Journal of Machine Learning Research 21 (2020) 1-36 Submitted 9/19; Revised 6/20; Published 9/20

The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Updates

Sebastian U. Stich SEBASTIAN.STICH@EPFL.CH
EPFL, Lausanne, Switzerland

Sai Praneeth Karimireddy SAI.KARIMIREDDY@EPFL.CH
EPFL, Lausanne, Switzerland

Editor: Mark Schmidt

Abstract

We analyze (stochastic) gradient descent (SGD) with delayed updates on smooth quasi-convex and non-convex functions and derive concise, non-asymptotic, convergence rates. We show that the rate of convergence in all cases consists of two terms: (i) a stochastic term which is not affected by the delay, and (ii) a higher order deterministic term which is only linearly slowed down by the delay. Thus, in the presence of noise, the effects of the delay become negligible after a few iterations and the algorithm converges at the same optimal rate as standard SGD. This result extends a line of research that showed similar results in the asymptotic regime or for strongly-convex quadratic functions only.

We further show similar results for SGD with more intricate form of delayed gradients—compressed gradients under error compensation and for local SGD where multiple workers perform local steps before communicating with each other. In all of these settings, we improve upon the best known rates.

These results show that SGD is robust to compressed and/or delayed stochastic gradient updates. This is in particular important for distributed parallel implementations, where asynchronous and communication efficient methods are the key to achieve linear speedups for optimization with multiple devices.

Keywords: Delayed Gradients, Error-Compensation, Error-Feedback, Gradient Compression, Local SGD, Machine Learning, Optimization, Stochastic Gradient Descent

1. Introduction

We consider the unconstrained optimization problem

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})\,, \tag{1}$$

for a quasi-convex (i.e. 1-quasar convex) or non-convex smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ and study a variety of stochastic gradient methods with delayed (or *stale*) updates. Stochastic gradient descent (SGD) methods (Robbins and Monro, 1951) generate a sequence $\{\mathbf{x}_t\}_{t\geq 0}$ of iterates for an arbitrary starting point $\mathbf{x}_0 \in \mathbb{R}^d$ and positive stepsizes $\{\gamma_t\}_{t\geq 0}$, by sequential updates of the form

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{g}_t\,, \qquad \text{where} \qquad \mathbf{g}_t = \nabla f(\mathbf{x}_t) + \boldsymbol{\xi}_t\,, \tag{SGD}$$

is a *stochastic gradient* for zero-mean noise terms $\{\boldsymbol{\xi}_t\}_{t\geq 0}$. When the noise is zero almost surely, then we recover the classic *gradient descent* method as a special case. SGD is the state of the art optimization method for many machine learning—especially deep learning—optimization problems (Bottou, 2010). In order to use the compute power of many parallel devices, it is essential to depart from the inherently serial updates as in (SGD). For instance, in mini-batch SGD (Dekel et al., 2012) several stochastic gradients are computed at the same iterate $\mathbf{x}_t$ (an operation which can be parallelized, but still requires synchronization among the devices). Fully asynchronous methods, where the devices operate completely independently, and e.g. write their updates to a shared memory (Niu et al., 2011) perform often better in practice, as the effect of stragglers (slow devices) is minimized. In an orthogonal line of work, gradient compression techniques have been developed with the aim to reduce the communication overhead between the devices (Alistarh et al., 2017). We analyze methods of both these types in this paper.

Stochastic gradient descent on μ-strongly convex functions has asymptotically the iteration complexity $\mathcal{O}\big(\frac{\sigma^2}{\mu\epsilon}\big)$ for sufficiently small $\epsilon \to 0$ (Polyak, 1990; Nemirovski et al., 2009) and where here σ^2 is an upper bound on the noise, $\mathbb{E}\,\|\boldsymbol{\xi}_t\|^2 \leq \sigma^2$, $\forall t \geq 0$ (we discuss more general bounds below). Chaturapruek et al. (2015) show that under certain regularity conditions, asynchronous SGD reaches the same asymptotic convergence rate as the standard (serial) SGD. For gradient compression techniques with *error compensation*—a technique first described in e.g. (Seide et al., 2014; Strom, 2015)—Stich et al. (2018) show that the asymptotic convergence rate $\mathcal{O}\big(\frac{G^2}{\mu\epsilon}\big)$, where G^2 denotes an upper bound on the second moment of the stochastic gradients, is attained for host of compression operators, such as e.g. sparsification, quantization or (biased) greedy selection. These two results show that asynchronous methods and gradient compression can both be used to hide communication overheads—asymptotically—*for free*. Thus they are very interesting techniques for distributed optimization. In this work we aim to derive tight (non-asymptotic) convergence rates to deepen our understanding of these schemes.

The starting point for our analysis is the recent work of (Arjevani et al., 2020) that studies the *delayed SGD* (D-SGD) algorithm for a fixed (integer) delay $\tau \geq 1$, given as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_{t-\tau}\mathbf{g}_{t-\tau}\,, \tag{D-SGD}$$

for $t \geq \tau$, and $\mathbf{x}_0 = \mathbf{x}_1 = \cdots = \mathbf{x}_\tau$ for the first iterations. Here $\mathbf{g}_{t-\tau} = \nabla f(\mathbf{x}_{t-\tau}) + \boldsymbol{\xi}_{t-\tau}$ is a stochastic gradient computed at $\mathbf{x}_{t-\tau}$, instead of at $\mathbf{x}_t$ as in the vanilla scheme (SGD). Arjevani et al. (2020) analyze (D-SGD) on convex quadratic functions. For μ-strongly convex, L-smooth quadratic function with minimum at $\mathbf{x}^\star$, they show that the suboptimality gap decreases as $\tilde{\mathcal{O}}\big(L\,\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \exp\big[-\frac{\mu T}{10 L\tau}\big] + \frac{\sigma^2}{\mu T}\big)$ after T iterations, i.e. the algorithm achieves iteration complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{L\tau}{\mu}\log\frac{1}{\epsilon}\big)$.[1] Again, we see that asymptotically, when $T \to \infty$ or $\epsilon \to 0$, the effect of the delay τ is negligible when $\sigma^2 > 0$. The delay τ only appears in the so-called optimization term that is only dominant for small σ^2 (and especially for deterministic delayed gradient descent where $\sigma^2 = 0$). The linear dependency on τ is optimal and cannot further be improved. These results were obtained with a technique based on generating functions—an approach that seems limited to quadratic functions. In this work we use the error-feedback framework to extend their results to general convex, and non-convex functions. Further, we also analyze more intricate forms of delays in the gradients—compressed gradients with error compensation, and local SGD.

[1] Following standard convention, the $\mathcal{O}$-notation hides constant factors, and the $\tilde{\mathcal{O}}$-notation hides constants and factors polylogarithmic in the problem parameters.

1.1 Main Contributions and Structure

Our main contributions are:

- In Section 4 we generalize the analysis of (Arjevani et al., 2020) of (D-SGD) to quasi-convex functions, and show the iteration complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{L\tau}{\mu}\log\frac{1}{\epsilon}\big)$ for strongly ($\mu > 0$) and $\mathcal{O}\big(\frac{L\tau\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\epsilon} + \frac{\sigma^2\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\epsilon^2}\big)$ for general ($\mu = 0$) quasi convex functions. The dependency on the problem parameters τ, L, μ is in tight up to logarithmic factors (for *accelerated* delayed gradient methods—which we do not consider here—these rates could be improved). Further, for arbitrary smooth non-convex functions, we show a iteration complexity of $\mathcal{O}\big(\frac{L\tau(f(\mathbf{x}_0)-f^\star)}{\epsilon} + \frac{L\sigma^2(f(\mathbf{x}_0)-f^\star)}{\epsilon^2}\big)$ for convergence to a stationary point i.e convergence of the squared gradient norm to zero.

- In Section 5 we generalize the analysis of (Stich et al., 2018) for SGD with gradient compression and error compensation to quasi-convex functions and show the iteration complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{L}{\mu\delta}\log\frac{1}{\epsilon}\big)$ for strongly ($\mu > 0$) and $\mathcal{O}\big(\frac{L\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\delta\epsilon} + \frac{\|\mathbf{x}_0-\mathbf{x}^\star\|^2\sigma^2}{\epsilon^2}\big)$ for general ($\mu = 0$) quasi convex functions. Here $\delta > 0$ is a parameter that measures the compression quality. For general smooth non-convex functions, we show an iteration complexity of $\mathcal{O}\big(\frac{L(f(\mathbf{x}_0)-f^\star)}{\delta\epsilon} + \frac{L\sigma^2(f(\mathbf{x}_0)-f^\star)}{\epsilon^2}\big)$ for convergence to a stationary point. This is the first analysis of these methods without the bounded gradient assumption and improves over all previous results. In particular, all previous results suffer from a quadratic dependence on δ whereas our rates have only a linear dependence.

- In Section 6 we derive complexity estimates for local SGD. This algorithm can be viewed as a special asynchronous stochastic gradient method and has become increasingly popular in recent years (Zinkevich et al., 2010; Zhang et al., 2016; Lin et al., 2020; Patel and Dieuleveut, 2019). Our complexity estimate improve previous results, but we do not believe that our bounds are tight.

We discuss the precise setting in Section 2 and highlight important cases. We further provide some key technical lemmas in Section 3.

1.2 Related Work

For in-depth discussion of SGD and its application in machine learning we refer to the book of (Bottou et al., 2018). Here we try to list the most closely related work by topic.

Asynchronous and delayed SGD. Asynchronous methods have been intensively studied over the last three decades, starting with (Bertsekas and Tsitsiklis, 1989). A large impact for distributed machine learning had the Hogwild! algorithm (Niu et al., 2011) that performs asynchronous (block)-coordinate updates on a shared parameter vector. Many early theoretical results for asynchronous methods depend on rigorous sparsity assumptions (Chaturapruek et al., 2015; Mania et al., 2017; Leblond et al., 2018).

An algorithm very similar to (D-SGD), but with arbitrary (instead of fixed) delays of at most τ iterations, was studied in (Agarwal and Duchi, 2011). For smooth convex functions they show a bound of $\mathcal{O}\big(\frac{\sigma}{\sqrt{T}} + \frac{\tau^2}{\sigma^2 T}\big)$ (in terms of σ, τ, T only, $\tau \geq 1$). Feyzmahdavian et al. (2016) improve the bound to $\mathcal{O}\big(\frac{\sigma}{\sqrt{T}} + \frac{\tau^2}{T}\big)$. Arjevani et al. (2020) show that for (D-SGD) the dependency on τ can be improved and show a bound $\mathcal{O}\big(\frac{\sigma}{\sqrt{T}} + \frac{\tau}{T}\big)$ on convex quadratic functions. Our work extends these results to a boarder class of functions. Our proof technique is different from theirs and also allows the analysis of more general delay models, for e.g. the variables delays as in (Agarwal and Duchi, 2011; Sra et al., 2016). For general non-convex smooth functions, Lian et al. (2015) show that after $\Omega\big(\frac{L\tau^2(f(\mathbf{x}_0)-f^\star)}{\sigma^2}\big)$ iterations the effect of the delay becomes negligible and we recover their result with our analysis.

Compressed gradient methods and error compensation. Convergence aspects of (stochastic) gradient descent with compressed gradients have been studied in various communities, and results for unbiased compression (perturbations) can be traced back to e.g. (Polyak, 1987). Jointly with the increase of the size of the deep learning models, the interest in gradient compression techniques has risen in the past years (Wen et al., 2017; Alistarh et al., 2017; Wangni et al., 2018). The convergence analysis of these methods, see e.g. (Alistarh et al., 2017), typically give rates of the form $\mathcal{O}\big(\frac{\omega\sigma^2}{\mu T}\big)$ where $\omega \geq 1$ is a parameter that measures the additional noise introduced by the (unbiased) compression operators. Despite the practical success of these methods, the linear slowdown in ω makes them less attractive from a theoretical point of view.

A different type of methods use error-correction, or other error-compensation mechanisms. A method of this type was for instance developed for a particular application in (Seide et al., 2014; Strom, 2015). Wu et al. (2018) analyze a method with error correction for quadratic functions, Stich et al. (2018) provide an analysis for strongly convex functions and a large class of compression operators, including biased compressors. As a key result they show that the optimal $\mathcal{O}\big(\frac{\sigma^2}{\mu T}\big)$ convergence rate can be attained, with the same asymptotic rate as the schemes without error compensation. These results were extended in (Karimireddy et al., 2019) to non-smooth and non-convex functions. Here we analyze this method in a more general setting, for instance without the bounded gradient assumption.

Local SGD. Local SGD (a.k.a. parallel SGD) is parallel version of SGD, where each device performs local updates of the form (SGD) in parallel on the local data, and the devices average their iterates after every τ updates. This is different from mini-batch SGD where the averaging happens after every iteration, but more closely related to mini-batch SGD with τ-times larger batchsizes on each device. This algorithm has attracted the attention of the community due to its application in federated learning (McMahan et al., 2017). Early analyses focused on variants with only one averaging step (McDonald et al., 2009; Zinkevich et al., 2010; Zhang et al., 2013; Shamir and Srebro, 2014; Godichon-Baggioni and Saadane, 2017; Jain et al., 2018). More practical are schemes that perform more frequent averaging of the parallel sequences (Zhang et al., 2016; Lin et al., 2020). Analyses have been developed for strongly convex (Patel and Dieuleveut, 2019; Stich, 2019a) and non-convex (Yu et al., 2018; Wang and Joshi, 2018) functions. These results show that local SGD can attain the optimal convergence rate of SGD when T is large enough compared to τ. For non-convex functions, the best bounds (with respect to only the parameter τ) are $T = \Omega(\tau^4)$ (Yu et al., 2018) and for strongly convex functions a better quadratic dependence $T = \Omega(\tau^2)$ is known (Patel and Dieuleveut, 2019; Stich, 2019a). Here we improve this to $T = \tilde{\Omega}(\tau)$ which is optimal up to logarithmic factors (we need $T \geq \tau$ to communicate at least once). However, a closer inspection of our bounds reveals that they are not yet tight in many cases.

They also do not match with the lower bounds (Arjevani and Shamir, 2015; Woodworth et al., 2018).

Proof techniques. Our proof consists of three parts: firstly (i), we follow closely the analysis in (Stich et al., 2018; Karimireddy et al., 2019) to derive a one-step progress estimate. This technique is based on ideas of the perturbed iterate analysis (Mania et al., 2017; Leblond et al., 2018) in combination with standard estimates (Nesterov, 2004). Secondly (ii), to derive the final complexity estimates and getting the optimal optimization terms in the rate, we use the technique from (Stich, 2019b) (there would be other options here, see for instance (Stich, 2020)). Thirdly (iii), whilst we follow similar techniques to estimate the error as in previous works, we split the error term in a bias and noise component. This allows to use bigger stepsizes: whilst e.g. Feyzmahdavian et al. (2016) had to use stepsizes $\mathcal{O}\bigl(\frac{1}{\tau^2}\bigr)$, we can use stepsizes $\mathcal{O}\bigl(\frac{1}{\tau}\bigr)$ (only showing the dependency on τ), similar as in (Arjevani et al., 2020).

