w_k\mathbf{E}\sigma_{2,k}^2\\
&\overset{(62)}{\leq} \frac{4G\tau\sigma_{2,0}^2}{1-\rho_1}\sum_{k=0}^K w_k(1-\rho_2)^k\\
&\quad+\frac{8GC_2}{(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K\sum_{l=0}^k w_k(1-\rho_2)^{k-l}r_l\\
&\overset{(79)}{\leq} \frac{4G\tau\sigma_{2,0}^2}{1-\rho_1}\sum_{k=0}^K\left(1+\frac{\rho_2}{2}\right)^{k+1}(1-\rho_2)^k\\
&\quad+\frac{8GC_2\tau}{(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K\sum_{l=0}^k\left(1+\frac{\rho_2}{2}\right)^{k-l}(1-\rho_2)^{k-l}w_l r_l\\
&\overset{(33)}{\leq} \frac{2G\tau(2+\rho_2)\sigma_{2,0}^2}{1-\rho_1}\sum_{k=0}^\infty\left(1-\frac{\rho_2}{2}\right)^k\\
&\quad+\frac{8GC_2\tau}{(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K\sum_{l=0}^k\left(1-\frac{\rho_2}{2}\right)^{k-l}w_l r_l\\
&\leq \frac{4G\tau(2+\rho_2)\sigma_{2,0}^2}{\rho_2(1-\rho_1)}\\
&\quad+\frac{8GC_2\tau}{(1-\rho_1)(1-\rho_2)}\left(\sum_{k=0}^K w_k r_k\right)\left(\sum_{k=0}^\infty\left(1-\frac{\rho_2}{2}\right)^k\right)\\
&= \frac{4G\tau(2+\rho_2)\sigma_{2,0}^2}{\rho_2(1-\rho_1)}\\
&\quad+\frac{16GC_2\tau}{\rho_2(1-\rho_1)(1-\rho_2)}\sum_{k=0}^K w_k r_k \qquad (84)
\end{aligned}
$$

Combining inequalities (81), (82), (83) and (84) we get

$$
\begin{aligned}
\sum_{k=0}^K\sum_{t=1}^\tau w_k\mathbf{E}\sigma_{1,k-t}^2 &\leq \frac{2\tau\left(2+\rho_1\right)\sigma_{1,0}^2}{\rho_1}+\frac{8\tau}{1-\rho_1}\left(\frac{C_1}{\rho_1}+\frac{2GC_2}{\rho_2(1-\rho_2)}\right)\sum_{k=0}^K w_k r_k\\
&\quad+\frac{4G\tau(2+\rho_2)\sigma_{2,0}^2}{\rho_2(1-\rho_1)}+\frac{2\tau D_2}{\rho}W_K. \qquad (85)
\end{aligned}
$$

Next, we derive

$$\sum_{k=0}^K\sum_{t=1}^\tau w_k\mathbf{E}\sigma_{2,k-t}^2 \leq \sum_{k=0}^K\sum_{t=1}^\tau 2w_{k-t}\mathbf{E}\sigma_{2,k-t}^2 \leq 2\tau\sum_{k=0}^K w_k\mathbf{E}\sigma_{2,k}^2$$

$$\overset{(62)}{\leq} 2\tau\sigma_{2,0}^2\sum_{k=0}^K w_k(1-\rho_1)^k + \frac{4C_2\tau}{1-\rho_2}\sum_{k=0}^K\sum_{l=0}^k w_k(1-\rho_2)^{k-l}r_l. \tag{86}$$

For the first term in the right-hand side of previous inequality we have

$$\begin{aligned}2\tau\sigma_{2,0}^2\sum_{k=0}^K w_k(1-\rho_2)^k &\overset{(79)}{\leq} 2\tau\sigma_{2,0}^2\sum_{k=0}^K\left(1+\frac{\rho_2}{2}\right)^{k+1}(1-\rho_2)^k\\ &\overset{(33)}{\leq} 2\tau\left(1+\frac{\rho_2}{2}\right)\sigma_{2,0}^2\sum_{k=0}^K\left(1-\frac{\rho_2}{2}\right)^k\\ &\leq \tau\left(2+\rho_2\right)\sigma_{2,0}^2\sum_{k=0}^\infty\left(1-\frac{\rho_2}{2}\right)^k \leq \frac{2\tau\left(2+\rho_2\right)\sigma_{2,0}^2}{\rho_2}.\end{aligned}$$

The second term in the right-hand side of (86) can be upper bounded in the following way:

$$\begin{aligned}\frac{4C_2\tau}{1-\rho_2}\sum_{k=0}^K\sum_{l=0}^k w_k(1-\rho_2)^{k-l}r_l &\overset{(79)}{\leq} \frac{4C_2\tau}{1-\rho_2}\sum_{k=0}^K\sum_{l=0}^k w_lr_l\left(1+\frac{\rho_2}{2}\right)^{k-l}(1-\rho_2)^{k-l}\\ &\overset{(33)}{\leq} \frac{4C_2\tau}{1-\rho_2}\sum_{k=0}^K\sum_{l=0}^k w_lr_l\left(1-\frac{\rho_2}{2}\right)^{k-l}\\ &\leq \frac{4C_2\tau}{1-\rho_2}\left(\sum_{k=0}^K w_kr_k\right)\left(\sum_{k=0}^\infty\left(1-\frac{\rho_2}{2}\right)^k\right)\\ &\leq \frac{8C_2\tau}{\rho_2(1-\rho_2)}\sum_{k=0}^K w_kr_k,\end{aligned}$$

hence

$$\sum_{k=0}^K\sum_{t=1}^\tau w_k\mathbf{E}\sigma_{2,k-t}^2 \overset{(86)}{\leq} \frac{2\tau\left(2+\rho_2\right)\sigma_{2,0}^2}{\rho_2} + \frac{8C_2\tau}{\rho_2(1-\rho_2)}\sum_{k=0}^K w_kr_k. \tag{87}$$

Plugging inequalities (80), (85) and (87) in (77) we obtain

$$\begin{aligned}3L\sum_{k=0}^K w_k\mathbf{E}\|e^k\|^2 &\leq 12\gamma^2L\tau\left(\hat{A}+\frac{2B_1'C_1}{\rho_1(1-\rho_1)}+\frac{2B_2'C_2}{\rho_2(1-\rho_2)}+\frac{4B_1'GC_2}{\rho_2(1-\rho_1)(1-\rho_2)}\right)\sum_{k=0}^K w_kr_k\\ &+\frac{6\gamma^2LB_1'\tau(2+\rho_1)}{\rho_1}\sigma_0^2+\frac{6\gamma^2\tau L(2+\rho_2)}{\rho_2}\left(\frac{2B_1'G}{1-\rho_1}+B_2'\right)\sigma_{2,0}^2\\ &+3\gamma^2\tau L\left(D_1'+\frac{2B_1'D_2}{\rho}\right)W_K.\end{aligned}$$

Taking into account that $\gamma \leq \frac{1}{4\sqrt{4L\tau\left(\hat{A}+\frac{2B_1'C_1}{\rho_1(1-\rho_1)}+\frac{2B_2'C_2}{\rho_2(1-\rho_2)}+\frac{4B_1'GC_2}{\rho_2(1-\rho_1)(1-\rho_2)}\right)}}$, $F_1 = \frac{6\gamma^2LB_1'\tau(2+\rho_1)}{\rho_1}$, $F_2 = \frac{6\gamma^2\tau L}{\rho_2}\left(\frac{2B_1'G(2+\rho_2)}{1-\rho_1}+B_2'\right)$ and $D_3 = 3\gamma\tau L\left(D_1'+\frac{2B_1'D_2}{\rho}\right)$ we get

$$3L\sum_{k=0}^K w_k\mathbf{E}\|e^k\|^2 \leq \frac{1}{4}\sum_{k=0}^K w_kr_k + F_1\sigma_{1,0}^2 + F_2\sigma_{2,0}^2 + \gamma D_3.$$

□

As a direct application of Lemma H.1 and Theorem 3.1 we get the following result.

**Theorem H.1.** *Let Assumptions 3.1 and 3.2 be satisfied, inequalities* (15), (16) *and* (17) *hold and*

$$\gamma \le \min\left\{\frac{1}{4(A'+C_1M_1+C_2M_2)}, \frac{1}{2\tau\mu}, \frac{1}{8\sqrt{L\tau\left(\hat{A}+\frac{2B_1'C_1}{\rho_1(1-\rho_1)}+\frac{2B_2'C_2}{\rho_2(1-\rho_2)}+\frac{4B_1'GC_2}{\rho_2(1-\rho_1)(1-\rho_2)}\right)}}\right\},$$

*where* $M_1 = \frac{4B_1'}{3\rho_1}$ *and* $M_2 = \frac{4\left(B_2'+\frac{4}{3}G\right)}{3\rho_2}$. *Then for all* $K \ge 0$ *we have*

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \le (1-\eta)^K\frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma} + 4\gamma\left(D_1'+MD_2+D_3\right)$$

*when* $\mu > 0$ *and*

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \le \frac{4(T^0+\gamma F_1\sigma_{1,0}^2+\gamma F_2\sigma_{2,0}^2)}{\gamma K} + 4\gamma\left(D_1'+MD_2+D_3\right)$$

*when* $\mu = 0$, *where* $\eta = \min\left\{\gamma\mu/2, \rho_1/4, \rho_2/4\right\}$, $T^k \overset{def}{=} \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$ *and*

$$F_1 = \frac{6\gamma^2 LB_1'\tau(2+\rho_1)}{\rho_1}, \quad F_2 = \frac{6\gamma^2\tau L(2+\rho_2)}{\rho_2}\left(\frac{2B_1'G}{1-\rho_1}+B_2'\right),$$

$$D_3 = 3\gamma\tau L\left(D_1'+\frac{2B_1'D_2}{\rho_1}\right).$$

**Algorithm 1** `DIANAsr` with Double Compression (`DIANAsr-DQ`)

**Input:** learning rates $\gamma > 0$, $\alpha \in (0, 1]$, initial vectors $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$

Set $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $g^{k-1}$ to all workers ▷ If $k = 0$, then broadcast $x^0$
  **for** $i = 1, \ldots, n$ in parallel **do**
    $x^k = x^{k-1} - \gamma g^{k-1}$ ▷ Ignore this line if $k = 0$
    Sample $g_i^{k,1} = \nabla f_{\xi_i^k}(x^k)$ satisfying Assumption I.1 independtently from other workers
    $\hat{\Delta}_i^k = g_i^{k,1} - h_i^k$
    Sample $\Delta_i^k \sim Q_1(\hat{\Delta}_i^k)$ indepently from other workers
    $g_i^{k,2} = h_i^k + \Delta_i^k$
    $h_i^{k+1} = h_i^k + \alpha\Delta_i^k$
  **end for**
  $g^{k,2} = \frac{1}{n}\sum_{i=1}^n g_i^{k,2} = h^k + \frac{1}{n}\sum_{i=1}^n \Delta_i^k$
  $h^{k+1} = \frac{1}{n}\sum\limits_{i=1}^n h_i^{k+1} = h^k + \alpha\frac{1}{n}\sum\limits_{i=1}^n \Delta_i^k$
  Sample $g^k \sim Q_2(g^{k,2})$
  $x^{k+1} = x^k - \gamma g^{k-1}$
**end for**

# I Special Cases: `SGD`

To illustrate the generality of our approach, we develop and analyse a new special case of `SGD` without error-feedback and show that in some cases, our framework recovers tighter rates than the framework from [11].

## I.1 `DIANA` with Arbitrary Sampling and Double Quantization

In this section we consider problem (1) with $f(x)$ being $\mu$-quasi strongly convex and $f_i(x)$ satisfying (3) where functions $f_{ij}(x)$ are differentiable, but not necessary convex. Following [13] we construct a stochastic reformulation of this problem:

$$f(x) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x) = \frac{1}{n}\sum_{i=1}^n f_{\xi_i}(x), \quad f_{\xi_i}(x) = \frac{1}{m}\sum_{j=1}^m \xi_{ij} f_{ij}(x), \tag{88}$$

where $\xi = (\xi_1^\top, \ldots, \xi_n^\top), \xi_i = (\xi_{i1}, \ldots, \xi_{im})^\top$ is a random vector with distribution $\mathcal{D}_i$ such that $\mathbf{E}_{\mathcal{D}_i}[\xi_{ij}] = 1$ for all $i \in [n], j \in [m]$ and the following assumption holds.

**Assumption I.1** (Expected smoothness)**.** *We assume that functions $f_1, \ldots, f_n$ are $\mathcal{L}$-smooth in expectation w.r.t. distributions $\mathcal{D}_1, \ldots, \mathcal{D}_n$, i.e., there exists constant $\mathcal{L} = \mathcal{L}(f, \mathcal{D}_1, \ldots, \mathcal{D}_n)$ such that*

$$\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right] \leq 2\mathcal{L}D_{f_i}(x, x^*) \tag{89}$$

*for all $i \in [n]$ and $x \in \mathbb{R}^d$.*

To solve this problem, we consider `DIANA` [37, 19] — a distributed stochastic method using unbiased compressions or *quantizations* for communication between workers and master. We start with the formal definition of quantization. In [37, 19] `DIANA` was analyzed under the assumption that stochastic gradients $g_i^k$ have uniformly bounded variances which is not very practical.

Therefore, we consider a slightly different method called `DIANAsr-DQ` which works with the stochastic reformulation (88) of problem (1)+(3), see Algorithm 1. Moreover, to illustrate the flexibility of our approach, we consider compression not only on the workers' side but also on the master side. To perform an update of `DIANAsr-DQ` master needs to gather quantized gradient differences $\Delta_i^k$ and the to broadcast quantized stochastic gradient $g^k$ to all workers. Clearly, in this case, only compressed vectors participate in communication.

In the concurrent work [40] the same method was independently proposed under the name of `Artemis`. However, our analysis is slightly more general: it is based on Assumption I.1 while in [40] authors assume $L$-cocoercivity of stochastic gradients almost surely. Next, a very similar approach was considered in [48], where authors present a method with error compensation on master and worker sides. Moreover, recently another method called `DORE` was developed in [35], which uses `DIANA`-trick on the worker side and error compensation on the master side. However, in these methods, compression operators are the same on both sides, despite the fact that gathering the information often costs much more than broadcasting. Therefore, the natural idea is in using different quantization for gathering and broadcasting, and it is what `DIANAsr-DQ` does. Moreover, we do not assume uniform boundedness of the second moment of the stochastic gradient like in [48], and we also do not assume uniform boundedness of the variance of the stochastic gradient like in [35]. Assumption I.1 is more natural and always holds for the problems (1)+(3) when $f_{ij}$ are convex and $L$-smooth for each $i \in [n]$, $j \in [m]$. In contrast, in the same setup, there exist such problems that the variance of the stochastic gradients is not uniformly upper bounded by any finite constant.

We assume that $Q_1$ and $Q_2$ satisfy (26) with parameters $\omega_1$ and $\omega_2$ respectively.

**Lemma I.1.** *Let Assumption I.1 be satisfied. Then, for all $k \geq 0$ we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k), \tag{90}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 2\mathcal{L}(1+\omega_2)\left(2+\frac{3\omega_1}{n}\right)\left(f(x^k)-f(x^*)\right) + \frac{3\omega_1(1+\omega_2)}{n}\sigma_k^2 + D_1', \tag{91}$$

*where* $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f(x^*)\|^2$ *and* $D_1' = \frac{(2+3\omega_1)(1+\omega_2)}{n^2}\sum\limits_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right]$.

*Proof.* First of all, we show inbiasedness of $g^k$:

$$\begin{aligned}\mathbf{E}\left[g^k \mid x^k\right] &\overset{(35),(26)}{=} \mathbf{E}\left[g^{k,2} \mid x^k\right] = h^k + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\Delta_i^k \mid x^k\right]\\ &\overset{(35),(26)}{=} h^k + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\hat{\Delta}_i^k \mid x^k\right]\\ &= h^k + \frac{1}{n}\sum_{i=1}^n\left(\nabla f_i(x^k) - h_i^k\right) = \nabla f(x^k).\end{aligned}$$

Next, to denote mathematical expectation w.r.t. the randomness coming from quantizations $Q_1$ and $Q_2$ at iteration $k$ we use $\mathbf{E}_{Q_1^k}[\cdot]$ and $\mathbf{E}_{Q_2^k}[\cdot]$ respectively. Using these notations and the definition of quantization we derive

$$\begin{aligned}\mathbf{E}_{Q_2^k}[\|g^k\|^2] &\overset{(34),(26)}{=} \|g^{k,1}\|^2 + \mathbf{E}_{Q_2^k}\left[\|g^{k,2}-g^{k,1}\|^2\right]\\ &\overset{(26)}{\leq} (1+\omega_2)\|g^{k,1}\|^2.\end{aligned}$$

Taking the conditopnal mathematical expectation $\mathbf{E}_{Q_1^k}[\cdot]$ from the both sides of previous inequality and using the independence of $\Delta_i^1, \ldots, \Delta_i^n$ we get

$$
\begin{aligned}
\mathbf{E}_{Q_1^k,Q_2^k}\left[\|g^k\|^2\right] &\overset{(35)}{=} (1+\omega_2)\mathbf{E}_{Q_1^k}\left[\|g^{k,1}\|^2\right] = (1+\omega_2)\mathbf{E}_{Q_1^k}\left[\left\|\frac{1}{n}\sum_{i=1}^n (h_i^k + \Delta_i^k)\right\|^2\right] \\
&\overset{(34)}{=} (1+\omega_2)\left\|\frac{1}{n}\sum_{i=1}^n \left(h_i^k + \hat{\Delta}_i^k\right)\right\|^2 + (1+\omega_2)\mathbf{E}_{Q_1^k}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\Delta_i^k - \hat{\Delta}_i^k)\right\|^2\right] \\
&= (1+\omega_2)\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*) + \nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \\
&\quad + \frac{(1+\omega_2)}{n^2}\sum_{i=1}^n \mathbf{E}_{Q_1^k}\left[\|\Delta_i^k - \hat{\Delta}_i^k\|^2\right] \\
&\overset{(31),(26)}{\leq} \frac{2(1+\omega_2)}{n}\sum_{i=1}^n \|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\|^2 \\
&\quad + 2(1+\omega_2)\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \\
&\quad + \frac{\omega_1(1+\omega_2)}{n^2}\sum_{i=1}^n \|\nabla f_{\xi_i^k}(x^k) - h_i^k\|^2 \\
&\overset{(31)}{\leq} \frac{2(1+\omega_2)}{n}\sum_{i=1}^n \|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\|^2 \\
&\quad + 2(1+\omega_2)\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \\
&\quad + \frac{3\omega_1(1+\omega_2)}{n^2}\sum_{i=1}^n \|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\|^2 \\
&\quad + \frac{3\omega_1(1+\omega_2)}{n^2}\sum_{i=1}^n \|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2 \\
&\quad + \frac{3\omega_1(1+\omega_2)}{n^2}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2.
\end{aligned}
$$

Finally, we take conditional mathematical expectation $\mathbf{E}[\cdot \mid x^k]$ from the both sides of the inequality above and use the independece of $\xi_1^k, \ldots, \xi_n^k$:

$$
\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\overset{(89)}{\leq} 2\mathcal{L}(1+\omega_2)\left(2 + \frac{3\omega_1}{n}\right)(f(x^k) - f(x^*)) + \frac{3\omega_1(1+\omega_2)}{n}\sigma_k^2 \\
&\quad + 2(1+\omega_2)\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{3\omega_1(1+\omega_2)}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&= 2\mathcal{L}(1+\omega_2)\left(2 + \frac{3\omega_1}{n}\right)(f(x^k) - f(x^*)) + \frac{3\omega_1(1+\omega_2)}{n}\sigma_k^2 \\
&\quad + \frac{(1+\omega_2)(2+3\omega_1)}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

□

**Lemma I.2.** *Let $f_i$ be convex and $L$-smooth, Assumption I.1 holds and $\alpha \le 1/(\omega_1+1)$. Then, for all $k \ge 0$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \le (1-\alpha)\sigma_k^2 + 2\alpha(3\mathcal{L}+4L)(f(x^k)-f(x^*)) + D_2, \tag{92}$$

*where* $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$ *and* $D_2 = \frac{3\alpha}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right]$.

*Proof.* For simplicity, we introduce new notation: $h_i^* \stackrel{\text{def}}{=} \nabla f_i(x^*)$. Using this we derive an upper bound for the second moment of $h_i^{k+1} - h_i^*$:

$$\begin{aligned}
\mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|h_i^k - h_i^* + \alpha\Delta_i^k\right\|^2 \mid x^k\right] \\
&\stackrel{(26)}{=} \|h_i^k - h_i^*\|^2 + 2\alpha\langle h_i^k - h_i^*, \nabla f_i(x^k) - h_i^k\rangle + \alpha^2\mathbf{E}\left[\|\Delta_i^k\|^2 \mid x^k\right] \\
&\stackrel{(26),(35)}{\le} \|h_i^k - h_i^*\|^2 + 2\alpha\langle h_i^k - h_i^*, \nabla f_i(x^k) - h_i^k\rangle \\
&\quad + \alpha^2(\omega_1+1)\mathbf{E}\left[\|\nabla f_{\xi_i^k}(x^k) - h_i^k\|^2 \mid x^k\right].
\end{aligned}$$

Using variance decomposition (34) and $\alpha \le 1/(\omega_1+1)$ we get

$$\begin{aligned}
\alpha^2(\omega_1+1)\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^k) - h_i^k\|^2\right] &\stackrel{(34)}{=} \alpha^2(\omega_1+1)\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^k) - \nabla f_i(x^k)\|^2\right] \\
&\quad + \alpha^2(\omega_1+1)\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\stackrel{(31)}{\le} 3\alpha\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\|^2\right] \\
&\quad + 3\alpha\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\quad + 3\alpha\|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\
&\quad + \alpha\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\stackrel{(11),(89)}{\le} 6\alpha(\mathcal{L}+L)D_{f_i}(x^k, x^*) + \alpha\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\quad + 3\alpha\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right]
\end{aligned}$$

Putting all together we obtain

$$\begin{aligned}
\mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] &\le \|h_i^k - h_i^*\|^2 + \alpha\left\langle \nabla f_i(x^k) - h_i^k, f_i(x^k) + h_i^k - 2h_i^*\right\rangle \\
&\quad + 6\alpha(\mathcal{L}+L)D_{f_i}(x^k, x^*) + 3\alpha\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\stackrel{(28)}{=} \|h_i^k - h_i^*\|^2 + \alpha\|\nabla f_i(x^k) - h_i^*\|^2 - \alpha\|h_i^k - h_i^*\|^2 \\
&\quad + 6\alpha(\mathcal{L}+L)D_{f_i}(x^k, x^*) + 3\alpha\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\stackrel{(11)}{\le} (1-\alpha)\|h_i^k - h_i^*\|^2 + \alpha(6\mathcal{L}+8L)D_{f_i}(x^k, x^*) \\
&\quad + 3\alpha\mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}$$

Summing up the above inequality for $i = 1, \ldots, n$ we derive

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] &\le \frac{1-\alpha}{n}\sum_{i=1}^n \|h_i^k - h_i^*\|^2 + \alpha(6\mathcal{L}+8L)(f(x^k) - f(x^*)) \\
&\quad + \frac{3\alpha}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}$$

□

**Theorem I.1.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1,\ldots,n$, $f(x)$ is $\mu$-quasi strongly convex and Assumption I.1 holds. Then* `DIANAsr-DQ` *satisfies Assumption 3.4 with*

$$A' = \mathcal{L}(1+\omega_2)\left(2+\frac{3\omega_1}{n}\right), \quad B_1' = \frac{3\omega_1(1+\omega_2)}{n},$$

$$D_1' = \frac{(2+3\omega_1)(1+\omega_2)}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right],$$

$$\sigma_{1,k}^2 = \sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2, \quad B_2' = 0, \quad \sigma_{2,k}^2 \equiv 0, \quad \rho_1 = \alpha, \quad \rho_2 = 1,$$

$$C_1 = \alpha(3\mathcal{L}+4L), \quad C_2 = 0, \quad D_2 = \frac{3\alpha}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right],$$

$$G = 0, \quad F_1 = F_2 = 0, \quad D_3 = 0,$$

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma \le \frac{1}{4(1+\omega_2)\left(\mathcal{L}\left(2+\frac{15\omega_1}{n}\right)+\frac{16L\omega_1}{n}\right)}, \quad \alpha \le \frac{1}{\omega+1}, \quad M_1 = \frac{4\omega_1(1+\omega_2)}{n\alpha}, \quad M_2 = 0$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}\right\}\right)^K \frac{4T^0}{\gamma} + 4\gamma\left(D_1' + M_1 D_2\right),$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4T^0}{\gamma K} + 4\gamma\left(D_1' + M_1 D_2\right)$$

*when $\mu = 0$, where $T^k \overset{def}{=} \|x^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2$.*

In other words, if

$$\gamma = \frac{1}{4(1+\omega_2)\left(\mathcal{L}\left(2+\frac{15\omega_1}{n}\right)+\frac{16L\omega_1}{n}\right)}, \quad \alpha = \frac{1}{\omega+1}$$

and $D_1 = 0$, i.e., $\nabla f_{\xi_i^k}(x^k) = \nabla f_i(x^k)$ almost surely, `DIANAsr-DQ` converges with the linear rate

$$\mathcal{O}\left(\left(\omega_1 + \frac{\mathcal{L}}{\mu}(1+\omega_2)\left(1+\frac{\omega_1}{n}\right)\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary I.1.** *Let the assumptions of Theorem I.1 hold and $\mu > 0$. Then after $K$ iterations of* `DIANAsq-DQ` *with the stepsize*

$$\begin{aligned}
\gamma_0 &= \frac{1}{4(1+\omega_2)\left(\mathcal{L}\left(2+\frac{15\omega_1}{n}\right)+\frac{16L\omega_1}{n}\right)}\\
\gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \frac{\mu^2K^2(\|x^0-x^*\|^2+M_1\gamma_0^2\sigma_{1,0}^2)}{D_1'+M_1D_2}\right\}\right)}{\mu K}\right\}, \quad M_1 = \frac{4\omega_1(1+\omega_2)}{n\alpha}
\end{aligned}$$

*and $\alpha = \frac{1}{\omega+1}$ we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(A'\|x^0 - x^*\|^2 \exp\left(-\min\left\{\frac{\mu}{A'}, \frac{1}{\omega_1}\right\}K\right) + \frac{D_1' + M_1D_2}{\mu K}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `DIANAsq-DQ` *requires*

$$\widetilde{\mathcal{O}}\left(\omega_1 + \frac{\mathcal{L}\left(1+\frac{\omega_1}{n}\right)(1+\omega_2)}{\mu} + \frac{(1+\omega_1)(1+\omega_2)}{n^2\mu\varepsilon}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary I.2.** *Let the assumptions of Theorem I.1 hold and $\mu = 0$. Then after $K$ iterations of* `DIANAsq-DQ` *with the stepsize*

$$\begin{aligned}\gamma_0 &= \frac{1}{4(1+\omega_2)\left(\mathcal{L}\left(2+\frac{15\omega_1}{n}\right)+\frac{16L\omega_1}{n}\right)}\\ \gamma &= \min\left\{\gamma_0, \sqrt{\frac{\|x^0-x^*\|^2}{M_1\sigma_{1,0}^2}}, \sqrt{\frac{\|x^0-x^*\|^2}{(D_1'+M_1D_2)K}}\right\}, \quad M_1 = \frac{4\omega_1(1+\omega_2)}{n\alpha}\end{aligned}$$

*and $\alpha = \frac{1}{\omega+1}$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2(1+\omega_2)\left(1+\frac{\omega_1}{n}\right)}{K} + \frac{R_0\sigma_{1,0}(1+\omega_1)\sqrt{1+\omega_2}}{\sqrt{n}K} + \frac{R_0\sqrt{(1+\omega_1)(1+\omega_2)D_{opt}}}{\sqrt{nK}}\right)$$

*where $R_0 = \|x^0 - x^*\|^2, D_{opt} = \frac{1}{n}\sum\limits_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2$. That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `DIANAsq-DQ` *requires*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2(1+\omega_2)\left(1+\frac{\omega_1}{n}\right)}{\varepsilon} + \frac{R_0\sigma_{1,0}(1+\omega_1)\sqrt{1+\omega_2}}{\sqrt{n}\varepsilon} + \frac{R_0^2(1+\omega_1)(1+\omega_2)D_{opt}}{n\varepsilon^2}\right)$$

*iterations.*

### I.2 Recovering Tight Complexity Bounds for `VR-DIANA`

In this section we consider the same problem (1)+(3) and variance reduced version of `DIANA` called `VR-DIANA` [19], see Algorithm 2. For simplicity we assume that each $f_{ij}$ is convex and $L$-smooth and $f_i$ is additionally $\mu$-strongly convex.

