Communication trade-offs for synchronized distributed SGD (Local-SGD) with large step size (with Appendix)

Aymeric DIEULEVEUT
MLO, EPFL, Lausanne, Switzerland
CMAP, Ecole Polytechnique, Palaiseau, France
aymeric.dieuleveut@polytechnique.edu

Kumar Kshitij PATEL
MLO, EPFL, Lausanne, Switzerland
TTIC-Toyota Technological Institute Chicago
kkpatel@ttic.edu

Abstract

Synchronous mini-batch SGD is state-of-the-art for large-scale distributed machine learning. However, in practice, its convergence is bottlenecked by slow communication rounds between worker nodes. A natural solution to reduce communication is to use the *"local-SGD"* model in which the workers train their model independently and synchronize every once in a while. This algorithm improves the computation-communication trade-off but its convergence is not understood very well. We propose a non-asymptotic error analysis, which enables comparison to *one-shot averaging* i.e., a single communication round among independent workers, and *mini-batch averaging* i.e., communicating at every step. We also provide adaptive lower bounds on the communication frequency for large step-sizes ($t^{-\alpha}$, $\alpha \in (1/2, 1)$) and show that *Local-SGD* reduces communication by a factor of $O\Big(\frac{\sqrt{T}}{P^{3/2}}\Big)$, with T the total number of gradients and P machines.

1 Introduction

We consider the minimization of an objective function which is accessible through unbiased estimates of its gradients. This problem has received attention from various communities over the last fifty years in optimization, stochastic approximation, and machine learning [1–7]. The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins-Monro algorithm [8], and some of its modifications based on averaging of the iterates [1, 2, 9]. For a convex differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, SGD iteratively updates an estimator $(\boldsymbol{v}_t)_{t\geq 0}$ for any $t \geq 1$

$$\boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta_t \boldsymbol{g}_t(\boldsymbol{v}_{t-1}), \tag{1}$$

where $(\eta_t)_{t\geq 0}$ is a deterministic sequence of positive scalars, referred to as the *learning rate* and $\boldsymbol{g}_t(\boldsymbol{v}_{t-1})$ is an oracle on the gradient of the function F at $\boldsymbol{v}_{t-1}$. We focus on objective functions that are both smooth and strongly convex [10]. While these assumptions might be restrictive in practice, they enable to provide a tight analysis of the error of SGD. In such a setting, two types of proofs have been used traditionally. On one hand, *Lyapunov*-type proofs rely on controlling the expected squared distance to the optimal point [11]. Such analysis suggests using *small* decaying steps, inversely proportional to the number of iterations (t^{-1}). On the other hand, studying the recursion as a stochastic process [1] enables to better capture the reduction of the noise through averaging. It results in optimal convergence rates for larger steps, typically scaling as $t^{-\alpha}$, $\alpha \in (1/2, 1)$ [10].

Over the past decade, the amount of available data has steadily increased: to adapt SGD to such situations, it has become necessary to *distribute* the workload between several machines, also referred to as *workers* [12–14]. For SGD, two extreme approaches have received attention: 1) workers run SGD independently and at the end aggregate their results, called *one-shot averaging* (**OSA**) [13, 15]

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

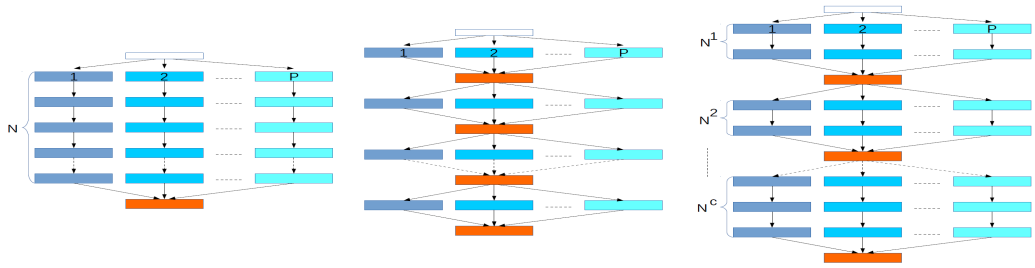

Figure 1: Schematic representation of one-shot averaging (left), mini-batch averaging (middle) and local-SGD (right). Vertical threads correspond to machines and orange boxes to communication rounds.

or *parameter mixing*, and 2) *mini-batch averaging* (**MBA**) [16–20], where workers communicate after every iteration: all gradients are thus computed at the *same* support point (iterate) and the algorithm is equivalent to using mini-batches of size P, with P the number of workers. While OSA requires only a single communication step, it typically does not perform very well in practice [21]. At the other extreme, MBA performs better in practice, but the number of communications equals the number of steps, which is a major burden, as communication is highly time consuming [22]. To optimize this computation-communication-convergence trade-off, we consider the *Local-SGD* framework: P workers run SGD iterations in parallel and communicate periodically. This framework encompasses *one-shot averaging* and *mini-batch averaging* as special cases (see Figure 1).

We make the following contributions:
1) We provide the first non-asymptotic analysis for local-SGD with large step sizes (typically scaling as $t^{-\alpha}$, for $\alpha \in (1/2; 1)$), in both on-line and finite horizon settings. Our assumptions encompass the ubiquitous *least-squares regression* and *logistic regression*.
2) Our comparison of the two extreme cases, OSA and MBA, underlines the communication trade-offs. While both of these algorithms are asymptotically equivalent for a fixed number of machines, mini-batch theoretically outperforms one-shot averaging when we consider the precise bias-variance split. In the regime where both the number of machines and gradients grow simultaneously we show that mini-batch SGD outperforms one-shot averaging.
3) Under three different sets of assumptions, we quantify the *frequency of communication* necessary for Local-SGD to be optimal (i.e., as good as mini-batch). Precisely, we show that the communication frequency can be reduced by as much as $O\left(\frac{\sqrt{T}}{P^{3/2}}\right)$, with T gradients and P workers. Moreover, our bounds suggest an adaptive communication frequency for logistic regression, which depending on the expected distance to the optimal point (a phenomenon observed by Zhang et al. [21]).
4) We support our analysis by experiments illustrating the behavior of the algorithms.