Follow up advances. Since the initial submission, Karimireddy et al. (2020); Woodworth et al. (2020) build upon our techniques to improve the results for local SGD. Woodworth et al. (2020) use an improved step-size to show that local SGD can sometimes converge faster, even beating large batch SGD. They also construct a lower bound example proving that their analysis is tight. Karimireddy et al. (2020) show that using two step-sizes (a global and local step-size) can give faster rates. Koloskova et al. (2020) analyze local SGD the setting where the data across the devices is heterogenous and Karimireddy et al. (2020) propose a new algorithm to overcome this heterogeneity.

2. Formal Setting

In this section we discuss our different settings and assumptions. For each of the problems studied in the later sections (delayed updates, compressed gradients, and local SGD) we analyze three cases: when f is a (i) strongly quasi-convex, (ii) general quasi-convex, or a (iii) arbitrary smooth (i.e. not comprised in classes (i) or (ii)) non-convex function.

We first examine the notion of quasi-convexity with respect to a minimizer $\mathbf{x}^\star \in \mathbb{R}^d$ of (1). This is a substantial relaxation of the standard convexity assumption, as the assumption also holds for certain non-convex functions, such as e.g. star convex functions. The definition coincides with the recently introduced class of $(1,\mu)$-quasar convex functions (Hinder et al., 2019). Our definition is slightly less general than the notion of quasi-convexity introduced in (Necoara et al., 2019), as we choose a particular minimizer $\mathbf{x}^\star$, the "quasar-convex point" and do not e.g. use projection on the set of minimizers as in (Necoara et al., 2019). However, extension of the analysis to these settings would be possible.

Assumption 1 (μ-quasi-convexity (w.r.t. $\mathbf{x}^\star$)) *The function $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable and* μ-quasi convex *for a constant $\mu \geq 0$* with respect to $\mathbf{x}^\star$*, that is*

$$f(\mathbf{x}) - f^\star + \frac{\mu}{2}\left\|\mathbf{x}-\mathbf{x}^\star\right\|^2 \leq \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\,, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{2}$$

Remark 1 *Note that f can be quasi-convex w.r.t. $\mathbf{x}^\star$ only if $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. If μ is possibly 0, we say the function is* general *quasi-convex. When $\mu > 0$, this assumption implies that such a $\mathbf{x}^\star$ is unique and we say the function is* strongly *quasi-convex.*

Remark 2 *For μ-strongly convex functions (under the standard definition), it holds*

$$f(\mathbf{x}) - f(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2 \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle \ , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Thus, by setting $\mathbf{y} = \mathbf{x}^\star$ we see that (2) *is more general. Interestingly,* (2) *also holds for non-convex functions. We give a few examples in Section 2.1.*

Remark 3 *For functions which satisfy the Polyak-Lojasiewicz condition (Karimi et al., 2016), we have*

$$\|\nabla f(\mathbf{x})\|^2 \geq 2\mu(f(\mathbf{x}) - f(\mathbf{x}^\star)) \,, \qquad \forall\, \mathbf{x} \in \mathbb{R}^d \,.$$

This is a weaker condition than quasi-convexity since (2) *implies that*

$$f(\mathbf{x}) - f^\star + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^\star\|^2 \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2 + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^\star\|^2 \ .$$

We will further assume that the gradients of the function f are Lipschitz, and hence that f is smooth.

Assumption 2 (L-smoothness) *The function $f \colon \mathbb{R}^d \to \mathbb{R}$ is differentiable and there exists a constant $L \geq 0$ such that*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \ , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \,. \tag{3}$$

We will next describe some implications of Assumption 2 which will be later useful.

Remark 4 *The Lipschitz gradient condition* (3) *implies that there exists a quadratic upper bound on f (Nesterov, 2004, Lemma 1.2.3):*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2 \ , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \,. \tag{4}$$

Further, minimizing both the left and right hand side with respect to $\mathbf{y}$ of (4) *yields*

$$\|\nabla f(\mathbf{x})\|^2 \leq 2L(f(\mathbf{x}) - f^\star) \,. \tag{5}$$

Finally, if f satisfies (4) *and is additionally convex, then (Nesterov, 2004, Theorem 2.1.5) shows*

$$\frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{6}$$

Remark 5 *The Assumptions 1 and 2 can only be satisfied together if $L \geq \mu$. This can be seen by combining* (2) *with* (4) *for $\mathbf{y} = \mathbf{x}^\star$:*

$$\frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^\star\|^2 \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle - f(\mathbf{x}) + f^\star \leq \frac{L}{2} \|\mathbf{x} - \mathbf{x}^\star\|^2 \ .$$

Lastly, we assume that the noise of the gradient oracle is bounded. Instead of assuming a uniform upper bound, we assume an upper bound of the following form:

Assumption 3 ((M,σ^2)-bounded noise) *For any $\mathbf{x}$, a gradient oracle of the form $\mathbf{g} = \nabla f(\mathbf{x}) + \boldsymbol{\xi}$ for a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$, and conditionally independent noise $\boldsymbol{\xi}$, there exists two constants $M, \sigma^2 \geq 0$, such that*

$$\mathbb{E}\left[\boldsymbol{\xi} \mid \mathbf{x}\right] = \mathbf{0}_d \,, \qquad\qquad \mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \leq M \left\|\nabla f(\mathbf{x})\right\|^2 + \sigma^2 \,. \tag{7}$$

We will next state a weaker variant which will be sufficient for quasi-convex functions.

Assumption 3* ((M,σ^2)-bounded noise) *For any $\mathbf{x}$, a gradient oracle of the form $\mathbf{g} = \nabla f(\mathbf{x}) + \boldsymbol{\xi}$ for a L-smooth quasi-convex function $f\colon \mathbb{R}^d \to \mathbb{R}$, and conditionally independent noise $\boldsymbol{\xi}$, there exists two constants $M, \sigma^2 \geq 0$, such that*

$$\mathbb{E}\left[\boldsymbol{\xi} \mid \mathbf{x}\right] = \mathbf{0}_d \,, \qquad\qquad \mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \leq 2LM(f(\mathbf{x}) - f^\star) + \sigma^2 \,. \tag{8}$$

Remark 6 *By combining Assumptions 2 and 3 it follows for any $\mathbf{x}$:*

$$\mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \leq M \left\|\nabla f(\mathbf{x})\right\|^2 + \sigma^2 \overset{(5)}{\leq} 2LM(f(\mathbf{x}) - f^\star) + \sigma^2.$$

This shows that (8) *in Assumption 3* is weaker than* (7) *in Assumption 3. Though we rely on* (7) *in our proofs for the sake of conciseness, it is straightforward to adapt our results to the weaker noise condition* (8) *for quasi-convex functions. We only need the stronger condition* (7) *for arbitrary non-convex functions.*

Remark 7 *Assumptions 2 and 3 together imply and an upper bound on the second moment of the gradient oracle of the form*

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}) + \boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] = \left\|\nabla f(\mathbf{x})\right\|^2 + \mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \overset{(5),(8)}{\leq} 2L(1+M)(f(\mathbf{x}) - f^\star) + \sigma^2 \,. \tag{9}$$

We will now discuss a few examples covered by our assumptions.

2.1 Key Settings Covered by the Quasi Convexity Assumption

Assumption 1 does clearly hold for convex and strongly convex functions, but interestingly also for certain non-convex functions. We also analyze general non-convex functions which do not satisfy Assumption 1, but only prove convergence to a stationary point.

Quasar convex functions. Functions satisfying Assumption 1 are variously called *quasi-strongly convex* (Necoara et al., 2019), *weakly strongly convex* (Karimi et al., 2016) or $(1,\mu)$-*(strongly) quasar-convex* in recent work by Hinder et al. (2019), extending a similar notion previously introduced in (Hardt et al., 2018). Our results can be extended to the more general (ν,μ)-quasar convex functions by following their techniques. The focus of this work is on delayed gradient updates and hence we leave such extensions for future work.

Star (strongly) convex functions. A notable class of functions satisfying Assumption 1 are differentiable star-convex functions. The function $f(x) = |x| \left(1 - e^{-|x|}\right)$ is smooth and star-convex, but not convex (Nesterov and Polyak, 2006). We verify Assumption 1 by observing $\langle \nabla f(x), x \rangle - f(x) = x^2 e^{-|x|} \geq 0$, that is, equation (2) holds for $\mu = 0$. More generally, smooth star convex functions can be constructed by extending an arbitrary smooth positive (but not necessarily convex) function $g \colon \mathbb{S}^{d-1} \to \mathbb{R}$ from the unit sphere sphere to $\mathbb{R}^d$ by e.g. setting

$$f(\mathbf{x}) = \|\mathbf{x}\| \cdot \left(1 - e^{-\|\mathbf{x}\|}\right) \cdot g\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) + \frac{\mu}{2} \|\mathbf{x}\|^2 .$$

For other examples and constructions see e.g. (Lee and Valiant, 2016).

2.2 Key Settings Covered by the General Noise Model

Our Assumption 3 on the noise generalizes the usual standard assumptions. The weaker Assumption 3* is more general as we highlight by discussing an inexhaustive list of cases covered.

Uniformly bounded gradients. A classical assumption in the analysis of stochastic gradient methods is to assume an uniform upper bound on the stochastic gradients, that is $\mathbb{E}\left[\|\nabla f(\mathbf{x}) + \boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \leq G^2$, for a parameter $G^2 \geq 0$, see e.g. (Nemirovski and Yudin, 1983; Nemirovski et al., 2009). This implies that $G^2 \geq \|\nabla f(\mathbf{x})\|^2 + \mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right]$ for any $\mathbf{x}$. Thus, even when $\mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] = 0$ we have $G^2 > 0$ in general, and thus this assumption is typically too loose to obtain good complexity estimates in the deterministic setting. In contrast Assumption 3 is satisfied with $M = 0$ and $\sigma^2 = 0$ in the deterministic setting.

Uniformly bounded noise. Much more fine grained is the uniformly bounded noise assumption, that is assuming $\mathbb{E}\left[\|\boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \leq \sigma^2$, see e.g. (Dekel et al., 2012). This setting recovers the deterministic analysis in the case $\sigma^2 = 0$ and is covered by setting $M = 0$ in Assumption 3. The uniformly bounded noise assumption appeared also in (Arjevani et al., 2020), thus we extend their analysis not only to a richer function class, but also to (moderately) more general noise models.

Strong-growth condition. Schmidt and Roux (2013) introduce the strong-growth condition where it is assumed that there exists a constant M such that $\mathbb{E}\left[\|\nabla f(\mathbf{x}) + \boldsymbol{\xi}\|^2 \mid \mathbf{x}\right] \leq M \|\nabla f(\mathbf{x})\|^2$ which translates to Assumption 3 with $\sigma^2 = 0$. Assumptions 3* with $\sigma^2 = 0$ is referred to as the *weak growth* condition in (Vaswani et al., 2018). These conditions are useful to study benign noise whose magnitude decreases as we get closer to the optimum, and in particular imply that the noise at the optimum is 0.

We now study two settings which particularly benefit from the weaker Assumption 3*, and so apply only to the quasi-convex setting (see Remark 6).

Finite-sum optimization. In finite sum optimization problems, the objective function can be written as $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ for components $f_i \colon \mathbb{R}^d \to \mathbb{R}$, and the gradient oracle is typically just the gradient of one component $\nabla f_i(\mathbf{x})$, where the index i is selected uniformly at random from $[n]$. If we assume that each component f_i is convex and satisfies the smoothness Assumption 2 with, for simplicity, the same constant L for each f_i, then we can observe:

$$\begin{aligned}\mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 &\leq \mathbb{E}_i\|\nabla f_i(\mathbf{x})\|^2 = \mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star) + \nabla f_i(\mathbf{x}^\star)\|^2 \\ &\leq 2\,\mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)\|^2 + 2\,\mathbb{E}_i\|\nabla f_i(\mathbf{x}^\star)\|^2 \\ &\overset{(6)}{\leq} \frac{4L}{n}\sum_{i=1}^n \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^\star) - \langle \nabla f_i(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle\right) + 2\,\mathbb{E}_i\|\nabla f_i(\mathbf{x}^\star)\|^2 \\ &= 4L(f(\mathbf{x}) - f^\star) + 2\,\mathbb{E}_i\|\nabla f_i(\mathbf{x}^\star)\|^2\,.\end{aligned}$$

Thus we see that smooth finite sum objectives naturally satisfy the weaker bounded noise Assumption 3* with parameters $M = 2$ and $\sigma^2 = 2\,\mathbb{E}_i\|\nabla f_i(\mathbf{x}^\star)\|^2$. These insights in the problem structure were first discussed in (Bach and Moulines, 2011) and refined in (Schmidt and Roux, 2013; Needell et al., 2016) and allowed to derive the first linear convergence rates for SGD on finite sum problems in the special case when $\sigma^2 = 0$. Sometimes this special setting is also referred to as the *interpolation setting* (Ma et al., 2018). Closely related is the refined notion of expected smoothness, see the discussions in (Gower et al., 2018, 2019).

Least-squares. A classic problem in the literature (Bach and Moulines, 2011) is the least squares minimization problem, where $f(\mathbf{x}) = \frac{1}{2}\,\mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\big[(b - \langle \mathbf{x}, \mathbf{a}\rangle)^2\big]$ measures the expected square loss over the data samples $(\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}$, sampled form a (unknown) distribution $\mathcal{D}$. With the notation $f_{(\mathbf{a},b)} := \frac{1}{2}(b - \langle \mathbf{x}, \mathbf{a}\rangle)^2$ the objective $f(\mathbf{x}) = \mathbb{E}_{(\mathbf{a},b)} f_{(\mathbf{a},b)}$ takes the standard form of a stochastic optimization problem. An unbiased stochastic gradient oracle for f is given by $\nabla f_{(\mathbf{a},b)}(\mathbf{x}) = -(b - \langle \mathbf{x}, \mathbf{a}\rangle)\cdot \mathbf{a}$. We observe that

$$\mathbb{E}_{(\mathbf{a},b)}\left\|\nabla f_{(\mathbf{a},b)}(\mathbf{x}) - \nabla f_{(\mathbf{a},b)}(\mathbf{x}^\star)\right\|^2 \leq \mathbb{E}_{(\mathbf{a},b)}\big[2\,\langle \mathbf{a}, \mathbf{a}\rangle \left(f_{(\mathbf{a},b)}(\mathbf{x}) - f_{(\mathbf{a},b)}(\mathbf{x}^\star)\right)\big]\,.$$

Thus, by assuming a bound on the fourth moment of $\mathbf{a}$, sometimes written in the form $\mathbb{E}\left[\langle \mathbf{a}, \mathbf{a}\rangle\, \mathbf{a}\mathbf{a}^\top\right] \preceq R^2\mathbf{A}$, for a number R^2 and Hessian $\mathbf{A} := \mathbb{E}\left[\mathbf{a}\mathbf{a}^\top\right]$ (cf. Bach and Moulines, 2011; Jain et al., 2018), we can further bound the right hand side by

$$\mathbb{E}_{(\mathbf{a},b)}\left\|\nabla f_{(\mathbf{a},b)}(\mathbf{x}) - \nabla f_{(\mathbf{a},b)}(\mathbf{x}^\star)\right\|^2 \leq 2R^2\left(f(\mathbf{x}) - f^\star\right)\,.$$

Following the same argumentation as outlined for the previous example, we see that least squares optimization under standard assumptions also satisfies the relaxed bounded noise Assumption 3*, with M proportional to $3(1 + R^2/L)$ and σ^2 estimating the noise at the optimum. We see that Assumption 3* in the form of (8) is slightly more general, though we like to point out that for least squares problems the analyses typically also make additional assumptions on the structure of covariance of the noise to get more fine-grained results (Dieuleveut et al., 2017; Jain et al., 2018), a refinement we do not consider here.