**Lemma I.3** (Lemmas 3, 5, 6 and 7 from [19])**.** *Let $\alpha \le \frac{1}{\omega+1}$. Then for all iterates $k \ge 0$ of Algorithm 2 the following inequalities hold:*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k), \tag{93}$$

$$\mathbf{E}\left[H^{k+1} \mid x^k\right] \le (1-\alpha)H^k + \frac{2\alpha}{m}D^k + 8\alpha Ln\left(f(x^k) - f(x^*)\right), \tag{94}$$

$$\mathbf{E}\left[D^{k+1} \mid x^k\right] \le \left(1 - \frac{1}{m}\right)D^k + 2Ln\left(f(x^k) - f(x^*)\right), \tag{95}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 2L\left(1 + \frac{4\omega+2}{n}\right)\left(f(x^k) - f(x^*)\right) + \frac{2\omega}{n^2}\frac{D^k}{m} + \frac{2(\omega+1)}{n^2}H^k, \tag{96}$$

*where $H^k = \sum\limits_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$ and $D^k = \sum\limits_{i=1}^n\sum\limits_{j=1}^m \|\nabla f_{ij}(w_{ij}^k) - \nabla f_{ij}(x^*)\|^2$.*

This lemma shows that `VR-DIANA` satisfies (15), (16) and (17). Applying Theorem F.1 we get the following result.

**Theorem I.2.** *Assume that $f_{ij}(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$ and $f_i(x)$ is $\mu$-strongly convex for all $i = 1, \ldots, n$. Then* `VR-DIANA` *satisfies Assumption 3.4 with*

$$A' = L\left(1 + \frac{4\omega+2}{n}\right), \quad B_1' = \frac{2(\omega+1)}{n}, \quad D_1' = 0,$$

$$\sigma_{1,k}^2 = H^k = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2, \quad B_2' = \frac{2\omega}{n},$$

$$\sigma_{2,k}^2 = D^k = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_{ij}^k) - \nabla f_{ij}(x^*)\|^2, \quad \rho_1 = \alpha, \quad \rho_2 = \frac{1}{m},$$

$$C_1 = 4\alpha L, \quad C_2 = \frac{L}{m}, \quad D_2 = 0, \quad G = 2, \quad F_1 = F_2 = 0, \quad D_3 = 0,$$

**Algorithm 2** `VR-DIANA` based on `LSVRG` (Variant 1), `SAGA` (Variant 2), [19]

**Input:** learning rates $\alpha > 0$ and $\gamma > 0$, initial vectors $x^0, h_1^0, \dots, h_n^0$, $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$

**for** $k = 0, 1, \dots$ **do**
  Sample random $u^k = \begin{cases} 1, & \text{with probability } \frac{1}{m} \\ 0, & \text{with probability } 1 - \frac{1}{m} \end{cases}$ ▷ only for Variant 1
  Broadcast $x^k$, $u^k$ to all workers
  **for** $i = 1, \dots, n$ in parallel **do** ▷ Worker side
    Pick $j_i^k$ uniformly at random from $[m]$
    $\mu_i^k = \frac{1}{m}\sum\limits_{j=1}^m \nabla f_{ij}(w_{ij}^k)$
    $g_i^k = \nabla f_{ij_i^k}(x^k) - \nabla f_{ij_i^k}(w_{ij_i^k}^k) + \mu_i^k$
    $\hat{\Delta}_i^k = Q(g_i^k - h_i^k)$
    $h_i^{k+1} = h_i^k + \alpha\hat{\Delta}_i^k$
    **for** $j = 1, \dots, m$ **do**
      $w_{ij}^{k+1} = \begin{cases} x^k, & \text{if } u^k = 1 \\ w_{ij}^k, & \text{if } u^k = 0 \end{cases}$ ▷ Variant 1 (L-SVRG): update epoch gradient if $u^k = 1$
      $w_{ij}^{k+1} = \begin{cases} x^k, & j = j_i^k \\ w_{ij}^k, & j \neq j_i^k \end{cases}$ ▷ Variant 2 (SAGA): update gradient table
    **end for**
  **end for**
  $h^{k+1} = h^k + \frac{\alpha}{n}\sum\limits_{i=1}^n \hat{\Delta}_i^k$ ▷ Gather quantized updates
  $g^k = \frac{1}{n}\sum\limits_{i=1}^n (\hat{\Delta}_i^k + h_i^k)$
  $x^{k+1} = x^k - \gamma g^k$
**end for**

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma \le \frac{3}{L\left(\frac{41}{3} + \frac{52\omega+35}{n}\right)}, \quad \alpha \le \frac{1}{\omega+1}, \quad M_1 = \frac{8(\omega+1)}{3n\alpha}, \quad M_2 = \frac{8\omega m}{3n} + \frac{32m}{9}$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}, \frac{1}{4m}\right\}\right)^K \frac{4T^0}{\gamma},$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4T^0}{\gamma K}$$

*when $\mu = 0$, where $T^k \stackrel{def}{=} \|x^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$.*

In other words, if $\mu > 0$ and

$$\gamma = \frac{3}{L\left(\frac{41}{3} + \frac{52\omega+35}{n}\right)}, \quad \alpha = \frac{1}{\omega+1},$$

then `VR-DIANA` converges with the linear rate

$$\mathcal{O}\left(\left(\omega + m + \kappa\left(1 + \frac{\omega}{n}\right)\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution which coincides with the rate obtained in [19]. We notice that the framework from [11] establishes slightly worse guarantee:

$$\mathcal{O}\left(\left(\omega + m + \kappa\left(1 + \frac{\omega}{n}\right)\frac{\max\{m, \omega+1\}}{m}\right)\ln\frac{1}{\varepsilon}\right)$$

This guarantee is strictly worse than our bound when $m \leq 1+\omega$. The key tool that helps us to improve the rate is two sequences of $\{\sigma_{1,k}^2\}_{k\geq 0}$, $\{\sigma_{2,k}^2\}_{k\geq 0}$ instead of one sequence $\{\sigma_k^2\}_{k\geq 0}$ as in [11].

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary I.3.** *Let the assumptions of Theorem I.2 hold and $\mu = 0$. Then after $K$ iterations of* `VR-DIANA` *with the stepsize*

$$\begin{aligned}
\gamma_0 &= \frac{3}{L\left(\frac{41}{3} + \frac{52\omega+35}{n}\right)} \\
\gamma &= \min\left\{\gamma_0, \sqrt{\frac{\|x^0 - x^*\|^2}{M_1\sigma_{1,0}^2 + M_2\sigma_{2,0}^2}}\right\}, \quad M_1 = \frac{8(\omega+1)}{3n\alpha}, \quad M_2 = \frac{8\omega m}{3n} + \frac{32m}{9}
\end{aligned}$$

*and $\alpha = \frac{1}{\omega+1}$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}\right)}{K} + \frac{R_0\sqrt{\frac{(1+\omega)^2}{n}\sigma_{1,0}^2 + \left(1+\frac{\omega}{n}\right)m\sigma_{2,0}^2}}{K}\right)$$

*where $R_0 = \|x^0 - x^*\|^2$. That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* `VR-DIANA` *requires*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}\right)}{\varepsilon} + \frac{R_0\sqrt{\frac{(1+\omega)^2}{n}\sigma_{1,0}^2 + \left(1+\frac{\omega}{n}\right)m\sigma_{2,0}^2}}{\varepsilon}\right)$$

*iterations.*

**Algorithm 3** `EC-SGDsr`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^k$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Sample $g_i^k = \nabla f_{\xi_i}(x^k)$
    $v_i^k = C(e_i^k + \gamma g_i^k)$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$
  $x^{k+1} = x^k - v^k$
**end for**

# J Special Cases: Error Compensated Methods

## J.1 `EC-SGDsr`

In this section we consider the same setup as in Section I.1 and assume additionally that $f_1, \ldots, f_n$ are $L$-smooth.

**Lemma J.1.** *For all $k \geq 0$ we have*

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2 \mid x^k\right] &\leq 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \\
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &\leq 6(\mathcal{L} + L)\left(f(x^k) - f(x^*)\right) + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \\
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\leq 4\mathcal{L}\left(f(x^k) - f(x^*)\right) + \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

*Proof.* Applying straightforward inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for $a, b \in \mathbb{R}^d$ we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*) + \nabla f_i(x^*)\|^2 \\
&\overset{(31)}{\leq} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \\
&\overset{(11)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2. \qquad (97)
\end{aligned}
$$

Similarly we obtain

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^k) - \nabla f_i(x^k)\|^2\right] \\
&\overset{(31)}{\leq} \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^k) - \nabla f_{\xi_i}(x^*)\|^2\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - \nabla f_i(x^k)\|^2 \\
&\overset{(11),(89)}{\leq} 6(\mathcal{L}+L)\left(f(x^k) - f(x^*)\right) \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

Next, using the independence of $\xi_1^k, \ldots, \xi_n^k$ we derive

$$
\begin{aligned}
\mathbf{E}\left[\left\|g^k\right\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*) + \nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\right\|^2 \mid x^k\right] \\
&\quad + 2\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(89)}{\leq} 4\mathcal{L}\left(f(x^k) - f(x^*)\right) + \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

□

Applying Theorem G.1 we get the following result.

**Theorem J.1.** *Assume that $f(x)$ is $\mu$-quasi strongly convex, $f_1, \ldots, f_n$ are $L$-smooth and Assumption I.1 holds. Then* `EC-SGDsr` *satisfies Assumption 3.3 with*

$$
A = 2L, \quad \widetilde{A} = 3(\mathcal{L}+L), \quad A' = 2\mathcal{L}, \quad B_1 = \widetilde{B}_1 = B_1' = B_2 = \widetilde{B}_2 = B_2' = 0,
$$

$$
D_1 = \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \quad \widetilde{D}_1 = \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0,
$$

$$
D_1' = \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \quad \rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad G = 0, \quad D_2 = 0,
$$

$$
F_1 = F_2 = 0, \quad D_3 = \frac{6L\gamma}{\delta}\left(\frac{2D_1}{\delta} + \widetilde{D}_1\right),
$$

*with $\gamma$ satisfying*

$$
\gamma \leq \min\left\{\frac{1}{8\mathcal{L}}, \frac{\delta}{4\sqrt{6L\left(4L + 3\delta(\mathcal{L}+L)\right)}}\right\}
$$

*and for all $K \geq 0$*

$$
\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + 4\gamma\left(D_1' + \frac{12L\gamma}{\delta^2}D_1 + \frac{6L\gamma}{\delta}\widetilde{D}_1\right)
$$

*when* $\mu > 0$ *and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4\|x^0 - x^*\|^2}{K\gamma} + 4\gamma\left(D_1' + \frac{12L\gamma}{\delta^2}D_1 + \frac{6L\gamma}{\delta}\widetilde{D}_1\right)$$

*when* $\mu = 0$.

In other words, EC-SGDsr converges with linear rate $\mathcal{O}\left(\left(\frac{\mathcal{L}}{\mu} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\mu\delta}\right)\ln\frac{1}{\varepsilon}\right)$ to the neighbourhood of the solution when $\mu > 0$. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary J.1.** *Let the assumptions of Theorem J.1 hold and* $\mu > 0$*. Then after* $K$ *iterations of* EC-SGDsr *with the stepsize*

$$\gamma = \min\left\{\frac{1}{8\mathcal{L}}, \frac{\delta}{4\sqrt{6L\left(4L + 3\delta(\mathcal{L}+L)\right)}}, \frac{\ln\left(\max\left\{2, \min\left\{\frac{\|x^0-x^*\|^2\mu^2K^2}{D_1'}, \frac{\delta\|x^0-x^*\|^2\mu^3K^3}{6L(2D_1/\delta + \widetilde{D}_1)}\right\}\right\}\right)}{\mu K}\right\}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\widetilde{\mathcal{O}}\left(\left(\mathcal{L} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta}\right)\|x^0 - x^*\|^2\exp\left(-\frac{\mu}{\mathcal{L} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta}}K\right) + \frac{D_1'}{\mu K} + \frac{L(\widetilde{D}_1 + D_1/\delta)}{\delta\mu^2K^2}\right).$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$ EC-SGDsr *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\mathcal{L}}{\mu} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta\mu} + \frac{D_1'}{\mu\varepsilon} + \frac{\sqrt{L(\widetilde{D}_1 + D_1/\delta)}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary J.2.** *Let the assumptions of Theorem J.1 hold and* $\mu = 0$*. Then after* $K$ *iterations of* EC-SGDsr *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{8\mathcal{L}}, \frac{\delta}{4\sqrt{6L\left(4L + 3\delta(\mathcal{L}+L)\right)}}\right\}$$

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{\|x^0 - x^*\|^2}{D_1'K}}, \sqrt[3]{\frac{\|x^0 - x^*\|^2\delta}{6L(2D_1/\delta + \widetilde{D}_1)K}}\right\}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{R_0^2\left(\mathcal{L} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta}\right)}{K} + \sqrt{\frac{R_0^2D_1'}{K}} + \frac{\sqrt[3]{LR_0^4(2D_1/\delta + \widetilde{D}_1)}}{(\delta K^2)^{1/3}}\right)$$

*where* $R_0 = \|x^0 - x^*\|^2$*. That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$ EC-SGDsr *requires*

$$\mathcal{O}\left(\frac{R_0^2\left(\mathcal{L} + \frac{L+\sqrt{\delta L\mathcal{L}}}{\delta}\right)}{\varepsilon} + \frac{R_0^2D_1'}{\varepsilon^2} + \frac{R_0^2\sqrt{L(2D_1/\delta + \widetilde{D}_1)}}{\sqrt{\delta\varepsilon^3}}\right)$$

*iterations.*

### J.2 EC-SGD

In this section we consider problem (1) with $f_i(x)$ satisfying (2) where functions $f_{\xi_i}(x)$ are differentiable and $L$-smooth almost surely in $\xi_i$, $i = 1, \ldots, n$.

**Algorithm 4** `EC-SGD`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^k$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Sample $g_i^k = \nabla f_{\xi_i}(x^k)$ independently from other workers
    $v_i^k = C(e_i^k + \gamma g_i^k)$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$
  $x^{k+1} = x^k - v^k$
**end for**

**Lemma J.2** (See also Lemmas 1,2 from [39])**.** *Assume that $f_{\xi_i}(x)$ are convex in $x$ for every $\xi_i$, $i = 1, \ldots, n$. Then for every $x \in \mathbb{R}^d$ and $i = 1, \ldots, n$*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)\|^2 &\leq 4L\left(f(x) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \\
\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla f_{\xi_i}(x) - \nabla f_i(x)\|^2 &\leq 12L\left(f(x) - f(x^*)\right) + \frac{3}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \\
\mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x)\right\|^2 &\leq 4L\left(f(x) - f(x^*)\right) + \frac{2}{n^2}\sum \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}$$

*If further $f(x)$ is $\mu$-strongly convex with $\mu > 0$ and possibly non-convex $f_i, f_{\xi_i}$, then for every $x \in \mathbb{R}^d$ and $i = 1, \ldots, n$*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)\|^2 &\leq 4L\kappa\left(f(x) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \\
\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla f_{\xi_i}(x) - \nabla f_i(x)\|^2 &\leq 12L\kappa\left(f(x) - f(x^*)\right) + \frac{3}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \\
\mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x)\right\|^2 &\leq 4L\kappa\left(f(x) - f(x^*)\right) + \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}$$

*where $\kappa = \frac{L}{\mu}$.*

*Proof.* We start with the case when functions $f_{\xi_i}(x)$ are convex in $x$ for every $\xi_i$. The first inequality follows from (97). Next, we derive

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla f_{\xi_i}(x) - \nabla f_i(x)\|^2 &\overset{(31)}{\leq} \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2 \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2 \\
&\quad + \frac{3}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - \nabla f_i(x)\|^2 \\
&\overset{(11)}{\leq} 12L\left(f(x) - f(x^*)\right) + \frac{3}{n}\sum_{i=1}^n \mathbf{E}\|\nabla f_{\xi_i}(x^*)\|^2.
\end{aligned}$$

Due to independence of $\xi_1^k, \ldots, \xi_n^k$ we get

$$
\begin{aligned}
\mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x)\right\|^2 &= \mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*) + \nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right] \\
&\quad + 2\mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \\
&\overset{(11)}{\leq} 4L\left(f(x) - f(x^*)\right) + \frac{2}{n^2}\sum \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

Next, we consider the second case: $f(x)$ is $\mu$-strongly convex with possibly non-convex $f_i, f_{\xi_i}$. In this case

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x)\|^2 &\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \\
&\overset{(10)}{\leq} 2L^2\|x - x^*\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \\
&\leq \frac{4L^2}{\mu}\left(f(x) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2
\end{aligned}
$$

where the last inequality follows from $\mu$-strong convexity of $f$. Similarly, we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_i(x)\|^2\right] &\overset{(31)}{\leq} \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - \nabla f_i(x)\|^2 \\
&\overset{(10)}{\leq} 6L^2\|x - x^*\|^2 \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\leq \frac{12L^2}{\mu}\left(f(x) - f(x^*)\right) \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

Finally, using independence of $\xi_1^k, \ldots, \xi_n^k$ we derive

$$
\begin{aligned}
\mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x)\right\|^2 &= \mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*) + \nabla f_{\xi_i}(x^*) - \nabla f_i(x^*))\right\|^2 \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right] \\
&\quad + 2\mathbf{E}_{\xi_1,\ldots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*))\right\|^2 \\
&\overset{(10)}{\leq} 2L^2\|x - x^*\|^2 + \frac{2}{n^2}\sum \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\leq \frac{4L^2}{\mu}(f(x) - f(x^*)) + \frac{2}{n^2}\sum \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}
$$

□

Applying Theorem G.1 we get the following result.

**Theorem J.2.** *Assume that $f_\xi(x)$ is convex and $L$-smooth in $x$ for every $\xi$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `EC-SGD` *satisfies Assumption 3.3 with*

$$A = A' = 2L, \quad \widetilde{A} = 6L, \quad B_1 = \widetilde{B}_1 = B_1' = B_2 = \widetilde{B}_2 = B_2' = 0,$$

$$D_1 = \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \quad \widetilde{D}_1 = \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0,$$

$$D_1' = \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \quad \rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad G = 0, \quad D_2 = 0,$$

$$F_1 = F_2 = 0, \quad D_3 = \frac{6L\gamma}{\delta}\left(\frac{2D_1}{\delta} + \widetilde{D}_1\right),$$

*with $\gamma$ satisfying*

$$\gamma \leq \frac{\delta}{8L\sqrt{6 + 9\delta}}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + 4\gamma\left(D_1' + \frac{12L\gamma}{\delta^2}D_1 + \frac{6L\gamma}{\delta}\widetilde{D}_1\right)$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4\|x^0 - x^*\|^2}{K\gamma} + 4\gamma\left(D_1' + \frac{12L\gamma}{\delta^2}D_1 + \frac{6L\gamma}{\delta}\widetilde{D}_1\right)$$

*when $\mu = 0$. If further $f(x)$ is $\mu$-strongly convex with $\mu > 0$ and possibly non-convex $f_i, f_{\xi_i}$, then* `EC-SGD` *satisfies Assumption 3.3 with*

$$A = A' = 2L\kappa, \quad \widetilde{A} = 6L\kappa, \quad B_1 = \widetilde{B}_1 = B_1' = B_2 = \widetilde{B}_2 = B_2' = 0,$$

$$D_1 = \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \quad \widetilde{D}_1 = \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0,$$

$$D_1' = \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right], \quad \rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad G = 0, \quad D_2 = 0,$$

$$F_1 = F_2 = 0, \quad D_3 = \frac{6L\gamma}{\delta}\left(\frac{2D_1}{\delta} + \widetilde{D}_1\right),$$

*with $\gamma$ satisfying*

$$\gamma \leq \min\left\{\frac{1}{8\kappa L}, \frac{\delta}{8L\sqrt{3\kappa(2+3\delta)}}\right\}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + 4\gamma\left(D_1' + \frac{12L\gamma}{\delta^2}D_1 + \frac{6L\gamma}{\delta}\widetilde{D}_1\right).$$

In other words, EC-SGD converges with linear rate $\mathcal{O}\left(\frac{\kappa}{\delta}\ln\frac{1}{\varepsilon}\right)$ to the neighbourhood of the solution when $f_\xi(x)$ are convex for each $\xi$ and $\mu > 0$. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary J.3.** *Let the assumptions of Theorem J.2 hold, $f_\xi(x)$ are convex for each $\xi$ and $\mu > 0$. Then after $K$ iterations of* EC-SGD *with the stepsize*

$$\gamma = \min\left\{\frac{\delta}{8L\sqrt{6+9\delta}}, \frac{\ln\left(\max\left\{2, \min\left\{\frac{\|x^0-x^*\|^2\mu^2K^2}{D_1'}, \frac{\delta\|x^0-x^*\|^2\mu^3K^3}{6L(2D_1/\delta+\widetilde{D}_1)}\right\}\right\}\right)}{\mu K}\right\}$$

*we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(\frac{L}{\delta}\|x^0 - x^*\|^2\exp\left(-\frac{\delta\mu}{L}K\right) + \frac{D_1'}{\mu K} + \frac{L(\widetilde{D}_1 + D_1/\delta)}{\delta\mu^2K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* EC-SGD *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\delta\mu} + \frac{D_1'}{\mu\varepsilon} + \frac{\sqrt{L(\widetilde{D}_1 + D_1/\delta)}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

**Corollary J.4.** *Let the assumptions of Theorem J.2 hold and $f(x)$ is $\mu$-strongly convex with $\mu > 0$ and possibly non-convex $f_i, f_{\xi_i}$. Then after $K$ iterations of* EC-SGD *with the stepsize*

$$\gamma = \min\left\{\frac{1}{8\kappa L}, \frac{\delta}{8L\sqrt{3\kappa(2+3\delta)}}, \frac{\ln\left(\max\left\{2, \min\left\{\frac{\|x^0-x^*\|^2\mu^2K^2}{D_1'}, \frac{\delta\|x^0-x^*\|^2\mu^3K^3}{6L(2D_1/\delta+\widetilde{D}_1)}\right\}\right\}\right)}{\mu K}\right\}$$

*we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\widetilde{\mathcal{O}}\left(\left(L\kappa + \frac{L\sqrt{\kappa}}{\delta}\right)\|x^0 - x^*\|^2\exp\left(-\min\left\{\frac{\delta\mu}{L\sqrt{\kappa}}, \frac{1}{\kappa^2}\right\}K\right) + \frac{D_1'}{\mu K} + \frac{L(\widetilde{D}_1 + D_1/\delta)}{\delta\mu^2K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* EC-SGD *requires*

$$\widetilde{\mathcal{O}}\left(\kappa^2 + \frac{\kappa^{3/2}}{\delta} + \frac{D_1'}{\mu\varepsilon} + \frac{\sqrt{L(\widetilde{D}_1 + D_1/\delta)}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary J.5.** *Let the assumptions of Theorem J.2 hold, $f_\xi(x)$ are convex for each $\xi$ and $\mu = 0$. Then after $K$ iterations of* EC-SGD *with the stepsize*

$$\gamma = \min\left\{\frac{\delta}{8L\sqrt{6+9\delta}}, \sqrt{\frac{\|x^0 - x^*\|^2}{D_1'K}}, \sqrt[3]{\frac{\|x^0 - x^*\|^2\delta}{6L(2D_1/\delta + \widetilde{D}_1)K}}\right\}$$

*we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta K} + \sqrt{\frac{R_0^2D_1'}{K}} + \frac{\sqrt[3]{LR_0^4(2D_1/\delta + \widetilde{D}_1)}}{(\delta K^2)^{1/3}}\right)$$

**Algorithm 5** `EC-GDstar` (see also [11])

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
    Broadcast $x^k$ to all workers
    **for** $i = 1, \ldots, n$ in parallel **do**
        $g_i^k = \nabla f_i(x^k) - \nabla f_i(x^*)$
        $v_i^k = C(e_i^k + \gamma g_i^k)$
        $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    **end for**
    $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$
    $x^{k+1} = x^k - v^k$
**end for**

*where $R_0 = \|x^0 - x^*\|^2$. That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ `EC-SGD` requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon} + \frac{R_0^2 D_1'}{\varepsilon^2} + \frac{R_0^2\sqrt{L(2D_1/\delta + \widetilde{D}_1)}}{\sqrt{\delta\varepsilon^3}}\right)$$

*iterations.*

### J.3 `EC-GDstar`

We assume that $i$-th node has access to the gradient of $f_i$ at the optimality, i.e., to the $\nabla f_i(x^*)$. It is unrealistic scenario but it gives some insights that we will use next in order to design the method that converges asymptotically to *the exact solution*.