The paper is organized as follows: in Section 2.1, we introduce the general setting, notations and algorithms, then in Section 2.2, we describe the related literature. Next, in Section 2.3, we describe assumptions made on the objective function. In Section 3, we provide our main results, their interpretation, consequence and comparison with other results. Results in the on-line setting and experiments are presented in the Appendix A.2 and Appendix B.

2 Algorithms and setting

We first introduce a couple of notations. We consider the finite dimensional Euclidean space $\mathbb{R}^d$ embedded with its canonical inner product $\langle \cdot, \cdot \rangle$. For any integer $\ell \in \mathbb{N}^*$, we denote by $[\ell]$ the set $\{1, \dots, \ell\}$. We consider a strongly-convex differentiable function $F : \mathbb{R}^d \to \mathbb{R}$. We denote $\boldsymbol{w}^\star := \operatorname{argmin}_{\boldsymbol{w}} F(\boldsymbol{w})$. With only one machine, *Serial-SGD* performs a sequence of updates according to Equation (1). In the next section, we describe Local-SGD, the object of this study.

2.1 Local-SGD algorithm

We consider P machines, each of them running SGD. Periodically, workers aggregate (i.e., average) their models and restart from the resulting model. We denote by C the number of communication steps. We define a *phase* as the time between two communication rounds. At *phase* $t \in [C]$, for any worker $p \in [P]$, we perform N^t *local steps* of SGD. Iterations are thus naturally indexed by

$(t,k) \in [C] \times [N^t]$. We consider the lexicographic order $\preccurlyeq$ on such pairs, which matches the order in which iterations are processed. Note that we assume the number of local steps to be the same over all machines p. While this assumption can be relaxed in practice, is facilitates our proof technique and notation. At any $k \in [N^t]$, we denote by $\boldsymbol{w}_{p,k}^t$ the model proposed by worker p, at phase t, after k local iterations. All machines initially start from the same point $\boldsymbol{w}_0$, that is for any $p \in [P]$, $\boldsymbol{w}_{p,0}^1 = \boldsymbol{w}_0$. The update rule is thus the following, for any $p \in [P], t \in [C], k \in [N^t]$:

$$\boldsymbol{w}_{p,k}^t = \boldsymbol{w}_{p,k-1}^t - \eta_k^t g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t). \tag{2}$$

Aggregation steps consist in averaging the final local iterates of a phase: for any $t \in [C]$, $\hat{\boldsymbol{w}}^t = \frac{1}{P}\sum_{p=1}^P \boldsymbol{w}_{p,N^t}^t$. At phase $t+1$, every worker $p \in [P]$ restarts from the averaged model: $\boldsymbol{w}_{p,0}^{t+1} := \hat{\boldsymbol{w}}^t$. Eventually, we want to control the excess risk of the Polyak-Ruppert (PR) averaged iterate:

$$\overline{\overline{\boldsymbol{w}}}^C = \frac{1}{\sum_{t=1}^C N^t}\sum_{t=1}^C N^t \overline{\boldsymbol{w}}^t = \frac{1}{P\sum_{t=1}^C N^t}\sum_{t=1}^C\sum_{p=1}^P\sum_{k=1}^{N^t} \boldsymbol{w}_{p,k}^t,$$

with $\overline{\boldsymbol{w}}^t = \frac{1}{PN^t}\sum_{k=1}^{N^t}\sum_{p=1}^P \boldsymbol{w}_{p,k}^t$. We use the notation $\overline{\boldsymbol{w}}$ to underline the fact that iterates are averaged over one phase and $\overline{\overline{\boldsymbol{w}}}$ when averaging is made over all iterations. All averaged iterates can be computed on-line.

The algorithm, called *local-SGD*, is thus parameterized by the number of machines P, communication steps C, local iterations $(N^t)_{t\in[C]}$, the starting point $\boldsymbol{w}_0$, the learning rate $(\eta_k^t)_{(t,k)\in[C]\times[N^t]}$, and the first order oracle on the gradient. Pseudo-code of the algorithm is given in the Appendix, in Fig. S5.

Link with classical algorithms. Special cases of Local-SGD correspond to *one-shot averaging* or *mini-batch averaging*. More precisely, for a total number of gradients T, with P workers, $C = T/P$ communication rounds, and $(N^t)_{t\in[C]} = (1,\dots,1)$, we realize an instance of P-mini-batch averaging (P-MBA). On the other hand, with P workers, $C = 1$ communication, and $(N^1) = T/P$, we realize an instance of one shot-averaging. Our goal is to get general convergence bounds for Local-SGD that recover classical bounds for both these settings when we choose the correct parameters. While comparing to Serial-SGD (which is also a particular case of the algorithm), would also be interesting, we focus here on the comparison between Local-SGD, *one-shot averaging* and *mini-batch averaging*. Indeed, the step size is generally increased for mini-batch with respect to Serial SGD, and the running efficiency of algorithms is harder to compare: we only focus on different algorithms that use the *same number of machines*.