3. Error-Feedback Framework

In this section we present a host of lemmas that will ease the presentation of the proofs in the subsequent sections.

3.1 Error-Compensated and Virtual Sequences

We will rewrite all algorithms that we consider here in the following, unified, notation, with auxiliary sequences $\{\mathbf{v}_t\}_{t\geq 0}$ that express the *applied updates*, and $\{\mathbf{e}_t\}_{t\geq 0}$ that aggregates the delayed information or synchronicity errors. We follow here the ideas from (Stich et al., 2018; Karimireddy et al., 2019) and consider algorithms in the form

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \mathbf{v}_t\,, \\ \mathbf{e}_{t+1} &= \mathbf{e}_t + \gamma_t \mathbf{g}_t - \mathbf{v}_t\,. \end{aligned} \tag{EC-SGD}$$

After T such updates, we output $\mathbf{x}^{\rm out} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ where $\mathbf{x}_t$ is chosen with probability proportional to w_t for some sequence of positive weights $\{w_t\}_{t=0}^{T-1}$.

For instance, for (D-SGD) we have the updates $\mathbf{v}_t = \gamma_{t-\tau}\mathbf{g}_{t-\tau}$ for $t \geq \tau$, and $\mathbf{v}_t = \mathbf{0}_d$ otherwise; with error terms $\mathbf{e}_t := \sum_{i=1}^{\tau} \gamma_{t-i}\mathbf{g}_{t-i}$ (here—for a light notation—we use the convention to only sum over positive indices).

For the analysis, it will be convenient to define a sequence of 'virtual' iterates $\{\tilde{\mathbf{x}}\}_{t\geq 0}$. That is, the iterates $\tilde{\mathbf{x}}_t$ never need to be actually computed, they only appear as a tool in the proof. Formally, we define

$$\tilde{\mathbf{x}}_t := \mathbf{x}_t - \mathbf{e}_t\,, \qquad \forall t \geq 0, \tag{10}$$

with $\tilde{\mathbf{x}}_0 := \mathbf{x}_0$ (note that $\mathbf{e}_0 = \mathbf{0}_d$). We observe that

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{x}_{t+1} - \mathbf{e}_{t+1} = (\mathbf{x}_t - \mathbf{v}_t) - (\mathbf{e}_t + \gamma_t\mathbf{g}_t - \mathbf{v}_t) = \tilde{\mathbf{x}}_t - \gamma_t \mathbf{g}_t\,. \tag{11}$$

3.2 A Descent Lemma For Quasi-Convex Functions

In the next lemma we derive a bound on the one-step progress for the virtual iterates $\tilde{\mathbf{x}}_t$ for quasi-convex functions. This proof combines standard techniques (Nesterov, 2004) with ideas from the perturbed iterate analysis (Mania et al., 2017; Leblond et al., 2018) and can be seen as an extension of (Stich et al., 2018, Lemma 3.1) to the more general setting considered in this work.

Lemma 8 *Let $\{\mathbf{x}_t, \mathbf{v}_t, \mathbf{e}_t\}_{t\geq 0}$ be defined as in* (EC-SGD) *with gradient oracle $\{\mathbf{g}_t\}_{t\geq 0}$ and objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ as in Assumptions 1–3. If $\gamma_t \leq \frac{1}{4L(1+M)}$, $\forall t \geq 0$, then for $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$ defined as in* (10),

$$\mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\right\|^2 \leq \left(1 - \frac{\mu\gamma_t}{2}\right)\mathbb{E}\left\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\right\|^2 - \frac{\gamma_t}{2}\,\mathbb{E}\left(f(\mathbf{x}_t) - f^\star\right) + \gamma_t^2\sigma^2 + 3L\gamma_t\,\mathbb{E}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2. \tag{12}$$

Proof We expand:

$$\left\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\right\|^2 \overset{(11)}{=} \left\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\right\|^2 - 2\gamma_t\left\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star\right\rangle + \gamma_t^2\left\|\mathbf{g}_t\right\|^2 + 2\gamma_t\left\langle \mathbf{g}_t, \mathbf{x}_t - \tilde{\mathbf{x}}_t\right\rangle\,,$$

and take expectation w.r.t. the random variable $\boldsymbol{\xi}_t$:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}_t}\Big[\left\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\right\|^2 \mid \mathbf{x}_t\Big] \overset{(9)}{\leq} &\left\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\right\|^2 - 2\gamma_t\left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\right\rangle + 2L(1+M)\gamma_t^2(f(\mathbf{x}_t) - f^\star) \\ &+ \gamma_t^2\sigma^2 + 2\gamma_t\left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \tilde{\mathbf{x}}_t\right\rangle\,. \end{aligned} \tag{13}$$

By Assumption 1:

$$-2\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle \overset{(2)}{\leq} -\mu \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2(f(\mathbf{x}_t) - f^\star)\,,$$

and by $2\langle \mathbf{a}, \mathbf{b} \rangle \leq \alpha \|\mathbf{a}\|^2 + \alpha^{-1} \|\mathbf{b}\|^2$ for $\alpha > 0$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,

$$2\langle \nabla f(\mathbf{x}_t), \tilde{\mathbf{x}}_t - \mathbf{x}_t \rangle \leq \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + 2L\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 \overset{(5)}{\leq} f(\mathbf{x}_t) - f^\star + 2L\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2\,.$$

And by $\|\mathbf{a} + \mathbf{b}\|^2 \leq (1+\beta)\|\mathbf{a}\|^2 + (1+\beta^{-1})\|\mathbf{b}\|^2$ for $\beta > 0$ (as a consequence of Jensen's inequality), we further observe

$$-\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \leq -\frac{1}{2}\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\|^2 + \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2\,.$$

Plugging all these inequalities together into (13) yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}_t}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu\gamma_t}{2}\right)\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\|^2 - \gamma_t(1 - 2L(1+M)\gamma_t)(f(\mathbf{x}_t) - f^\star) \\ + \gamma_t^2\sigma^2 + \gamma_t(2L+\mu)\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2\,.\end{aligned}$$

The claim follows by the choice $\gamma_t \leq \frac{1}{4L(1+M)}$ and $L \geq \mu$. ■

3.3 A Descent Lemma for Non-Convex Functions

We now turn our attention to non-convex functions and prove a descent lemma for $\tilde{\mathbf{x}}_t$. This proof follows the template of (Ghadimi and Lan, 2013; Karimireddy et al., 2019) mildly extending the techniques to the general noise condition studied here.

Lemma 9 *Let $\{\mathbf{x}_t, \mathbf{v}_t, \mathbf{e}_t\}_{t\geq 0}$ be defined as in* (EC-SGD) *with gradient oracle $\{\mathbf{g}_t\}_{t\geq 0}$ and objective function $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying Assumptions 2 and 3. If $\gamma_t \leq \frac{1}{2L(1+M)}$, $\forall t \geq 0$, then for $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$ defined as in* (10),

$$\mathbb{E}[f(\tilde{\mathbf{x}}_{t+1})] \leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \frac{\gamma_t}{4}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma_t^2 L\sigma^2}{2} + \frac{\gamma_t L^2}{2}\mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2\,. \tag{14}$$

Proof We begin using the definition of $\tilde{\mathbf{x}}_{t+1}$ and the smoothness of f,

$$f(\tilde{\mathbf{x}}_{t+1}) \overset{(4)}{\leq} f(\tilde{\mathbf{x}}_t) - \gamma_t\langle \nabla f(\tilde{\mathbf{x}}_t), \mathbf{g}_t \rangle + \frac{\gamma_t^2 L}{2}\|\mathbf{g}_t\|^2\,.$$

Taking expectation with respect to $\boldsymbol{\xi}_t$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}_t}[f(\tilde{\mathbf{x}}_{t+1})|\mathbf{x}_t] &\overset{(7)}{\leq} f(\tilde{\mathbf{x}}_t) - \gamma_t\langle \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t)\rangle + \frac{\gamma_t^2 L(1+M)}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma_t^2 L\sigma^2}{2} \\ &= f(\tilde{\mathbf{x}}_t) - \gamma_t\left(1 - \frac{\gamma_t L(1+M)}{2}\right)\|\nabla f(\mathbf{x}_t)\|^2 \\ &\quad + \gamma_t\langle \nabla f(\mathbf{x}_t) - \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t)\rangle + \frac{\gamma_t^2 L\sigma^2}{2}\,.\end{aligned}$$

Again using $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2} \|\mathbf{a}\|^2 + \frac{1}{2} \|\mathbf{b}\|^2$, we can simplify the expression as follows

$$\begin{aligned}\langle \nabla f(\mathbf{x}_t) - \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t) \rangle &\leq \frac{1}{2} \|\nabla f(\mathbf{x}_t) - \nabla f(\tilde{\mathbf{x}}_t)\|^2 + \frac{1}{2} \|\nabla f(\mathbf{x}_t)\|^2 \\ &\overset{(3)}{\leq} \frac{L^2}{2} \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{1}{2} \|\nabla f(\mathbf{x}_t)\|^2 \,.\end{aligned}$$

Plugging this back, we get our result that

$$\mathbb{E}_{\boldsymbol{\xi}_t}[f(\tilde{\mathbf{x}}_{t+1}) | \mathbf{x}_t] \leq f(\tilde{\mathbf{x}}_t) - \frac{\gamma_t (1 - \gamma_t L(1+M))}{2} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma_t^2 L \sigma^2}{2} + \frac{\gamma_t L^2}{2} \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 \,.$$

Noting that $\gamma_t \leq \frac{1}{2L(1+M)}$ implies $\frac{\gamma_t(1-\gamma_t L(1+M))}{2} \leq \frac{\gamma_t}{4}$ yields the lemma. ■

3.4 Stepsizes

We will study SGD with constant stepsizes and slowly decreasing stepsizes in this paper. It will become handy to formalize 'slowly decreasing' in the following way.

Definition 10 (τ-slow sequences) *The sequence $\{a_t\}_{t\geq 0}$ of positive values is* τ-slow decreasing *for parameter $\tau \geq 1$ if*

$$a_{t+1} \leq a_t \,, \qquad \forall t \geq 0, \qquad \textit{and,} \qquad a_{t+1}\left(1 + \frac{1}{2\tau}\right) \geq a_t \,, \qquad \forall t \geq 0.$$

The sequence $\{a_t\}_{t\geq 0}$ is τ-slow increasing *if $\{a_t^{-1}\}_{t\geq 0}$ is τ-slow decreasing.*

Example 1 *The sequences $\{u_t := (\kappa + t)^2\}_{t\geq 0}$, $\{v_t := \kappa + t\}_{t \geq 0}$ and $\{w_t := \left(1 - \frac{1}{4c\tau}\right)^{-t}\}_{t\geq 0}$, for $\kappa \geq 8\tau$, $c \geq 1$, and $\tau \geq 1$, are examples of τ-slow increasing sequences.*

Proof First, consider the sequence $\{u_t\}_{t\geq 0}$. The condition $u_{t+1} \geq u_t$ is easily verified. Furthermore, $\frac{u_{t+1}}{u_t} = \frac{(\kappa+t+1)^2}{(\kappa+t)^2} \leq \frac{(\kappa+1)^2}{\kappa^2} = \frac{u_1}{u_0}$, so it suffices to check:

$$u_1 \left(1 + \frac{1}{2\tau}\right)^{-1} \leq (\kappa + 1)^2 \left(1 - \frac{1}{4\tau}\right) = \kappa^2 + \kappa \underbrace{\left(2 - \frac{\kappa}{4\tau}\right)}_{\leq 0} + \underbrace{\left(1 - \frac{\kappa}{2\tau}\right)}_{\leq 0} - \frac{1}{4\tau} \leq \kappa^2 = u_0 \,,$$

where the first inequality is due to $\left(1 + \frac{x}{2}\right)^{-1} \leq 1 - \frac{x}{4}$, for $0 \leq x \leq 1$. This can be verified by $\left(1 + \frac{x}{2}\right)\left(1 - \frac{x}{4}\right) = 1 + \frac{x(2-x)}{8} \geq 1 + \frac{x}{8} \geq 1$ for $0 \leq x \leq 1$. This shows that $\{u_t\}_{t\geq 0}$ is τ-slow increasing. It follows, that $\{v_t = \sqrt{u_t}\}_{t\geq 0}$ satisfies $v_{t+1} \leq v_t\left(1 + \frac{1}{2\tau}\right)^{1/2} \leq v_t\left(1 + \frac{1}{2\tau}\right)$ and hence is also τ-slow increasing. For the last sequence we verify that

$$w_{t+1}^{-1}\left(1 + \frac{1}{2\tau}\right) = w_t^{-1}\left(1 - \frac{1}{4c\tau}\right)\left(1 + \frac{1}{2\tau}\right) \geq w_t^{-1}\left(1 - \frac{1}{4\tau}\right)\left(1 + \frac{1}{2\tau}\right) \geq w_t^{-1} \,,$$

in agreement with Definition 10. ■

Remark 11 *For stepsizes of the form $\{\gamma_t = \frac{c}{\kappa + t}\}_{t\geq 0}$ for a constant $c \geq 0$, $\kappa \geq 8\tau$ and $\tau \geq 1$ it follows from Example 1 that $\{\gamma_t\}_{t\geq 0}$ and $\{\gamma_t^2\}_{t\geq 0}$ are τ-slow decreasing sequences. Constant stepsizes, $\{\gamma_t = \gamma\}_{t\geq 0}$ for a constant $\gamma > 0$ are τ-slow decreasing for any $\tau \geq 1$.*

3.5 Technical Lemmas for Deriving the Complexity Estimates

The next two technical lemmas we borrow from (Stich, 2019b). For completeness we include the proofs as well. Lemma 14 is an extension to deal with arbitrary non-convex functions or function which are quasi-convex with $\mu = 0$.