Assume that $f(x)$ is $\mu$-quasi strongly convex and each $f_i$ is convex and $L$-smooth. By definition of $g_i^k$ it trivially follows that

$$g^k = \frac{1}{n}\sum_{i=1}^n g_i^k = \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x^k) - \nabla f_i(x^*)\right) = \nabla f(x^k) - \nabla f(x^*) = \nabla f(x^k),$$

$g_i^k = \bar{g}_i^k$, and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \|g_i^k\|^2 &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\ &\overset{(11)}{\le} \frac{2L}{n}\sum_{i=1}^n \left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^*\rangle\right) = 2L\left(f(x^k) - f(x^*)\right), \\ \|g^k\|^2 &= \|\nabla f(x^k)\|^2 \overset{(11)}{\le} 2L\left(f(x^k) - f(x^*)\right).\end{aligned}$$

Applying Theorem G.1 we get the following result.

**Theorem J.3.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$ and $f(x)$ is $\mu$-quasi strongly convex. Then `EC-GDstar` satisfies Assumption 3.3 with*

$$A = A' = L, \quad \widetilde{A} = 0, \quad B_1 = B_2 = \widetilde{B}_1 = \widetilde{B}_2 = B_1' = B_2' = 0,$$
$$D_1 = \widetilde{D}_1 = D_1' = 0, \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0,$$
$$\rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad G = 0, \quad D_2 = 0, \quad F_1 = F_2 = 0, \quad D_3 = 0,$$

*with $\gamma$ satisfying*

$$\gamma \le \frac{\delta}{8L\sqrt{3}}$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma},$$

**Algorithm 6** `EC-SGD-DIANA`

**Input:** learning rates $\gamma > 0$, $\alpha \in (0,1]$, initial vectors $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1,\ldots,n$

Set $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$

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

  Broadcast $x^k, h^k$ to all workers

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

    Sample $\hat{g}_i^k$ such that $\mathbf{E}[\hat{g}_i^k \mid x^k] = \nabla f_i(x^k)$ and $\mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \le \widetilde{D}_{1,i}$ independently from other workers

    $g_i^k = \hat{g}_i^k - h_i^k + h^k$

    $v_i^k = C(e_i^k + \gamma g_i^k)$

    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$

    $h_i^{k+1} = h_i^k + \alpha Q(\hat{g}_i^k - h_i^k)$

  **end for**

  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$, $h^{k+1} = \frac{1}{n}\sum\limits_{i=1}^n h_i^{k+1} = h^k + \alpha\frac{1}{n}\sum\limits_{i=1}^n Q(\hat{g}_i^k - h_i^k)$

  $x^{k+1} = x^k - v^k$

**end for**

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4\|x^0 - x^*\|^2}{K\gamma}$$

*when $\mu = 0$.*

In other words, `EC-GDstar` converges with linear rate $\mathcal{O}\left(\frac{\kappa}{\delta}\ln\frac{1}{\varepsilon}\right)$ to the exact solution when $\mu > 0$ removing the drawback of `EC-SGD` and `EC-GD`. If $\mu = 0$ then the rate of convergence is $\mathcal{O}\left(\frac{L\|x^0 - x^*\|^2}{\delta\varepsilon}\right)$. However, `EC-GDstar` relies on the fact that $i$-th node knows $\nabla f_i(x^*)$ which is not realistic.

### J.4 `EC-SGD-DIANA`

In this section we present a new method that converges to the exact optimum asymptotically but does not need to know $\nabla f_i(x^*)$ and instead of this it learns the gradients at the optimum. This method is inspired by another method called `DIANA` (see [37, 19]).

We notice that master needs to gather only $C(e_i^k + \gamma g_i^k)$ and $Q(\hat{g}_i^k - h_i^k)$ from all nodes in order to perform an update.

**Lemma J.3.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1,\ldots,n$. Then, for all $k \ge 0$ we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k), \tag{98}$$

$$\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 \le 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2, \tag{99}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \le \widetilde{D}_1, \tag{100}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 2L\left(f(x^k) - f(x^*)\right) + \frac{\widetilde{D}_1}{n} \tag{101}$$

*where $\widetilde{D}_1 = \frac{1}{n}\sum_{i=1}^n \widetilde{D}_{1,i}$ and $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f(x^*)\|^2$.*

*Proof.* First of all, we show unbiasedness of $g^k$:

$$\mathbf{E}\left[g^k \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[g_i^k \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x^k) - h_i^k + h^k\right) = \nabla f(x^k).$$

Next, we derive the upper bound for $\|\bar{g}_i^k\|^2$:

$$\begin{aligned}
\|\bar{g}_i^k\|^2 &= \|\nabla f_i(x^k) - h_i^k - h^k\|^2 \\
&\overset{(31)}{\leq} 2\|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + 2\left\|h_i^k - \nabla f_i(x^*) - \left(h^k + \nabla f(x^*)\right)\right\|^2 \\
&\overset{(11)}{\leq} 4L\left(f_i(x^k) - \nabla f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^*\rangle\right) \\
&\quad +2\left\|h_i^k - \nabla f_i(x^*) - \left(h^k + \nabla f(x^*)\right)\right\|^2.
\end{aligned}$$

Summing up previous inequality for $i = 1, \ldots, n$ we get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 &\leq 4L(f(x^k) - f(x^*)) + \frac{2}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^*) - \left(\frac{1}{n}\sum_{i=1}^n (h_i^k - \nabla f_i(x^*))\right)\right\|^2 \\
&\overset{(34)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|h_i^k - \nabla f(x^*)\|^2. \qquad (102)
\end{aligned}$$

Using the unbiasedness of $\hat{g}_i^k$ we derive

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \leq \frac{1}{n}\sum_{i=1}^n \widetilde{D}_{1,i} = \widetilde{D}_1.$$

Finally, we obtain the upper bound for the second moment of $g^k$ using the independence of $\hat{g}_1^k, \ldots, \hat{g}_n^k$:

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\overset{(34)}{=} \|\nabla f(x^k)\|^2 + \mathbf{E}\left[\|g^k - \nabla f(x^k)\|^2\right] \\
&\overset{(11)}{\leq} 2L(f(x^k) - f(x^*)) + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\hat{g}_i^k - \nabla f_i(x^k))\right\|^2 \mid x^k\right] \\
&= 2L(f(x^k) - f(x^*)) + \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\left\|\hat{g}_i^k - \nabla f_i(x^k)\right\|^2 \mid x^k\right] \\
&\leq 2L(f(x^k) - f(x^*)) + \frac{1}{n^2}\sum_{i=1}^n \widetilde{D}_{1,i}.
\end{aligned}$$

□

**Lemma J.4.** *Let assumptions of Lemma J.3 hold and $\alpha \leq 1/(\omega+1)$. Then, for all $k \geq 0$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq (1-\alpha)\sigma_k^2 + 2L\alpha(f(x^k) - f(x^*)) + \alpha^2(\omega+1)\widetilde{D}_1, \qquad (103)$$

*where $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$ and $\widetilde{D}_1 = \frac{1}{n}\sum_{i=1}^n \widetilde{D}_{1,i}$.*

*Proof.* For simplicity, we introduce new notation: $h_i^* \overset{\text{def}}{=} \nabla f_i(x^*)$. Using this we derive an upper bound for the second moment of $h_i^{k+1} - h_i^*$:

$$\begin{aligned}
\mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|h_i^k - h_i^* + \alpha Q(\hat{g}_i^k - h_i^k)\right\|^2 \mid x^k\right] \\
&\overset{(26)}{=} \|h_i^k - h_i^*\|^2 + 2\alpha\langle h_i^k - h_i^*, \nabla f_i(x^k) - h_i^k\rangle \\
&\quad +\alpha^2\mathbf{E}\left[\|Q(\hat{g}_i^k - h_i^k)\|^2 \mid x^k\right] \\
&\overset{(26),(35)}{\leq} \|h_i^k - h_i^*\|^2 + 2\alpha\langle h_i^k - h_i^*, \nabla f_i(x^k) - h_i^k\rangle \\
&\quad +\alpha^2(\omega+1)\mathbf{E}\left[\|\hat{g}_i^k - h_i^k\|^2 \mid x^k\right].
\end{aligned}$$

Using variance decomposition (34) and $\alpha \leq 1/(\omega+1)$ we get

$$\begin{aligned}
\alpha^2(\omega+1)\mathbf{E}\left[\|\hat{g}_i^k - h_i^k\|^2 \mid x^k\right] &\overset{(34)}{=} \alpha^2(\omega+1)\mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \\
&\quad +\alpha^2(\omega+1)\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\leq \alpha^2(\omega+1)\widetilde{D}_{1,i} + \alpha\|\nabla f_i(x^k) - h_i^k\|^2.
\end{aligned}$$

Putting all together we obtain

$$\begin{aligned}
\mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] &\leq \|h_i^k - h_i^*\|^2 + \alpha\left\langle \nabla f_i(x^k) - h_i^k, f_i(x^k) + h_i^k - 2h_i^*\right\rangle + \alpha^2(\omega+1)\widetilde{D}_{1,i} \\
&\overset{(28)}{=} \|h_i^k - h_i^*\|^2 + \alpha\|\nabla f_i(x^k) - h_i^*\|^2 - \alpha\|h_i^k - h_i^*\|^2 + \alpha^2(\omega+1)\widetilde{D}_{1,i} \\
&\overset{(11)}{\leq} (1-\alpha)\|h_i^k - h_i^*\|^2 + 2L\alpha\left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^*\rangle\right) \\
&\quad +\alpha^2(\omega+1)\widetilde{D}_{1,i}.
\end{aligned}$$

Summing up the above inequality for $i = 1, \ldots, n$ we derive

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] \leq \frac{1-\alpha}{n}\sum_{i=1}^n \|h_i^k - h_i^*\|^2 + 2L\alpha(f(x^k) - f(x^*)) + \frac{\alpha^2(\omega+1)}{n}\sum_{i=1}^n \widetilde{D}_{1,i}.$$

□

Applying Theorem G.1 we get the following result.

**Theorem J.4.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `EC-SGD-DIANA` *satisfies Assumption 3.3 with*

$$A = 2L, \quad \widetilde{A} = 0, \quad A' = L, \quad B_1 = 2, \quad \widetilde{D}_1 = \frac{1}{n}\sum_{i=1}^n \widetilde{D}_{1,i}, \quad \sigma_{1,k}^2 = \sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2,$$

$$B_1' = B_2' = B_2 = \widetilde{B}_1 = \widetilde{B}_2 = 0, \quad \sigma_{2,k}^2 \equiv 0, \quad \rho_1 = \alpha, \quad \rho_2 = 1, \quad C_1 = L\alpha, \quad C_2 = 0, \quad D_1 = 0,$$

$$D_2 = \alpha^2(\omega+1)\widetilde{D}_1, \quad D_1' = \frac{D_1}{n}, \quad G = 0,$$

$$F_1 = \frac{96L\gamma^2}{\delta^2\alpha\left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}\right\}\right)}, \quad F_2 = 0, \quad D_3 = \frac{6L\gamma}{\delta}\left(\frac{4\alpha(\omega+1)}{\delta} + 1\right)\widetilde{D}_1,$$

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma \leq \min\left\{\frac{1}{4L}, \frac{\delta\sqrt{1-\alpha}}{8L\sqrt{6(3-\alpha)}}\right\}, \quad \alpha \leq \frac{1}{\omega+1}, \quad M_1 = M_2 = 0$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}\right\}\right)^K \frac{4(\|x^0 - x^*\|^2 + \gamma F_1\sigma_0^2)}{\gamma} + 4\gamma\left(D_1' + D_3\right),$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(\|x^0 - x^*\|^2 + \gamma F_1\sigma_0^2)}{\gamma K} + 4\gamma\left(D_1' + D_3\right)$$

*when $\mu = 0$.*

In other words, if

$$\gamma = \min\left\{\frac{1}{4L}, \frac{\delta\sqrt{1-\alpha}}{8L\sqrt{6(3-\alpha)}}\right\}, \quad \alpha = \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\}$$

and $\widetilde{D}_1 = 0$, i.e., $\hat{g}_i^k = \nabla f_i(x^k)$ almost surely (this is the setup of `EC-GD-DIANA`), `EC-SGD-DIANA` converges with the linear rate

$$\mathcal{O}\left(\left(\omega + \frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution in the case when $\mu > 0$.

**Corollary J.6.** *Let the assumptions of Theorem J.4 hold and $\mu > 0$. Then after $K$ iterations of* `EC-SGD-DIANA` *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{4L}, \frac{\delta\sqrt{1-\alpha}}{8L\sqrt{6(3-\alpha)}}\right\}, \quad R_0 = \|x^0 - x^*\|, \quad \tilde{F}_1 = \frac{784L\gamma^2}{7\delta^2\alpha},$$

$$\gamma = \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\frac{n\left(R_0^2+\tilde{F}_1\gamma_0\sigma_{1,0}^2\right)\mu^2K^2}{\widetilde{D}_1}, \frac{\delta\left(R_0^2+\tilde{F}_1\gamma_0\sigma_{1,0}^2\right)\mu^3K^3}{6L\widetilde{D}_1\left(4\alpha(\omega+1)/\delta+1\right)}\right\}\right\}\right)}{\mu K}\right\},$$

*and $\alpha \le \frac{1}{\omega+1}$ we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(\frac{L}{\delta}R_0^2\exp\left(-\min\left\{\frac{\delta\mu}{L}, \alpha\right\}K\right) + \frac{\widetilde{D}_1}{n\mu K} + \frac{L\widetilde{D}_1\left(\alpha(\omega+1)/\delta+1\right)}{\delta\mu^2K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `EC-SGD-DIANA` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\alpha} + \frac{L}{\delta\mu} + \frac{D_1}{n\mu\varepsilon} + \frac{\sqrt{L\widetilde{D}_1\left(\alpha(\omega+1)/\delta+1\right)}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

*In particular, if $\alpha = \frac{1}{\omega+1}$, then to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `EC-SGD-DIANA` *requires*

$$\widetilde{\mathcal{O}}\left(\omega + \frac{L}{\delta\mu} + \frac{\widetilde{D}_1}{n\mu\varepsilon} + \frac{\sqrt{L\widetilde{D}_1}}{\delta\mu\sqrt{\varepsilon}}\right) \text{ iterations,}$$

*and if $\alpha = \frac{\delta}{\omega+1}$, then to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `EC-SGD-DIANA` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\omega+1}{\delta} + \frac{L}{\delta\mu} + \frac{\widetilde{D}_1}{n\mu\varepsilon} + \frac{\sqrt{L\widetilde{D}_1}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary J.7.** *Let the assumptions of Theorem J.4 hold and $\mu = 0$. Then after $K$ iterations of* `EC-SGD-DIANA` *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{4L}, \frac{\delta\sqrt{1-\alpha}}{8L\sqrt{6(3-\alpha)}}\right\}, \quad R_0 = \|x^0 - x^*\|,$$

$$\gamma = \min\left\{\gamma_0, \sqrt[3]{\frac{R_0^2\delta^2\alpha\left(1-\min\left\{\frac{\gamma_0\mu}{2}, \frac{\alpha}{4}\right\}\right)}{96L\sigma_0^2}}, \sqrt{\frac{nR_0^2}{\widetilde{D}_1K}}, \sqrt[3]{\frac{\delta R_0^2}{6L\widetilde{D}_1\left(\frac{4\alpha(\omega+1)}{\delta}+1\right)K}}\right\},$$

*and $\alpha \le \frac{1}{\omega+1}$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta K} + \frac{\sqrt[3]{LR_0^4\sigma_0^2}}{K\sqrt[3]{\delta^2\alpha}} + \sqrt{\frac{R_0^2\widetilde{D}_1}{nK}} + \sqrt[3]{\frac{LR_0^4\widetilde{D}_1\left(\frac{\alpha(\omega+1)}{\delta}+1\right)}{\delta K^2}}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `EC-SGD-DIANA` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon} + \frac{\sqrt[3]{LR_0^4\sigma_0^2}}{\varepsilon\sqrt[3]{\delta^2\alpha}} + \frac{R_0^2\widetilde{D}_1}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\widetilde{D}_1\left(\frac{\alpha(\omega+1)}{\delta}+1\right)}}{\sqrt{\delta\varepsilon^3}}\right)$$

**Algorithm 7** `EC-SGDsr-DIANA`

**Input:** learning rates $\gamma > 0$, $\alpha \in (0,1]$, initial vectors $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$
Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
Set $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^k, h^k$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Sample $\hat{g}_i^k = \nabla f_{\xi_i^k}(x^k)$ satisfying Assumption I.1 independtently from other workers
    $g_i^k = \hat{g}_i^k - h_i^k + h^k$
    $v_i^k = C(e_i^k + \gamma g_i^k)$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    $h_i^{k+1} = h_i^k + \alpha Q(\hat{g}_i^k - h_i^k)$ ▷ $Q(\cdot)$ is calculated independtly from other workers
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$, $h^{k+1} = \frac{1}{n}\sum\limits_{i=1}^n h_i^{k+1} = h^k + \alpha\frac{1}{n}\sum\limits_{i=1}^n Q(\hat{g}_i^k - h_i^k)$
  $x^{k+1} = x^k - v^k$
**end for**

*iterations. In particular, if* $\alpha = \frac{1}{\omega+1}$, *then to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ `EC-SGD-DIANA` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon} + \frac{\sqrt[3]{LR_0^4(\omega+1)\sigma_0^2}}{\varepsilon\sqrt[3]{\delta^2}} + \frac{R_0^2\widetilde{D}_1}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\widetilde{D}_1}}{\delta\sqrt{\varepsilon^3}}\right) \text{ iterations,}$$

*and if* $\alpha = \frac{\delta}{\omega+1}$, *then to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ `EC-SGD-DIANA` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon} + \frac{\sqrt[3]{LR_0^4(\omega+1)\sigma_0^2}}{\delta\varepsilon} + \frac{R_0^2\widetilde{D}_1}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\widetilde{D}_1}}{\sqrt{\delta\varepsilon^3}}\right) \text{ iterations.}$$

### J.5 `EC-SGDsr-DIANA`

In this section we consider the same setup as in Section I.1 and consider `EC-SGD-DIANA` adjusted to this setup. The resulting algorithm is called `EC-SGDsr-DIANA`, see

**Lemma J.5.** *Let Assumption I.1 be satisfied and $f_i$ be convex and $L$-smooth for all $i \in [n]$. Then, for all $k \ge 0$ we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k), \tag{104}$$

$$\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 \le 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2, \tag{105}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \le 6(\mathcal{L} + L)\left(f(x^k) - f(x^*)\right) + \widetilde{D}_1, \tag{106}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 4\mathcal{L}\left(f(x^k) - f(x^*)\right) + D_1' \tag{107}$$

*where* $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f(x^*)\|^2$, $\widetilde{D}_1 = \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right]$ *and* $D_1' = \frac{2}{n^2}\sum\limits_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right]$.

*Proof.* First of all, we show unbiasedness of $g^k$:

$$\mathbf{E}\left[g^k \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[g_i^k \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x^k) - h_i^k + h^k\right) = \nabla f(x^k).$$

Following the same steps as in the proof of (102) we derive (105). Next, we establish (106):

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^k) - \nabla f_i(x^k)\|^2\right] \\
&\overset{(31)}{\leq} \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\|^2\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - \nabla f_i(x^k)\|^2 \\
&\overset{(11),(89)}{\leq} 6(\mathcal{L}+L)\left(f(x^k) - f(x^*)\right) \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}$$

Finally, we obtain the upper bound for the second moment of $g^k$ using the independence of $\xi_1^k, \ldots, \xi_n^k$:

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*) + \nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*))\right\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x^k) - \nabla f_{\xi_i^k}(x^*)\|^2 \mid x^k\right] \\
&\quad + 2\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*))\right\|^2 \mid x^k\right] \\
&\overset{(89)}{\leq} 4\mathcal{L}\left(f(x^k) - f(x^*)\right) + \frac{2}{n^2}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right].
\end{aligned}$$

□

**Lemma J.6.** *Let $f_i$ be convex and $L$-smooth, Assumption I.1 holds and $\alpha \leq 1/(\omega+1)$. Then, for all $k \geq 0$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq (1-\alpha)\sigma_k^2 + 2\alpha(3\mathcal{L}+4L)(f(x^k) - f(x^*)) + D_2, \tag{108}$$

*where $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$ and $D_2 = \alpha^2(\omega+1)\widetilde{D}_1$.*

*Proof.* The proof is identical to the proof of Lemma I.2 up to the following changes in the notation: $\omega_1 = \omega$, $\Delta_i^k = Q(\hat{g}_i^k - h_i^k)$ and $\hat{\Delta}_i^k = \hat{g}_i^k - h_i^k$. □

Applying Theorem G.1 we get the following result.

**Theorem J.5.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$, $f(x)$ is $\mu$-quasi strongly convex and Assumption I.1 holds. Then* `EC-SGDsr-DIANA` *satisfies Assumption 3.3 with*

$$A = 2L, \quad \widetilde{A} = 3(\mathcal{L}+L), \quad A' = 2\mathcal{L}, \quad B_1 = 2, \quad \widetilde{D}_1 = \frac{3}{n}\sum_{i=1}^n \mathbf{E}_{\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right],$$

$$\sigma_{1,k}^2 = \sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2, \quad D_1 = 0, \quad D_1' = \frac{2}{3n}\widetilde{D}_1, \quad D_2 = \alpha^2(\omega+1)\widetilde{D}_1$$

$$\widetilde{B}_1 = B_1' = B_2' = B_2 = \widetilde{B}_2 = 0, \quad \sigma_{2,k}^2 \equiv 0, \quad \rho_1 = \alpha, \quad \rho_2 = 1, \quad C_1 = 2\alpha(3\mathcal{L}+4L), \quad C_2 = 0,$$

$$G = 0, \quad F_1 = \frac{96L\gamma^2}{\delta^2\alpha\left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}\right\}\right)}, \quad F_2 = 0, \quad D_3 = \frac{6L\gamma}{\delta}\left(\frac{4\alpha(\omega+1)}{\delta} + 1\right)\widetilde{D}_1,$$

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma \leq \min\left\{\frac{1}{4\mathcal{L}}, \frac{\delta}{4\sqrt{6L\left(4L + 3\delta(\mathcal{L}+L) + \frac{16(3\mathcal{L}+4L)}{1-\alpha}\right)}}\right\}, \quad \alpha \leq \frac{1}{\omega+1}, \quad M_1 = M_2 = 0.$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}\right\}\right)^K \frac{4(\|x^0 - x^*\|^2 + \gamma F_1 \sigma_0^2)}{\gamma} + 4\gamma\left(D_1' + D_3\right),$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(\|x^0 - x^*\|^2 + \gamma F_1 \sigma_0^2)}{\gamma K} + 4\gamma\left(D_1' + D_3\right)$$

*when $\mu = 0$.*

Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution in the case when $\mu > 0$.