2.2 Related Work

On Stochastic Gradient Descent. Bounds on the excess risk of SGD for convex functions have been widely studied: most proofs rely on controlling the decay of the mean squared distance $\mathbb{E}[\|\boldsymbol{v}_t - \boldsymbol{w}^\star\|^2]$, which results in an upper bound on the mean excess of risk $\mathbb{E}[F(\bar{\boldsymbol{v}}_t) - F(\boldsymbol{w}^\star)]$ [23, 24]. This upper bound is composed of a "bias" term that depends on the initial condition, and a "variance" term that involves either an upper bound on the *norm* of the noisy gradient (in the non-smooth case), or an upper bound on the *variance* of the noisy gradient in the smooth case [5, 11]. In the strongly convex case such an approach advocates for the use of *small* step sizes, scaling as $(\mu t)^{-1}$. However, in practice, this is not a very satisfying result, as the constant μ is typically unknown, and convergence is very sensitive to ill-conditioning. On the other hand, in the smooth and strongly-convex case, the classical analysis by Polyak and Juditsky [1], relies on an explicit decomposition of the stochastic process $(\bar{\boldsymbol{v}}_t - \boldsymbol{w}^\star)_{t\geq 1}$: the effect of averaging on the noise term is better taken into account, and this analysis thus suggests to use larger steps, and results in the optimal rate for $\eta_t \propto t^{-\alpha}$, with $\alpha \in (0;1)$. This type of analysis has been successfully used recently [10, 15, 25, 26].

For quadratic functions, larger steps can be used, as pointed by Bach and Moulines [27]. Indeed, even with *non-decaying* step size, the averaged process converges to the optimal point. Several studies focus on understanding properties of SGD for quadratic functions: a detailed non-asymptotic analysis is provided by Défossez and Bach [28], acceleration under the additive noise oracle (see Assumption **A**4 below) is studied by Dieuleveut et al. [29] (without this assumption by Jain et al. [30]), and Jain et al. [20] analyze the effects of mini-batch and tail averaging.

One shot averaging. In this approach, the P-independent workers compute several steps of stochastic gradient descent, and a unique communication step is used to average the different models [13, 31, 32].

Zinkevich et al. [13] show a reduction of the variance when multiple workers are used, but neither consider the Polyak-Ruppert averaged iterate as the final output, nor provide non-asymptotic rates. Zhang et al. [33] provide the first non-asymptotic results for OSA but their dependence on constants (like strong convexity constant μ, moment bounds, etc.) is worse; as well as their single machine convergence bound [34] is not truly non-asymptotic (like for e.g., Bach and Moulines [10]). More importantly, their results hold only for small learning rates like $\frac{c}{\mu t}$. Rosenblatt and Nadler [35] have also discussed the asymptotic equivalence of OSA with vanilla-SGD by providing an analysis up to the second order terms. Further, Jain et al. [20] have provided non-asymptotic results for least-square regression using similar Polyak-Juditsky analysis of the stochastic process, while our results apply to more general problems. Their approach encompasses one shot averaging and the effect of tail averaging, that we do not consider here. Recently, Godichon and Saadane [15] proposed an approach similar to ours (but only for one shot averaging). However, their result relies on an asymptotic bound, namely $\mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2] \leq C_1 \eta_t$ (as in Rakhlin et al. [34]), while our analysis is purely non-asymptotic and we also improve the upper bound on the noise term which results from the analysis.

Mini-batch averaging. Mini-batch averaging has been studied by Dekel et al. [16], Takáč et al. [17]. These papers show an improvement in the variance of the process, and make comparisons to SGD. It has been found that increasing the mini-batch size often leads to increasing generalization errors, which limits their distributivity [36]. Jain et al. [20] have provided upper bounds on learning-rate and mini-batch size for optimal performance. Recently, large mini-batches have been leveraged successfully in deep learning as in [37–39] by properly tuning learning rates, etc.

Local-SGD. Zhang et al. [21] empirically show that Local-SGD performs well. They also provide a theoretical guarantee on the variance of the process, however, they assume the variance of the estimated gradients to be uniformly upper bounded (Assumption **A**4 below). Such an assumption is restrictive in practice, for example it is not satisfied for least squares regression. In a simultaneous work, Stich [40] has provided an analysis for local-SGD. The limitation with their analysis is that they also assume bounded gradients and use a small step size scaling as $\frac{c}{\mu t}$. More importantly, their analysis doesn't extend to the extreme case of one-shot averaging like ours. Lin et al. [41] have experimentally shown that Local-SGD is better than the synchronous mini-batch techniques, in terms of overcoming the large communication bottleneck. Recently, Yu et al. [42] have given convergence rates for the non-convex synchronous and a stale synchronous settings.

We have summarized the major limitations of some of these analyses in Table S3, given in Appendix I. Our motivation is to get away with some of these restrictive assumptions, and provide tight upper bounds for the above three averaging schemes. In the following section, we present the set of assumptions under which our analysis is conducted.

2.3 Assumptions

We first make the following classical assumptions on the objective function $F : \mathbb{R}^d \to \mathbb{R}$. In the following, we use different subsets of these assumptions:

A1 (Strong convexity) *The function F is strongly-convex with convexity constant $\mu > 0$.*

A2 (Smoothness and regularity) *The function F is three times continuously differentiable with second and third uniformly bounded derivatives: $\sup_{\boldsymbol{w}\in\mathbb{R}^d} \left|\left|\left|F^{(2)}(\boldsymbol{w})\right|\right|\right| < L$, and $\sup_{\boldsymbol{w}\in\mathbb{R}^d} \left|\left|\left|F^{(3)}(\boldsymbol{w})\right|\right|\right| < M$. Especially F is L-smooth.*

Q1 (Quadratic function) *There exists a positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, such that the function F is the quadratic function $\boldsymbol{w} \mapsto \|\Sigma^{1/2}(\boldsymbol{w} - \boldsymbol{w}^\star)\|^2/2$.*

If **Q**1 is satisfied, then Assumptions **A**1, **A**2 are satisfied, and L and μ are respectively the largest and smallest eigenvalues of Σ. At any iteration $(t, k) \in [C] \times [N^t]$, any machine can query an unbiased estimator of the gradient $g^t_{p,k}(\boldsymbol{w})$ at a point $\boldsymbol{w}$. Formally, we make the following assumption:

A3 (Oracle on the gradient) *We observe unbiased estimators of the gradient $g^t_{p,k+1}(\boldsymbol{w})$: for any $(t, k) \in [C] \times [N^t]$ and $\boldsymbol{w} \in \mathbb{R}^d$, $\mathbb{E}[g^t_{p,k+1}(\boldsymbol{w}^t_{p,k})|\boldsymbol{w}^t_{p,k}] = F'(\boldsymbol{w}^t_{p,k})$. Moreover, for any* fixed *$\boldsymbol{w}$ the functions $(g^t_{p,k})_{(t,k)}(\boldsymbol{w})$ are i.i.d. . (See Appendix A.1 for a more formal statement.)*

In Proposition 3, we make the additional, stronger assumption that the variance of gradient estimates is uniformly upper bounded, a standard assumption in the SGD literature, see e.g. Zhang et al. [21]:

A4 (Uniformly bounded variance) *The variance of the error, $\mathbb{E}[\|g^t_{p,k}(\boldsymbol{w}^t_{p,k}) - F'(\boldsymbol{w}^t_{p,k})\|^2]$ is uniformly upper bounded by σ^2_∞, a constant which does not depend on the iteration.*

Assumption **A**4 is for example true if the sequence of random vectors $(g^t_{p,k+1}(\boldsymbol{w}^t_{p,k}) - F'(\boldsymbol{w}^t_{p,k}))_{t\in[C],k\in[N^t],p\in[P]}$ is i.i.d.. This setting is referred to as the semi-stochastic setting [29].

We also consider the following conditions on the regularity of the gradients, for $p \geq 2$:

A5 (Cocoercivity of the random gradients) *For any $p \in [P]$, $t \in [C]$, $k \in [N^t]$, $g^t_{p,k}$ is almost surely L-co-coercive (with the same constant as in **A**2): that is, for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, $L\langle g^t_{p,k}(\boldsymbol{w}_1) - g^t_{p,k}(\boldsymbol{w}_2)\boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \geq \|g^t_{p,k}(\boldsymbol{w}_1) - g^t_{p,k}(\boldsymbol{w}_2)\|^2$.*

Almost sure L-co-coercivity [43] is for example satisfied if for any (p,k,t), there exist a random function $f^t_{p,k}$ which is a.s. convex and L-smooth and such that $g^t_{p,k} = (f^t_{p,k})'$. Finally, we assume the fourth order moment of the random gradients at $\boldsymbol{w}^\star$ to be well defined:

A6 (Finite variance at $\boldsymbol{w}^\star$) *$\exists \sigma \geq 0$, s.t. for any $t, k, p \in [C]\times[N^t]\times[P]$, $\mathbb{E}[\|g^t_{p,k}(\boldsymbol{w}^\star)\|^4] \leq \sigma^4$.*

It must be noted that **A**6 is a much weaker assumption than **A**4, for e.g., least-square regression satisfies former but not latter. Most of these assumptions are classical in machine learning. SGD for least squares regression satisfies **Q**1, **A**3, **A**5 and **A**6. On the other hand, SGD for logistic regression satisfies **A**1, **A**2, **A**3 and **A**4. Our main result Theorem 6 (lower bounding the frequency of communications) applies to both these sets of assumptions. In Appendix C.3 we further detail how these assumptions apply in machine learning.

Learning rate. We always assume that for any $t \in [C], k \in [N^t]$, the learning rate satisfies $2\eta^t_k L \leq 1$. We consider two different types of learning rates:
1) in the *finite horizon* (FH) case, the step size $(\eta^t_k)_{(t,k)\in[C]\times[N^t]}$ is a constant η, that can depend on the number of iterations eventually performed by the algorithm; 2) in the *on-line* case, the sequence of step size is a subsequence of a universal sequence $(\tilde{\eta}_\ell)_{\ell\geq 0}$. Moreover, in our analysis, when using decaying learning rate, the step size only depends on the number of iterations processed in the past: $\eta^t_k = \tilde{\eta}_{\{\sum_{t'=1}^{t-1} N^{t'}+k\}}$. Especially, the step size at iteration (t,k) does not depend on the machine.

Though both of these approaches are often considered to be nearly equivalent [44, 45], fundamental differences exist in their convergence properties. The *on-line* case is harder to analyze, but ultimately provides a better convergence rate. However as the behavior is easier to interpret in the finite horizon case, we postpone results for on-line setting to Appendix A.2. In the following section, we present our main results.

3 Main Results

Sketch of the proof. We follow the approach by Polyak and Juditsky, which relies on the following decomposition: for any $p \in [P], t \in [C], k \in [N^t]$, Equation (2) is trivially equivalent to: $\eta^t_k F''(\boldsymbol{w}^\star)(\boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^\star) = \boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^t_{p,k} - \eta^t_k[g^t_{p,k}(\boldsymbol{w}^t_{p,k-1}) - F'(\boldsymbol{w}^t_{p,k-1})] - \eta^t_k[F'(\boldsymbol{w}^t_{p,k-1}) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^\star)]$. We have added and subtracted a first order Taylor expansion around the optimal value $\boldsymbol{w}^\star$ of the gradient. Thus, using the definition of $\overline{\overline{\boldsymbol{w}}}^C$:

$$F''(\boldsymbol{w}^\star)\left(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star\right) = \frac{1}{P\sum_{t=1}^C N^t}\sum_{t=1}^{C}\sum_{p=1}^{P}\sum_{k=1}^{N^t}\left(\frac{\boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^t_{p,k}}{\eta^t_k} - \left[g^t_{p,k}(\boldsymbol{w}^t_{p,k-1}) - F'(\boldsymbol{w}^t_{p,k-1})\right]\right.$$
$$\left. - \left[F'(\boldsymbol{w}^t_{p,k-1}) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^\star)\right]\right). \tag{3}$$

In other words, the error can be decomposed into three terms: the first one mainly depends on the *initial condition*, the second one is a *noise term*: it is the mean of centered random variables (as $\mathbb{E}[g^t_{p,k}(\boldsymbol{w}^t_{p,k-1}) - F'(\boldsymbol{w}^t_{p,k-1})] = 0$), and the third is a *residual term* that accounts for the fact that the function is not quadratic (if F is quadratic, then $F'(\boldsymbol{w}^t_{p,k-1}) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^\star) = 0$).