Lemma 12 ((Stich, 2019b, Lemma 7)) *For decreasing stepsizes* $\left\{\gamma_t := \frac{2}{a(\kappa+t)}\right\}_{t\geq 0}$*, and weights* $\{w_t := (\kappa + t)\}_{t\geq 0}$ *for parameters* $\kappa \geq 1$*, it holds for every non-negative sequence* $\{r_t\}_{t\geq 0}$ *and any* $a > 0$*,* $c \geq 0$ *that*

$$\Psi_T := \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{w_t}{\gamma_t}\left(1 - a\gamma_t\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + c\gamma_t w_t\right) \leq \frac{a\kappa^2 r_0}{T^2} + \frac{4c}{aT}\,,$$

where $W_T := \sum_{i=0}^{T} w_t$*.*

Proof We start by observing that

$$\frac{w_t}{\gamma_t}\left(1 - a\gamma_t\right) r_t = \frac{a}{2}(\kappa + t)(\kappa + t - 2) r_t = \frac{a}{2}\left((\kappa + t - 1)^2 - 1\right) r_t \leq \frac{a}{2}(\kappa + t - 1)^2 r_t\,. \quad (15)$$

By plugging in the definitions of γ_t and w_t in Ψ_T, we end up with a telescoping sum:

$$\Psi_T \overset{(15)}{\leq} \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{a}{2}(\kappa + t - 1)^2 r_t - \frac{a}{2}(\kappa + t)^2 r_{t+1}\right) + \sum_{t=0}^{T}\frac{2c}{aW_T} \leq \frac{a(\kappa - 1)^2 r_0}{2W_T} + \frac{2c(T + 1)}{aW_T}.$$

The lemma now follows with $(\kappa - 1)^2 \leq \kappa^2$, and $W_T = \sum_{t=0}^{T}(\kappa + t) = \frac{(2\kappa+T)(T+1)}{2} \geq \frac{T(T+1)}{2} \geq \frac{T^2}{2}$. ■

Lemma 13 ((Stich, 2019b, Lemma 2)) *For every non-negative sequence* $\{r_t\}_{t\geq 0}$ *and any parameters* $d \geq a > 0$*,* $c \geq 0$*,* $T \geq 0$*, there exists a constant* $\gamma \leq \frac{1}{d}$*, such that for constant stepsizes* $\{\gamma_t = \gamma\}_{t\geq 0}$ *and weights* $w_t := (1 - a\gamma)^{-(t+1)}$ *it holds*

$$\Psi_T := \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{w_t}{\gamma_t}\left(1 - a\gamma_t\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + c\gamma_t w_t\right) = \tilde{\mathcal{O}}\left(dr_0 \exp\left[-\frac{aT}{d}\right] + \frac{c}{aT}\right)\,.$$

Proof By plugging in the values for γ_t and w_t, we observe that we again end up with a telescoping sum and estimate

$$\Psi_T = \frac{1}{\gamma W_T}\sum_{t=0}^{T}\left(w_{t-1} r_t - w_t r_{t+1}\right) + \frac{c\gamma}{W_T}\sum_{t=0}^{t} w_t \leq \frac{r_0}{\gamma W_T} + c\gamma \leq \frac{r_0}{\gamma}\exp\left[-a\gamma T\right] + c\gamma\,,$$

where we used the estimate $W_T \geq w_T \geq (1 - a\gamma)^{-T} \geq \exp[a\gamma T]$ for the last inequality. The lemma now follows by carefully tuning γ. See the proof of Theorem 2 in (Stich, 2019b). ■

Lemma 14 *For every non-negative sequence $\{r_t\}_{t\geq 0}$ and any parameters $d \geq 0$, $c \geq 0$, $T \geq 0$, there exists a constant $\gamma \leq \frac{1}{d}$, such that for constant stepsizes $\{\gamma_t = \gamma\}_{t\geq 0}$ it holds:*

$$\Psi_T := \frac{1}{T+1}\sum_{t=0}^{T}\left(\frac{r_t}{\gamma_t} - \frac{r_{t+1}}{\gamma_t} + c\gamma_t\right) \leq \frac{dr_0}{T+1} + \frac{2\sqrt{cr_0}}{\sqrt{T+1}}\,.$$

Proof For constant stepsizes $\gamma_t = \gamma$ we can derive the estimate

$$\Psi_T = \frac{1}{\gamma(T+1)}\sum_{t=0}^{T}(r_t - r_{t+1}) + c\gamma \leq \frac{r_0}{\gamma(T+1)} + c\gamma\,.$$

We distinguish two cases (similar as in (Arjevani et al., 2020)): if $\frac{r_0}{c(T+1)} \leq \frac{1}{d^2}$, then we chose the stepsize $\gamma = \left(\frac{r_0}{c(T+1)}\right)^{1/2}$ and get

$$\Psi_T \leq \frac{2\sqrt{cr_0}}{\sqrt{T+1}}\,,$$

on the other hand, if $\frac{r_0}{c(T+1)} > \frac{1}{d^2}$, then we choose $\gamma = \frac{1}{d}$ and get

$$\Psi_T \leq \frac{dr_0}{T+1} + \frac{c}{d} \leq \frac{dr_0}{T+1} + \frac{\sqrt{cr_0}}{\sqrt{T+1}}\,.$$

These two bounds show the lemma. ■

3.6 Technical Lemma for Splitting the Bias and Noise Terms

For arbitrary vectors $\mathbf{a}_1, \dots, \mathbf{a}_\tau \in \mathbb{R}^d$ we have the inequality $\|\sum_{i=1}^{\tau}\mathbf{a}_i\|^2 \leq \tau\sum_{i=1}^{\tau}\|\mathbf{a}_i\|^2$ (as an application of Jensen's inequality). This bound is tight in general (consider $\mathbf{a}_1 = \mathbf{a}_2 = \dots = \mathbf{a}_\tau$). However, for independent zero-mean random variables $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_\tau$ we get the much tighter estimate $\mathbb{E}\,\|\sum_{i=1}^{\tau}\boldsymbol{\xi}_i\|^2 \leq \sum_{i=1}^{\tau}\mathbb{E}\,\|\boldsymbol{\xi}_i\|^2$. The following lemma combines these two estimates.

Lemma 15 *Suppose we have a martingale sequence $\mathbf{a}_1 + \boldsymbol{\xi}_1, \dots, \mathbf{a}_\tau + \boldsymbol{\xi}_\tau$ such that for any $1 \leq t \leq \tau$, we have $\mathbb{E}[\mathbf{a}_t + \boldsymbol{\xi}_t \,|\, \mathcal{F}_{t-1}] = \mathbf{a}_t$ conditioned on filtration $\mathcal{F}_{t-1}$. Then for any $\beta > 0$,*

$$\mathbb{E}\left\|\sum_{i=1}^{\tau}\mathbf{a}_i + \boldsymbol{\xi}_i\right\|^2 \leq (1+\beta)\tau\sum_{i=1}^{\tau}\mathbb{E}\,\|\mathbf{a}_i\|^2 + (1+\beta^{-1})\sum_{i=1}^{\tau}\mathbb{E}\,\|\boldsymbol{\xi}_i\|^2\,. \tag{16}$$

Proof We can simplify as follows:

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{\tau}\mathbf{a}_i + \boldsymbol{\xi}_i\right\|^2 &\leq (1+\beta)\,\mathbb{E}\left\|\sum_{i=1}^{\tau}\mathbf{a}_i\right\|^2 + (1+\beta^{-1})\,\mathbb{E}\left\|\sum_{i=1}^{\tau}\boldsymbol{\xi}_i\right\|^2 \\ &\leq (1+\beta)\tau\sum_{i=1}^{\tau}\mathbb{E}\,\|\mathbf{a}_i\|^2 + (1+\beta^{-1})\,\mathbb{E}\left\|\sum_{i=1}^{\tau}\boldsymbol{\xi}_i\right\|^2,\end{aligned}$$

where we used $\|\sum_{i=1}^{\tau}\mathbf{a}_i\|^2 \leq \tau\sum_{i=1}^{\tau}\|\mathbf{a}_i\|^2$, and $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\beta)\|\mathbf{a}\|^2 + (1+\beta^{-1})\|\mathbf{b}\|^2$ for the inequalities. The last term can further be simplified as

$$\mathbb{E}_{\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_\tau}\left\|\sum_{i=1}^{\tau}\boldsymbol{\xi}_i\right\|^2 = \sum_{i,j\in[\tau]}\mathbb{E}_{\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_\tau}[\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_j] = \sum_{i=1}^{\tau}\mathbb{E}_{\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_\tau}[\boldsymbol{\xi}_i^\top\boldsymbol{\xi}_i]\,.$$

The cross terms (when $i \neq j$) in the last equality evaluate to zero since $\{\boldsymbol{\xi}_t\}$ form a martingale difference sequence. Taking a full expectation over both sides yields the lemma. ■

4. SGD with Delayed Updates

In this section we analyze SGD with delayed updates (D-SGD) and extend the results of (Arjevani et al., 2020) from quadratic convex functions to general smooth functions (both quasi-convex and non-convex) through a different proof technique.

SGD with delayed updates, in the form as introduced in (Arjevani et al., 2020), can be cast in the (EC-SGD) framework by setting

$$\mathbf{v}_t = \begin{cases}\gamma_{t-\tau}\mathbf{g}_{t-\tau}\,, & \text{if } t \geq \tau,\\ \mathbf{0}_d\,, & \text{if } t < \tau,\end{cases} \qquad\qquad \mathbf{e}_t := \sum_{i=1}^{\tau}\gamma_{t-i}\mathbf{g}_{t-i}\,, \tag{17}$$

where here $\tau \geq 1$ is an integer *delay*. We use here (and throughout this paper) the convention to only sum over non-negative indices $(t-i) \geq 0$, i.e. the sum in (17) consists of $\min\{\tau, t\}$ terms. We prove the following rates of convergence:

Theorem 16 *Let $\{\mathbf{x}_t\}_{t\geq 0}$ denote the iterates of delayed stochastic gradient descent* (D-SGD) *with constant stepsize $\{\gamma_t = \gamma\}_{t\geq 0}$ on a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ under assumptions Assumptions 2 and 3. Then, if f*

- *satisfies Assumption 1 for $\mu > 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(\tau+M)}$ (chosen as in Lemma 13) such that*

$$\mathbb{E}\, f(\mathbf{x}^{\mathrm{out}}) - f^\star = \tilde{\mathcal{O}}\left(L(\tau+M)\|\mathbf{x}_0-\mathbf{x}^\star\|^2\exp\left[-\frac{\mu T}{10L(\tau+M)}\right] + \frac{\sigma^2}{\mu T}\right),$$

  where the output $\mathbf{x}^{\mathrm{out}} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ is chosen to be $\mathbf{x}_t$ with probability proportional to $(1-\mu\gamma/2)^{-t}$.

- *satisfies Assumption 1 for $\mu = 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(\tau+M)}$ (chosen as in Lemma 14) such that*

$$\mathbb{E}\, f(\mathbf{x}^{\mathrm{out}}) - f^\star = \mathcal{O}\left(\frac{L(\tau+M)\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{T} + \frac{\sigma\|\mathbf{x}_0-\mathbf{x}^\star\|}{\sqrt{T}}\right),$$

  where the output $\mathbf{x}^{\mathrm{out}} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ is chosen uniformly at random from the iterates $\{\mathbf{x}_t\}_{t=0}^{T-1}$.

- *is an arbitrary non-convex function, then there exists a stepsize* $\gamma \leq \frac{1}{10L(\tau+M)}$ *(chosen as in Lemma 14), such that*

$$\mathbb{E}\left\|\nabla f(\mathbf{x}^{\rm out})\right\|^2 = \mathcal{O}\left(\frac{L(\tau+M)(f(\mathbf{x}_0)-f^\star)}{T} + \sigma\sqrt{\frac{L(f(\mathbf{x}_0)-f^\star)}{T}}\right) .$$

where the output $\mathbf{x}^{\rm out} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ *is chosen uniformly at random from the iterates* $\{\mathbf{x}_t\}_{t=0}^{T-1}$.

Remark 17 *Proving convergence for a randomly picked iterate* $\mathbf{x}^{\rm out} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ *is equivalent to show convergence of a (weighted) average of the output criterion, e.g.* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E} f(\mathbf{x}_t) - f^\star$ *for general quasi-convex functions (and a weighted average for strongly-quasi convex functions). If in addition convexity holds, our results show convergence of* $\mathbb{E} f(\bar{\mathbf{x}}_T) - f^\star$*, where* $\bar{\mathbf{x}}_T := \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t$ *is a (weighted) average of the iterates.*

Remark 18 *We only consider constant stepsizes in Theorem 16. We can also prove convergence for decreasing stepsize, for instance if* f *satisfies Assumptions 1–3 for* $\mu > 0$ *and stepsizes are chosen as* $\bigl\{\gamma_t = \frac{4}{\mu(\kappa+t)}\bigr\}_{t\geq 0}$ *with* $\kappa = \frac{40L(\tau+M)}{\mu}$*, then it holds*

$$\mathbb{E} f(\mathbf{x}^{\rm out}) - f^\star = \mathcal{O}\left(\frac{L(\tau+M)^2\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2}{\mu T^2} + \frac{\sigma^2}{\mu T}\right) ,$$

where the output $\mathbf{x}^{\rm out} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ *is chosen to be* $\mathbf{x}_t$ *with probability proportional to* $(\kappa+t)$*. This rate is dominated by the first claim in Theorem 16 when ignoring logarithmic terms (hidden in* $\tilde{\mathcal{O}}$*). Similar statements can be proven in the settings of Theorem 22 and 24 below, but are omitted for brevity.*

Remark 19 *By imposing additionally* $T = \Omega\bigl(\frac{L(\tau+M)}{\mu}\bigr)$ *the constant in front of the exponential term in Theorem 16 could be improved to* μ *instead of* $L(\tau+M)$ *(see e.g. Karimireddy et al., 2020, Lemma 1).*

Theorem 16 shows that only the (asymptotically faster decaying) optimization terms (the terms depending on the initial error $\|\mathbf{x}_0 - \mathbf{x}^\star\|$) are impacted by the delay τ; the stochastic terms (the terms depending on σ), are unaffected by the delay. This means that stochastic delayed gradient descent converges for any constant delay τ asymptotically at the same rate as stochastic gradient descent without delays. This makes delayed gradient updates a powerful technique to e.g. hide communication cost in distributed environments.

Arjevani et al. (2020) prove for strongly convex quadratic functions and uniformly bounded noise ($M = 0$) an upper bound of $\tilde{\mathcal{O}}\bigl(L\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 \exp\bigl[-\frac{\mu T}{10L\tau}\bigr] + \frac{\sigma^2}{\mu T}\bigr)$. We see that the statistical term precisely matches with our result, for the fast decaying exponential term only the prefactors (L vs. $L\tau$) are in a slight mismatch(this might be an artifact of our proof technique, see also Remark 19). Both results imply a $\tilde{O}\bigl(\frac{\sigma^2}{\mu\epsilon} + \tau\frac{L}{\mu}\bigr)$ iteration complexity to reach a target accuracy $\epsilon > 0$. The effect of the delay τ becomes negligible if the target accuracy is smaller than $\tilde{O}\bigl(\frac{\sigma^2}{L\tau}\bigr)$, or when the number of iterations T is sufficiently larger than $\tilde{\Omega}\bigl(\frac{L\tau}{\mu}\bigr)$ (cf. the analogous discussion in (Arjevani et al., 2020)). This is a very mild condition, as we must have $T \geq \tau$ to even observe a single stochastic gradient at the starting point $\mathbf{x}_0$.

Arjevani et al. (2020, Theorem 3) derive a lower bound of $\tilde{\Omega}\big(\tau\sqrt{L/\mu}\big)$ in the deterministic ($\sigma^2 = 0$) setting. It follows that the linear dependence on the delay τ is optimal and cannot further be improved. Complexity estimates with the square root of the condition number $\frac{L}{\mu}$ are only reached for *accelerated* gradient methods, so it is no surprise that our (non-accelerated) gradient methods does not reach the same complexity.

For the general quasi-convex case ($\mu = 0$), our upper bound matches precisely the bound in (Arjevani et al., 2020), but extends the analysis to non-quadratic functions under Assumption 1. From the iteration complexity $\mathcal{O}\big(\frac{L\tau\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\epsilon} + \frac{\|\mathbf{x}_0-\mathbf{x}^\star\|^2\sigma^2}{\epsilon^2}\big)$ we see that the effect of the delay becomes negligible if the number of iterations T is sufficiently larger than $\Omega\big(\frac{L^2\tau^2\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\sigma^2}\big)$. Note the quadratic dependence on τ in the previous condition, as opposed to the strongly convex case where the dependence was only linear. With *accelerated* stochastic methods (e.g. (Ghadimi and Lan, 2012)), one could hope to attain an improved iteration complexity of $\mathcal{O}\big(\tau\sqrt{\frac{L\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\epsilon}} + \frac{\|\mathbf{x}_0-\mathbf{x}^\star\|^2\sigma^2}{\epsilon^2}\big)$. Such a result would imply that after $\Omega\big(\tau^{4/3}\big(\frac{L\|\mathbf{x}_0-\mathbf{x}^\star\|}{\sigma}\big)^{2/3}\big)$ iterations, the effect of the delay would be negligible. Obtaining such an improvement as well as studying its optimality is left for future work.