**Corollary J.8.** *Let the assumptions of Theorem J.5 hold and $\mu > 0$. Then after $K$ iterations of* `EC-SGDsr-DIANA` *with the stepsize*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{4\mathcal{L}}, \frac{\delta}{4\sqrt{6L\left(4L + 3\delta(\mathcal{L}+L) + \frac{16(3\mathcal{L}+4L)}{1-\alpha}\right)}}\right\}, \\
R_0 &= \|x^0 - x^*\|, \quad \tilde{F}_1 = \frac{96L\gamma_0^2}{\delta^2\alpha\left(1 - \min\left\{\frac{\gamma_0\mu}{2}, \frac{\alpha}{4}\right\}\right)}, \\
\gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\frac{3n\left(R_0^2 + \tilde{F}_1\gamma_0\sigma_{1,0}^2\right)\mu^2K^2}{2\widetilde{D}_1}, \frac{\delta\left(R_0^2 + \tilde{F}_1\gamma_0\sigma_{1,0}^2\right)\mu^3K^3}{6L\widetilde{D}_1\left(\frac{4\alpha(\omega+1)}{\delta} + 1\right)}\right\}\right\}\right)}{\mu K}\right\},
\end{aligned}$$

*and $\alpha \leq \frac{1}{\omega+1}$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\widetilde{\mathcal{O}}\left(\left(\mathcal{L} + \frac{\sqrt{L\mathcal{L}}}{\delta}\right)R_0^2 \exp\left(-\min\left\{\frac{\mu}{\mathcal{L} + \frac{\sqrt{L\mathcal{L}}}{\delta}}, \alpha\right\}K\right) + \frac{\widetilde{D}_1}{n\mu K} + \frac{L\widetilde{D}_1\left(\frac{\alpha(\omega+1)}{\delta} + 1\right)}{\delta\mu^2K^2}\right)$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* `EC-SGDsr-DIANA` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\alpha} + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\widetilde{D}_1}{n\mu\varepsilon} + \frac{\sqrt{L\widetilde{D}_1\left(\frac{\alpha(\omega+1)}{\delta} + 1\right)}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

*In particular, if $\alpha = \frac{1}{\omega+1}$, then to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* `EC-SGDsr-DIANA` *requires*

$$\widetilde{\mathcal{O}}\left(\omega + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\widetilde{D}_1}{n\mu\varepsilon} + \frac{\sqrt{L\widetilde{D}_1}}{\delta\mu\sqrt{\varepsilon}}\right) \text{ iterations,}$$

*and if $\alpha = \frac{\delta}{\omega+1}$, then to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* `EC-SGDsr-DIANA` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\omega+1}{\delta} + \frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}}}{\delta\mu} + \frac{\widetilde{D}_1}{n\mu\varepsilon} + \frac{\sqrt{L\widetilde{D}_1}}{\mu\sqrt{\delta\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary J.9.** *Let the assumptions of Theorem J.5 hold and $\mu = 0$. Then after $K$ iterations of* EC-SGDsr-DIANA *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{4\mathcal{L}}, \frac{\delta}{4\sqrt{6L\left(4L + 3\delta(\mathcal{L}+L) + \frac{16(3\mathcal{L}+4L)}{1-\alpha}\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,$$

$$\gamma = \min\left\{\gamma_0, \sqrt[3]{\frac{R_0^2\delta^2\alpha\left(1-\min\left\{\frac{\gamma_0\mu}{2}, \frac{\alpha}{4}\right\}\right)}{96L\sigma_0^2}}, \sqrt{\frac{3nR_0^2}{2\widetilde{D}_1K}}, \sqrt[3]{\frac{\delta R_0^2}{6L\widetilde{D}_1\left(\frac{4\alpha(\omega+1)}{\delta}+1\right)K}}\right\},$$

*and $\alpha \le \frac{1}{\omega+1}$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2}{K} + \frac{\sqrt{\mathcal{L}L}R_0^2}{\delta K} + \frac{\sqrt[3]{LR_0^4\sigma_0^2}}{K\sqrt[3]{\delta^2\alpha}} + \sqrt{\frac{R_0^2\widetilde{D}_1}{nK}} + \sqrt[3]{\frac{LR_0^4\widetilde{D}_1\left(\frac{\alpha(\omega+1)}{\delta}+1\right)}{\delta K^2}}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* EC-SGDsr-DIANA *requires*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2}{\varepsilon} + \frac{\sqrt{\mathcal{L}L}R_0^2}{\delta\varepsilon} + \frac{\sqrt[3]{LR_0^4\sigma_0^2}}{\varepsilon\sqrt[3]{\delta^2\alpha}} + \frac{R_0^2\widetilde{D}_1}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\widetilde{D}_1\left(\frac{\alpha(\omega+1)}{\delta}+1\right)}}{\sqrt{\delta\varepsilon^3}}\right)$$

*iterations. In particular, if $\alpha = \frac{1}{\omega+1}$, then to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* EC-SGDsr-DIANA *requires*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2}{\varepsilon} + \frac{\sqrt{\mathcal{L}L}R_0^2}{\delta\varepsilon} + \frac{\sqrt[3]{LR_0^4(\omega+1)\sigma_0^2}}{\varepsilon\sqrt[3]{\delta^2}} + \frac{R_0^2\widetilde{D}_1}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\widetilde{D}_1}}{\delta\sqrt{\varepsilon^3}}\right) \text{ iterations,}$$

*and if $\alpha = \frac{\delta}{\omega+1}$, then to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* EC-SGDsr-DIANA *requires*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2}{\varepsilon} + \frac{\sqrt{\mathcal{L}L}R_0^2}{\delta\varepsilon} + \frac{\sqrt[3]{LR_0^4(\omega+1)\sigma_0^2}}{\delta\varepsilon} + \frac{R_0^2\widetilde{D}_1}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\widetilde{D}_1}}{\sqrt{\delta\varepsilon^3}}\right) \text{ iterations.}$$

### J.6 EC-LSVRG

In this section we consider problem (1) with $f(x)$ being $\mu$-quasi strongly convex and $f_i(x)$ satisfying (3) where functions $f_{ij}(x)$ are convex and $L$-smooth. For this problem we propose a new method called EC-LSVRG which takes for the origin another method called LSVRG (see [17, 31]).

**Lemma J.7.** *For all $k \ge 0$, $i \in [n]$ we have*

$$\bar{g}_i^k = \mathbf{E}\left[g_i^k \mid x^k\right] = \nabla f_i(x^k) \tag{109}$$

*and*

$$\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 \le 4L\left(f(x^k) - f(x^*)\right) + D_1, \tag{110}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \le 12L\left(f(x^k) - f(x^*)\right) + 3\sigma_k^2, \tag{111}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2 \tag{112}$$

*where $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$ and $D_1 = \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2$.*

**Algorithm 8** `EC-LSVRG`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$

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

  Broadcast $x^k$ to all workers

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

    Pick $l$ uniformly at random from $[m]$

    Set $g_i^k = \nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k)$

    $v_i^k = C(e_i^k + \gamma g_i^k)$

    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$

    $w_i^{k+1} = \begin{cases} x^k, & \text{with probability } p, \\ w_i^k, & \text{with probability } 1-p \end{cases}$

  **end for**

  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$

  $x^{k+1} = x^k - v^k$

**end for**

*Proof.* First of all, we derive unbiasedness of $g_i^k$:

$$\mathbf{E}\left[g_i^k \mid x^k\right] = \frac{1}{m}\sum_{j=1}^m \left(\nabla f_{ij}(x^k) - \nabla f_{ij}(w_i^k) + \nabla f_i(w_i^k)\right) = \nabla f_i(x^k).$$

Next, we get an upper bound for $\frac{1}{n}\sum\limits_{i=1}^n \|\bar{g}_i^k\|^2$:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k)\|^2 \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \\
&\overset{(11)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2.
\end{aligned}$$

Using (109) we establish the following inequality:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &\overset{(31)}{\leq} \frac{3}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\|^2 \mid x^k\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - \nabla f_i(x^k)\|^2 \\
&\overset{(11),(34)}{\leq} 12L\left(f(x^k) - f(x^*)\right) + \frac{3}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2.
\end{aligned}$$

Finally, we derive (112):

$$
\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\|^2 \mid x^k\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&= \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(x^k) - \nabla f_{ij}(x^*)\|^2 \\
&\quad + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*) - \frac{1}{m}\sum_{j=1}^m \left(\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\right)\right\|^2 \\
&\overset{(11),(34)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\right\|^2.
\end{aligned}
$$

□

**Lemma J.8.** *For all $k \geq 0$, $i \in [n]$ we have*

$$
\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq (1-p)\sigma_k^2 + 2Lp\left(f(x^k) - f(x^*)\right), \tag{113}
$$

*where $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

*Proof.* By definition of $w_i^{k+1}$ we get

$$
\begin{aligned}
\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] &= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{E}\left[\|\nabla f_{ij}(w_i^{k+1}) - \nabla f_{ij}(x^*)\|^2 \mid x^k\right] \\
&= \frac{1-p}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2 + \frac{p}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(x^k) - \nabla f_{ij}(x^*)\|^2 \\
&\overset{(11)}{\leq} (1-p)\sigma_k^2 + \frac{2Lp}{nm}\sum_{i=1}^n\sum_{j=1}^m D_{f_{ij}}(x^k, x^*) \\
&= (1-p)\sigma_k^2 + 2Lp\left(f(x^k) - f(x^*)\right).
\end{aligned}
$$

□

Applying Theorem G.1 we get the following result.

**Theorem J.6.** *Assume that $f(x)$ is $\mu$-quasi strongly convex and functions $f_{ij}$ are convex and $L$-smooth for all $i \in [n], j \in [m]$. Then* `EC-LSVRG` *satisfies Assumption 3.3 with*

$$
A = 2L, \quad \widetilde{A} = 12L, \quad A' = 2L, \quad B_1 = \widetilde{B}_1 = B_1' = B_2 = 0, \quad D_1 = \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2,
$$

$$
D_1' = \widetilde{D}_1 = 0, \quad \widetilde{B}_2 = 3, \quad B_2' = 2, \quad \sigma_{1,k}^2 \equiv 0, \quad C_1 = 0,
$$

$$
\sigma_{2,k}^2 = \sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2, \quad \rho_1 = 1, \quad \rho_2 = p, \quad C_2 = Lp, \quad D_2 = 0,
$$

$$
G = 0, \quad F_1 = 0, \quad F_2 = \frac{72L\gamma^2}{\delta p\left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{p}{4}\right\}\right)}, \quad D_3 = \frac{12L\gamma}{\delta^2} D_1,
$$

*with $\gamma$ satisfying*

$$\gamma \le \min\left\{\frac{1}{24L}, \frac{\delta}{8L\sqrt{3\left(2+3\delta\left(2+\frac{1}{1-p}\right)\right)}}\right\}, \quad M_2 = \frac{4}{p}.$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{p}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma} + \frac{48L\gamma^2}{\delta^2}D_1$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma K} + \frac{48L\gamma^2}{\delta^2}D_1$$

*when $\mu = 0$, where $T^k \overset{def}{=} \|x^k - x^*\|^2 + M_2\gamma^2\sigma_k^2$.*

In other words, `EC-LSVRG` converges with linear rate $\mathcal{O}\left(\left(\frac{1}{p} + \frac{\kappa}{\delta\sqrt{1-p}}\right)\ln\frac{1}{\varepsilon}\right)$ to the neighbourhood of the solution. If $m \ge 2$ then taking $p = \frac{1}{m}$ we get that in expectation the sample complexity of one iteration of `EC-LSVRG` is $\mathcal{O}(1)$ gradients calculations per node as for `EC-SGDsr` with standard sampling and the rate of convergence to the neighbourhood becomes $\mathcal{O}\left(\left(m + \frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$. We notice that the size of this neighbourhood is typically smaller than for `EC-SGDsr`, but still the method fails to converge to the exact solution with linear rate. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution in the case when $\mu > 0$.

**Corollary J.10.** *Let the assumptions of Theorem J.6 hold and $\mu > 0$. Then after $K$ iterations of* `EC-LSVRG` *with the stepsize*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{24L}, \frac{\delta}{8L\sqrt{3\left(2+3\delta\left(2+\frac{1}{1-p}\right)\right)}}\right\},\\
\tilde{T}^0 &= \|x^0 - x^*\|^2 + M_2\gamma_0^2\sigma_0^2, \quad \tilde{F}_2 = \frac{72L\gamma_0^2}{\delta p\left(1 - \min\left\{\frac{\gamma_0\mu}{2}, \frac{p}{4}\right\}\right)},\\
\gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \frac{\delta^2\left(\tilde{T}^0 + \tilde{F}_2\gamma_0\sigma_0^2\right)\mu^3K^3}{48LD_1}\right\}\right)}{\mu K}\right\},
\end{aligned}$$

*and $p = \frac{1}{m}$, $m \ge 2$ we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(\frac{L}{\delta}\left(\tilde{T}^0 + \tilde{F}_2\gamma_0\sigma_0^2\right)\exp\left(-\min\left\{\frac{\delta\mu}{L}, \frac{1}{m}\right\}K\right) + \frac{LD_1}{\delta^2\mu^2K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `EC-LSVRG` *requires*

$$\widetilde{\mathcal{O}}\left(m + \frac{L}{\delta\mu} + \frac{\sqrt{LD_1}}{\delta\mu\sqrt{\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary J.11.** *Let the assumptions of Theorem J.6 hold and $\mu = 0$. Then after $K$ iterations of* `EC-LSVRG` *with the stepsize*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{24L}, \frac{\delta}{8L\sqrt{3\left(2+3\delta\left(2+\frac{1}{1-p}\right)\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,\\
\gamma &= \min\left\{\gamma_0, \sqrt{\frac{R_0^2 p}{4\sigma_0^2}}, \sqrt[3]{\frac{R_0^2\delta p\left(1 - \min\left\{\frac{\gamma_0\mu}{2}, \frac{p}{4}\right\}\right)}{72L\sigma_0^2}}, \sqrt[3]{\frac{\delta^2R_0^2}{12LD_1K}}\right\},
\end{aligned}$$

**Algorithm 9** `EC-LSVRGstar`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^k$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Pick $l$ uniformly at random from $[m]$
    Set $g_i^k = \nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k) - \nabla f_i(x^*)$
    $v_i^k = C(e_i^k + \gamma g_i^k)$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    $w_i^{k+1} = \begin{cases} x^k, & \text{with probability } p,\\ w_i^k, & \text{with probability } 1-p \end{cases}$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$
  $x^{k+1} = x^k - v^k$
**end for**

*and* $p = \frac{1}{m}$, $m \geq 2$ *we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta K} + \frac{\sqrt{mR_0^2\sigma_0^2}}{K} + \frac{\sqrt[3]{LR_0^4 m\sigma_0^2}}{\sqrt[3]{\delta}K} + \frac{\sqrt[3]{LR_0^4}}{(\delta K)^{2/3}}\sqrt[3]{\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2}\right).$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$ `EC-LSVRG` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon} + \frac{\sqrt{mR_0^2\sigma_0^2}}{\varepsilon} + \frac{\sqrt[3]{LR_0^4 m\sigma_0^2}}{\sqrt[3]{\delta}\varepsilon} + \frac{R_0^2}{\delta\varepsilon^{3/2}}\sqrt{\frac{L}{n}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2}\right)$$

*iterations.*

### J.7 `EC-LSVRGstar`

In the setup of Section J.6 we now assume that $i$-th node has an access to the $\nabla f_i(x^*)$. Under this unrealistic assumption we construct the method called `EC-LSVRGstar` that asymptotically converges to the exact solution.

**Lemma J.9.** *For all* $k \geq 0$, $i \in [n]$ *we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \tag{114}$$

*and*

$$\frac{1}{n}\sum_{i=1}^n\|\bar{g}_i^k\|^2 \leq 2L\left(f(x^k) - f(x^*)\right), \tag{115}$$

$$\frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2, \tag{116}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2, \tag{117}$$

*where* $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.

*Proof.* First of all, we derive unbiasedness of $g^k$:

$$\begin{aligned}
\mathbf{E}\left[g^k \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k) - \nabla f_i(x^*) \mid x^k\right]\\
&= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left(\nabla f_{ij}(x^k) - \nabla f_{ij}(w_i^k) + \nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\\
&= \nabla f(x^k) + \frac{1}{n}\sum_{i=1}^n\left(-\nabla f_i(w_i^k) + \nabla f_i(w_i^k)\right) - \nabla f(x^*) = \nabla f(x^k).
\end{aligned}$$

Next, we get an upper bound for $\frac{1}{n}\sum\limits_{i=1}^n \|\bar{g}_i^k\|^2$:

$$\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \overset{(11)}{\leq} 2L\left(f(x^k) - f(x^*)\right).$$

Since the variance of random vector is not greater than its second moment we obtain:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &\overset{(34)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\|^2 \mid x^k\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(11),(34)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2.
\end{aligned}$$

Inequality (117) trivially follows from the inequality above by Jensen's inequality and convexity of $\|\cdot\|^2$. □

**Lemma J.10.** *For all $k \geq 0$, $i \in [n]$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq (1-p)\sigma_k^2 + 2Lp\left(f(x^k) - f(x^*)\right), \tag{118}$$

*where $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

*Proof.* The proof of this lemma is identical to the proof of Lemma J.8. □

Applying Theorem G.1 we get the following result.

**Theorem J.7.** *Assume that $f(x)$ is $\mu$-quasi strongly convex and functions $f_{ij}$ are convex and $L$-smooth for all $i \in [n], j \in [m]$. Then* `EC-LSVRGstar` *satisfies Assumption 3.3 with*

$$A = L, \quad \widetilde{A} = A' = 2L, \quad B_1 = \widetilde{B}_1 = B_1' = B_2 = 0, \quad \widetilde{B}_2 = B_2' = 2, \quad D_1 = D_1' = 0,$$

$$\sigma_{1,k}^2 \equiv 0,, \quad C_1 = 0, \quad \sigma_{2,k}^2 = \sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2, \quad \rho_1 = 1,$$

$$\rho_2 = p, \quad C_2 = Lp, \quad D_2 = 0, \quad G = 0, \quad F_1 = 0, \quad F_2 = \frac{48L\gamma^2(2+p)}{\delta p}, \quad D_3 = 0,$$

*with $\gamma$ satisfying*

$$\gamma \leq \min\left\{\frac{3}{56L}, \frac{\delta}{8L\sqrt{3\left(1+\delta\left(1+\frac{2}{1-p}\right)\right)}}\right\}, \quad M_2 = \frac{8}{3p}.$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{p}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma}$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma K}$$

*when $\mu = 0$, where $T^k \overset{def}{=} \|x^k - x^*\|^2 + M_2\gamma^2\sigma_k^2$.*

In other words, EC-LSVRGstar converges with linear rate $\mathcal{O}\left(\left(\frac{1}{p}+\frac{\kappa}{\delta\sqrt{1-p}}\right)\ln\frac{1}{\varepsilon}\right)$ exactly to the solution when $\mu>0$. If $m\geq 2$ then taking $p=\frac{1}{m}$ we get that in expectation the sample complexity of one iteration of EC-LSVRGstar is $\mathcal{O}(1)$ gradients calculations per node as for EC-SGDsr with standard sampling and the rate of convergence becomes $\mathcal{O}\left(\left(m+\frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$.

Applying Lemma D.3 we get the complexity result in the case when $\mu=0$.

**Corollary J.12.** *Let the assumptions of Theorem J.7 hold and $\mu=0$. Then after $K$ iterations of* EC-LSVRGstar *with the stepsize*

$$\gamma_0 = \min\left\{\frac{3}{56L}, \frac{\delta}{8L\sqrt{3\left(1+\delta\left(1+\frac{2}{1-p}\right)\right)}}\right\}, \quad R_0=\|x^0-x^*\|,$$

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{3pR_0^2}{8\sigma_0^2}}, \sqrt[3]{\frac{R_0^2\delta p\left(1-\min\left\{\frac{\gamma_0\mu}{2},\frac{p}{4}\right\}\right)}{72L\sigma_0^2}}\right\},$$

*and $p=\frac{1}{m}$, $m\geq 2$ we have $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta K}+\frac{\sqrt{R_0^2m\sigma_0^2}}{K}+\frac{\sqrt[3]{LR_0^4m\sigma_0^2}}{\sqrt[3]{\delta}K}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]\leq\varepsilon$* EC-LSVRGstar *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon}+\frac{\sqrt{R_0^2m\sigma_0^2}}{\varepsilon}+\frac{\sqrt[3]{LR_0^4m\sigma_0^2}}{\sqrt[3]{\delta}\varepsilon}\right)$$

*iterations.*

However, such convergence guarantees are obtained under very restrictive assumption: the method requires to know vectors $\nabla f_i(x^*)$.

### J.8 EC-LSVRG-DIANA

In the setup of Section J.6 we construct a new method called EC-LSVRG-DIANA which does not require to know $\nabla f_i(x^*)$ and has linear convergence to the exact solution. As in EC-SGD-DIANA the master needs to gather only $C(e_i^k+\gamma g_i^k)$ and $Q(\hat{g}_i^k-h_i^k)$ from all nodes in order to perform an update.