Controlling different terms in Equation (3). The variance of the noise $g^t_{p,k}(\boldsymbol{w}^t_{p,k-1}) - F'(\boldsymbol{w}^t_{p,k-1})$ and the residual term both directly depend on the distance $\|\boldsymbol{w}^t_{p,k-1} - \boldsymbol{w}^\star\|^2$. The proof is thus composed of two aspects: (1) we first provide a tight control for this quantity, with or without communication: in the following propositions, this corresponds to an upper bound on $\mathbb{E}[\|\boldsymbol{w}^t_{p,k} - \boldsymbol{w}^\star\|^2]$ [1], (2) we provide the subsequent upper bound on $\mathbb{E}[\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\|^2]$.

[1] more precisely, on $\mathbb{E}[\|\tilde{\boldsymbol{w}}^t - \boldsymbol{w}^\star\|^2]$ and $\mathbb{E}[\|\boldsymbol{w}^1_{p,k} - \boldsymbol{w}^\star\|^2]$ for MBA and OSA respectively.

We first compare the convergence in the two extreme situations, *i.e.*, for *Mini-batch averaging* (MBA) and *One-shot averaging* (OSA) for *finite horizon* setting, and then provide these results for local-SGD.

3.1 Results for MBA and OSA, Finite Horizon setting

First we assume the step size η_k^t to be a constant η at every iteration for any $t \in [C], k \in [N^t]$. Our first contribution is to provide *non-asymptotic* convergence rates for *MBA* and *OSA*, that allow a simple comparison. For the benefit of presentation, we define following quantities: $Q_{bias} = 1 + \frac{M^2\eta}{\mu}\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\|^2 + \frac{L^2\eta}{\mu P}$, $Q_{1,var}(X) = \frac{L^2\eta}{\mu} + \frac{P}{X\eta\mu}$, $Q_{2,var}(X) = \frac{M^2XP\eta^2\sigma^2}{\mu^2}$.

In the following, we use the $\precsim$ notation to denote inequality up to an absolute constant. Recall that for MBA, the total number of gradients processed is $T = PC$, while it is $T = PN$ for OSA. We have the following results respectively for MBA and OSA:

Proposition 1 (Mini-batch Averaging) *Under Assumptions A1, A2, A3, A5, A6, we have the following bound for mini-batch SGD: for any $t \in [C]$,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] \leq (1-\eta\mu)^t \left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2\eta}{P}\frac{1-(1-\eta\mu)^t}{\mu}, \tag{4}$$

$$\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\right\|^2\right] \precsim \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{\eta^2C^2}Q_{bias} + \frac{\sigma^2}{T}\Big(1 + \frac{Q_{1,var}(C)}{P} + \frac{Q_{2,var}(C)}{P^2}\Big). \tag{5}$$

Proposition 2 (One-shot Averaging) *Under Assumptions A1, A2, A3, A5, A6, we have the following bound for one shot averaging: $p \in [P], t = 1, k \in [N]$,*

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^1 - \boldsymbol{w}^\star\right\|^2\right] \leq (1-\eta\mu)^k \left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\eta\frac{1-(1-\eta\mu)^k}{\mu}, \tag{6}$$

$$\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\right\|^2\right] \precsim \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{\eta^2N^2}Q_{bias} + \frac{\sigma^2}{T}\big(1 + Q_{1,var}(N) + Q_{2,var}(N)\big). \tag{7}$$

Interpretation, fixed P. Using mini-batch naturally reduces the variance of the process $(\boldsymbol{w}_{p,k}^t)_{p\in[P],t\in[C],k\in[N^t]}$. Equations (4) and (6) show that the speed at which the initial condition is forgotten remains the same, but that the variance of the local process is reduced by a factor P.

Equations (5) and (7) show that the convergence depends on an *initial condition* term and a *variance term*. For a fixed number of machines P, and a step size scaling as $\eta = X^{-\alpha}$, $0.5 < \alpha < 1$, $X \in \{N, C\}$, the speed at which the *initial condition* is forgotten is asymptotically dictated by $Q_{bias}/(\eta X)^2$ where $X \in \{N, C\}$, for *both algorithms* (if we use the same number of gradients for both algorithms, naturally, $N = C$.) As for the variance term, it scales as σ^2T^{-1} as $T \to \infty$, as the remaining terms $Q_{var}(X)$ asymptotically vanish for $\eta = X^{-\alpha}$. It reduces with the total number T of gradients used in the process. Interestingly, this term is *the same* for the two extreme cases (MBA and OSA): it does not depend on the number of communication rounds. This phenomenon is often described as *"the noise is the noise and SGD doesn't care"* (for asynchronous SGD, [46]). Though we recover this asymptotic equivalence here, our belief is that this asymptotic point of view is typically misleading as the asymptotic regime is not always reached, and the residual terms do then matter.

Indeed, the lower order terms do have a dependence on the number of communication rounds: when the number of communications increases, the overall effect of the noise is reduced. More precisely, since $Q_{var}(N) = Q_{var}(C)$ the remaining terms are respectively P or P^2 times smaller for mini-batch. This provides a theoretical explanation of why mini-batch SGD outperforms one shot averaging in practice. It also highlights the weakness of an asymptotic analysis: the dominant term might be equivalent, without reflecting the actual behavior of the algorithm. Disregarding communication aspects, mini-batch SGD is in that sense *optimal*.

Note that for quadratic functions, $Q_{2,var} = 0$ as $M = 0$. The conditions on the step size can thus be relaxed, and the asymptotic rates described above would be valid for any step size satisfying $\eta \leq \mu$ [20]. Extension to the on-line setting, eventually leading to a better convergence rate, is given in Proposition S7 in AppendixA.2.