For arbitary non-convex functions, we obtain an iteration complexity of $\mathcal{O}\big(\frac{L\tau(f(\mathbf{x}_0)-f^\star)}{\epsilon} + \frac{L\sigma^2(f(\mathbf{x}_0)-f^\star)}{\epsilon^2}\big)$ to reach $\mathbb{E}\left\|\nabla f(\mathbf{x}^{\text{out}})\right\|^2 \leq \epsilon$. Thus, when the number of iterations is larger than $\Omega\big(\frac{L\tau^2(f(\mathbf{x}_0)-f^\star)}{\sigma^2}\big)$, the first term becomes small and the effect of the delay becomes negligible. This matches with previous analysis of bounded delayed methods for non-convex functions (Lian et al., 2015). Like in the general quasi-convex case previously studied, the dependence on τ here is quadratic. However, unlike in the general quasi-convex case, we believe that this is optimal.

For smooth deterministic non-convex functions ($\sigma^2 = 0$), Carmon et al. (2017) show a lower bound of $\Omega(1/\epsilon)$ for *exact* gradient methods to reach an ϵ stationary point. Thus, unlike in the convex case, the rate cannot be improved using acceleration. We believe that one can prove the iteration complexity of $\mathcal{O}(\tau/\epsilon)$ shown here is optimal by combining the techniques of Arjevani et al. (2020, Theorem 3) and Carmon et al. (2017). Further, (Arjevani et al., 2019) show that the second stochastic term $\mathcal{O}(\sigma^2/\epsilon^2)$ is also optimal. Together, this shows that our rates are unimprovable for general smooth non-convex functions.

Our analysis here focuses on (D-SGD), where the delays are exactly τ. However, it will become clear form the proof that our analysis applies to more general settings, for instance as in (Feyzmahdavian et al., 2016); it suffices that τ denotes an upper bound on the largest delay.

We provide the proof of Theorem 16 in Section 4.2 below. But first, we would like to add a few comments on mini-batch SGD.

4.1 Mini-Batch SGD

Mini-batch SGD (Dekel et al., 2012) is a standard algorithm for distributed optimization, especially large scale machine learning. Each update step is only performed after accumulating a mini-batch of τ stochastic gradients, all computed with respect to the same point:

$$\mathbf{x}_{t+1} = \begin{cases} \mathbf{x}_t - \frac{\gamma}{\tau}\sum_{i=0}^{\tau-1}\big(\nabla f(\mathbf{x}_{t-i}) + \boldsymbol{\xi}_{t-i}\big)\,, & \text{if } \tau|(t+1), \\ \mathbf{x}_t, & \text{otherwise.} \end{cases}$$

We can equivalently write the updates in the (EC-SGD) framework:

$$\mathbf{v}_t = \begin{cases} \frac{\gamma}{\tau}\sum_{i=0}^{\tau-1}\big(\nabla f(\mathbf{x}_{t-i}) + \boldsymbol{\xi}_{t-i}\big)\,, & \text{if } \tau|(t+1), \\ \mathbf{0}_d, & \text{otherwise,} \end{cases} \qquad \mathbf{e}_t = \gamma \sum_{i=\lfloor t/\tau\rfloor\cdot\tau}^{t-1}\big(\nabla f(\mathbf{x}_i) + \boldsymbol{\xi}_i\big)\,.$$

This means, $\mathbf{x}_{b\tau+0} = \mathbf{x}_{b\tau+1} = \mathbf{x}_{(b+1)\tau-1}$ and the error $\mathbf{e}_{b\tau} = \mathbf{0}_d$ for every integer $b \geq 0$. Mini-batch SGD can be seen as a synchronous version of (D-SGD) in the sense that instead of applying the updates with the constant delay τ, the method waits for τ gradients to be computed and applies them in a combined update.

Our proof technique also applies to mini-batch SGD (though we will not give the computations in detail here—the proof follows in close analogy to the proof of Theorem 16) and gives a complexity (number of stochastic gradient computations[2]) on quasi-convex functions of $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \tau\frac{L}{\mu}\big)$ when $\mu > 0$ and $\mathcal{O}\big(\frac{L\tau\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\epsilon} + \frac{\|\mathbf{x}_0-\mathbf{x}^\star\|^2\sigma^2}{\epsilon^2}\big)$ when $\mu = 0$. For general non-convex functions we obtain iteration complexity of $\mathcal{O}\big(\frac{L\tau(f(\mathbf{x}_0)-f^\star)}{\epsilon} + \frac{L\sigma^2(f(\mathbf{x}_0)-f^\star)}{\epsilon^2}\big)$. These bounds match with the known bounds for this method (up to logarithmic factors) (see e.g. Dekel et al., 2012; Bottou et al., 2018; Stich, 2019b). Moreover, we see that the complexity of the delayed method matches with mini-batch SGD.

4.2 Proof of Theorem 16

We start with a key lemma where we derive an upper bound on $\mathbb{E}\,\|\mathbf{e}_t\|^2$.

Lemma 20 *Let $\{\mathbf{e}_t\}_{t\geq 0}$ be defined as in* (17) *for stepsizes $\{\gamma_t\}_{t\geq 0}$ with $\gamma_t \leq \frac{1}{10L(\tau+M)}$, $\forall t \geq 0$ and $\{\gamma_t^2\}_{t\geq 0}$ τ-slow decaying. Then*

$$\mathbb{E}\left[3L\,\|\mathbf{e}_t\|^2\right] \leq \frac{1}{10L\tau}\sum_{i=1}^{\tau}\mathbb{E}\,\|\nabla f(\mathbf{x}_{t-i})\|^2 + 2\gamma_t\sigma^2\,. \tag{18}$$

Furthermore, for any τ-slow increasing sequence $\{w_t\}_{t\geq 0}$ of non-negative values it holds:

$$3L\sum_{t=0}^{T} w_t\,\mathbb{E}\,\|\mathbf{e}_t\|^2 \leq \frac{1}{5L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\,\|\nabla f(\mathbf{x}_{t-i})\|^2\right) + 2\sigma^2\sum_{t=0}^{T} w_t\gamma_t\,. \tag{19}$$

[2] Sometimes the complexity bounds for mini-batch SGD are written in terms of iterations, each comprising τ stochastic gradient computations (one mini-batch). The complexity estimates in number of *iterations* translate to $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\tau\epsilon} + \frac{L}{\mu}\big)$, $\mathcal{O}\big(\frac{L\|\mathbf{x}_0-\mathbf{x}^\star\|^2}{\epsilon} + \frac{\|\mathbf{x}_0-\mathbf{x}^\star\|^2\sigma^2}{\tau\epsilon^2}\big)$ and $\mathcal{O}\big(\frac{L(f(\mathbf{x}_0)-f^\star)}{\epsilon} + \frac{L\sigma^2(f(\mathbf{x}_0)-f^\star)}{\tau\epsilon^2}\big)$ respectively.

Proof We start with the first claim. By definition and Lemma 15 from above (with $\beta = \frac{1}{2}$), we have the bound:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{e}_t\right\|^2 &= \mathbb{E}\left\|\sum_{i=1}^{\tau} \gamma_{t-i}\big(\nabla f(\mathbf{x}_{t-i}) + \boldsymbol{\xi}_{t-i})\big)\right\|^2 \\
&\overset{(16)}{\leq} \frac{3}{2}\min\{\tau,t\}\sum_{i=1}^{\tau}\gamma_{t-i}^2\left\|\nabla f(\mathbf{x}_{t-i})\right\|^2 + 3\sum_{i=1}^{\tau}\gamma_{t-i}^2\,\mathbb{E}\left\|\boldsymbol{\xi}_{t-i}\right\|^2 \\
&\overset{(7)}{\leq} \frac{3}{2}(\tau+M)\sum_{i=1}^{\tau}\gamma_{t-i}^2\left\|\nabla f(\mathbf{x}_{t-i})\right\|^2 + 3\sigma^2\sum_{i=1}^{\tau}\gamma_{t-i}^2\,.
\end{aligned}$$

For $i \leq \tau$ we have the upper bound $\gamma_{t-i}^2 \leq \gamma_t^2\big(1+\frac{1}{2\tau}\big)^\tau \leq \gamma_t^2 \exp\left[\frac{\tau}{2\tau}\right] \leq 2\gamma_t^2$, as $1 + x \leq e^x$, $\forall x \in \mathbb{R}$. Thus we can simplify:

$$\mathbb{E}\left\|\mathbf{e}_t\right\|^2 \leq \gamma_t^2(\tau+M)\sum_{i=1}^{\tau}\Big(3\left\|\nabla f(\mathbf{x}_{t-i})\right\|^2 + 6\min\{\tau,t\}\sigma^2\Big)\,.$$

By observing that the choice $\gamma_t \leq \frac{1}{10L(\tau+M)}$ implies $\big(3L\cdot 3(\tau+M)\gamma_t^2\big) \leq \frac{1}{10L(\tau+M)} \leq \frac{1}{10L\tau}$ and $\big(3L\cdot 6\tau\gamma_t\big) \leq \frac{2L\tau}{L(\tau+M)} \leq 2$ we show the first claim.

For the second claim, we observe that for τ-slow increasing $\{w_t\}_{t\geq 0}$ we have $w_t \leq w_{t-i}\big(1+\frac{1}{2\tau}\big)^i \leq w_{t-i}\big(1+\frac{1}{2\tau}\big)^\tau \leq w_{t-i}\exp\left[\frac{1}{2}\right] \leq 2w_{t-i}$ for every $0 \leq i \leq \tau$. Thus we can estimate

$$\begin{aligned}
3L\sum_{t=0}^{T} w_t\,\mathbb{E}\left\|\mathbf{e}_t\right\|^2 &\overset{(18)}{\leq} \frac{1}{5L}\sum_{t=0}^{T}\frac{w_t}{2\tau}\sum_{i=1}^{\tau}\Big(\mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i})\right\|^2\Big) + 2\sigma^2\sum_{t=0}^{T} w_t\gamma_t \\
&\leq \frac{1}{5L}\sum_{t=0}^{T}\frac{1}{\tau}\sum_{i=1}^{\tau} w_{t-i}\Big(\mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i})\right\|^2\Big) + 2\sigma^2\sum_{t=0}^{T} w_t\gamma_t \\
&\leq \frac{1}{5L}\sum_{t=0}^{T} w_t\Big(\mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i})\right\|^2\Big) + 2\sigma^2\sum_{t=0}^{T} w_t\gamma_t\,.
\end{aligned}$$

This concludes the proof. ■

This lemma, together with the estimate on the one step progress derived in Lemma 8 for the convex case and Lemma 9 for the non-convex case, allows us to obtain a recursive description of the suboptimality. To obtain the complexity estimates, we follow closely the technique outlined in (Stich, 2019b).

Proof of Theorem 16 We split the proof in two parts. First dealing with the three cases under the quasi-convexity Assumption 1, and then addressing the remaining case without Assumption 1.

Quasi convex functions (claims 1–2, and Remark 18). Lemma 20 together with (5) gives

$$3L\sum_{t=0}^{T} w_t\,\mathbb{E}\left\|\mathbf{e}_t\right\|^2 \leq \frac{2}{5}\sum_{t=0}^{T} w_t\left(\mathbb{E}\,f(\mathbf{x}_{t-i}) - f^\star\right) + 2\sigma^2\sum_{t=0}^{T} w_t\gamma_t\,.$$

We would like to remark that we could have written Lemma 20 directly in the form as given above, relying on only Assumption 3* instead of the stronger Assumption 3 (see Remark 6). However, this would be insufficient for proving convergence to stationary points for arbitrary non-convex functions (claim 4) as we will see later.

Observe that the conditions of Lemma 8 are satisfied. With the notation $r_t := \mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\right\|^2$ and $s_t := \mathbb{E} f(\mathbf{x}_t) - f^\star$ we thus have for any $w_t > 0$:

$$\frac{w_t}{2} s_t \overset{(12)}{\leq} \frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + \gamma_t w_t \sigma^2 + 3 w_t L\, \mathbb{E}\left\|\mathbf{e}_t\right\|^2 .$$

Suppose now—we show this below—that the conditions of Lemma 20 are satisfied. With this lemma we have:

$$\frac{1}{2}\sum_{t=0}^T w_t s_t \leq \sum_{t=0}^T \left(\frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + 3\gamma_t w_t \sigma^2 \right) + \frac{2}{5}\sum_{t=0}^T w_t s_t .$$

This can be rewritten as

$$\frac{1}{W_T}\sum_{t=0}^T w_t s_t \leq \frac{10}{W_T}\sum_{t=0}^T \left(\frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + 3\gamma_t w_t \sigma^2 \right) =: \Xi_T .$$

All that is remaining is to check that the conditions of Lemma 20 are indeed satisfied, and to derive an estimate on Ξ_T. For this, we discuss the two cases of the theorem separately.

For the first claim (constant stepsize, $\mu > 0$), the conditions of Lemma 20 are easy to check, as $\gamma \leq \frac{1}{10L(\tau+M)}$ by definition. We further observe that $\left(1 - \frac{\mu\gamma}{2}\right) \geq \left(1 - \frac{\mu}{20L(\tau+M)}\right) \geq \left(1 - \frac{1}{8\tau}\right)$ and by Example 1 it follows that weights chosen as $w_t = (1 - \frac{\mu\gamma}{2})^{-(t+1)}$ are 2τ-slow increasing (and hence also τ-slow increasing). The claim follows by Lemma 20, Lemma 13 and observing that the indicated sampling probability for choosing $\mathbf{x}^{\text{out}}$ out of $\{\mathbf{x}_t\}_{t=0}^{T-1}$ matches with the chosen weights w_t.

For the second claim (constant stepsize, $\mu = 0$), we invoke Lemma 14 with weights $\{w_t = 1\}_{t\geq 0}$.

Finally, for the stepsizes $\gamma_t = \frac{4}{\mu(\kappa+t)}$ as used in Remark 18, we observe $\gamma_t \leq \gamma_0 = \frac{4}{\mu\kappa} \leq \frac{1}{10L(\tau+M)}$, by the choice of κ. In Remark 11 we have further shown that $\{\gamma_t^2\}_{t\geq 0}$ is τ slow decreasing, as $\kappa \geq 8\tau$ (and $\tau \geq 1$). Furthermore, in Example 1 we have show that the weights $\{w_t = \kappa + t\}_{t\geq 0}$ are 2τ-slow increasing for $\kappa \geq 16\tau$ (and hence they are also τ-slow increasing). Thus, the conditions of Lemma 20 are indeed satisfied and the technical Lemma 12 provides the claimed upper bound on Ξ_T.