**Lemma J.11.** *Assume that $f_{ij}(x)$ is convex and $L$-smooth for all $i=1,\ldots,n$, $j=1,\ldots,m$. Then, for all $k\geq 0$ we have*

$$\mathbf{E}\left[g^k\mid x^k\right] = \nabla f(x^k), \tag{119}$$

$$\frac{1}{n}\sum_{i=1}^n\|\bar{g}_i^k\|^2 \leq 4L\left(f(x^k)-f(x^*)\right)+2\sigma_{1,k}^2, \tag{120}$$

$$\frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\|g_i^k-\bar{g}_i^k\|^2\mid x^k\right] \leq 6L\left(f(x^k)-f(x^*)\right)+3\sigma_{1,k}^2+3\sigma_{2,k}^2, \tag{121}$$

$$\mathbf{E}\left[\|g^k\|^2\mid x^k\right] \leq 4L\left(f(x^k)-f(x^*)\right)+2\sigma_{2,k}^2 \tag{122}$$

*where*

$$\sigma_{1,k}^2=\frac{1}{n}\sum_{i=1}^n\|h_i^k-\nabla f(x^*)\|^2, \quad \sigma_{2,k}^2=\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2.$$

*Proof.* First of all, we show unbiasedness of $g^k$:

$$\begin{aligned}\mathbf{E}\left[g^k\mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\hat{g}_i^k-h_i^k+h^k\mid x^k\right]\\ &= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left(\nabla f_{ij}(x^k)-\nabla f_{ij}(w_i^k)+\nabla f_i(w_i^k)-h_i^k+h^k\right)=\nabla f(x^k).\end{aligned}$$

**Algorithm 10** `EC-LSVRG-DIANA`

**Input:** learning rates $\gamma > 0$, $\alpha \in (0, 1]$, initial vectors $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
Set $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$
**for** $k = 0, 1, \ldots$ **do**
    Broadcast $x^k, h^k$ to all workers
    **for** $i = 1, \ldots, n$ in parallel **do**
        Pick $l$ uniformly at random from $[m]$
        Set $\hat{g}_i^k = \nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k)$
        $g_i^k = \hat{g}_i^k - h_i^k + h^k$
        $v_i^k = C(e_i^k + \gamma g_i^k)$
        $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
        $h_i^{k+1} = h_i^k + \alpha Q(\hat{g}_i^k - h_i^k)$
        $w_i^{k+1} = \begin{cases} x^k, & \text{with probability } p, \\ w_i^k, & \text{with probability } 1-p \end{cases}$
    **end for**
    $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$, $h^{k+1} = \frac{1}{n}\sum\limits_{i=1}^n h_i^{k+1} = h^k + \alpha\frac{1}{n}\sum\limits_{i=1}^n Q(\hat{g}_i^k - h_i^k)$
    $x^{k+1} = x^k - v^k$
**end for**

Next, we derive the upper bound for $\frac{1}{n}\sum\limits_{i=1}^n \|\bar{g}_i^k\|^2$:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - h_i^k + h^k\|^2 \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + \frac{2}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^*) - \left(h^k - \nabla f(x^*)\right)\right\|^2 \\
&\overset{(11),(34)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2.
\end{aligned}$$

Since the variance of random vector is not greater than its second moment we obtain:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &\overset{(34)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2 \mid x^k\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k) - h_i^k + h^k\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} \frac{3}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\right\|^2 \mid x^k\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\quad + \frac{3}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^*) - \left(h^k - \nabla f(x^*)\right)\right\|^2 \\
&\overset{(11),(34)}{\leq} 6L\left(f(x^k) - f(x^*)\right) + \frac{3}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2 \\
&\quad + \frac{3}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^*)\right\|^2.
\end{aligned}$$

Finally, we obtain an upper boud for the second moment of $g^k$:

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\|^2 \mid x^k\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&= \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(x^k) - \nabla f_{ij}(x^*)\|^2 \\
&\quad + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*) - \frac{1}{m}\sum_{j=1}^m \left(\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\right)\right\|^2 \\
&\overset{(11),(34)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\right\|^2.
\end{aligned}$$

□

**Lemma J.12.** *Assume that $\alpha \leq 1/(\omega+1)$. Then, for all $k \geq 0$ we have*

$$\mathbf{E}\left[\sigma_{1,k+1}^2 \mid x^k\right] \leq (1-\alpha)\sigma_{1,k}^2 + 6L\alpha(f(x^k) - f(x^*)) + 2\alpha\sigma_{2,k}^2, \tag{123}$$

$$\mathbf{E}\left[\sigma_{2,k+1}^2 \mid x^k\right] \leq (1-p)\sigma_{k,2}^2 + 2Lp\left(f(x^k) - f(x^*)\right) \tag{124}$$

*where $\sigma_{1,k}^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$ and $\sigma_{2,k}^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

*Proof.* First of all, we derive an upper bound for the second moment of $h_i^{k+1} - h_i^*$:

$$\begin{aligned}
\mathbf{E}\left[\|h_i^{k+1} - h_i^*\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|h_i^k - h_i^* + \alpha Q(\hat{g}_i^k - h_i^k)\right\|^2 \mid x^k\right] \\
&\overset{(26)}{=} \|h_i^k - h_i^*\|^2 + 2\alpha\langle h_i^k - h_i^*, \nabla f_i(x^k) - h_i^k\rangle \\
&\quad + \alpha^2\mathbf{E}\left[\|Q(\hat{g}_i^k - h_i^k)\|^2 \mid x^k\right] \\
&\overset{(26),(35)}{\leq} \|h_i^k - h_i^*\|^2 + 2\alpha\langle h_i^k - h_i^*, \nabla f_i(x^k) - h_i^k\rangle \\
&\quad + \alpha^2(\omega+1)\mathbf{E}\left[\|\hat{g}_i^k - h_i^k\|^2 \mid x^k\right].
\end{aligned}$$

Using variance decomposition (34) and $\alpha \leq 1/(\omega+1)$ we get

$$\begin{aligned}
\alpha^2(\omega+1)\mathbf{E}\left[\|\hat{g}_i^k - h_i^k\|^2 \mid x^k\right] &\overset{(34)}{=} \alpha^2(\omega+1)\mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] + \alpha^2(\omega+1)\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\leq \alpha\mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] + \alpha\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\overset{(31)}{\leq} 2\alpha\mathbf{E}\left[\left\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(x^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\quad + 2\alpha\mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\quad + \alpha\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\overset{(34)}{\leq} 2\alpha\mathbf{E}\left[\left\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\right\|^2 \mid x^k\right] \\
&\quad + 2\alpha\mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*)\right\|^2 \mid x^k\right] + \alpha\|\nabla f_i(x^k) - h_i^k\|^2 \\
&\overset{(11)}{\leq} 4L\alpha D_{f_i}(x^k, x^*) + \frac{2\alpha}{m}\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2 \\
&\quad + \alpha\|\nabla f_i(x^k) - h_i^k\|^2
\end{aligned}$$

Putting all together we obtain

$$\begin{aligned}
\mathbf{E}\left[\|h_i^{k+1}-h_i^*\|^2 \mid x^k\right] &\le \|h_i^k-h_i^*\|^2+\alpha\left\langle\nabla f_i(x^k)-h_i^k, f_i(x^k)+h_i^k-2h_i^*\right\rangle\\
&\quad +4L\alpha D_{f_i}(x^k,x^*)+\frac{2\alpha}{m}\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2\\
&\overset{(28)}{=} \|h_i^k-h_i^*\|^2+\alpha\|\nabla f_i(x^k)-h_i^*\|^2-\alpha\|h_i^k-h_i^*\|^2\\
&\quad +4L\alpha D_{f_i}(x^k,x^*)+\frac{2\alpha}{m}\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2\\
&\overset{(11)}{\le} (1-\alpha)\|h_i^k-h_i^*\|^2+6L\alpha D_{f_i}(x^k,x^*)\\
&\quad +\frac{2\alpha}{m}\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2.
\end{aligned}$$

Summing up the above inequality for $i=1,\ldots,n$ we derive

$$\mathbf{E}\left[\sigma_{1,k+1}^2 \mid x^k\right] \le (1-\alpha)\sigma_{1,k}^2+6L\alpha(f(x^k)-f(x^*))+2\alpha\sigma_{2,k}^2.$$

Similarly to the proof of Lemma J.8 we get

$$\begin{aligned}
\mathbf{E}\left[\sigma_{2,k+1}^2 \mid x^k\right] &= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\mathbf{E}\left[\|\nabla f_{ij}(w_i^{k+1})-\nabla f_{ij}(x^*)\|^2 \mid x^k\right]\\
&= \frac{1-p}{nm}\sum_{i=1}^n\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2\\
&\quad +\frac{p}{nm}\sum_{i=1}^n\sum_{j=1}^m\|\nabla f_{ij}(x^k)-\nabla f_{ij}(x^*)\|^2\\
&\overset{(11)}{\le} (1-p)\sigma_{2,k}^2+\frac{2Lp}{nm}\sum_{i=1}^n\sum_{j=1}^m D_{f_{ij}}(x^k,x^*)\\
&= (1-p)\sigma_{2,k}^2+2Lp\left(f(x^k)-f(x^*)\right).
\end{aligned}$$

□

Applying Theorem G.1 we get the following result.

**Theorem J.8.** *Assume that $f_{ij}(x)$ is convex and $L$-smooth for all $i=1,\ldots,n$, $j=1,\ldots,m$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `EC-LSVRG-DIANA` *satisfies Assumption 3.3 with*

$$A=A'=2L,\quad B_1'=B_2=0,\quad B_1=B_2'=2,\quad D_1=\widetilde{D}_1=D_1'=D_2=D_3=0,$$

$$\widetilde{A}=3L,\quad \widetilde{B}_1=\widetilde{B}_2=3,\quad \sigma_{1,k}^2=\frac{1}{n}\sum_{i=1}^n\|h_i^k-\nabla f_i(x^*)\|^2,\quad \rho_1=\alpha,$$

$$\sigma_{2,k}^2=\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2,\quad \rho_2=p,\quad C_1=3L\alpha,\quad C_2=Lp,$$

$$G=2,\quad F_1=\frac{24L\gamma^2\left(\frac{4}{\delta}+3\right)}{\delta\alpha\left(1-\min\left\{\frac{\gamma\mu}{2},\frac{\alpha}{4},\frac{p}{4}\right\}\right)},\quad F_2=\frac{24L\gamma^2\left(\frac{4}{1-\alpha}\left(\frac{4}{\delta}+3\right)+3\right)}{\delta p\left(1-\min\left\{\frac{\gamma\mu}{2},\frac{\alpha}{4},\frac{p}{4}\right\}\right)},$$

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma\le\min\left\{\frac{9}{296L},\frac{\delta}{4L\sqrt{6\left(4+3\delta+\frac{2}{1-\alpha}\left(3+\frac{4}{1-p}\right)(4+3\delta)+\frac{6\delta}{1-p}\right)}}\right\},\quad \alpha\le\frac{1}{\omega+1}$$

*with* $M_1 = 0$ *and* $M_2 = \frac{8}{3p} + \frac{32}{9p}$ *and for all* $K \geq 0$

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}, \frac{p}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_1 \sigma_{1,0}^2 + \gamma F_2 \sigma_{2,0}^2)}{\gamma},$$

*when* $\mu > 0$ *and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(T^0 + \gamma F_1 \sigma_{1,0}^2 + \gamma F_2 \sigma_{2,0}^2)}{K\gamma}$$

*when* $\mu = 0$*, where* $T^k \stackrel{def}{=} \|x^k - x^*\|^2 + M_2 \gamma^2 \sigma_{2,k}^2$.

In other words, if $p = 1/m$, $m \geq 2$ and

$$\gamma = \min\left\{\frac{9}{296L}, \frac{\delta}{4L\sqrt{6\left(4 + 3\delta + \frac{2}{1-\alpha}\left(3 + \frac{4}{1-p}\right)(4 + 3\delta) + \frac{6\delta}{1-p}\right)}}\right\}, \quad \alpha = \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\},$$

then `EC-LSVRG-DIANA` converges with the linear rate

$$\mathcal{O}\left(\left(\omega + m + \frac{\kappa}{\delta}\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution when $\mu > 0$.

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary J.13.** *Let the assumptions of Theorem J.8 hold and* $\mu = 0$*. Then after* $K$ *iterations of* `EC-LSVRG-DIANA` *with the stepsize*

$$\gamma_0 = \min\left\{\frac{9}{296L}, \frac{\delta}{4L\sqrt{6\left(4 + 3\delta + \frac{2}{1-\alpha}\left(3 + \frac{4}{1-p}\right)(4 + 3\delta) + \frac{6\delta}{1-p}\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,$$

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{9pR_0^2}{56\sigma_{2,0}^2}}, \sqrt[3]{\frac{R_0^2}{\frac{24L\left(\frac{4}{\delta}+3\right)}{\delta\alpha\left(1-\min\left\{\frac{\gamma_0\mu}{2},\frac{\alpha}{4},\frac{p}{4}\right\}\right)}\sigma_{1,0}^2 + \frac{24L\left(\frac{4}{1-\alpha}\left(\frac{4}{\delta}+3\right)+3\right)}{\delta p\left(1-\min\left\{\frac{\gamma_0\mu}{2},\frac{\alpha}{4},\frac{p}{4}\right\}\right)}\sigma_{2,0}^2}}\right\},$$

*and* $p = \frac{1}{m}$*,* $m \geq 2$*,* $\alpha = \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\}$ *we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta K} + \frac{\sqrt{R_0^2 m \sigma_{2,0}^2}}{K} + \frac{\sqrt[3]{LR_0^4((\omega+1)\sigma_{1,0}^2 + m\sigma_{2,0}^2)}}{\delta^{2/3}K}\right).$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$ `EC-LSVRG-DIANA` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\delta\varepsilon} + \frac{\sqrt{R_0^2 m \sigma_{2,0}^2}}{\varepsilon} + \frac{\sqrt[3]{LR_0^4((\omega+1)\sigma_{1,0}^2 + m\sigma_{2,0}^2)}}{\delta^{2/3}\varepsilon}\right)$$

*iterations.*

Table 4: Complexity of SGD methods with delayed updates established in this paper. Symbols: $\varepsilon$ = error tolerance; $\delta$ = contraction factor of compressor $\mathcal{C}$; $\omega$ = variance parameter of compressor $\mathcal{Q}$; $\kappa = {}^{L}/_{\mu}$; $\mathcal{L}$ = expected smoothness constant; $\sigma_*^2$ = variance of the stochastic gradients in the solution; $\zeta_*^2$ = average of $\|\nabla f_i(x^*)\|^2$; $\sigma^2$ = average of the uniform bounds for the variances of stochastic gradients of workers; $\mathcal{M}_{2,q} = (\omega+1)\sigma^2 + \omega\zeta_*^2$; $\sigma_q^2 = (1+\omega)\left(1+\frac{\omega}{n}\right)\sigma^2$. †`D-QGDstar` is a special case of `D-QSGDstar` where each worker $i$ computes the full gradient $\nabla f_i(x^k)$; ‡`D-GD-DIANA` is a special case of `D-SGD-DIANA` where each worker $i$ computes the full gradient $\nabla f_i(x^k)$.

| Problem | Method | Alg # | Citation | Sec # | Rate (constants ignored) |
|---|---|---|---|---|---|
| (1)+(3) | `D-SGDsr` | Alg 15 | **new** | K.5 | $\widetilde{\mathcal{O}}\left(\frac{\mathcal{L}+\sqrt{L^2\tau^2+L\mathcal{L}\tau}}{\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\tau\sigma_*^2}}{\mu\sqrt{n\varepsilon}}\right)$ |
| (1)+(2) | `D-SGD` | Alg 11 | [46] | K.1 | $\widetilde{\mathcal{O}}\left(\tau\kappa + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\tau\sigma_*^2}}{\mu\sqrt{n\varepsilon}}\right)$ |
| (1)+(2) | `D-QSGD` | Alg 12 | **new** | K.2 | $\widetilde{\mathcal{O}}\left(\kappa\left(\tau+\frac{\omega}{n}\right) + \frac{\mathcal{M}_{2,q}}{n\mu\varepsilon} + \frac{\sqrt{L\tau\mathcal{M}_{2,q}}}{\mu\sqrt{n\varepsilon}}\right)$ |
| (1)+(2) | `D-QSGDstar` | Alg 13 | **new** | K.3 | $\widetilde{\mathcal{O}}\left(\kappa\left(\tau+\frac{\omega}{n}\right) + \frac{\sigma^2}{n\mu\varepsilon} + \frac{\sqrt{L\tau\sigma^2}}{\mu\sqrt{n\varepsilon}}\right)$ |
| (1)+(2) | `D-QGDstar`† | Alg 13 | **new** | K.3 | $\mathcal{O}\left(\kappa\left(\tau+\frac{\omega}{n}\right)\log\frac{1}{\varepsilon}\right)$ |
| (1)+(2) | `D-SGD-DIANA` | Alg 14 | **new** | K.4 | $\widetilde{\mathcal{O}}\left(\omega + \kappa\left(\tau+\frac{\omega}{n}\right) + \frac{\sigma^2}{n\mu\varepsilon} + \frac{\sqrt{L\tau\sigma_q^2}}{\mu\sqrt{n\varepsilon}}\right)$ |
| (1)+(2) | `D-GD-DIANA`‡ | Alg 14 | **new** | K.4 | $\mathcal{O}\left(\left(\omega + \kappa\left(\tau+\frac{\omega}{n}\right)\right)\log\frac{1}{\varepsilon}\right)$ |
| (1)+(3) | `D-LSVRG` | Alg 16 | **new** | K.6 | $\mathcal{O}\left((m+\kappa\tau)\log\frac{1}{\varepsilon}\right)$ |
| (1)+(3) | `D-QLSVRG` | Alg 17 | **new** | K.7 | $\widetilde{\mathcal{O}}\left(m + \kappa\left(\tau+\frac{\omega}{n}\right) + \frac{\zeta_*^2}{n\mu\varepsilon} + \frac{\sqrt{L\tau\zeta_*^2}}{\mu\sqrt{n\varepsilon}}\right)$ |
| (1)+(3) | `D-QLSVRGstar` | Alg 18 | **new** | K.8 | $\mathcal{O}\left(\left(m + \kappa\left(\tau+\frac{\omega}{n}\right)\right)\log\frac{1}{\varepsilon}\right)$ |
| (1)+(3) | `D-LSVRG-DIANA` | Alg 19 | **new** | K.9 | $\mathcal{O}\left(\left(\omega + m + \kappa\left(\tau+\frac{\omega}{n}\right)\right)\log\frac{1}{\varepsilon}\right)$ |

**Algorithm 11** `D-SGD`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
    Broadcast $x^k$ to all workers
    **for** $i = 1, \ldots, n$ in parallel **do**
        Sample $g_i^k = \nabla f_{\xi_i}(x^k) - \nabla f_i(x^*)$
        $v_i^k = \begin{cases}\gamma g_i^{k-\tau}, & \text{if } k \geq \tau,\\ 0, & \text{if } k < \tau\end{cases}$
        $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    **end for**
    $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k = \frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x^{k-\tau})$
    $x^{k+1} = x^k - v^k$
**end for**

# K Special Cases: Delayed Updates Methods

## K.1 `D-SGD`

In this section we consider the same setup as in Section J.2. We notice that vectors $e_i^k$ appear only in the analysis and there is no need to compute them. Moreover, we use $\nabla f_i(x^*)$ in the definition of $g_i^k$ which is problematic at the firt glance. Indeed, workers do not know $\nabla f_i(x^*)$. However, since $0 = \nabla f(x^*) = \frac{1}{n}\nabla f_i(x^*)$ and master node uses averages of $g_i^k$ for the updates one can ignore $\nabla f_i(x^*)$ in $g_i^k$ in the implementation of `D-SGD` and get exactly the same method. We define $g_i^k$ in such a way only for the theoretical analysis.

**Lemma K.1** (see also Lemmas 1,2 from [39])**.** *Assume that $f_{\xi_i}(x)$ are convex in $x$ for every $\xi_i$, $i = 1,\ldots,n$. Then for every $x \in \mathbb{R}^d$ and $i = 1,\ldots,n$*

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 4L(f(x^k) - f(x^*)) + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right]. \tag{125}$$

*If further $f(x)$ is $\mu$-quasi strongly convex with possibly non-convex $f_i, f_{\xi_i}$ and $\mu > 0$, then for every $x \in \mathbb{R}^d$ and $i = 1,\ldots,n$*

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 4L\kappa(f(x^k) - f(x^*)) + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right], \tag{126}$$

*where $\kappa = \frac{L}{\mu}$.*

*Proof.* By definition of $g^k$ we have

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i}(x^k) - \nabla f_{\xi_i}(x^*) + \nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(31)}{\le} 2\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i}(x^k) - \nabla f_{\xi_i}(x^*)\right)\right\|^2 \mid x^k\right] \\
&\quad + 2\underbrace{\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*))\right\|^2\right]}_{\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x^*)\right]} \\
&\overset{(31)}{\le} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i}(x^k) - \nabla f_{\xi_i}(x^*)\|^2 \mid x^k\right] \\
&\quad + \frac{2}{n^2}\sum_{i=1}^n \underbrace{\mathbf{E}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right]}_{\mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right]},
\end{aligned} \tag{127}$$

where in the last inequality we use independence of $\nabla f_{\xi_i}(x^*)$, $i = 1,\ldots,n$. Using this we derive inequality (125):

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\overset{(127),(11)}{\le} \frac{4L}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_{\xi_i}}(x^k, x^*) \mid x^k\right] + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right] \\
&= \frac{4L}{n}\sum_{i=1}^n D_{f_i}(x^k, x^*) + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right] \\
&= 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right].
\end{aligned}$$

Next, if $f(x)$ is $\mu$-quasi strongly convex, but $f_i, f_{\xi_i}$ are not necessary convex, we obtain

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\overset{(127),(10)}{\le} \frac{2L^2}{n}\sum_{i=1}^n \|x^k - x^*\|^2 + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right] \\
&\overset{(9)}{\le} \frac{4L^2}{\mu}\left(f(x^k) - f(x^*)\right) + \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right].
\end{aligned}$$

□

**Theorem K.1.** *Assume that $f_\xi(x)$ is convex in $x$ for every $\xi$. Then* D-SGD *satisfies Assumption 3.4 with*

$$A' = 2L, \quad B_1' = B_2' = 0, \quad D_1' = \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right], \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0$$

$$\rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad D_2 = 0$$

$$F_1 = F_2 = 0, \quad D_3 = \frac{6\gamma\tau L}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right]$$

*with $\gamma$ satisfying*

$$\gamma \leq \frac{1}{8L\sqrt{2\tau(\tau+2)}}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + \frac{8\gamma}{n^2}(1 + 3L\gamma\tau)\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right]$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4\|x^0 - x^*\|^2}{\gamma K} + \frac{8\gamma}{n^2}(1 + 3L\gamma\tau)\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right]$$

*when $\mu = 0$. If further $f_i(x)$ are $\mu$-strongly convex with possibly non-convex $f_{\xi_i}$ and $\mu > 0$, then* D-SGD *satisfies Assumption 3.4 with*

$$A' = 2\kappa L, \quad B_1' = B_2' = 0, \quad D_1' = \frac{2}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right], \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0,$$

$$\rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad D_2 = 0, \quad G = 0,$$

$$F_1 = F_2 = 0, \quad D_3 = \frac{6\gamma\tau L}{n^2}\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right]$$

*with $\gamma$ satisfying*

$$\gamma \leq \min\left\{\frac{1}{8\kappa L}, \frac{1}{8L\sqrt{2\tau(\tau+2\kappa)}}\right\}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + \frac{8\gamma}{n^2}(1 + 3L\gamma\tau)\sum_{i=1}^n \mathrm{Var}\left[\nabla f_{\xi_i}(x^*)\right].$$

In other words, D-SGD converges with linear rate $\mathcal{O}\left(\tau\kappa \ln \frac{1}{\varepsilon}\right)$ to the neighbourhood of the solution when $\mu > 0$. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary K.1.** *Let the assumptions of Theorem K.1 hold, $f_\xi(x)$ are convex for each $\xi$ and $\mu > 0$. Then after $K$ iterations of* D-SGD *with the stepsize*

$$\gamma = \min\left\{\frac{1}{8L\sqrt{2\tau(\tau+2)}}, \frac{\ln\left(\max\left\{2, \min\left\{\frac{\|x^0-x^*\|^2\mu^2K^2}{D_1'}, \frac{\|x^0-x^*\|^2\mu^3K^3}{3\tau L D_1}\right\}\right\}\right)}{\mu K}\right\}$$

*we have*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(L\tau\|x^0 - x^*\|^2 \exp\left(-\frac{\mu}{\tau L}K\right) + \frac{D_1'}{\mu K} + \frac{L\tau D_1'}{\mu^2 K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* D-SGD *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\tau L}{\mu} + \frac{D_1'}{\mu\varepsilon} + \frac{\sqrt{L\tau D_1'}}{\mu\sqrt{\varepsilon}}\right) \text{ iterations.}$$

**Corollary K.2.** *Let the assumptions of Theorem K.1 hold and $f(x)$ is $\mu$-strongly convex with $\mu > 0$ and possibly non-convex $f_i, f_{\xi_i}$. Then after $K$ iterations of* D-SGD *with the stepsize*

$$\gamma = \min\left\{\frac{1}{8\kappa L}, \frac{1}{8L\sqrt{2\tau(\tau+2\kappa)}}, \frac{\ln\left(\max\left\{2, \min\left\{\frac{\|x^0-x^*\|^2\mu^2K^2}{D_1'}, \frac{\|x^0-x^*\|^2\mu^3K^3}{L\tau D_1'}\right\}\right\}\right)}{\mu K}\right\}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\widetilde{\mathcal{O}}\left(L\left(\kappa + \tau\sqrt{\kappa}\right)\|x^0 - x^*\|^2 \exp\left(-\min\left\{\frac{\mu}{\tau L\sqrt{\kappa}}, \frac{1}{\kappa^2}\right\}K\right) + \frac{D_1'}{\mu K} + \frac{L\tau D_1'}{\mu^2K^2}\right).$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ D-SGD *requires*

$$\widetilde{\mathcal{O}}\left(\kappa^2 + \tau\kappa^{3/2} + \frac{D_1'}{\mu\varepsilon} + \frac{\sqrt{L\tau D_1'}}{\mu\sqrt{\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.3.** *Let the assumptions of Theorem K.1 hold, $f_\xi(x)$ are convex for each $\xi$ and $\mu = 0$. Then after $K$ iterations of* D-SGD *with the stepsize*

$$\gamma = \min\left\{\frac{1}{8L\sqrt{2\tau(\tau+2)}}, \sqrt{\frac{\|x^0 - x^*\|^2}{D_1'K}}, \sqrt[3]{\frac{\|x^0 - x^*\|^2}{3L\tau D_1'K}}\right\}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{\tau LR_0^2}{K} + \sqrt{\frac{R_0^2\tau D_1'}{K}} + \frac{\sqrt[3]{LR_0^4\tau D_1'}}{K^{2/3}}\right)$$

*where $R_0 = \|x^0 - x^*\|$. That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ D-SGD *requires*

$$\mathcal{O}\left(\frac{\tau LR_0^2}{\varepsilon} + \frac{R_0^2D_1'}{\varepsilon^2} + \frac{R_0^2\sqrt{L\tau D_1'}}{\varepsilon^{3/2}}\right)$$

*iterations.*

### K.2 D-QSGD

In this section we show how one can combine delayed updates with quantization using our scheme.

**Lemma K.2.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$. Then, for all $k \ge 0$ we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k),$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 2L\left(1 + \frac{2\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + \frac{(\omega+1)D}{n} + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2$$

*where* $D = \frac{1}{n}\sum_{i=1}^n D_i$.