Interpretation, $P, T \to \infty$. When both the total number of gradients used T and the number of machines P are allowed to grow simultaneously, the asymptotic regime is not necessarily the same for MBA and OSA, as remaining terms are not always negligible. For example, if fixing $\eta = X^{-2/3}$, $X \in \{N, C\}$ (we chose $\alpha = 2/3$ to balance $Q_{1,var}$ and $Q_{2,var}$), the variance term would be controlled by $\sigma^2 T^{-1}(1 + \frac{P}{\mu C^{1/3}})$. Thus, unless $P \leq \mu C^{1/3}$, MBA could outperform OSA by a factor as large as P.

Novelty and proofs. Both Propositions 1 and 2 are proved in the Appendix G. Importantly, Equations (4) and (6) respectively imply Equations (5) and (7) under the stated conditions: this is the reason why we only focus on proving equations similar to Equations (4) and (6) for Local-SGD.

Proposition 1 is similar to the analysis of *Serial-SGD* for large step size, but with a reduction in the variance proportional to the number of machines. Such a result is derived from the analysis by Dieuleveut et al. [25], combining the approach of Bach and Moulines [27] with the correct upper bound for smooth strongly convex SGD [47], and controlling similarly higher order moments. While this result is expected, we have not found it under such a simple form in the literature. Proposition 2 follows a similar approach, we combine the proof for mini-batch with a control of the iterates of each of the machines. This is closely related to Godichon and Saadane [15], but we preserve a non-asymptotic approach.

Remark: link with convergence in function values. As we use Equation (3) as a starting point, we provide convergence results on the Mahalanobis distance $\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\|^2$: it is the natural quantity in such a setting [10, 15, 27]. These results could be translated into function value convergence $F(\overline{\overline{\boldsymbol{w}}}^C) - F(\boldsymbol{w}^\star)$, using the inequality $F(\overline{\overline{\boldsymbol{w}}}^C) - F(\boldsymbol{w}^\star) \leq L\mu^{-2}\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\|^2$ but the dependence on μ would be pessimistic and sub-optimal. However, a similar approach has been used by Bach [44], under a slightly different set of assumptions (including self-concordance, e.g., for logistic regression), recovering optimal rates. Extension to such a set of assumptions, which relies on tracking other quantities, is an important direction.

While the "classical proof", which provides rates for function values directly (with smoothness, or with uniformly bounded gradients) has a better dependence on μ, one cannot easily obtain a noise reduction when averaging between machines. Similarly, there is no proof showing that one-shot averaging is asymptotically optimal that relies only on function values. In other words, these proofs do not adequately capture the noise reduction due to averaging. Moreover, such proof techniques relying on function values typically involve a small step size $1/(\mu t)$ (because the noise reduction is captured inefficiently). Such step size performs poorly in practice (initial condition is forgotten slowly), and μ is unknown.

In conclusion, though they do not directly result in optimal dependence on μ for function values, we believe our approach allows to correctly capture the effect of the noise, and is thus suitable for capturing the effect of Local-SGD.

Comparing upper bounds: Our analysis relies on upper bounds: one should handle comparison with cautions. Nevertheless, we think our analysis is tight enough to provide good insights, especially because the bound for OSA averaging nearly matches the bound for MBA (contrary to Stich [40]). Moreover, the bounds given above are tight in the following senses, see Appendix A.3 for details:
(i) the bias term in equations (5) and (7) is clearly *exact* in the simple case of a quadratic one dimensional function, in the absence of noise: it is normal that in such a situation, MBA and OSA converge similarly: each of the P independent machines computes the same recursion!
(ii) the bound for the variance, scaling as $(PN)^{-1}$ for any $\eta \propto N^{-\alpha}, 0.5 < \alpha < 1$, matches the statistical minimax rate [48] for least squares regression: from the statistical point of view, if we are only given NP independent observations, then no estimator can have an error uniformly lower than $\sigma^2(PN)^{-1}$.

Optimizing over the step size in Eqs (5) and (7) results in a somehow disappointing observation: the rate for $\eta \propto N^{-\alpha}, 0.5 < \alpha < 1$[2] is dictated by the bias and scales as $O((\eta N)^{-2})$, which is slow (but tight, see point (i) above). This is unfortunately unavoidable *with constant step sizes*: the convergence rate with decaying steps is much faster in the on-line setting[3], but bounds are much harder to read see

[2] A good step size is unlikely to be larger than $1/\sqrt{N}$: such "very large" LR (which is rarely used in practice) does not perform well for non-quadratic functions (note that for quadratic, the $NP\eta^2$ vanishes, and a constant η would get a rate $1/N^2 + 1/PN$).

[3] the bias decreases as $1/N^2$ instead of $1/(\eta N)^2$ (see Prop.S7).

Sec. A.2. In other words, bounds in Propositions 1 and 2 are *tight*, but *slower* than in on-line setting. As *all the trade-offs regarding communications are preserved* (our main focus), we chose to highlight the results in finite horizon in the main text.

Conclusion: for a fixed or limited number of machines, asymptotically, the convergence rate is similar for OSA and MBA. However, non-asymptotically, or when the number of machines also increases, the dominant terms can be as much as P^2 times smaller for MBA. In the following we provide conditions for Local-SGD to perform as well as MBA (while requiring much fewer communication rounds).

3.2 Convergence of Local-SGD, Finite Horizon setting

For local-SGD we first consider the case of a quadratic function, under the assumption that the noise has a uniformly upper bounded variance. While this set of assumptions is not realistic, it allows an intuitive presentation of the results. Similar results for settings encompassing LSR and LR follow. We provide a bound on the moment of an iterate after the communication step $\hat{\boldsymbol{w}}^t$ (i.e., the restart point of the next phase), and on the second order moment of any iterate. For $t \in [C]$, we denote $\boldsymbol{N}_1^t := \sum_{t'=1}^{t} N^{t'}$.