Arbitrary smooth non-convex functions (claim 3). For the non-convex case, Lemma 9 gives us the progress of one step. Using notation $r_t := 4\,\mathbb{E}(f(\tilde{\mathbf{x}}_t) - f^\star)$, $s_t := \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2$, $c = 4L\sigma^2$, and $w_t = 1$ we have

$$\begin{aligned}
\frac{1}{4W_T}\sum_{t=0}^T w_t s_t &\overset{(14)}{\leq} \frac{1}{W_T}\sum_{t=0}^T w_t \left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{8} \right) + \frac{L^2}{2W_T}\sum_{t=0}^T w_t\, \mathbb{E}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2 \\
&\overset{(19)}{\leq} \frac{1}{W_T}\sum_{t=0}^T w_t \left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{8} \right) + \frac{L^2}{2W_T}\sum_{t=0}^T \left(\frac{1}{15L^2} w_t s_t + \frac{w_t \gamma_t c}{4L^2} \right) .
\end{aligned}$$

The above equation can be simplified as:

$$\frac{1}{5W_T}\sum_{t=0}^{T} w_t s_t \leq \frac{1}{W_T}\sum_{t=0}^{T} w_t \left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{4}\right) .$$

We can now invoke Lemma 14 with weights $\{w_t = 1\}_{t\geq 0}$ to finish the proof. ■

5. Error Compensated SGD with Arbitrary Compressors

In this section we analyze SGD with error-feedback (or error-compensation) and generalize and improve the results of (Stich et al., 2018; Karimireddy et al., 2019) through the more refined analysis developed here. This method is of particular importance in distributed optimization to reduce communication costs, but we consider only the single worker case here.

We consider algorithms that take the following form in the (EC-SGD) framework:

$$\mathbf{v}_t := \mathcal{C}(\mathbf{e}_t + \gamma_t \mathbf{g}_t)\,, \qquad\qquad \mathbf{e}_{t+1} := \mathbf{e}_t + \gamma_t \mathbf{g}_t - \mathbf{v}_t\,, \tag{20}$$

where here $\mathcal{C}$ denotes a δ-*approximate compressor* or δ-compressor for short.

Definition 21 (δ-approximate compressor) *A random operator $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ that satisifes for a parameter $\delta > 0$:*

$$\mathbb{E}_{\mathcal{C}}\|\mathbf{x} - \mathcal{C}(\mathbf{x})\|^2 \leq (1-\delta)\,\|\mathbf{x}\|^2\ , \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{21}$$

In contrast to (D-SGD) studied in the previous section, we do not precisely know the explicit structure of $\mathbf{e}_t$ here (e.g. if it can be written as the sum of τ stochastic gradient estimators). Instead, for δ-approximate compressors, we only know an upper bound on the squared norm $\|\mathbf{e}_t\|^2$. The notion (20) comprises a much richer class of algorithms. For illustration—and to highlight the connection the previous section—consider the compressor $\mathcal{C}_\tau$, defined for a parameter $\tau \geq 1$ as:

$$\mathcal{C}_\tau(\mathbf{x}) = \begin{cases} \mathbf{0}_d\,, & \text{with probability } 1 - \frac{1}{\tau}\,, \\ \mathbf{x}\,, & \text{with probability } \frac{1}{\tau}\,. \end{cases}$$

This operator is a $\delta = \frac{1}{\tau}$ compressor. Moreover, $\mathbf{e}_t$ can be written as the sum of (in expectation) τ stochastic gradients. Thus, we would expect the algorithm (D-SGD) to behave similarly as algorithm (20) with a $\delta = \frac{1}{\tau}$ compressor. Indeed, this intuition is true and Theorem 22 can be seen as a generalization of Theorem 16 with τ replaced by $\frac{2}{\delta}$. The following operator

$$\left[\mathcal{S}_\delta(\mathbf{x})\right]_i = \begin{cases} 0\,, & \text{with probability } 1-\delta\,, \\ [\mathbf{x}]_i\,, & \text{with probability } \delta\,, \end{cases}$$

where $[\mathbf{x}]_i := \langle \mathbf{x}, \mathbf{e}_i\rangle$, is also a δ-approximate compressor. This shows that this notion not only comprises delayed gradients, but also delayed (atomic) block-coordinates updates.

We refer e.g. to (Alistarh et al., 2018; Stich et al., 2018; Cordonnier, 2018; Karimireddy et al., 2019) for the discussion of key examples, including sparsification and quantization (that we have not discussed here). However, it is important to note that whilst our framework (EC-SGD) covers distributed SGD implementations in general (see for instance Section 4.1), we here analyze error compensated SGD only for the special case of a single-machine implementation and our results do no apply to a fully distributed optimization setting as for instance considered in (Cordonnier, 2018). Fully compressed communication in such a distributed setting involves each machine maintaining a separate error vector, and hence does not directly fit in our framework (EC-SGD).

Theorem 22 *Let $\{\mathbf{x}_t\}_{t\geq 0}$ denote the iterates of the error compensated stochastic gradient descent* (20) *with constant stepsize $\{\gamma_t = \gamma\}_{t\geq 0}$ and with a δ-approximate compressor on a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ under assumptions Assumptions 2 and 3. Then, if f*

- *satisfies Assumption 1 for $\mu > 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(2/\delta+M)}$ (chosen as in Lemma 13) such that*

$$\mathbb{E}\, f(\mathbf{x}^{\text{out}}) - f^\star = \tilde{\mathcal{O}}\left(L(1/\delta+M)\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 \exp\left[-\frac{\mu T}{10L(2/\delta+M)}\right] + \frac{\sigma^2}{\mu T}\right),$$

  where the output $\mathbf{x}^{\text{out}} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ is chosen to be $\mathbf{x}_t$ with probability proportional to $(1 - \mu\gamma/2)^{-t}$.

- *satisfies Assumption 1 for $\mu = 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(2/\delta+M)}$ (chosen as in Lemma 14) such that*

$$\mathbb{E}\, f(\mathbf{x}^{\text{out}}) - f^\star = \mathcal{O}\left(\frac{L(1/\delta+M)\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2}{T} + \frac{\sigma\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|}{\sqrt{T}}\right),$$

  where the output $\mathbf{x}^{\text{out}} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ is chosen uniformly at random from the iterates $\{\mathbf{x}_t\}_{t=0}^{T-1}$.

- *is an arbitrary non-convex function, then there exists a stepsize $\gamma \leq \frac{1}{10L(1/\delta+M)}$ (chosen as in Lemma 14), such that*

$$\mathbb{E}\left\|\nabla f(\mathbf{x}^{\text{out}})\right\|^2 = \mathcal{O}\left(\frac{L(1/\delta+M)(f(\mathbf{x}_0) - f^\star)}{T} + \sigma\sqrt{\frac{L(f(\mathbf{x}_0) - f^\star)}{T}}\right).$$

  where the output $\mathbf{x}^{\text{out}} \in \{\mathbf{x}_t\}_{t=0}^{T-1}$ is chosen uniformly at random from the iterates $\{\mathbf{x}_t\}_{t=0}^{T-1}$.

In analogy to Theorem 16, this result shows that the stochastic terms (the ones depending on σ) in the rate are not affected by the δ parameter. Stich et al. (2018) proved under the bounded gradient assumption for strongly convex functions an upper bound of $\mathcal{O}\big(\frac{\mu}{\delta^3 T^3}\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 + \frac{LG^2}{\mu^2\delta^2T^2} + \frac{G^2}{\mu T}\big)$ which implies an iteration complexity of $\Omega\big(\frac{G^2}{\mu\epsilon} + \frac{G\sqrt{L}}{\mu\delta\sqrt{\epsilon}}\big)$. Our second result implies iteration complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{L}{\mu\delta}\big)$ which is strictly better, as $\sigma^2 \leq G^2$ in general. Moreover, as in the previous section we notice that the impact of the compression becomes negligible if ϵ is smaller than $\tilde{\mathcal{O}}\big(\frac{\delta\sigma^2}{L}\big)$ or $T = \tilde{\Omega}\big(\frac{L}{\mu\delta}\big)$. As only for $T = \Omega\big(\frac{1}{\delta}\big)$ the first gradient is fully received (in expectation), this is again a very mild condition that implies *compression for free*, i.e. without increasing the iteration complexity.

By the same arguments as in the previous section and considering the operator $\mathcal{C}_\tau$ from above, we see that the linear dependency on $\frac{1}{\delta}$ cannot further be improved in general. The general quasi convex setting ($\mu = 0$) was not studied in previous work.

Under the bounded gradient assumption, Karimireddy et al. (2019) prove that that the gradient norm converges at a rate of $\mathcal{O}\big(G\sqrt{\frac{(f(\mathbf{x}_0)-f^\star)}{T}} + \frac{LG^2}{\delta^2 T}\big)$. This translates to an iteration complexity of $\mathcal{O}\big(\frac{L(f(\mathbf{x}_0)-f^\star)}{\delta^2\epsilon} + \frac{LG^2(f(\mathbf{x}_0)-f^\star)}{\epsilon^2}\big)$. In contrast, our rates give an iteration complexity of $\mathcal{O}\big(\frac{L(f(\mathbf{x}_0)-f^\star)}{\delta\epsilon} + \frac{L\sigma^2(f(\mathbf{x}_0)-f^\star)}{\epsilon^2}\big)$. Thus we improve in two regards: first our rates replace the second moment bound with a variance bound, and second we obtain a linear dependence on δ instead of the quadratic dependence in (Karimireddy et al., 2019).

Our improved rates are partially due by relaxing the bounded gradient assumption, and also a more careful bound on the error term.

5.1 Proof of Theorem 22

We follow a similar structure as in the proof of Theorem 16. We first derive an an upper bound on $\mathbb{E}\,\|\mathbf{e}_t\|^2$.

Lemma 23 *Let $\mathbf{e}_t$ be as in* (20) *for a δ-approximate compressor $\mathcal{C}$ and stepsizes $\{\gamma_t\}_{t\geq 0}$ with $\gamma_{t+1} \leq \frac{1}{10L(2/\delta+M)}$, $\forall t \geq 0$ and $\{\gamma_t^2\}_{t\geq 0}$ $\frac{2}{\delta}$-slow decaying. Then*

$$\mathbb{E}\left[3L\,\|\mathbf{e}_{t+1}\|^2\right] \leq \frac{\delta}{64L}\sum_{i=0}^{t}\left(1-\frac{\delta}{4}\right)^{t-i}\left(\mathbb{E}\,\|\nabla f(\mathbf{x}_{t-i})\|^2\right) + \gamma_t\sigma^2\,. \tag{22}$$

Furthermore, for any $\frac{4}{\delta}$-slow increasing non-negative sequence $\{w_t\}_{t\geq 0}$ it holds:

$$3L\sum_{t=0}^{T} w_t\,\mathbb{E}\,\|\mathbf{e}_t\|^2 \leq \frac{1}{8L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\,\|\nabla f(\mathbf{x}_{t-i})\|^2\right) + \sigma^2\sum_{t=0}^{T} w_t\gamma_t\,.$$

Proof We start with the first claim. By definition of the error sequence

$$\mathbb{E}_{\mathcal{C}}\|\mathbf{e}_{t+1}\|^2 \overset{(20)}{=} \mathbb{E}_{\mathcal{C}}\|\mathbf{e}_t + \gamma_t\mathbf{g}_t - \mathcal{C}(\mathbf{e}_t+\gamma_t\mathbf{g}_t)\|^2 \overset{(21)}{\leq} (1-\delta)\,\|\mathbf{e}_t+\gamma_t\mathbf{g}_t\|^2\,.$$

We further simplify with Assumption 3:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}_t}\|\mathbf{e}_{t+1}\|^2 &\leq (1-\delta)\,\mathbb{E}_{\boldsymbol{\xi}_t}\|\mathbf{e}_t + \gamma_t(\nabla f(\mathbf{x}_t)+\boldsymbol{\xi}_t)\|^2\\ &\overset{(7)}{=} (1-\delta)\,\|\mathbf{e}_t+\gamma_t\nabla f(\mathbf{x}_t)\|^2 + \gamma_t^2(1-\delta)\,\mathbb{E}_{\boldsymbol{\xi}_t}\|\boldsymbol{\xi}_t\|^2\\ &\overset{(7)}{\leq} (1-\delta)\,\|\mathbf{e}_t+\gamma_t\nabla f(\mathbf{x}_t)\|^2 + \gamma_t^2(1-\delta)\left(M\,\|\nabla f(\mathbf{x}_t)\|^2+\sigma^2\right)\,.\end{aligned}$$

With $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\beta)\,\|\mathbf{a}\|^2 + (1+\beta^{-1})\,\|\mathbf{b}\|^2$, for any $\beta \geq 0$, we continue:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}_t}\|\mathbf{e}_{t+1}\|^2 &\leq (1-\delta)(1+\beta)\,\|\mathbf{e}_t\|^2 + \gamma_t^2(1-\delta)(1+1/\beta)\,\|\nabla f(\mathbf{x}_t)\|^2\\ &\qquad + \gamma_t^2(1-\delta)\left(M\,\|\nabla f(\mathbf{x}_t)\|^2+\sigma^2\right)\\ &\overset{(5)}{\leq} (1-\delta)(1+\beta)\,\|\mathbf{e}_t\|^2 + \gamma_t^2(1+1/\beta+M)\,\|\nabla f(\mathbf{x}_t)\|^2 + \gamma_t^2\sigma^2\,,\end{aligned}$$

where we used the smoothness Assumption 2 and dropped—for convenience—one superfluous $(1-\delta)$ factor in the last inequality. By unrolling the recurrence, and picking $\beta = \frac{\delta}{2(1-\delta)}$, such that $(1+1/\beta) = (2-\delta)/\delta \leq 2/\delta$, and $(1-\delta)(1+\beta) \leq (1-\delta/2)$ we find

$$\begin{aligned}\mathbb{E}\left\|\mathbf{e}_{t+1}\right\|^2 &\leq \sum_{i=0}^t \gamma_i^2 \left[(1-\delta)(1+\beta)\right]^{t-i} \Big((1+1/\beta+M)\,\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + \sigma^2\Big) \\ &\leq \sum_{i=0}^t \gamma_i^2 \left(1-\frac{\delta}{2}\right)^{t-i} \left(\left(\frac{2}{\delta}+M\right)\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + \sigma^2\right).\end{aligned}$$

For $\frac{2}{\delta}$-slow decreasing $\{\gamma_t^2\}_{t\geq 0}$ it holds $\gamma_i^2 \leq \gamma_t^2\big(1+\frac{\delta}{4}\big)^{t-i}$. As $(1-\delta/2)(1+\delta/4) \leq (1-\delta/4)$, we continue:

$$\begin{aligned}\mathbb{E}\left\|\mathbf{e}_{t+1}\right\|^2 &\leq \sum_{i=0}^t \gamma_t^2 \left(1+\frac{\delta}{4}\right)^{t-i}\left(1-\frac{\delta}{2}\right)^{t-i}\left(\frac{2}{\delta}+M\right)\left\|\nabla f(\mathbf{x}_i)\right\|^2 + \gamma_t^2\sum_{i=0}^t\left(1+\frac{\delta}{4}\right)^{t-i}\sigma^2 \\ &\leq \gamma_t^2 \sum_{i=0}^t \left(1-\frac{\delta}{4}\right)^{t-i}\left(\frac{2}{\delta}+M\right)\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + \gamma_t^2\frac{4\sigma^2}{\delta}.\end{aligned}$$

By observing that the choice of the stepsize $\gamma_t \leq \frac{1}{10L(2/\delta+M)}$ implies $\big(3L\cdot(2/\delta+M)\gamma_t^2\big) \leq \frac{\delta}{64L}$ and $\big(3L\cdot 4/\delta\gamma_t\big) \leq 1$ we prove the first claim.