*Proof.* First of all, we show unbiasedness of $g^k$:

$$\begin{aligned}\mathbf{E}\left[g^k \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[g_i^k \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\mathbf{E}_Q\left[Q(\hat{g}_i^k) - \nabla f_i(x^*)\right] \mid x^k\right] \\ &\overset{(26)}{=} \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x^k) - \nabla f_i(x^*)\right) = \nabla f(x^k),\end{aligned}$$

**Algorithm 12** `D-QSGD`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^{k-\tau}$ to all workers
  **for** $i = 1, \ldots, n$ **do**
    Sample $\hat{g}_i^{k-\tau}$ independently from other nodes such that $\mathbf{E}[\hat{g}_i^{k-\tau} \mid x^{k-\tau}] = \nabla f_i(x^{k-\tau})$ and $\mathbf{E}\left[\|\hat{g}_i^{k-\tau} - \nabla f_i(x^{k-\tau})\|^2 \mid x^{k-\tau}\right] \leq D_i$
    $g_i^{k-\tau} = Q(\hat{g}_i^{k-\tau}) - \nabla f_i(x^*)$ (quantization is performed independently from other nodes)
    $v_i^k = \gamma g_i^{k-\tau}$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k = \frac{\gamma}{n}\sum_{i=1}^n g_i^{k-\tau} = \frac{\gamma}{n}\sum_{i=1}^n Q(\hat{g}_i^{k-\tau})$
  $x^{k+1} = x^k - v^k$
**end for**

where $\mathbf{E}_Q\left[\cdot\right]$ denotes mathematical expectation w.r.t. the randomness coming only from the quantization. Next, we derive the upper bound for the second moment of $g^k$:

$$\begin{aligned}
\mathbf{E}_Q\left[\|g^k\|^2\right] &= \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(Q(\hat{g}_i^k) - \nabla f_i(x^*)\right)\right\|^2\right] \\
&\overset{(34)}{=} \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(Q(\hat{g}_i^k) - \hat{g}_i^k\right)\right\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2. \qquad (128)
\end{aligned}$$

Since $Q(\hat{g}_1^k), \ldots, Q(\hat{g}_n^k)$ are independent quantizations, we get

$$\begin{aligned}
\mathbf{E}_Q\left[\|g^k\|^2\right] &\overset{(128)}{\leq} \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_Q\left[\|Q(\hat{g}_i^k) - \hat{g}_i^k\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \\
&\overset{(26)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \|\hat{g}_i^k\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2.
\end{aligned}$$

Taking conditional expectation $\mathbf{E}\left[\cdot \mid x^k\right]$ from the both sides of the previous inequality we obtain

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\leq \frac{\omega}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k\|^2 \mid x^k\right] + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(34)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^k)\|^2 + \frac{\omega}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \\
&\quad + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x^k) - \nabla f_i(x^*)\right)\right\|^2}_{\|\nabla f(x^k) - \nabla f(x^*)\|^2} + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^k)\right)\right\|^2 \mid x^k\right].
\end{aligned}$$

It remains to estimate terms in the second and the third lines of the previous inequality:

$$\begin{aligned}
\frac{\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^k)\|^2 &\overset{(31)}{\leq} \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \\
&\overset{(11)}{\leq} \frac{4\omega L}{n}\left(f(x^k) - f(x^*)\right) + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, \\
\frac{\omega}{n}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] &\leq \frac{\omega}{n^2}\sum_{i=1}^n D_i = \frac{\omega D}{n}, \\
\|\nabla f(x^k) - \nabla f(x^*)\|^2 &\overset{(11)}{\leq} 2L\left(f(x^k) - f(x^*)\right), \\
\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^k)\right)\right\|^2 \mid x^k\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \\
&\leq \frac{1}{n^2}\sum_{i=1}^n D_i = \frac{D}{n}.
\end{aligned}$$

Putting all together we get

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 2L\left(1 + \frac{2\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + \frac{(\omega+1)D}{n} + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2.$$

□

**Theorem K.2.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `D-QSGD` *satisfies Assumption 3.4 with*

$$A' = L\left(1 + \frac{2\omega}{n}\right), \quad B_1' = B_2' = 0, \quad D_1' = \frac{(\omega+1)D}{n} + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2,$$
$$\sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0, \quad \rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad D_2 = 0$$
$$F_1 = F_2 = 0, \quad G = 0, \quad D_3 = \frac{3\gamma\tau L}{n}\left((\omega+1)D + \frac{2\omega}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2\right)$$

*with $\gamma$ satisfying*

$$\gamma \leq \min\left\{\frac{1}{4L(1 + {}^{2\omega}/_{n})}, \frac{1}{8L\sqrt{2\tau\left(\tau + 1 + {}^{2\omega}/_{n}\right)}}\right\}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + \gamma\left(D_1' + D_3\right)$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4\|x^0 - x^*\|^2}{\gamma K} + \gamma\left(D_1' + D_3\right)$$

*when $\mu = 0$.*

In other words, `D-QSGD` converges with the linear rate

$$\mathcal{O}\left(\left(\kappa\left(1 + \frac{\omega}{n}\right) + \kappa\sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)\ln\frac{1}{\varepsilon}\right)$$

to the neighbourhood of the solution when $\mu > 0$. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary K.4.** *Let the assumptions of Theorem K.2 hold, $f_\xi(x)$ are convex for each $\xi$ and $\mu > 0$. Then after $K$ iterations of* `D-QSGD` *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{4L(1+{}^{2\omega}/_{n})}, \frac{1}{8L\sqrt{2\tau\left(\tau+1+{}^{2\omega}/_{n}\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,$$

$$\gamma = \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\frac{R_0^2\mu^2K^2}{D_1'}, \frac{R_0^2\mu^3K^3}{3\tau LD_1'}\right\}\right\}\right)}{\mu K}\right\}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\widetilde{\mathcal{O}}\left(LR_0^2\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)\exp\left(-\frac{\mu}{L\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)}K\right)+\frac{D_1'}{\mu K}+\frac{L\tau D_1'}{\mu^2K^2}\right).$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ `D-QSGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\left(1+\frac{\omega}{n}\right)+\frac{L}{\mu}\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}+\frac{D_1'}{\mu\varepsilon}+\frac{\sqrt{L\tau D_1'}}{\mu\sqrt{\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.5.** *Let the assumptions of Theorem K.2 hold and $\mu = 0$. Then after $K$ iterations of* `D-QSGD` *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{4L(1+{}^{2\omega}/_{n})}, \frac{1}{8L\sqrt{2\tau\left(\tau+1+{}^{2\omega}/_{n}\right)}}\right\},$$

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{\|x^0-x^*\|^2}{D_1'K}}, \sqrt[3]{\frac{\|x^0-x^*\|^2}{3L\tau D_1'K}}\right\}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}\right)}{K}+\frac{LR_0^2\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}}{K}+\sqrt{\frac{R_0^2D_1'}{K}}+\frac{\sqrt[3]{LR_0^4\tau D_1'}}{K^{2/3}}\right)$$

*where $R_0 = \|x^0 - x^*\|$. That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$ `D-QSGD` *requires*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}\right)}{\varepsilon}+\frac{LR_0^2\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}}{\varepsilon}+\frac{R_0^2D_1'}{\varepsilon^2}+\frac{R_0^2\sqrt{L\tau D_1'}}{\varepsilon^{3/2}}\right)$$

*iterations.*

### K.3 `D-QSGDstar`

As we saw in Section K.2 `D-QSGD` fails to converge to the exact optimum asymptotically even if $\hat{g}_i^k = \nabla f_i(x^k)$ for all $i = 1, \ldots, n$ almost surely, i.e., all $D_i = 0$ for all $i = 1, \ldots, n$. As for `EC-GDstar` we assume now that $i$-th worker has an access to $\nabla f_i(x^*)$. Using this one can construct the method with delayed updates that converges asymptotically to the exact solution when the full gradients are available.

**Lemma K.3.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$. Then, for all $k \ge 0$ we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k), \tag{129}$$

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \le 2L\left(1+\frac{\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + \frac{(\omega+1)D}{n} \tag{130}$$

*where* $D = \frac{1}{n}\sum_{i=1}^n D_i$.

**Algorithm 13** `D-QSGDstar`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^{k-\tau}$ to all workers
  **for** $i = 1, \ldots, n$ **do**
    Sample $\hat{g}_i^{k-\tau}$ independently from other nodes such that $\mathbf{E}[\hat{g}_i^{k-\tau} \mid x^{k-\tau}] = \nabla f_i(x^{k-\tau})$ and $\mathbf{E}\left[\|\hat{g}_i^{k-\tau} - \nabla f_i(x^{k-\tau})\|^2 \mid x^{k-\tau}\right] \le D_i$
    $g_i^{k-\tau} = Q(\hat{g}_i^{k-\tau} - \nabla f_i(x^*))$ (quantization is performed independently from other nodes)
    $v_i^k = \gamma g_i^{k-\tau}$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k, \quad g^k = \frac{1}{n}\sum_{i=1}^n g_i^k, \quad v^k = \frac{1}{n}\sum_{i=1}^n v_i^k = \frac{\gamma}{n}\sum_{i=1}^n g_i^{k-\tau} = \frac{\gamma}{n}\sum_{i=1}^n Q(\hat{g}_i^{k-\tau} - \nabla f_i(x^*))$
  $x^{k+1} = x^k - v^k$
**end for**

*Proof.* First of all, we show unbiasedness of $g^k$:

$$\begin{aligned}\mathbf{E}\left[g^k \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[g_i^k \mid x^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\mathbf{E}_Q\left[Q(\hat{g}_i^k - \nabla f_i(x^*))\right] \mid x^k\right] \\ &\overset{(26)}{=} \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x^k) - \nabla f_i(x^*)\right) = \nabla f(x^k),\end{aligned}$$

where $\mathbf{E}_Q[\cdot]$ denotes mathematical expectation w.r.t. the randomness coming only from the quantization. Next, we derive the upper bound for the second moment of $g^k$:

$$\begin{aligned}\mathbf{E}_Q\left[\|g^k\|^2\right] &= \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(Q\left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right)\right\|^2\right] \\ &\overset{(34)}{=} \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(Q\left(\hat{g}_i^k - \nabla f_i(x^*)\right) - \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right)\right\|^2\right] \\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2. \end{aligned} \tag{131}$$

Since $Q\left(\hat{g}_1^k - \nabla f_1(x^*)\right), \ldots, Q\left(\hat{g}_n^k - \nabla f_n(x^*)\right)$ are independent quantizations, we get

$$\begin{aligned}\mathbf{E}_Q\left[\|g^k\|^2\right] &\overset{(131)}{\le} \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_Q\left[\left\|Q\left(\hat{g}_i^k - \nabla f_i(x^*)\right) - \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2\right] \\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \\ &\overset{(26)}{\le} \frac{\omega}{n^2}\sum_{i=1}^n \|\hat{g}_i^k - \nabla f_i(x^*)\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2.\end{aligned}$$

Taking conditional expectation $\mathbf{E}\left[\cdot \mid x^k\right]$ from the both sides of the previous inequality we obtain

$$
\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\leq \frac{\omega}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^*)\|^2 \mid x^k\right] + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(34)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + \frac{\omega}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \\
&\quad + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla f_i(x^k) - \nabla f_i(x^*)\right)\right\|^2}_{\|\nabla f(x^k) - \nabla f(x^*)\|^2} + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{g}_i^k - \nabla f_i(x^k)\right)\right\|^2 \mid x^k\right].
\end{aligned}
$$

It remains to estimate terms in the second and the third lines of the previous inequality:

$$
\begin{aligned}
\frac{\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 &\overset{(11)}{\leq} \frac{2\omega L}{n}\left(f(x^k) - f(x^*)\right), \\
\frac{\omega}{n}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] &\leq \frac{\omega}{n^2}\sum_{i=1}^n D_i = \frac{\omega D}{n}, \\
\|\nabla f(x^k) - \nabla f(x^*)\|^2 &\overset{(11)}{\leq} 2L\left(f(x^k) - f(x^*)\right), \\
\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{g}_i^k - \nabla f_i(x^k)\right)\right\|^2 \mid x^k\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^k)\|^2 \mid x^k\right] \\
&\leq \frac{1}{n^2}\sum_{i=1}^n D_i = \frac{D}{n}.
\end{aligned}
$$

Putting all together we get

$$
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 2L\left(1 + \frac{\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + \frac{(\omega+1)D}{n}.
$$

□

**Theorem K.3.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `D-QSGDstar` *satisfies Assumption 3.4 with*

$$
A' = L\left(1 + \frac{\omega}{n}\right), \quad B_1' = B_2' = 0, \quad D_1' = \frac{(\omega+1)D}{n}, \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0,
$$
$$
\rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad D_2 = 0, \quad G = 0,
$$
$$
F_1 = F_2 = 0, \quad D_3 = \frac{3\gamma\tau L(\omega+1)D}{n}
$$

*with $\gamma$ satisfying*

$$
\gamma \leq \min\left\{\frac{1}{4L(1 + \omega/n)}, \frac{1}{8L\sqrt{\tau\left(\tau + 1 + \omega/n\right)}}\right\}.
$$

*and for all $K \geq 0$*

$$
\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + 4\gamma\left(D_1' + D_3\right)
$$

*when $\mu > 0$ and*

$$
\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4\|x^0 - x^*\|^2}{\gamma K} + 4\gamma\left(D_1' + D_3\right)
$$

*when $\mu = 0$.*

In other words, D-QSGDstar converges with the linear rate

$$\mathcal{O}\left(\left(\tau+\kappa\left(1+\frac{\omega}{n}\right)+\kappa\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution when $\mu>0$ and $D=0$, i.e., $\hat{g}_i^k=\nabla f_i(x^k)$ for all $i=1,\ldots,n$ almost surely. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary K.6.** *Let the assumptions of Theorem K.3 hold and $\mu>0$. Then after $K$ iterations of* D-QSGDstar *with the stepsize*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{4L(1+\omega/n)},\frac{1}{8L\sqrt{\tau\left(\tau+1+\omega/n\right)}}\right\}, \quad R_0=\|x^0-x^*\|,\\ \gamma &= \min\left\{\gamma_0,\frac{\ln\left(\max\left\{2,\min\left\{\frac{nR_0^2\mu^2K^2}{D},\frac{nR_0^2\mu^3K^3}{3\tau LD}\right\}\right\}\right)}{\mu K}\right\}\end{aligned}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]$ *of order*

$$\widetilde{\mathcal{O}}\left(LR_0^2\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)\exp\left(-\frac{\mu}{L\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)}K\right)+\frac{D}{n\mu K}+\frac{L\tau D}{n\mu^2K^2}\right).$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]\le\varepsilon$ D-QSGDstar *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\left(1+\frac{\omega}{n}\right)+\frac{L}{\mu}\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}+\frac{D}{n\mu\varepsilon}+\frac{\sqrt{L\tau D}}{\mu\sqrt{n\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu=0$.

**Corollary K.7.** *Let the assumptions of Theorem K.3 hold and $\mu=0$. Then after $K$ iterations of* D-QSGDstar *with the stepsize*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{4L(1+2\omega/n)},\frac{1}{8L\sqrt{\tau\left(\tau+1+\omega/n\right)}}\right\},\\ \gamma &= \min\left\{\gamma_0,\sqrt{\frac{n\|x^0-x^*\|^2}{DK}},\sqrt[3]{\frac{n\|x^0-x^*\|^2}{3L\tau DK}}\right\}\end{aligned}$$

*we have* $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}\right)}{K}+\frac{LR_0^2\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}}{K}+\sqrt{\frac{R_0^2D}{nK}}+\frac{\sqrt[3]{LR_0^4\tau D}}{n^{1/3}K^{2/3}}\right)$$

*where $R_0=\|x^0-x^*\|$. That is, to achive* $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]\le\varepsilon$ D-QSGDstar *requires*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}\right)}{\varepsilon}+\frac{LR_0^2\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}}{\varepsilon}+\frac{R_0^2D}{n\varepsilon^2}+\frac{R_0^2\sqrt{L\tau D}}{\sqrt{n}\varepsilon^{3/2}}\right)$$

*iterations.*

### K.4 D-SGD-DIANA

In this section we present a practical version of D-QSGDstar: D-SGD-DIANA.

**Lemma K.4** (Lemmas 1 and 2 from [19])**.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i=1,\ldots,n$ and $\alpha\le 1/(\omega+1)$. Then, for all $k\ge 0$ we have*

$$\mathbf{E}\left[g^k\mid x^k\right] = \nabla f(x^k), \tag{132}$$

$$\mathbf{E}\left[\|g^k\|^2\mid x^k\right] \le 2L\left(1+\frac{2\omega}{n}\right)\left(f(x^k)-f(x^*)\right)+\frac{2\omega\sigma_k^2}{n}+\frac{(\omega+1)D}{n} \tag{133}$$

$$\mathbf{E}\left[\sigma_{k+1}^2\mid x^k\right] \le (1-\alpha)\sigma_k^2+2L\alpha\left(f(x^k)-f(x^*)\right)+\alpha D \tag{134}$$

*where $\sigma_k^2=\frac{1}{n}\sum_{i=1}^n\|h_i^k-\nabla f_i(x^*)\|^2$ and $D=\frac{1}{n}\sum_{i=1}^n D_i$.*

**Algorithm 14** `D-SGD-DIANA`

**Input:** learning rates $\gamma > 0, \alpha \in (0,1]$, initial vectors $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
Set $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$
**for** $k = 0, 1, \ldots$ **do**
    Broadcast $x^{k-\tau}$ to all workers
    **for** $i = 1, \ldots, n$ **do**
        Sample $\hat{g}_i^{k-\tau}$ independently from other nodes such that $\mathbf{E}[\hat{g}_i^{k-\tau} \mid x^{k-\tau}] = \nabla f_i(x^{k-\tau})$ and $\mathbf{E}\left[\|\hat{g}_i^{k-\tau} - \nabla f_i(x^{k-\tau})\|^2 \mid x^{k-\tau}\right] \le D_i$
        $\hat{\Delta}_i^{k-\tau} = Q(\hat{g}_i^{k-\tau} - h_i^{k-\tau})$ (quantization is performed independently from other nodes)
        $g_i^{k-\tau} = h_i^{k-\tau} + \hat{\Delta}_i^{k-\tau}$
        $v_i^k = \gamma g_i^{k-\tau}$
        $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
        $h_i^{k-\tau+1} = h_i^{k-\tau} + \alpha\hat{\Delta}_i^{k-\tau}$
    **end for**
    $h^{k-\tau} = \frac{1}{n}\sum_{i=1}^n h_i^{k-\tau},\ e^k = \frac{1}{n}\sum_{i=1}^n e_i^k,\ g^k = \frac{1}{n}\sum_{i=1}^n g_i^k,\ v^k = \frac{1}{n}\sum_{i=1}^n v_i^k = \frac{\gamma}{n}\sum_{i=1}^n g_i^{k-\tau} = \gamma h^{k-\tau} + \frac{\gamma}{n}\sum_{i=1}^n \hat{\Delta}_i^{k-\tau}$
    $x^{k+1} = x^k - v^k$
    $h^{k-\tau+1} = h^{k-\tau} + \frac{\alpha}{n}\sum_{i=1}^n \hat{\Delta}_i^{k-\tau}$
**end for**

**Theorem K.4.** *Assume that $f_i(x)$ is convex and $L$-smooth for all $i = 1, \ldots, n$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `D-SGD-DIANA` *satisfies Assumption 3.4 with*

$$A' = L\left(1 + \frac{2\omega}{n}\right), \quad B_1' = \frac{2\omega}{n}, \quad D_1' = \frac{(\omega+1)D}{n}, \quad \sigma_{1,k}^2 = \sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2,$$

$$B_2' = 0, \quad \rho_1 = \alpha, \quad \rho_2 = 1, \quad C_1 = L\alpha, \quad C_2 = 0, \quad D_2 = \frac{\alpha(\omega+1)D}{n}, \quad G = 0,$$

$$F_1 = \frac{12\gamma^2 L\omega\tau(2+\alpha)}{n\alpha}, \quad F_2 = 0, \quad D_3 = 3\gamma\tau L\left(1 + \frac{4\omega}{n}\right)\frac{(\omega+1)D}{n}$$

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma \le \min\left\{\frac{1}{4L(1 + {}^{14\omega}/{}_{3n})}, \frac{1}{8L\sqrt{2\tau\left(1 + \tau + {}^{2\omega}/{}_{n} + {}^{4\omega}/{}_{n(1-\alpha)}\right)}}\right\}, \quad \alpha \le \frac{1}{\omega+1}, \quad M_1 = \frac{8\omega}{3n\alpha}$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_1\sigma_0^2)}{\gamma} + 4\gamma\left(D_1' + M_1 D_2 + D_3\right)$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4(T^0 + \gamma F_1\sigma_0^2)}{\gamma K} + 4\gamma\left(D_1' + M_1 D_2 + D_3\right)$$

*when $\mu = 0$, where $T^k \overset{def}{=} \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_k^2$.*

In other words, if

$$\gamma \le \min\left\{\frac{1}{4L(1 + {}^{14\omega}/{}_{3n})}, \frac{1}{8L\sqrt{2\tau\left(1 + \tau + {}^{10\omega}/{}_{n}\right)}}\right\}, \quad \alpha \le \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\}$$

then `D-SGD-DIANA` converges with the linear rate

$$\mathcal{O}\left(\left(\omega + \kappa\left(1 + \frac{\omega}{n}\right) + \kappa\sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution when $\mu > 0$. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary K.8.** *Let the assumptions of Theorem K.4 hold and $\mu > 0$. Then after $K$ iterations of* D-SGD-DIANA *with the stepsize*

$$
\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{4L(1+{14\omega}/{3n})}, \frac{1}{8L\sqrt{2\tau\left(1+\tau+{10\omega}/{n}\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,\\
\tilde{F}_1 &= \frac{12L\omega\tau(2+\alpha)\gamma_0^2}{n\alpha}, \quad \tilde{T}^0 = R_0^2 + M_1\gamma_0^2\sigma_0^2,\\
\gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\frac{(\tilde{T}^0+\gamma_0\tilde{F}_1\sigma_0^2)\mu^2K^2}{D_1'+M_1D_2}, \frac{(\tilde{T}^0+\gamma_0\tilde{F}_1\sigma_0^2)\mu^3K^3}{3\tau L\left(D_1'+\frac{2B_1'D_2}{\alpha}\right)}\right\}\right\}\right)}{\mu K}\right\}
\end{aligned}
$$

*and $\alpha \leq \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\}$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$
\widetilde{\mathcal{O}}\left(LR_0^2\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)\exp\left(-\min\left\{\frac{\mu}{L\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)}, \frac{1}{1+\omega}\right\}K\right)\right)
+\widetilde{\mathcal{O}}\left(\frac{D_1'+M_1D_2}{\mu K} + \frac{\tau L\left(D_1'+\frac{B_1'D_2}{\alpha}\right)}{\mu^2K^2}\right).
$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* D-SGD-DIANA *requires*

$$
\widetilde{\mathcal{O}}\left(\omega + \frac{L}{\mu}\left(1+\frac{\omega}{n}\right) + \frac{L}{\mu}\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)} + \frac{(\omega+1)\left(1+\frac{\omega}{n}\right)D}{n\mu\varepsilon} + \frac{\sqrt{L\tau(\omega+1)\left(1+\frac{\omega}{n}\right)D}}{\mu\sqrt{n\varepsilon}}\right)
$$

*iterations.*

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.9.** *Let the assumptions of Theorem K.4 hold and $\mu = 0$. Then after $K$ iterations of* D-SGD-DIANA *with the stepsize*

$$
\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{4L(1+{14\omega}/{3n})}, \frac{1}{8L\sqrt{2\tau\left(1+\tau+{10\omega}/{n}\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,\\
\gamma &= \min\left\{\gamma_0, \sqrt{\frac{R_0^2}{M_1\sigma_0^2}}, \sqrt[3]{\frac{R_0^2n\alpha}{12L\omega\tau(2+\alpha)\sigma_0^2}}, \sqrt{\frac{R_0^2}{(D_1'+M_1D_2)K}}, \sqrt[3]{\frac{R_0^2}{3\tau L\left(D_1'+\frac{2B_1'D_2}{\alpha}\right)K}}\right\}
\end{aligned}
$$

*we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$
\mathcal{O}\left(\frac{L\left(1+\frac{\omega}{n}\right)R_0^2}{K} + \frac{L\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}R_0^2}{K} + \frac{\sqrt{R_0^2\omega(1+\omega)\sigma_0^2}}{\sqrt{n}K} + \frac{\sqrt[3]{R_0^4L\tau\omega(1+\omega)\sigma_0^2}}{\sqrt[3]{n}K}\right)
+\mathcal{O}\left(\sqrt{\frac{(1+\omega)\left(1+\frac{\omega}{n}\right)R_0^2D}{nK}} + \frac{\sqrt[3]{R_0^4\tau L(1+\omega)\left(1+\frac{\omega}{n}\right)D}}{n^{1/3}K^{2/3}}\right).
$$

**Algorithm 15** `D-SGDsr`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^{k-\tau}$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Sample $g_i^{k-\tau} = \nabla f_{\xi_i}(x^{k-\tau}) - \nabla f_i(x^*)$
    $v_i^k = \gamma g_i^{k-\tau}$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k = \frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i}(x^{k-\tau})$
  $x^{k+1} = x^k - v^k$
**end for**

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$ `D-SGD-DIANA` *requires*

$$\mathcal{O}\left(\frac{L\left(1+\frac{\omega}{n}\right)R_0^2}{\varepsilon} + \frac{L\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}R_0^2}{\varepsilon} + \frac{\sqrt{R_0^2\omega(1+\omega)\sigma_0^2}}{\sqrt{n}\varepsilon} + \frac{\sqrt[3]{R_0^4L\tau\omega(1+\omega)\sigma_0^2}}{\sqrt[3]{n}\varepsilon}\right)$$
$$+\mathcal{O}\left(\frac{(1+\omega)\left(1+\frac{\omega}{n}\right)R_0^2D}{n\varepsilon^2} + \frac{R_0^2\sqrt{\tau L(1+\omega)\left(1+\frac{\omega}{n}\right)D}}{n^{1/2}\varepsilon^{3/2}}\right) \quad \textit{iterations.}$$

### K.5 `D-SGDsr`

In this section we consider the same settings as in Section J.1, but this time we consider delayed updates. Moreover, in this section we need slightly weaker assumption.