Proposition 3 (Local-SGD: Quadratic Functions with Bounded Noise) *Under Assumptions **Q** 1, **A**3, **A**4, we have the following bound for Local-SGD: for any $p \in [P], t \in [C], k \in [N^t]$,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right] \leq (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}} \left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2 \eta}{P} \frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{\mu}$$

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star\right\|^2\right] \leq (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}+k} \left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \sigma_\infty^2 \eta \left(\underbrace{\frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{P\mu}}_{\text{long term reduced variance}} + \underbrace{\frac{1-(1-\eta\mu)^k}{\mu}}_{\text{local iteration variance}} \right).$$

To prove such a result, we use the classical technique, and introduce a *ghost* sequence $\breve{\boldsymbol{w}}_k^t := \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{w}_{p,k}^t$, and recursively control $\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2$. We conclude by remarking that $\breve{\boldsymbol{w}}_{N^t}^t = \hat{\boldsymbol{w}}^t$. This proof is given in Appendix D.2.

Interpretation. The variance bound for the iterates "just after" communication, $\hat{w}^t$ exactly behaves as in mini-batch case: the initialization term decays linearly with the number of local steps, and the variance is reduced proportionally to the number of workers P. On the other hand, the bound on the iterates $\boldsymbol{w}_{p,k}^t$ shows that the variance of this process is composed of a "long term" reduced variance, that accumulates through phases, and is increasingly converging to $\frac{\sigma_\infty^2 \eta}{P\mu}$ and of an extra variance $\eta\sigma_\infty^2 \frac{1-(1-\eta\mu)^k}{\mu}$, that increases within the phase, and is upper bounded by $\sigma_\infty^2 \eta^2 k$.

In the case of constant step size, the iterates of serial SGD converge to a limit distribution π_η that depends on the step size [25]. Here, the iterates after communication (or the mini-batch iterates) converge to a distribution with reduced variance $\pi_{\eta/P}$, thus local iterates periodically restart from a distribution with reduced variance, then slowly "diverge" to the distribution with large variance. If the number of local iterations is small enough, the iterates keep a reduced variance. More precisely, we have the following result.

Corollary 4 *If for all $t \in [C]$, $N^t \leq (\mu\eta P)^{-1}$, then the second order moment of $\boldsymbol{w}_{p,k}^t$ admits the same upper bound as the mini-batch iterate $\hat{\boldsymbol{w}}_{MB}^{\boldsymbol{N}_1^{t-1}+k}$ (Equation (4)) up to a constant factor of 2. As a consequence, Equation (5) is still valid, and Local-SGD performs "optimally".*

Interpretation. This result shows that if the algorithm communicates often enough, the convergence of the Polyak Ruppert iterate $\overline{\overline{\boldsymbol{w}}}^C$ is as good as in the mini-batch case, thus it is "optimal". Moreover, the minimal number of communication rounds is easy to define: the maximal number of local steps N^t decays as the number of workers and the step size increases. This bound implies that more communication steps are necessary when more machines are used. Note that $(\eta P)^{-1}$ is a large number, as a typical value for η is inversely proportional to (a power of) the number of local steps for e.g., $(\sum_{t'=1}^{t} N^{t'})^{-\alpha}$, $\alpha \in (1/2; 1)$.

Example 5 *With constant number of local steps $N^t = N$, and learning rate $\eta = c(NC)^{-1/2}$ in order to obtain an optimal $O(\sigma^2 T^{-1})$ parallel variance[4] rate, local-SGD communicates $O(\sqrt{NC}/(P\mu))$ times less as compared to mini-batch averaging.*

We believe that this is the first result (with Stich [40]) that shows a communication reduction proportional to a power of the number of local steps of a local solver (i.e., $O(\sqrt{NC})$), compared to mini-batch averaging. In the following, we alternatively relax the bounded variance assumption **A**4 and the quadratic assumption **Q**1, and show similar results for Local-SGD. This allows us to successively cover the cases of least squares regression (LSR) and logistic regression (LR).

Theorem 6 *Under either of the following sets of assumptions, the convergence of the Polyak Ruppert iterate $\overline{\overline{\boldsymbol{w}}}^C$ is as good as in the mini-batch case, up to a constant:*
*(i) Assume **Q**1, **A**3, **A**5, **A**6, and for any $t \in [C]$, $N^t \leq (\mu\eta P)^{-1}$ and $\mu\eta^2 \boldsymbol{N}_1^t = O(1)$.*
*(ii) Assume **A**1, **A**2, **A**3, **A**4, and for any $t \in [C]$, $N^t \leq \inf\left((\eta P M \mathbb{E}[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|])^{-1}, (\mu\eta P)^{-1}\right)$.*

These results are derived from Proposition S16 and Proposition S20 which generalize Proposition 3. Those results are proved in Appendix D and E and constitute the main technical challenge of the paper.

Interpretation. We note that in both of these situations, the optimal rates can be achieved if the communications happen often enough, and beyond such a number of communication rounds, there is no substantial improvement in the convergence. This result corresponds to the effect observed in practice [21]. The first set of assumption is valid for LSR, the second for LR. In the first case, the maximal number of local steps before communication is upper bounded by the same ratio as in Corollary 4, but the "constant" that appears is $\exp(\mu\eta^2 \boldsymbol{N}_1^t)$, so we need this quantity to be small (which is typically always satisfied in practice) in order to be optimal w.r.t. mini-batch averaging. A similar result as Theorem 5 can be provided reducing the communication by a factor of $O(\frac{\sqrt{NC}}{P\mu})$.