For the second claim, we observe that for $\frac{4}{\delta}$-slow increasing $\{w_t\}_{t\geq 0}$ we have $w_t \leq w_{t-i}\big(1+\frac{\delta}{8}\big)^i$. Hence,

$$\begin{aligned}3L\sum_{t=0}^T w_t\,\mathbb{E}\left\|\mathbf{e}_t\right\|^2 &\overset{(22)}{\leq} \frac{\delta}{64L}\sum_{t=0}^T\sum_{i=0}^{t-1} w_t\left(1-\frac{\delta}{4}\right)^{t-i}\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + w_t\gamma_{t-1}\sigma^2 \\ &\leq \frac{\delta}{64L}\sum_{t=0}^T\sum_{i=0}^{t-1} w_i\left(1+\frac{\delta}{8}\right)^{t-i}\left(1-\frac{\delta}{4}\right)^{t-i}\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + \sigma^2\sum_{t=0}^T w_t\gamma_t \\ &\leq \frac{\delta}{64L}\sum_{t=0}^T\sum_{i=0}^{t-1} w_i\left(1-\frac{\delta}{8}\right)^{t-i}\mathbb{E}\left\|\nabla f(\mathbf{x}_i)\right\|^2 + \sigma^2\sum_{t=0}^T w_t\gamma_t \\ &\leq \frac{\delta}{64L}\sum_{t=0}^T w_t\,\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2\sum_{i=0}^\infty\left(1-\frac{\delta}{8}\right)^i + \sigma^2\sum_{t=0}^T w_t\gamma_t.\end{aligned}$$

Observing $\sum_{i=0}^\infty(1-\delta/8)^i \leq \frac{8}{\delta}$ concludes the proof. ■

With this lemma we can now complete the proof analogously to the proof of Theorem 16, as we outline next.

Proof of Theorem 22 We observe that by setting $\tau := \frac{2}{\delta}$ in Theorem 22 and Lemma 23, respectively, we fall back in the setting of Theorem 16 and Lemma 20. Thus the proof follows along the same lines and we will not repeat it here. There is only one small caveat: Lemma 23 requires 2τ-slow increasing weights, instead of τ-slow increasing weights as in Lemma 20. However, it can easily be checked that this condition is satisfied (see also the remarks in the proof of Theorem 16). ■

6. Local SGD with Infrequent Communication

In this section we analyze local SGD (parallel SGD) with the developed tools. We follow closely (Stich, 2019a) and provide an analysis without the bounded gradient assumption.

The local SGD algorithm evolves $K \geq 1$ sequences $\left\{\mathbf{x}_t^k\right\}_{t\geq 0}^{k\in[K]}$ in parallel, for an integer K. The sequences are synchronized every $\tau \geq 1$ iterations, in the following way:

$$\mathbf{x}_{t+1}^k = \begin{cases} \frac{1}{K}\sum_{k=1}^K \left(\mathbf{x}_t^k - \gamma_t \mathbf{g}_t^k\right) & \text{if } \tau|(t+1), \\ \mathbf{x}_t^k - \gamma_t \mathbf{g}_t^k & \text{otherwise.} \end{cases} \tag{23}$$

Extending our notion in a natural way, we denote by $\mathbf{g}_t^k$ the gradient oracle on worker k at iteration t, with $\mathbf{g}_t^k = \nabla f(\mathbf{x}_k^t) + \boldsymbol{\xi}_t^k$ and we assume Assumption 3 for the noise $\boldsymbol{\xi}_t^k$.
Using a similar proof technique as in the previous sections, we can derive the following complexity estimates.

Theorem 24 *Let $\left\{\mathbf{x}_t^k\right\}_{t\geq 0}^{k\in[K]}$ denote the iterates of local SGD (23) with constant stepsize $\{\gamma_t = \gamma\}_{t\geq 0}$ on a differentiable function $f \colon \mathbb{R}^d \to \mathbb{R}$ under assumptions Assumptions 2 and 3. Then, if f*

- *satisfies Assumption 1 for $\mu > 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(\tau K + M)}$ (chosen as in Lemma 13) such that*

$$\mathbb{E} f(\mathbf{x}^{\text{out}}) - f^\star = \tilde{\mathcal{O}}\left(L(\tau K + M)\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 \exp\left[-\frac{\mu T}{10L(\tau K + M)}\right] + \frac{\sigma^2}{\mu K T}\right),$$

  where the output $\mathbf{x}^{\text{out}} \in \{\mathbf{x}_t^k\}_{t-1\in[T]}^{k\in[K]}$ is chosen to be $\mathbf{x}_t^k$ with probability proportional to $(1-\mu\gamma/2)^{-t}$ (uniformly over $k \in [K]$).
- *satisfies Assumption 1 for $\mu = 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(\tau K + M)}$ (chosen as in Lemma 14) such that*

$$\mathbb{E} f(\mathbf{x}^{\text{out}}) - f^\star = \mathcal{O}\left(\frac{L(\tau K + M)\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2}{T} + \frac{\sigma\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|}{\sqrt{KT}}\right),$$

  where the output $\mathbf{x}^{\text{out}} \in \{\mathbf{x}_t^k\}_{t-1\in[T]}^{k\in[K]}$ is chosen uniformly at random from the iterates $\{\mathbf{x}_t^k\}_{t-1\in[T]}^{k\in[K]}$.
- *is an arbitrary non-convex function, then there exists a stepsize $\gamma \leq \frac{1}{10L(\tau K + M)}$ (chosen as in Lemma 14), such that*

$$\mathbb{E}\left\|\nabla f(\mathbf{x}^{\text{out}})\right\|^2 = \mathcal{O}\left(\frac{L(\tau K + M)(f(\mathbf{x}_0) - f^\star)}{T} + \sigma\sqrt{\frac{L(f(\mathbf{x}_0) - f^\star)}{KT}}\right).$$

  where the output $\mathbf{x}^{\text{out}} \in \{\mathbf{x}_t^k\}_{t-1\in[T]}^{k\in[K]}$ is chosen uniformly at random from the iterates $\{\mathbf{x}_t^k\}_{t-1\in[T]}^{k\in[K]}$.

Stich (2019a) shows that under the bounded gradient assumption, local SGD can converge on strongly convex functions at the optimal statistical rate $\mathcal{O}\big(\frac{\sigma^2}{\mu K T}\big)$ given $T = \Omega(\frac{L\tau^2 K}{\mu})$. Here we study a more general setting and prove an iteration complexity of $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu K \epsilon} + \frac{L(\tau K + M)}{\mu}\big)$, i.e. the same ϵ dependency for the statistical term, but much milder dependency on the the optimization term. Local SGD achieves optimal $\mathcal{O}\big(\frac{\sigma^2}{\mu K \epsilon}\big)$ iteration complexity if $T = \tilde{\Omega}\big(\frac{L\tau K}{\mu}\big)$. This improves the previously known bound of $T = \Omega(\tau^2 K)$ with quadratic dependence on τ to the linear $\tilde{\Omega}(\tau K)$ dependence. We now compare these estimates to the complexity of mini-batch SGD. To make the comparison and discussion of results easier to follow, we will express all complexity bounds in this paragraph in terms of total stochastic gradient computations, i.e. Theorem 24 gives for local SGD a complexity estimate of $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{K^2 L \tau}{\mu}\big)$ when $M = 0$ and $\mu > 0$. Two settings are of particular interest: (i) first, we consider mini-batch SGD with batch size K, that has oracle complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{KL}{\mu}\big)$ (recalling the results from Section 4.1). We observe that local SGD reaches the same statistical term with τ-times less communication, but a worse optimization term. Another interesting setting is the comparison to (ii) mini-batch SGD with much larger batch size τK but with the same number of communication rounds. This algorithm has oracle complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{KL\tau}{\mu}\big)$. When $K = \mathcal{O}(1)$, local SGD has the same complexity as mini-batch SGD under this setting.

Besides these positive observations, the bound provided here does not seem to be optimal (especially the dependency on K). For instance, we see that the estimate becomes vacuous when $\tau = T$. However, we know that we should at least expect (oracle) complexity $\tilde{\mathcal{O}}\big(\frac{\sigma^2}{\mu\epsilon} + \frac{KL}{\mu}\big)$ in this case (convergence of each individual sequence). This indicates that our balancing of the statistical and the optimization term is not optimal. In fact, currently the best known lower bound for the oracle complexity of local SGD is $\tilde{\Omega}\big(\frac{\sigma^2}{\mu\epsilon} + K\sqrt{\frac{L}{\mu}}\big)$ by Woodworth et al. (2018). Even ignoring the dependence on the condition number (which is improvable via acceleration), our rates are off by $\mathcal{O}(\tau K)$.

Patel and Dieuleveut (2019) present a very detailed analysis of local SGD on strongly convex and smooth functions, not only considering the convergence in function value as we do here, but by providing more refined analysis on the behavior of the iterates, following (Bach and Moulines, 2011; Dieuleveut et al., 2017). However, they consider only polynomial (Polyak-Ruppert) averaging and do thus not recover the exponential decaying dependency on the initial bias in the complexity estimates. Combining their estimates with the exponential averaging might be an interesting future direction.

6.1 Proof of Theorem 24

Analogously to the virtual iterate $\tilde{\mathbf{x}}_t$ in the previous proofs, we define here a (virtual) averaged iterate $\tilde{\mathbf{x}}_t$, by setting $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$ and

$$\tilde{\mathbf{x}}_{t+1} := \tilde{\mathbf{x}}_t - \frac{\gamma_t}{K} \sum_{k=1}^{K} \mathbf{g}_t^k\,, \qquad \forall t \geq 0. \tag{24}$$

We need a slightly adapted version of Lemmas 8 and 9.

Lemma 25 *Let* $\{\mathbf{x}_t^k\}_{t\geq 0}^{k\in[K]}$ *be defined as in* (23) *with gradient oracles* $\{\mathbf{g}_t^k\}_{t\geq 0}^{k\in[K]}$ *and objective function* $f\colon \mathbb{R}^d \to \mathbb{R}$ *as in Assumptions 1–3. If* $\gamma_t \leq \frac{K}{4L(K+M)}$, $\forall t \geq 0$, *then for* $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$ *defined as in* (24),

$$\mathbb{E}\,\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 \leq \Big(1 - \frac{\mu\gamma_t}{2}\Big)\,\mathbb{E}\,\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\|^2 - \frac{\gamma_t}{2K}\sum_{k=1}^K \mathbb{E}\,(f(\mathbf{x}_t^k) - f^\star) + \frac{\gamma_t^2\sigma^2}{K} + \frac{3L\gamma_t}{K}\sum_{k=1}^K \mathbb{E}\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 .$$

Lemma 26 *Let* $\{\mathbf{x}_t^k\}_{t\geq 0}^{k\in[K]}$ *be defined as in* (23) *with gradient oracles* $\{\mathbf{g}_t^k\}_{k\in[K],t\geq 0}$ *and a smooth possibly non-convex function* $f\colon \mathbb{R}^d \to \mathbb{R}$ *satisfying Assumptions 2 and 3. If* $\gamma_t \leq \frac{K}{2L(K+M)}$, $\forall t \geq 0$, *then for* $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$ *defined as in* (24),

$$\mathbb{E}\,f(\tilde{\mathbf{x}}_{t+1}) \leq \mathbb{E}\,f(\tilde{\mathbf{x}}_t) - \frac{\gamma_t}{4K}\sum_{k=1}^K \mathbb{E}\left\|\nabla f(\mathbf{x}_t^k)\right\|^2 + \frac{\gamma_t^2 L\sigma^2}{2K} + \frac{L^2\gamma_t}{2K}\sum_{k=1}^K \mathbb{E}\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 .$$

The proofs lemmas 25 and 26 are very similar to those of Lemmas 8 and 9. We defer them to the appendix.

Similar as in the previous sections, we will now first derive an upper bound on $\mathbb{E}\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2$ in Lemma 27 below. The theorem then follows analogously to the proofs of Theorem 16 and 22 and we omit it here.

Lemma 27 *Let* $\{\tilde{\mathbf{x}}_t\}_{t\geq 0}$, $\{\mathbf{x}_t^k\}_{t\geq 0}^{k\in[K]}$ *be defined as above and stepsizes* $\{\gamma_t\}_{t\geq 0}$ *with* $\gamma_t \leq \frac{1}{10L(\tau K+M)}$, $\forall t \geq 0$ *and* $\{\gamma_t^2\}_{t\geq 0}$ *is* τ*-slow decaying. Then*

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[3L\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|\right]^2 \leq \frac{1}{10L\tau K}\sum_{k=1}^K\sum_{i=0}^{\tau-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i}^k)\right\|^2 + 2\frac{\gamma_t\sigma^2}{K} . \tag{25}$$

Furthermore, for any τ*-slow increasing non-negative sequence* $\{w_t\}_{t\geq 0}$ *it holds:*

$$\frac{1}{K}\sum_{k=1}^K\sum_{t=0}^T \mathbb{E}\left[3L\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|\right]^2 \leq \frac{1}{5LK}\sum_{k=1}^K\sum_{t=0}^T w_t\,\mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i}^k)\right\|^2 + 2\frac{\sigma^2}{K}\sum_{t=0}^T w_t\gamma_t .$$

Proof We start with the first claim. By definition and Lemma 15 from above, we have the bound:

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 &= \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_{\lfloor t/\tau\rfloor\tau} - (\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{\lfloor t/\tau\rfloor\tau})\right\|^2 \leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_{\lfloor t/\tau\rfloor\tau}\right\|^2 \\
&\leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\sum_{i=0}^{\tau-1}\gamma_{t-i}\big(\nabla f(\mathbf{x}_{t-i}^k) + \boldsymbol{\xi}_{t-i}^k\big)\right\|^2 \\
&\overset{(16)}{\leq} \frac{3\tau}{2K}\sum_{k=1}^K\sum_{i=0}^{\tau-1}\gamma_{t-i}^2\,\mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i}^k)\right\|^2 + \frac{3}{K}\sum_{k=1}^K\sum_{i=0}^{\tau-1}\gamma_{t-i}^2\,\mathbb{E}\left\|\boldsymbol{\xi}_{t-i}^k\right\|^2 \\
&\overset{(7)}{\leq} \frac{3(\tau+M)}{2K}\sum_{k=1}^K\sum_{i=0}^{\tau-1}\gamma_{t-i}^2\,\mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i}^k)\right\|^2 + 3\sigma^2\sum_{i=0}^{\tau-1}\gamma_{t-i}^2 ,
\end{aligned}$$

where we used $\mathbb{E}\left\|X-\mathbb{E}\,X\right\|^2 \leq \mathbb{E}\left\|X\right\|^2$, for random variable X, for the first inequality. For $i \leq \tau$ we have the upper bound $\gamma_{t-i}^2 \leq \gamma_t^2\bigl(1+\frac{1}{2\tau}\bigr)^\tau \leq \gamma_t^2 \exp\bigl[\frac{\tau}{2\tau}\bigr] \leq 2\gamma_t^2$, as $1+x \leq e^x$, $\forall x \in \mathbb{R}$. Thus we can simplify:

$$\mathbb{E}\left\|\mathbf{e}_t\right\|^2 \leq \gamma_t^2 \left(\frac{3(\tau+M)}{K} \sum_{k=1}^{K} \sum_{i=0}^{\tau-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_{t-i}^k)\right\|^2 + 6\tau\sigma^2 \right) . \tag{26}$$

By observing that the choice $\gamma_t \leq \frac{1}{10L(\tau K+M)}$ implies $\bigl(3L \cdot 3(\tau+M)\gamma_t^2\bigr) \leq \frac{1}{10L(\tau K+M)} \leq \frac{1}{10L\tau K} \leq \frac{1}{10L\tau}$ and $\bigl(3L \cdot 6\tau\gamma_t\bigr) \leq \frac{2L\tau}{L(\tau K+M)} \leq \frac{2}{K}$ we show the first claim. The second claim follows analogously to the proof in Lemma 20 from (25). This concludes the proof. ■

We note that our proof can recover the subsequent result of (Woodworth et al., 2020) by using (26) directly without the further simplification, and choosing a better step size γ_t as in (Woodworth et al., 2020).