**Assumption K.1** (Expected smoothness)**.** *We assume that function $f$ is $\mathcal{L}$-smooth in expectation w.r.t. distribution $\mathcal{D}$, i.e., there exists constant $\mathcal{L} = \mathcal{L}(f, \mathcal{D})$ such that*

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] \leq 2\mathcal{L}\left(f(x) - f(x^*)\right) \tag{135}$$

*for all $i \in [n]$ and $x \in \mathbb{R}^d$.*

**Lemma K.5.** *For all $k \geq 0$ we have*

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 4\mathcal{L}\left(f(x^k) - f(x^*)\right) + 2\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^*)\|^2\right]. \tag{136}$$

*Proof.* Applying straightforward inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for $a, b \in \mathbb{R}^d$ we get

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla f_{\xi_i}(x^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(31)}{\leq} 2\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^k) - \nabla f_\xi(x^*)\|^2\right] + 2\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right] \\
&\overset{(135)}{\leq} 4\mathcal{L}\left(f(x^k) - f(x^*)\right) + 2\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(x^*)\|^2\right].
\end{aligned}$$

□

**Theorem K.5.** *Assume that $f(x)$ is $\mu$-quasi strongly convex, $L$-smooth and Assumption K.1 holds. Then* `D-SGDsr` *satisfies Assumption 3.4 with*

$$A' = 2\mathcal{L}, \quad B_1' = B_2' = 0, \quad D_1' = 2\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2, \quad \sigma_{1,k}^2 \equiv \sigma_{2,k}^2 \equiv 0$$
$$\rho_1 = \rho_2 = 1, \quad C_1 = C_2 = 0, \quad D_2 = 0, \quad G = 0,$$
$$F_1 = F_2 = 0, \quad D_3 = 6\gamma\tau L\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2$$

*with $\gamma$ satisfying*

$$\gamma \le \min\left\{\frac{1}{8\mathcal{L}}, \frac{1}{8\sqrt{L\tau\left(L\tau + 2\mathcal{L}\right)}}\right\}$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \frac{\gamma\mu}{2}\right)^K \frac{4\|x^0 - x^*\|^2}{\gamma} + 8\gamma(1 + 3\gamma\tau L)\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4\|x^0 - x^*\|^2}{\gamma K} + 8\gamma(1 + 3\gamma\tau L)\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2$$

*when $\mu = 0$.*

In other words, `D-SGDsr` converges with linear rate $\mathcal{O}\left(\left(\frac{\mathcal{L}}{\mu} + \frac{\sqrt{L\mathcal{L}\tau + L^2\tau^2}}{\mu}\right)\ln\frac{1}{\varepsilon}\right)$ to the neighbourhood of the solution when $\mu > 0$. Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary K.10.** *Let the assumptions of Theorem K.5 hold and $\mu > 0$. Then after $K$ iterations of* `D-SGDsr` *with the stepsize*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{8\mathcal{L}}, \frac{1}{8\sqrt{L\tau\left(L\tau + 2\mathcal{L}\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,\\ \gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\frac{R_0^2\mu^2K^2}{D_1'}, \frac{R_0^2\mu^3K^3}{3\tau L D_1'}\right\}\right\}\right)}{\mu K}\right\}\end{aligned}$$

*we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\widetilde{\mathcal{O}}\left(R_0^2\left(\mathcal{L} + \sqrt{L^2\tau^2 + L\mathcal{L}\tau}\right)\exp\left(-\frac{\mu}{\tau L}K\right) + \frac{\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}{\mu K} + \frac{L\tau\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}{\mu^2 K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `D-SGDsr` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\mathcal{L} + \sqrt{L^2\tau^2 + L\mathcal{L}\tau}}{\mu} + \frac{\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}{\mu\varepsilon} + \frac{\sqrt{L\tau\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}}{\mu\sqrt{\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.11.** *Let the assumptions of Theorem K.5 hold and $\mu = 0$. Then after $K$ iterations of* `D-SGDsr` *with the stepsize*

$$\gamma = \min\left\{\frac{1}{8\mathcal{L}}, \frac{1}{8\sqrt{L\tau\left(L\tau + 2\mathcal{L}\right)}}, \sqrt{\frac{\|x^0 - x^*\|^2}{D_1' K}}, \sqrt[3]{\frac{\|x^0 - x^*\|^2}{3L\tau D_1' K}}\right\}$$

*we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2}{K} + \frac{\sqrt{L^2\tau^2 + L\mathcal{L}\tau}R_0^2}{K} + \sqrt{\frac{R_0^2\tau\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}{K}} + \frac{\sqrt[3]{LR_0^4\tau\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}}{K^{2/3}}\right)$$

*where $R_0 = \|x^0 - x^*\|$. That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \varepsilon$* `D-SGDsr` *requires*

$$\mathcal{O}\left(\frac{\mathcal{L}R_0^2}{\varepsilon} + \frac{\sqrt{L^2\tau^2 + L\mathcal{L}\tau}R_0^2}{\varepsilon} + \frac{R_0^2\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}{\varepsilon^2} + \frac{R_0^2\sqrt{L\tau\mathbf{E}_{\mathcal{D}}\|\nabla f_\xi(x^*)\|^2}}{\varepsilon^{3/2}}\right)$$

*iterations.*

**Algorithm 16** D-LSVRG

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$

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

  Broadcast $x^{k-\tau}$ to all workers

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

    Pick $l$ uniformly at random from $[m]$

    Set $g_i^{k-\tau} = \nabla f_{il}(x^{k-\tau}) - \nabla f_{il}(w_i^{k-\tau}) + \nabla f_i(w_i^{k-\tau})$

    $v_i^k = \gamma g_i^{k-\tau}$

    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$

    $w_i^{k-\tau+1} = \begin{cases} x^{k-\tau}, & \text{with probability } p, \\ w_i^{k-\tau}, & \text{with probability } 1-p \end{cases}$

  **end for**

  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$

  $x^{k+1} = x^k - v^k$

**end for**

### K.6 D-LSVRG

In the same settings as in Section J.6 we now consider a new method called D-LSVRG which is another modification of LSVRG that works with delayed updates.

**Lemma K.6.** *For all $k \geq 0$, $i \in [n]$ we have*

$$\mathbf{E}\left[g_i^k \mid x^k\right] = \nabla f_i(x^k) \tag{137}$$

*and*

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2, \tag{138}$$

*where* $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.

*Proof.* First of all, we derive unbiasedness of $g_i^k$:

$$\mathbf{E}\left[g_i^k \mid x^k\right] = \frac{1}{m}\sum_{j=1}^m \left(\nabla f_{ij}(x^k) - \nabla f_{ij}(w_i^k) + \nabla f_i(w_i^k)\right) = \nabla f_i(x^k).$$

Next, we estimate the second moment of $g^k$:

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{il}(x^k) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k)\right)\right\|^2\right] \\
&= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{il}(x^k) - \nabla f_{il}(x^*) + \nabla f_{il}(x^*) - \nabla f_{il}(w_i^k) + \nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2\right] \\
&\overset{(31)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k) - \nabla f_{il}(x^*)\|^2 \mid x^k\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right\|^2 \mid x^k\right] \\
&\overset{(34)}{\leq} \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(x^k) - \nabla f_{ij}(x^*)\|^2 + \frac{2}{n}\mathbf{E}\left[\left\|\nabla f_{il}(w_i^k) - \nabla f_{il}(x^*)\right\|^2 \mid x^k\right] \\
&\overset{(11)}{\leq} \frac{4L}{nm}\sum_{i=1}^n\sum_{j=1}^m D_{f_{ij}}(x^k, x^*) + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2 \\
&= 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2.
\end{aligned}$$

□

**Lemma K.7.** *For all $k \geq 0$, $i \in [n]$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq (1-p)\sigma_k^2 + 2Lp\left(f(x^k) - f(x^*)\right), \tag{139}$$

*where* $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.

*Proof.* The proof is identical to the proof of Lemma J.8. □

**Theorem K.6.** *Assume that $f(x)$ is $\mu$-quasi strongly convex and functions $f_{ij}$ are convex and $L$-smooth for all $i \in [n], j \in [m]$. Then* `D-LSVRG` *satisfies Assumption 3.4 with*

$$A' = 2L, \quad B_1' = 0, \quad B_2' = 2, \quad D_1' = 0, \quad \sigma_{2,k}^2 = \sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2,$$

$$\sigma_{1,k}^2 \equiv 0, \quad \rho_1 = 1, \quad \rho_2 = p, \quad C_1 = 0, \quad C_2 = Lp, \quad D_2 = 0,$$

$$G = 0, \quad F_1 = 0, \quad F_2 = \frac{12\gamma^2 L\tau(2+p)}{p}, \quad D_3 = 0$$

*with $\gamma$ satisfying*

$$\gamma \leq \min\left\{\frac{3}{56L}, \frac{1}{8L\sqrt{\tau\left(2+\tau+{}^4\!/\!(1-p)\right)}}\right\}, \quad M_2 = \frac{8}{3p}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{p}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_2 \sigma_0^2)}{\gamma}$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma K}$$

*when $\mu = 0$, where* $T^k \stackrel{def}{=} \|\tilde{x}^k - x^*\|^2 + M_2\gamma^2\sigma_k^2$.

In other words, `D-LSVRG` converges with linear rate $\mathcal{O}\left(\left(\frac{1}{p} + \kappa\sqrt{\tau\left(\tau + \frac{1}{(1-p)}\right)}\right)\ln\frac{1}{\varepsilon}\right)$ to the exact solution when $\mu > 0$. If $m \geq 2$ then taking $p = \frac{1}{m}$ we get that in expectation the sample complexity of one iteration of `D-LSVRG` is $\mathcal{O}(1)$ gradients calculations per node as for `D-SGDsr` with standard sampling and the rate of convergence to the exact solution becomes $\mathcal{O}\left((m + \kappa\tau)\ln\frac{1}{\varepsilon}\right)$.

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.12.** *Let the assumptions of Theorem K.6 hold and $\mu = 0$. Then after $K$ iterations of* `D-LSVRG` *with the stepsize*

$$\gamma = \min\left\{\frac{3}{56L}, \frac{1}{8L\sqrt{\tau\left(2+\tau+{}^4\!/\!(1-p)\right)}}, \sqrt{\frac{\|x^0 - x^*\|^2}{M_2\sigma_0^2}}, \sqrt[3]{\frac{\|x^0 - x^*\|^2 p}{12L\tau(2+p)\sigma_0^2}}\right\}$$

*and $p = \frac{1}{m}$, $m \geq 2$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{L\tau R_0^2}{K} + \frac{\sqrt{R_0^2 m \sigma_0^2}}{K} + \frac{\sqrt[3]{R_0^4 L\tau\sigma_0^2}}{K}\right)$$

*where $R_0 = \|x^0 - x^*\|$. That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* `D-LSVRG` *requires*

$$\mathcal{O}\left(\frac{L\tau R_0^2}{\varepsilon} + \frac{\sqrt{R_0^2 m\sigma_0^2}}{\varepsilon} + \frac{\sqrt[3]{R_0^4 L\tau\sigma_0^2}}{\varepsilon}\right)$$

*iterations.*

**Algorithm 17** `D-QLSVRG`

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^{k-\tau}$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Pick $l$ uniformly at random from $[m]$
    Set $\hat{g}_i^{k-\tau} = \nabla f_{il}(x^{k-\tau}) - \nabla f_{il}(w_i^{k-\tau}) + \nabla f_i(w_i^{k-\tau})$
    Set $g_i^{k-\tau} = Q(\hat{g}_i^{k-\tau})$ (quantization is performed independently from other nodes)
    $v_i^k = \gamma g_i^{k-\tau}$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    $w_i^{k-\tau+1} = \begin{cases} x^{k-\tau}, & \text{with probability } p, \\ w_i^{k-\tau}, & \text{with probability } 1-p \end{cases}$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$
  $x^{k+1} = x^k - v^k$
**end for**

### K.7 `D-QLSVRG`

In this section we add a quantization to `D-LSVRG`.

**Lemma K.8.** *For all $k \geq 0$, $i \in [n]$ we have*

$$\mathbf{E}\left[g_i^k \mid x^k\right] = \nabla f_i(x^k)$$

*and*

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 4L\left(1 + \frac{2\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + 2\left(1 + \frac{2\omega}{n}\right)\sigma_k^2 + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2,$$

*where $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

*Proof.* First of all, we derive unbiasedness of $g_i^k$:

$$\begin{aligned}\mathbf{E}\left[g_i^k \mid x^k\right] &\overset{(35)}{=} \mathbf{E}\left[\mathbf{E}_Q\left[Q(\hat{g}_i^k)\right] \mid x^k\right] \overset{(26)}{=} \mathbf{E}\left[\hat{g}_i^k \mid x^k\right] \\ &= \frac{1}{m}\sum_{j=1}^m \left(\nabla f_{ij}(x^k) - \nabla f_{ij}(w_i^k) + \nabla f_i(w_i^k)\right) = \nabla f_i(x^k).\end{aligned}$$

Next, we estimate the second moment of $g^k$:

$$\begin{aligned}\mathbf{E}_Q\left[\|g^k\|^2\right] &= \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n Q(\hat{g}_i^k)\right\|^2\right] \\ &\overset{(34)}{=} \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(Q(\hat{g}_i^k) - \hat{g}_i^k\right)\right\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n \hat{g}_i^k\right\|^2.\end{aligned}$$

Since quantization on nodes is performed independently we can decompose the first term from the last row of the previous inequality into the sum of variances:

$$\begin{aligned}\mathbf{E}_Q\left[\|g^k\|^2\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_Q\left\|Q(\hat{g}_i^k) - \hat{g}_i^k\right\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \hat{g}_i^k\right\|^2 \\ &\overset{(26)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \|\hat{g}_i^k\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \\ &\overset{(31)}{\leq} \left(1 + \frac{2\omega}{n}\right)\frac{1}{n}\sum_{i=1}^n \|\hat{g}_i^k - \nabla f_i(x^*)\|^2 + \frac{2\omega}{n^2}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2.\end{aligned}$$

Taking conditional mathematical expectation $\mathbf{E}\left[\cdot \mid x^k\right]$ from the both sides of previous inequality we get

$$
\begin{aligned}
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\le \left(1+\frac{2\omega}{n}\right)\frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{il}(x^k)-\nabla f_{il}(x^*)\|^2 \mid x^k\right]\\
&\quad + \left(1+\frac{2\omega}{n}\right)\frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k)-\nabla f_{il}(x^*) - \left(\nabla f_i(w_i^k)-\nabla f_i(x^*)\right)\right\|^2 \mid x^k\right]\\
&\quad + \frac{2\omega}{n^2}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2\\
&\le \left(1+\frac{2\omega}{n}\right)\frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(x^k)-\nabla f_{ij}(x^*)\|^2\\
&\quad + \left(1+\frac{2\omega}{n}\right)\frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{il}(w_i^k)-\nabla f_{il}(x^*)\right\|^2 \mid x^k\right] + \frac{2\omega}{n^2}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2\\
&\overset{(11)}{\le} \left(1+\frac{2\omega}{n}\right)\frac{4L}{nm}\sum_{i=1}^n\sum_{j=1}^m D_{f_{ij}}(x^k,x^*)\\
&\quad + \left(1+\frac{2\omega}{n}\right)\frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\right\|^2 + \frac{2\omega}{n^2}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2\\
&= 4L\left(1+\frac{2\omega}{n}\right)\left(f(x^k)-f(x^*)\right) + 2\left(1+\frac{2\omega}{n}\right)\sigma_k^2 + \frac{2\omega}{n^2}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2.
\end{aligned}
$$

□

**Lemma K.9.** *For all $k \ge 0$, $i\in[n]$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \le (1-p)\sigma_k^2 + 2Lp\left(f(x^k)-f(x^*)\right), \tag{140}$$

*where $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2$.*

*Proof.* The proof is identical to the proof of Lemma J.8. □

**Theorem K.7.** *Assume that $f(x)$ is $\mu$-quasi strongly convex and functions $f_{ij}$ are convex and $L$-smooth for all $i\in[n], j\in[m]$. Then* `D-QLSVRG` *satisfies Assumption 3.4 with*

$$A' = 2L\left(1+\frac{2\omega}{n}\right),\quad B_1'=0,\quad B_2' = 2\left(1+\frac{2\omega}{n}\right),\quad D_1' = \frac{2\omega}{n^2}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2,\quad \sigma_{1,0}^2\equiv 0,$$

$$\sigma_{2,k}^2 = \sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^*)\|^2,\quad \rho_1=1,\quad \rho_2=p,\quad C_2=Lp,\quad D_2=0,$$

$$C_1=0,\quad G=0,\quad F_1=0,\quad F_2=\frac{12\gamma^2L\tau\left(1+\frac{2\omega}{n}\right)(2+p)}{p},\quad D_3=\frac{6\gamma\tau L\omega}{n^2}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2$$

*with $\gamma$ satisfying*

$$\gamma \le \min\left\{\frac{3}{56L(1+{}^{2\omega}/_{n})}, \frac{1}{8L\sqrt{\tau\left(\tau+2\left(1+{}^{2\omega}/_{n}\right)\left(1+{}^{2}/_{(1-p)}\right)\right)}}\right\},\quad M_2 = \frac{8\left(1+\frac{2\omega}{n}\right)}{3p}$$

*and for all $K\ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \le \left(1-\min\left\{\frac{\gamma\mu}{2},\frac{p}{4}\right\}\right)^K\frac{4(T^0+\gamma F_2\sigma_0^2)}{\gamma} + 4\gamma\left(D_1'+D_3\right)$$

*when $\mu>0$ and*

$$\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \le \frac{4(T^0+\gamma F_2\sigma_0^2)}{\gamma K} + 4\gamma\left(D_1'+D_3\right)$$

*when $\mu=0$, where $T^k \overset{def}{=} \|\tilde{x}^k-x^*\|^2 + M_2\gamma^2\sigma_k^2$.*

In other words, D-QLSVRG converges with linear rate

$$\mathcal{O}\left(\left(\frac{1}{p}+\kappa\left(1+\frac{\omega}{n}\right)+\kappa\sqrt{\tau\left(\tau+\left(1+\frac{\omega}{n}\right)\left(1+\frac{1}{(1-p)}\right)\right)}\right)\ln\frac{1}{\varepsilon}\right)$$

to neighbourhood the solution when $\mu > 0$. If $m \geq 2$ then taking $p = \frac{1}{m}$ we get that in expectation the sample complexity of one iteration of D-QLSVRG is $\mathcal{O}(1)$ gradients calculations per node as for D-QSGDsr with standard sampling and the rate of convergence to the neighbourhood of the solution becomes

$$\mathcal{O}\left(\left(m+\kappa\left(1+\frac{\omega}{n}\right)+\kappa\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)\ln\frac{1}{\varepsilon}\right).$$

Applying Lemma D.2 we establish the rate of convergence to $\varepsilon$-solution.

**Corollary K.13.** *Let the assumptions of Theorem K.7 hold, $f_\xi(x)$ are convex for each $\xi$ and $\mu > 0$. Then after $K$ iterations of* D-QLSVRG *with the stepsize*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{3}{56L(1+2\omega/n)}, \frac{1}{8L\sqrt{\tau\left(\tau+2\left(1+2\omega/n\right)\left(1+2/(1-p)\right)\right)}}\right\}, \quad R_0 = \|x^0-x^*\|,\\ \gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\frac{R_0^2\mu^2K^2}{D_1'}, \frac{R_0^2\mu^3K^3}{3\tau LD_1'}\right\}\right\}\right)}{\mu K}\right\}\end{aligned}$$

*and $p = \frac{1}{m}$, $m \geq 2$ we have $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]$ of order*

$$\widetilde{\mathcal{O}}\left(LR_0^2\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)\exp\left(-\frac{\mu}{L\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)}K\right)+\frac{D_1'}{\mu K}+\frac{L\tau D_1'}{\mu^2K^2}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \leq \varepsilon$* D-QLSVRG *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\left(1+\frac{\omega}{n}\right)+\frac{L}{\mu}\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}+\frac{D_1'}{\mu\varepsilon}+\frac{\sqrt{L\tau D_1'}}{\mu\sqrt{\varepsilon}}\right) \text{ iterations.}$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.14.** *Let the assumptions of Theorem K.7 hold and $\mu = 0$. Then after $K$ iterations of* D-QLSVRG *with the stepsize*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{3}{56L(1+2\omega/n)}, \frac{1}{8L\sqrt{\tau\left(\tau+2\left(1+2\omega/n\right)\left(1+2/(1-p)\right)\right)}}\right\}, \quad R_0 = \|x^0-x^*\|,\\ \gamma &= \min\left\{\gamma_0, \sqrt{\frac{R_0^2}{M_2\sigma_0^2}}, \sqrt[3]{\frac{R_0^2p}{12L\tau\left(1+\frac{2\omega}{n}\right)(2+p)}}, \sqrt{\frac{R_0^2}{D_1'K}}, \sqrt[3]{\frac{R_0^2}{3L\tau D_1'K}}\right\}\end{aligned}$$

*and $p = \frac{1}{m}$, $m \geq 2$ we have $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)}{K}+\frac{\sqrt{R_0^2m\left(1+\frac{\omega}{n}\right)\sigma_0^2}}{K}+\frac{\sqrt[3]{R_0^4L\tau m\left(1+\frac{\omega}{n}\right)}}{K}\right)$$
$$+\mathcal{O}\left(\sqrt{\frac{R_0^2D_1'}{K}}+\frac{\sqrt[3]{LR_0^4\tau D_1'}}{K^{2/3}}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K)-f(x^*)\right] \leq \varepsilon$* D-QLSVRG *requires*

$$\mathcal{O}\left(\frac{LR_0^2\left(1+\frac{\omega}{n}+\sqrt{\tau\left(\tau+\frac{\omega}{n}\right)}\right)}{\varepsilon}+\frac{\sqrt{R_0^2m\left(1+\frac{\omega}{n}\right)\sigma_0^2}}{\varepsilon}+\frac{\sqrt[3]{R_0^4L\tau m\left(1+\frac{\omega}{n}\right)}}{\varepsilon}\right)$$
$$+\mathcal{O}\left(\frac{R_0^2D_1'}{\varepsilon^2}+\frac{R_0^2\sqrt{L\tau D_1'}}{\varepsilon^{3/2}}\right)$$

*iterations.*

**Algorithm 18** D-QLSVRGstar

**Input:** learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^{k-\tau}$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Pick $l$ uniformly at random from $[m]$
    Set $\hat{g}_i^{k-\tau} = \nabla f_{il}(x^{k-\tau}) - \nabla f_{il}(w_i^{k-\tau}) + \nabla f_i(w_i^{k-\tau})$
    Set $g_i^{k-\tau} = Q(\hat{g}_i^{k-\tau} - \nabla f_i(x^*))$ (quantization is performed independently from other nodes)
    $v_i^k = \gamma g_i^{k-\tau}$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    $w_i^{k-\tau+1} = \begin{cases} x^{k-\tau}, & \text{with probability } p, \\ w_i^{k-\tau}, & \text{with probability } 1-p \end{cases}$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k$
  $x^{k+1} = x^k - v^k$
**end for**

### K.8 D-QLSVRGstar

Now we assume that $i$-th node has an access to $\nabla f_i(x^*)$ and modify D-QLSVRG in order to get convergence asymptotically to the exact optimum.