7. Conclusion

We leverage the error-feedback framework to analyze the effect of different forms of delayed updated in a unified manner. We prove that the effects of such delays is negligible for SGD in the presence of noise. This finding comes as no surprise, as it agrees with previous results for SGD with delayed updates or with gradient compression (Chaturapruek et al., 2015; Arjevani et al., 2020; Stich et al., 2018; Karimireddy et al., 2019). We improve on these previous work by providing a tighter non-asymptotic convergence analysis in a more general setting. While our analysis matches with known lower bounds in some settings, in others (such as the local SGD) still leaves a gap. A further limitation of the analysis is that in its current form it is restricted to only unconstrained objectives. Overcoming these limitations, as well as incorporating acceleration, are fruitful avenues for future research. Here. we studied three forms of delays in well prescribed theoretical forms. Similar results can be derived for asynchronous methods with atomic updates under more general conditions (i.e. variable bounded, instead of fixed delays, block-coordinate updates, etc.), as well as a combination of the three delays studied here (e.g. local SGD with compressed communication). These results could be worked out in future work if there is concrete need dictated by practice.

Acknowledgments

We thank Martin Jaggi for comments and pointing out related notions of quasi-convexity, and Thijs Vogels for his comments on this manuscript. We acknowledge funding from SNSF grant 200021_175796, as well as a Google Focused Research Award.

Appendix A. Deferred Proofs

A.1 Proof of Lemma 25

Proof of Lemma 25 We expand:

$$\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 \overset{(24)}{=} \|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\|^2 - \frac{2\gamma_t}{K}\sum_{k=1}^K \left\langle \mathbf{g}_t^k, \mathbf{x}_t^k - \mathbf{x}^\star \right\rangle + \frac{\gamma_t^2}{K^2}\left\|\sum_{k=1}^K \mathbf{g}_t^k\right\|^2 + \frac{2\gamma_t}{K}\sum_{k=1}^K \left\langle \mathbf{g}_t^k, \tilde{\mathbf{x}}_t - \mathbf{x}_t^k \right\rangle ,$$

By using independence,

$$\mathbb{E}_{\boldsymbol{\xi}_t^1,\dots,\boldsymbol{\xi}_t^k}\left\|\sum_{k=1}^K \mathbf{g}_t^k\right\|^2 = \left\|\sum_{k=1}^K \nabla f(\mathbf{x}_t^k)\right\|^2 + \sum_{k=1}^K \mathbb{E}\left\|\boldsymbol{\xi}_t^k\right\|^2 \overset{(8)}{\leq} 2L(K+M)\sum_{k=1}^K (f(\mathbf{x}_t^k) - f^\star) + K\sigma^2 .$$

Thus we can take expectation above:

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 \mid \tilde{\mathbf{x}}_t\right] \overset{(9)}{\leq} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 &- \frac{2\gamma_t}{K}\sum_{k=1}^K \left\langle \nabla f(\mathbf{x}_t^k), \mathbf{x}_t^k - \mathbf{x}^\star \right\rangle \\ &+ \frac{2L(K+M)\gamma_t^2}{K^2}\sum_{k=1}^K (f(\mathbf{x}_t^k) - f^\star) \qquad (27) \\ &+ \frac{\gamma_t^2\sigma^2}{K} + \frac{2\gamma_t}{K}\sum_{k=1}^K \left\langle \nabla f(\mathbf{x}_t^k), \tilde{\mathbf{x}}_t^k - \mathbf{x}_t \right\rangle .\end{aligned}$$

By Assumption 1:

$$-2\left\langle \nabla f(\mathbf{x}_t^k), \mathbf{x}_t^k - \mathbf{x}^\star \right\rangle \overset{(2)}{\leq} -\mu\left\|\mathbf{x}_t^k - \mathbf{x}^\star\right\|^2 - 2(f(\mathbf{x}_t^k) - f^\star) ,$$

and by $2\langle \mathbf{a}, \mathbf{b}\rangle \leq \alpha\|\mathbf{a}\|^2 + \alpha^{-1}\|\mathbf{b}\|^2$ for $\alpha > 0$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,

$$2\left\langle \nabla f(\mathbf{x}_t^k), \tilde{\mathbf{x}}_t - \mathbf{x}_t^k \right\rangle \leq \frac{1}{2L}\left\|\nabla f(\mathbf{x}_t^k)\right\|^2 + 2L\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 \overset{(5)}{\leq} f(\mathbf{x}_t^k) - f^\star + 2L\left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 .$$

And by $\|\mathbf{a} + \mathbf{b}\|^2 \leq (1+\beta)\|\mathbf{a}\|^2 + (1+\beta^{-1})\|\mathbf{b}\|^2$ for $\beta > 0$ (as a consequence of Jensen's inequality), we further observe

$$-\left\|\mathbf{x}_t^k - \mathbf{x}^\star\right\|^2 \leq -\frac{1}{2}\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\|^2 + \left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 .$$

Plugging all these inequalities together into (27) yields

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 \mid \tilde{\mathbf{x}}_t\right] \leq \left(1 - \frac{\mu\gamma_t}{2}\right)\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\|^2 &- \frac{\gamma_t(K - 2L(K+M)\gamma_t)}{K^2}\sum_{k=1}^K (f(\mathbf{x}_t^k) - f^\star) \\ &+ \frac{\gamma_t^2\sigma^2}{K} + \frac{\gamma_t(2L+\mu)}{K}\sum_{k=1}^K \left\|\mathbf{x}_t^k - \tilde{\mathbf{x}}_t\right\|^2 .\end{aligned}$$

The claim follows by the choice $\gamma_t \leq \frac{K}{4L(K+M)}$ and $L \geq \mu$. ■

A.2 Proof of Lemma 26

Proof of Lemma 26 We begin using the definition of $\tilde{\mathbf{x}}_{t+1}$ and the smoothness of f

$$f(\tilde{\mathbf{x}}_{t+1}) \overset{(24)}{\leq} f(\tilde{\mathbf{x}}_t) - \frac{\gamma_t}{K}\sum_{k=1}^K \left\langle \nabla f(\tilde{\mathbf{x}}_t), \mathbf{g}_t^k \right\rangle + \frac{\gamma_t^2 L}{2K^2}\left\| \sum_{k=1}^K \mathbf{g}_t^k \right\|^2 .$$

With Assumption 3 on the noise and using independence we have

$$\mathbb{E}_{\boldsymbol{\xi}_t^1,\dots,\boldsymbol{\xi}_t^k}\left\| \sum_{k=1}^K \mathbf{g}_t^k \right\|^2 \overset{(7)}{=} \left\| \sum_{k=1}^K \nabla f(\mathbf{x}_t^k) \right\|^2 + \sum_{k=1}^K \mathbb{E}\left\| \boldsymbol{\xi}_t^k \right\|^2 \overset{(5),(7)}{\leq} (K+M)\sum_{k=1}^K \left\| \nabla f(\mathbf{x}_t^k) \right\|^2 + K\sigma^2 .$$

Thus we proceed by taking expectation on both sides as follows

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}_t^1,\dots,\boldsymbol{\xi}_t^k} f(\tilde{\mathbf{x}}_{t+1})|\mathbf{x}_t &\leq f(\tilde{\mathbf{x}}_t) - \frac{\gamma_t}{K}\sum_{k=1}^K \left\langle \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t^k) \right\rangle + \frac{\gamma_t^2 L}{2K^2}\mathbb{E}_{\boldsymbol{\xi}_t^1,\dots,\boldsymbol{\xi}_t^k}\left\| \sum_{k=1}^K \mathbf{g}_t^k \right\|^2 \\
&\leq f(\tilde{\mathbf{x}}_t) - \frac{\gamma_t}{K}\sum_{k=1}^K \left\langle \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t^k) \right\rangle + \frac{\gamma_t^2 L(K+M)}{2K^2}\sum_{k=1}^K \left\| \nabla f(\mathbf{x}_t^k) \right\|^2 + \frac{L\gamma_t^2\sigma^2}{2K} \\
&= f(\tilde{\mathbf{x}}_t) - \left(\frac{\gamma_t}{K} - \frac{\gamma_t^2 L(K+M)}{2K^2} \right)\sum_{k=1}^K \left\| \nabla f(\mathbf{x}_t^k) \right\|^2 \\
&\qquad + \frac{\gamma_t}{K}\sum_{k=1}^K \left\langle \nabla f(\mathbf{x}_t^k) - \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t^k) \right\rangle + \frac{L\gamma_t^2\sigma^2}{2K} .
\end{aligned}$$

Note that Cauchy-Schwarz and Jensen inequalities together give $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{\beta}{2}\|\mathbf{a}\|^2 + \frac{1}{2\beta}\|\mathbf{b}\|^2$ for any $\beta > 0$. Using this observation with $\beta = 1$ we can proceed as

$$\begin{aligned}
\sum_{k=1}^K \left\langle \nabla f(\mathbf{x}_t^k) - \nabla f(\tilde{\mathbf{x}}_t), \nabla f(\mathbf{x}_t^k) \right\rangle &\leq \sum_{k=1}^K \frac{1}{2}\left\| \nabla f(\mathbf{x}_t^k) - \nabla f(\tilde{\mathbf{x}}_t) \right\|^2 + \sum_{k=1}^K \frac{1}{2}\left\| \nabla f(\mathbf{x}_t^k) \right\|^2 \\
&\overset{(3)}{\leq} \frac{L^2}{2}\sum_{k=1}^K \left\| \mathbf{x}_t^k - \tilde{\mathbf{x}}_t \right\|^2 + \frac{1}{2}\sum_{k=1}^K \left\| \nabla f(\mathbf{x}_t^k) \right\|^2 .
\end{aligned}$$

Plugging this back, we get our result that

$$\mathbb{E}_{\boldsymbol{\xi}_t}[f(\tilde{\mathbf{x}}_{t+1})|\mathbf{x}_t] \leq f(\tilde{\mathbf{x}}_t) - \gamma_t\left(\frac{1}{2K} - \frac{\gamma_t L(K+M)}{2K^2} \right)\sum_{k=1}^K \left\| \nabla f(\mathbf{x}_t^k) \right\|^2 + \frac{\gamma_t^2 L\sigma^2}{2K} + \frac{\gamma_t L^2}{2K}\sum_{k=1}^K \left\| \mathbf{x}_t^k - \tilde{\mathbf{x}}_t \right\|^2 .$$

Noting that $\gamma_t \leq \frac{K}{2L(K+M)}$ implies $\gamma_t\left(\frac{1}{2K} - \frac{\gamma_t L(K+M)}{2K^2} \right) \leq \frac{\gamma_t}{4K}$ yields the lemma. ■

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems 24*, pages 873–881. Curran Associates, Inc., 2011.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems 30*, pages 1709–1720. Curran Associates, Inc., 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems 31*, pages 5977–5987. Curran Associates, Inc., 2018.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems 28*, pages 1756–1764. Curran Associates, Inc., 2015.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1902.04686*, 2019.

Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *PMLR*, pages 111–132, 2020.

Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459. Curran Associates, Inc., 2011.

D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, 1989.

L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2017.

Sorathan Chaturapruek, John C Duchi, and Christopher Ré. Asynchronous stochastic convex optimization: the noise is in the noise and SGD don't care. In *Advances in Neural Information Processing Systems 28*, pages 1531–1539. Curran Associates, Inc., 2015.

Jean-Baptiste Cordonnier. Convex optimization using sparsified stochastic gradient descent with memory. Master thesis (Adv: S. U. Stich, M. Jaggi), EPFL, 2018.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1):165–202, January 2012. ISSN 1532-4435.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18 (101):1–51, 2017.

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, Dec 2016. ISSN 0018-9286.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Antoine Godichon-Baggioni and Sofiane Saadane. On the rates of convergence of parallelized averaged stochastic gradient algorithms. *arXiv preprint arXiv:1710.07926*, 2017.

Robert M. Gower, Peter Richtárik, and Francis Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arXiv preprint arXiv:1805.02632*, 2018.

Robert M. Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 2019.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Oliver Hinder, Aaron Sidford, and Nimit S. Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *arXiv preprint arXiv:1906.11985*, 2019.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, pages 795–811, Berlin, Heidelberg, 2016. Springer-Verlag.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3252–3261. PMLR, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for on-device federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020.

Remi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *Journal of Machine Learning Research*, 19(81):1–68, 2018.

J. C. H. Lee and P. Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614, 2016.

Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, editors. *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015. ISCA.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. *International Conference on Learning Representations (ICLR)*, 2020.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3325–3334. PMLR, 2018.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

Ryan McDonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems 22*, pages 1231–1239. Curran Associates, Inc., 2009.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, May 2019. ISSN 1436-4646.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1): 549–573, Jan 2016. ISSN 1436-4646.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Springer Science & Business Media.* Springer US, Boston, MA, 2004.

Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 693–701. Curran Associates Inc., 2011.

Kumar Kshitij Patel and Aymeric Dieuleveut. Communication trade-offs for synchronized distributed SGD with large step size. In *Advances in Neural Information Processing Systems 32*, pages 13601–13612. Curran Associates, Inc., 2019.

B. T. Polyak. New method of stochastic approximation type. *Autom. Remote Control*, 51 (7):937–946, 1990.

Boris T. Polyak. *Introduction to Optimization.* OptimizationSoftware, Inc., 1987.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In Li et al. (2015), pages 1058–1062.

O. Shamir and N. Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 850–857, 2014.

Suvrit Sra, Adams Wei Yu, Mu Li, and Alex Smola. Adadelay: Delay adaptive distributed stochastic optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 957–965. PMLR, 2016.

Sebastian U. Stich. Local SGD converges fast and communicates little. *International Conference on Learning Representations (ICLR)*, 2019a.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232v2*, 2019b.

Sebastian U. Stich. On communication compression for distributed optimization on heterogeneous data. *arXiv preprint arXiv:2009.02388*, 2020.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems 31*, pages 4447–4458. Curran Associates, Inc., 2018.

Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In Li et al. (2015), pages 1488–1492.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems 31*, pages 1306–1316. Curran Associates, Inc., 2018.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems 30*, pages 1509–1519. Curran Associates, Inc., 2017.

Blake E Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *Advances in Neural Information Processing Systems 31*, pages 8496–8506. Curran Associates, Inc., 2018.

Blake E. Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch sgd? In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020.

Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5325–5333. PMLR, 2018.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted SGD for non-convex optimization with faster convergence and less communication. *arXiv preprint arXiv:1807.06629*, 2018.

Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel SGD: When does averaging help? *arXiv preprint arXiv:1904.11325*, 2016.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J. Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*, pages 2595–2603. Curran Associates, Inc., 2010.