**Lemma K.10.** *For all $k \geq 0$, $i \in [n]$ we have*

$$\mathbf{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \tag{141}$$

*and*

$$\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] \leq 2L\left(1 + \frac{\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + 2\left(1 + \frac{\omega}{n}\right)\sigma_k^2, \tag{142}$$

*where $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

*Proof.* First of all, we derive unbiasedness of $g_i^k$:

$$\begin{aligned}
\mathbf{E}\left[g^k \mid x^k\right] &\overset{(35)}{=} \mathbf{E}\left[\mathbf{E}_Q\left[\frac{1}{n}\sum_{i=1}^n Q(\hat{g}_i^k - \nabla f_i(x^*))\right] \mid x^k\right] \overset{(26)}{=} \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right) \mid x^k\right] \\
&= \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m \left(\nabla f_{ij}(x^k) - \nabla f_{ij}(w_i^k) + \nabla f_i(w_i^k)\right) = \nabla f(x^k).
\end{aligned}$$

Next, we estimate the second moment of $g^k$:

$$\begin{aligned}
\mathbf{E}_Q\left[\|g^k\|^2\right] &= \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n Q(\hat{g}_i^k - \nabla f_i(x^*))\right\|^2\right] \\
&\overset{(34)}{=} \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(Q(\hat{g}_i^k - \nabla f_i(x^*)) - \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right)\right\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n \hat{g}_i^k - \nabla f_i(x^*)\right\|^2.
\end{aligned}$$

Since quantization on nodes is performed independently we can decompose the first term from the last row of the previous inequality into the sum of variances:

$$\begin{aligned}
\mathbf{E}_Q\left[\|g^k\|^2\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_Q\left\|Q(\hat{g}_i^k - \nabla f_i(x^*)) - \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \hat{g}_i^k - \nabla f_i(x^*)\right\|^2 \\
&\overset{(26)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \|\hat{g}_i^k - \nabla f_i(x^*)\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \\
&\overset{(31)}{\leq} \left(1 + \frac{\omega}{n}\right)\frac{1}{n}\sum_{i=1}^n \|\hat{g}_i^k - \nabla f_i(x^*)\|^2.
\end{aligned}$$

Taking conditional mathematical expectation $\mathbf{E}\left[\cdot \mid x^k\right]$ from the both sides of previous inequality and using the bound

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^*)\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2$$

implicitly obtained in the proof of Lemma K.8 we get (142). □

**Lemma K.11.** *For all $k \geq 0$, $i \in [n]$ we have*

$$\mathbf{E}\left[\sigma_{k+1}^2 \mid x^k\right] \leq (1-p)\sigma_k^2 + 2Lp\left(f(x^k) - f(x^*)\right), \tag{143}$$

*where* $\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^n \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.

*Proof.* The proof is identical to the proof of Lemma J.8. □

**Theorem K.8.** *Assume that $f(x)$ is $\mu$-quasi strongly convex and functions $f_{ij}$ are convex and $L$-smooth for all $i \in [n], j \in [m]$. Then* `D-QLSVRGstar` *satisfies Assumption 3.4 with*

$$A' = 2L\left(1 + \frac{2\omega}{n}\right), \quad B_1' = 0, \quad B_2' = 2\left(1 + \frac{2\omega}{n}\right), \quad D_1' = 0, \quad \sigma_{1,0}^2 \equiv 0,$$

$$\sigma_{2,k}^2 = \sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2, \quad \rho_1 = 1, \quad \rho_2 = p, \quad C_2 = Lp, \quad D_2 = 0,$$

$$C_1 = 0, \quad G = 0, \quad F_1 = 0, \quad F_2 = \frac{12\gamma^2 L\tau\left(1 + \frac{2\omega}{n}\right)(2+p)}{p}, \quad D_3 = 0$$

*with $\gamma$ satisfying*

$$\gamma \leq \min\left\{\frac{3}{56L(1 + {}^{2\omega}/{}_n)}, \frac{1}{8L\sqrt{\tau\left(\tau + 2\left(1 + {}^{2\omega}/{}_n\right)\left(1 + {}^{2}/{}_{(1-p)}\right)\right)}}\right\}, \quad M_2 = \frac{8\left(1 + \frac{2\omega}{n}\right)}{3p}$$

*and for all $K \geq 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{p}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma}$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{4(T^0 + \gamma F_2\sigma_0^2)}{\gamma K}$$

*when $\mu = 0$, where $T^k \overset{def}{=} \|\tilde{x}^k - x^*\|^2 + M_2\gamma^2\sigma_k^2$.*

In other words, `D-QLSVRGstar` converges with linear rate

$$\mathcal{O}\left(\left(\frac{1}{p} + \kappa\left(1 + \frac{\omega}{n}\right) + \kappa\sqrt{\tau\left(\tau + \left(1 + \frac{\omega}{n}\right)\left(1 + \frac{1}{(1-p)}\right)\right)}\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution when $\mu > 0$. If $m \geq 2$ then taking $p = \frac{1}{m}$ we get that in expectation the sample complexity of one iteration of `D-QLSVRGstar` is $\mathcal{O}(1)$ gradients calculations per node as for `D-QSGDsr` with standard sampling and the rate of convergence to the exact solution becomes

$$\mathcal{O}\left(\left(m + \kappa\left(1 + \frac{\omega}{n}\right) + \kappa\sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)\ln\frac{1}{\varepsilon}\right).$$

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.15.** *Let the assumptions of Theorem K.8 hold and $\mu = 0$. Then after $K$ iterations of* `D-QLSVRGstar` *with the stepsize*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{3}{56L(1+2\omega/n)}, \frac{1}{8L\sqrt{\tau\left(\tau + 2\left(1+2\omega/n\right)\left(1+2/(1-p)\right)\right)}}\right\}, \quad R_0 = \|x^0 - x^*\|,\\
\gamma &= \min\left\{\gamma_0, \sqrt{\frac{R_0^2}{M_2\sigma_0^2}}, \sqrt[3]{\frac{R_0^2 p}{12L\tau\left(1+\frac{2\omega}{n}\right)(2+p)}}\right\}
\end{aligned}$$

*and $p = \frac{1}{m}$, $m \geq 2$ we have $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ of order*

$$\mathcal{O}\left(\frac{LR_0^2\left(1 + \frac{\omega}{n} + \sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)}{K} + \frac{\sqrt{R_0^2 m\left(1+\frac{\omega}{n}\right)\sigma_0^2}}{K} + \frac{\sqrt[3]{R_0^4 L\tau m\left(1+\frac{\omega}{n}\right)}}{K}\right).$$

*That is, to achive $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$* `D-QLSVRGstar` *requires*

$$\mathcal{O}\left(\frac{LR_0^2\left(1 + \frac{\omega}{n} + \sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)}{\varepsilon} + \frac{\sqrt{R_0^2 m\left(1+\frac{\omega}{n}\right)\sigma_0^2}}{\varepsilon} + \frac{\sqrt[3]{R_0^4 L\tau m\left(1+\frac{\omega}{n}\right)}}{\varepsilon}\right)$$

*iterations.*

However, such convergence guarantees are obtained under very restrictive assumption: the method requires to know vectors $\nabla f_i(x^*)$.

### K.9 `D-LSVRG-DIANA`

In the setup of Section K.6 we construct a new method with delayed updates and quantization called `D-LSVRG-DIANA` which does not require to know $\nabla f_i(x^*)$ and has linear convergence to the exact solution.

**Lemma K.12.** *Assume that $f_{ij}(x)$ is convex and $L$-smooth for all $i = 1,\ldots,n$, $j = 1,\ldots,m$. Then, for all $k \geq 0$ we have*

$$\begin{aligned}
\mathbf{E}\left[g^k \mid x^k\right] &= \nabla f(x^k), &(144)\\
\mathbf{E}\left[\|g^k\|^2 \mid x^k\right] &\leq 4L\left(1+\frac{2\omega}{n}\right)\left(f(x^k) - f(x^*)\right) + \frac{2\omega}{n}\sigma_{1,k}^2 + 2\left(1+\frac{2\omega}{n}\right)\sigma_{2,k}^2 &(145)
\end{aligned}$$

*where $\sigma_{1,k}^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f(x^*)\|^2$ and $\sigma_{2,k}^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

*Proof.* First of all, we show unbiasedness of $g^k$:

$$\begin{aligned}
\mathbf{E}\left[g^k \mid x^k\right] &\overset{(35)}{=} h^k + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\mathbf{E}_Q\left[\hat{\Delta}_i^k\right] \mid x^k\right] \overset{(26)}{=} h^k + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\hat{g}_i^k - h_i^k \mid x^k\right]\\
&= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left(\nabla f_{ij}(x^k) - \nabla f_{ij}(w_i^k) + \nabla f_i(w_i^k)\right) = \nabla f(x^k).
\end{aligned}$$

**Algorithm 19** D-LSVRG-DIANA

**Input:** learning rates $\gamma > 0$, $\alpha \in (0,1]$, initial vectors $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$

Set $e_i^0 = 0$ for all $i = 1, \ldots, n$
Set $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$
**for** $k = 0, 1, \ldots$ **do**
  Broadcast $x^{k-\tau}$ to all workers
  **for** $i = 1, \ldots, n$ in parallel **do**
    Pick $l$ uniformly at random from $[m]$
    Set $\hat{g}_i^{k-\tau} = \nabla f_{il}(x^{k-\tau}) - \nabla f_{il}(w_i^{k-\tau}) + \nabla f_i(w_i^{k-\tau})$
    $\hat{\Delta}_i^{k-\tau} = Q(\hat{g}_i^{k-\tau} - h_i^{k-\tau})$ (quantization is performed independently from other nodes)
    $g_i^{k-\tau} = h_i^{k-\tau} + \hat{\Delta}_i^{k-\tau}$
    $v_i^k = \gamma g_i^{k-\tau}$
    $e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$
    $h_i^{k-\tau+1} = h_i^{k-\tau} + \alpha \hat{\Delta}_i^{k-\tau}$
  **end for**
  $e^k = \frac{1}{n}\sum_{i=1}^n e_i^k$, $g^k = \frac{1}{n}\sum_{i=1}^n g_i^k = h^k + \frac{1}{n}\sum\limits_{i=1}^n \hat{\Delta}_i^k$, $v^k = \frac{1}{n}\sum_{i=1}^n v_i^k = \gamma h^{k-\tau} + \frac{\gamma}{n}\sum_{i=1}^n \hat{\Delta}_i^{k-\tau}$
  $h^{k-\tau+1} = \frac{1}{n}\sum\limits_{i=1}^n h_i^{k-\tau+1} = h^{k-\tau} + \alpha\frac{1}{n}\sum\limits_{i=1}^n \hat{\Delta}_i^{k-\tau}$
  $x^{k+1} = x^k - v^k$
**end for**

Next, we derive the upper bound for the second moment of $g^k$:

$$
\begin{aligned}
\mathbf{E}_Q\left[\|g^k\|^2\right] &= \mathbf{E}_Q\left[\left\|h^k + \frac{1}{n}\sum_{i=1}^n \hat{\Delta}_i^k\right\|^2\right] \\
&\overset{(34)}{=} \mathbf{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{\Delta}_i^k - \hat{g}_i^k + h_i^k\right)\right\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n \hat{g}_i^k\right\|^2.
\end{aligned}
$$

Since quantization on nodes is performed independently we can decompose the first term from the last row of the previous inequality into the sum of variances:

$$
\begin{aligned}
\mathbf{E}_Q\left[\|g^k\|^2\right] &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_Q\left[\|\hat{\Delta}_i^k - \hat{g}_i^k + h_i^k\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{g}_i^k - \nabla f_i(x^*)\right)\right\|^2 \\
&\overset{(26),(31)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \|\hat{g}_i^k - h_i^k\|^2 + \frac{1}{n}\sum_{i=1}^n \|\hat{g}_i^k - \nabla f_i(x^*)\|^2 \\
&\overset{(31)}{\leq} \left(1 + \frac{2\omega}{n}\right)\frac{1}{n}\sum_{i=1}^n \|\hat{g}_i^k - \nabla f_i(x^*)\|^2 + \frac{2\omega}{n^2}\sum_{i=1}^n \|h_i^k - f_i(x^*)\|^2.
\end{aligned}
$$

Taking mathematical expectation $\mathbf{E}\left[\cdot \mid x^k\right]$ from the both sides of the previous inequality and using the bound

$$
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\hat{g}_i^k - \nabla f_i(x^*)\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2
$$

implicitly obtained in the proof of Lemma K.8 we get (145). □

**Lemma K.13.** *Assume that $\alpha \leq 1/(\omega+1)$. Then, for all $k \geq 0$ we have*

$$
\mathbf{E}\left[\sigma_{1,k+1}^2 \mid x^k\right] \leq (1-\alpha)\sigma_{1,k}^2 + 6L\alpha(f(x^k) - f(x^*)) + 2\alpha\sigma_{2,k}^2,
$$

$$\mathbf{E}\left[\sigma_{2,k+1}^2 \mid x^k\right] \le (1-p)\sigma_{k,2}^2 + 2Lp\left(f(x^k) - f(x^*)\right)$$

*where* $\sigma_{1,k}^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$ *and* $\sigma_{2,k}^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.

*Proof.* The proof is identical to the proof of Lemma J.12. □

**Theorem K.9.** *Assume that $f_{ij}(x)$ is convex and $L$-smooth for all $i = 1,\ldots,n$, $j = 1,\ldots,m$ and $f(x)$ is $\mu$-quasi strongly convex. Then* `D-LSVRG-DIANA` *satisfies Assumption 3.4 with*

$$A' = 2L\left(1 + \frac{2\omega}{n}\right), \quad B_1' = \frac{2\omega}{n}, \quad B_2' = 2\left(1 + \frac{2\omega}{n}\right), \quad D_1' = 0,$$

$$\sigma_{1,k}^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2, \quad \sigma_{2,k}^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2,$$

$$\rho_1 = \alpha, \quad \rho_2 = p, \quad C_1 = 3L\alpha, \quad C_2 = Lp, \quad D_2 = 0, \quad G = 2,$$

$$F_1 = \frac{12\gamma^2 L\omega\tau(2+\alpha)}{n\alpha}, \quad F_2 = \frac{12\gamma^2\tau L(2+p)}{p}\left(\frac{4\omega}{n(1-\alpha)} + 1 + \frac{2\omega}{n}\right), \quad D_3 = 0$$

*with $\gamma$ and $\alpha$ satisfying*

$$\gamma \le \min\left\{\frac{1}{8L\left(\frac{37}{9} + \frac{24\omega}{3n}\right)}, \frac{1}{8L\sqrt{\tau\left(2 + \tau + \frac{4}{1-p} + \frac{4\omega}{n}\left(1 + \frac{3}{1-\alpha} + \frac{2}{1-p} + \frac{4}{(1-\alpha)(1-p)}\right)\right)}}\right\},$$

$$\alpha \le \frac{1}{\omega+1}, \quad M_1 = \frac{8\omega}{3n\alpha}, \quad M_2 = \frac{8\left(7 + \frac{6\omega}{n}\right)}{9p}.$$

*and for all $K \ge 0$*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\alpha}{4}, \frac{p}{4}\right\}\right)^K \frac{4(T^0 + \gamma F_1\sigma_{1,0}^2 + \gamma F_2\sigma_{2,0}^2)}{\gamma}$$

*when $\mu > 0$ and*

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \le \frac{4(T^0 + \gamma F_1\sigma_{1,0}^2 + \gamma F_2\sigma_{2,0}^2)}{\gamma K}$$

*when $\mu = 0$, where* $T^k \stackrel{def}{=} \|\tilde{x}^k - x^*\|^2 + M_1\gamma^2\sigma_{1,k}^2 + M_2\gamma^2\sigma_{2,k}^2$.

In other words, if $m \ge 2$, $p = 1/m$, $\alpha = \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\}$ and

$$\gamma \le \min\left\{\frac{1}{8L\left(\frac{37}{9} + \frac{24\omega}{3n}\right)}, \frac{1}{8L\sqrt{\tau\left(2 + \tau + \frac{4}{1-p} + \frac{4\omega}{n}\left(1 + \frac{3}{1-\alpha} + \frac{2}{1-p} + \frac{4}{(1-\alpha)(1-p)}\right)\right)}}\right\},$$

`D-LSVRG-DIANA` converges with the linear rate

$$\mathcal{O}\left(\left(\omega + m + \kappa\left(1 + \frac{\omega}{n}\right) + \kappa\sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)\ln\frac{1}{\varepsilon}\right)$$

to the exact solution when $\mu > 0$.

Applying Lemma D.3 we get the complexity result in the case when $\mu = 0$.

**Corollary K.16.** *Let the assumptions of Theorem K.9 hold and $\mu = 0$. Then after $K$ iterations of* `D-LSVRG-DIANA` *with the stepsize*

$$\gamma_0 = \min\left\{\frac{1}{8L\left(\frac{37}{9} + \frac{24\omega}{3n}\right)}, \frac{1}{8L\sqrt{\tau\left(2 + \tau + \frac{4}{1-p} + \frac{4\omega}{n}\left(1 + \frac{3}{1-\alpha} + \frac{2}{1-p} + \frac{4}{(1-\alpha)(1-p)}\right)\right)}}\right\},$$

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{R_0^2}{M_1\sigma_{1,0}^2 + M_2\sigma_{2,0}^2}}, \sqrt[3]{\frac{R_0^2}{12\tau L\left(\frac{\omega(2+\alpha)}{n\alpha} + \frac{2+p}{p}\left(1 + \frac{2\omega}{n} + \frac{4\omega}{n(1-\alpha)}\right)\right)}}\right\},$$

*where* $R_0 = \|x^0 - x^*\|$, $\alpha = \min\left\{\frac{1}{\omega+1}, \frac{1}{2}\right\}$ *and* $p = \frac{1}{m}$, $m \geq 2$ *we have* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right]$ *of order*

$$\mathcal{O}\left(\frac{LR_0^2\left(1 + \frac{\omega}{n} + \sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)}{K} + \frac{\sqrt{R_0^2\omega(\omega+1)\sigma_{1,0}^2}}{\sqrt{n}K} + \frac{\sqrt{R_0^2 m\left(1 + \frac{\omega}{n}\right)\sigma_{2,0}^2}}{K}\right)$$
$$+\mathcal{O}\left(\frac{\sqrt[3]{R_0^4\tau L\omega(\omega+1)\sigma_{1,0}^2}}{\sqrt[3]{n}K} + \frac{\sqrt[3]{R_0^4\tau Lm\left(1 + \frac{\omega}{n}\right)\sigma_{2,0}^2}}{K}\right)$$

*That is, to achive* $\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \varepsilon$ `D-LSVRG-DIANA` *requires*

$$\mathcal{O}\left(\frac{LR_0^2\left(1 + \frac{\omega}{n} + \sqrt{\tau\left(\tau + \frac{\omega}{n}\right)}\right)}{\varepsilon} + \frac{\sqrt{R_0^2\omega(\omega+1)\sigma_{1,0}^2}}{\sqrt{n}\varepsilon} + \frac{\sqrt{R_0^2 m\left(1 + \frac{\omega}{n}\right)\sigma_{2,0}^2}}{\varepsilon}\right)$$
$$+\mathcal{O}\left(\frac{\sqrt[3]{R_0^4\tau L\omega(\omega+1)\sigma_{1,0}^2}}{\sqrt[3]{n}\varepsilon} + \frac{\sqrt[3]{R_0^4\tau Lm\left(1 + \frac{\omega}{n}\right)\sigma_{2,0}^2}}{\varepsilon}\right)$$

*iterations.*

Table 5: The parameters for which the methods from Tables 1 and 4 satisfy Assumption 3.4. The meaning of the expressions appearing in the table, as well as their justification is defined in details in the Sections J and K. Symbols: $\varepsilon$ = error tolerance; $\delta$ = contraction factor of compressor $\mathcal{C}$; $\omega$ = variance parameter of compressor $\mathcal{Q}$; $\kappa = {}^L\!/\!_\mu$; $\mathcal{L}$ = expected smoothness constant; $\sigma_*^2$ = variance of the stochastic gradients in the solution; $\zeta_*^2$ = average of $\|\nabla f_i(x^*)\|^2$; $\sigma^2$ = average of the uniform bounds for the variances of stochastic gradients of workers.

| Method | $A'$ | $B_1'$ | $B_2'$ | $\rho_1$ | $\rho_2$ | $C_1$ | $C_2$ | $F_1,\ F_2$ | $G$ | $D_1',\ D_2,\ D_3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| EC-SGDsr | $2\mathcal{L}$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $\frac{2\sigma_*^2}{n},\ 0,\ \frac{6L\gamma}{\delta}\left(\frac{4\zeta_*^2}{\delta}+3\sigma_*^2\right)$ |
| EC-SGD | $2L$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $\frac{2\sigma_*^2}{n},\ 0,\ \frac{12L\gamma}{\delta}\left(\frac{2\zeta_*^2}{\delta}+\sigma_*^2\right)$ |
| EC-GDstar | $L$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $0,\ 0,\ 0$ |
| EC-SGD-DIANA | $L$ | 0 | 0 | $\alpha$ | 1 | $L\alpha$ | 0 | $\frac{96L\gamma^2}{\delta^2\alpha(1-\eta)},\ 0$ | 0 | $\frac{\sigma^2}{n},\ \alpha^2(\omega+1)\sigma^2,\ \frac{6L\gamma}{\delta}\left(\frac{4\alpha(\omega+1)}{\delta}+1\right)\sigma^2$ |
| EC-SGDsr-DIANA | $2\mathcal{L}$ | 0 | 0 | $\alpha$ | 1 | $2\alpha(3\mathcal{L}+4L)$ | 0 | $\frac{96L\gamma^2}{\delta^2\alpha(1-\eta)},\ 0$ | 0 | $\frac{2\sigma_*^2}{n},\ \alpha^2(\omega+1)\sigma_*^2,\ \frac{18L\gamma}{\delta}\left(\frac{4\alpha(\omega+1)}{\delta}+1\right)\sigma_*^2$ |
| EC-LSVRG | $2L$ | 0 | 2 | 1 | $p$ | 0 | $Lp$ | $0,\ \frac{72L\gamma^2}{\delta p(1-\eta)}$ | 0 | $0,\ 0,\ \frac{24L\gamma}{\delta^2}\zeta_*^2$ |
| EC-LSVRGstar | $2L$ | 0 | 2 | 1 | $p$ | 0 | $Lp$ | $0,\ \frac{48L\gamma^2}{\delta p}$ | 0 | $0,\ 0,\ 0$ |
| EC-LSVRG-DIANA | $2L$ | 0 | 2 | $\alpha$ | $p$ | $3L\alpha$ | $Lp$ | $\frac{24L\gamma^2\left(\frac{4}{\delta}+3\right)}{\delta\alpha(1-\eta)},\ \frac{24L\gamma^2\left(\frac{4}{1-\alpha}\left(\frac{4}{\delta}+3\right)+3\right)}{\delta p(1-\eta)}$ | 2 | $0,\ 0,\ 0$ |
| D-SGDsr | $2\mathcal{L}$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $\frac{2\sigma_*^2}{n},\ 0,\ \frac{6L\tau\gamma\sigma_*^2}{n}$ |
| D-SGD | $2L$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $\frac{2\sigma_*^2}{n},\ 0,\ \frac{6L\tau\gamma\sigma_*^2}{n}$ |
| D-QSGD | $L\left(1+\frac{2\omega}{n}\right)$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $\frac{(\omega+1)\sigma^2}{n}+\frac{2\omega\zeta_*^2}{n},\ 0,\ \frac{3\gamma\tau L}{n}\left((\omega+1)\sigma^2+2\omega\zeta_*^2\right)$ |
| D-QSGDstar | $L\left(1+\frac{\omega}{n}\right)$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $\frac{(\omega+1)\sigma^2}{n},\ 0,\ \frac{3\gamma\tau L(\omega+1)\sigma^2}{n}$ |
| D-QGDstar | $L\left(1+\frac{\omega}{n}\right)$ | 0 | 0 | 1 | 1 | 0 | 0 | $0,\ 0$ | 0 | $0,\ 0,\ 0$ |
| D-SGD-DIANA | $L\left(1+\frac{2\omega}{n}\right)$ | $\frac{2\omega}{n}$ | 0 | $\alpha$ | 1 | $L\alpha$ | 0 | $\frac{12\gamma^2L\omega\tau(2+\alpha)}{n\alpha},\ 0$ | 0 | $\frac{(\omega+1)\sigma^2}{n},\ \frac{\alpha(\omega+1)\sigma^2}{n},\ 3\gamma\tau L\left(1+\frac{4\omega}{n}\right)\frac{(\omega+1)\sigma^2}{n}$ |
| D-LSVRG | $2L$ | 0 | 2 | 1 | $p$ | 0 | $Lp$ | $0,\ \frac{12\gamma^2L\tau(2+p)}{np}$ | 0 | $0,\ 0,\ 0$ |
| D-QLSVRG | $2L\left(1+\frac{2\omega}{n}\right)$ | 0 | $2\left(1+\frac{2\omega}{n}\right)$ | 1 | $p$ | 0 | $Lp$ | $0,\ \frac{12\gamma^2L\tau\left(1+\frac{2\omega}{n}\right)\tau(2+p)}{p}$ | 0 | $\frac{2\omega\zeta_*^2}{n},\ 0,\ \frac{6\gamma\tau L\omega\zeta_*^2}{n}$ |
| D-QLSVRGstar | $2L\left(1+\frac{2\omega}{n}\right)$ | 0 | $2\left(1+\frac{2\omega}{n}\right)$ | 1 | $p$ | 0 | $Lp$ | $0,\ \frac{12\gamma^2L\left(1+\frac{2\omega}{n}\right)\tau(2+p)}{p}$ | 0 | $0,\ 0,\ 0$ |
| D-LSVRG-DIANA | $2L\left(1+\frac{2\omega}{n}\right)$ | $\frac{2\omega}{n}$ | $2\left(1+\frac{2\omega}{n}\right)$ | $\alpha$ | $p$ | $3L\alpha$ | $Lp$ | $\frac{12\gamma^2L\omega\tau(2+\alpha)}{n\alpha},\ \frac{12\gamma^2\tau L(2+p)}{p}\left(1+\frac{2\omega(3-\alpha)}{n(1-\alpha)}\right)$ | 0 | $0,\ 0,\ 0$ |