In the second case, the maximal number of local steps is smaller than before, by a factor μ^{-1}, but the allowed maximal number of local steps can increase along with the epochs, as $\mathbb{E}[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|]$ is typically decaying. This adaptive communication frequency has been observed to work well in practice [21] and also explored in [49], in a setting without PR averaging. Assuming optimization on a compact space with radius R for instance, one can obtain a $O(\frac{\sqrt{NC}}{P^2})$ times improvement in communication, similar to Theorem 5.

Though they may reflect the actual behavior of the algorithm, such results might be difficult to use directly in practice, as μ is unknown. However, as it is not the limiting factor in Theorem 6.2, an estimation of $\mathbb{E}[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|]$ could allow us to use adaptive phases lengths to minimize communications.

4 Conclusion

Stochastic approximation and distributed optimization are both very densely studied research areas. However, in practice most distributed applications stick to bulk synchronous mini-batch SGD. While the algorithm has desirable convergence properties, it suffers from a huge communication bottleneck. In this paper we have analyzed a natural generalization of mini-batch averaging, Local-SGD. Our analysis is non-asymptotic, which helps us to better understand the exact communication trade-offs. We give feasible lower bounds on communication frequency which significantly reduce the need for communication, while providing similar non-asymptotic convergence as mini-batch averaging. Our results apply to common loss functions, and use large step sizes. Further, our analysis unifies and extends all the scattered results for one-shot averaging, mini-batch averaging and Local-SGD, providing an intuitive understanding of their behavior.

While they provide some intuition and are believed to be tight, our comparisons are based on upper bounds. Proving corresponding lower bounds is an interesting and important open direction. Also, it would also be interesting to study observable quantities to predict an adaptive communication frequency and to relax some of the technical assumptions required by the analysis. The on-line case, experiments, proofs, additional materials and a review of distributed optimization follow in the appendix.

[4] in online setting, the same example would hold, resulting in a $O(\frac{\sigma^2}{T})$ convergence **rate** (not only variance).

Acknowledgements

We would like to acknowledge Sai Praneeth Reddy, Sebastian Stich, Martin Jaggi and Nathan Srebro for helpful comments and discussions at various stages of this project.

References

[1] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[2] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[3] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

[4] Y. Nesterov and J. P. Vial. Confidence Level Solutions for Stochastic Programming. *Automatica*, 44(6):1559–1568, 2008. ISSN 0005-1098. doi: 10.1016/j.automatica.2008.01.017. URL http://dx.doi.org/10.1016/j.automatica.2008.01.017.

[5] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM J. on Optimization*, 19(4):1574–1609, 2009. ISSN 1052-6234. doi: 10.1137/070704277. URL http://dx.doi.org/10.1137/070704277.

[6] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

[7] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the conference on machine learning (ICML)*, 2004.

[8] H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407, 1951.

[9] O. Shamir and T. Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. *Proceedings of the 30^th International Conference on Machine Learning*, 2013.

[10] F. Bach and E. Moulines. Non-asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 451–459, USA, 2011. Curran Associates Inc. ISBN 978-1-61839-599-3. URL http://dl.acm.org/citation.cfm?id=2986459.2986510.

[11] P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *International Conference on Machine Learning (ICML)*, pages 1–9, 2015.

[12] O. Delalleau and Y. Bengio. Parallel stochastic gradient descent. 2007.

[13] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

[14] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011.

[15] A. B. Godichon and S. Saadane. On the rates of convergence of Parallelized Averaged Stochastic Gradient Algorithms. *ArXiv e-prints*, 2017.

[16] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

[17] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro. Mini-batch primal and dual methods for svms. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages III–1022. JMLR. org, 2013.

[18] M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670. ACM, 2014.

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

[20] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing Stochastic Approximation Through Mini-Batching and Tail-Averaging. *ArXiv e-prints*, 2016.

[21] J. Zhang, C. De Sa, I. Mitliagkas, and C. Ré. Parallel SGD: When does averaging help? *ArXiv e-prints*, 2016.

[22] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang. The zipml framework for training models with end-to-end low precision: The cans, the cannots, and a little bit of deep learning. *arXiv preprint arXiv:1611.05402*, 2016.

[23] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an O(1/t) rate for the stochastic projected subgradient method. ArXiv e-prints 1212.2002, 2012.

[24] A. Rakhlin, O. Shamir, and K. Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *ArXiv e-prints*, 2011.

[25] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Annals of Statistics*, 2018.

[26] S. Gadat and F. Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *ArXiv e-prints*, 2017.

[27] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[28] A. Défossez and F. Bach. Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2015.

[29] A. Dieuleveut, N. Flammarion, and F. Bach. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. *Journal of Machine Learning research*, 2016.

[30] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent. *arXiv preprint arXiv:1704.08227*, 2017.

[31] R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

[32] R. McDonald, K. Hall, and G. Mann. Distributed training strategies for the structured perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 456–464. Association for Computational Linguistics, 2010.

[33] Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

[34] A. Rakhlin, O. Shamir, K. Sridharan, et al. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*. Citeseer, 2012.

[35] J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016. doi: 10.1093/imaiai/iaw013. URL http://dx.doi.org/10.1093/imaiai/iaw013.

[36] M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670. ACM, 2014.

[37] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ArXiv e-prints*, 2016.

[38] Y. You, I. Gitman, and B. Ginsburg. Large Batch Training of Convolutional Networks. *ArXiv e-prints*, 2017.

[39] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *ArXiv e-prints*, 2017.

[40] S. U. Stich. Local SGD Converges Fast and Communicates Little. *ICLR 2019*, 2019.

[41] T. Lin, S. U. Stich, and M. Jaggi. Don't Use Large Mini-Batches, Use Local SGD. *ArXiv e-prints*, 2018.

[42] H. Yu, S. Yang, and S. Zhu. Parallel Restarted SGD for Non-Convex Optimization with Faster Convergence and Less Communication. *ArXiv e-prints*, 2018.

[43] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

[44] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, 2014.

[45] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 2016. doi: 10.1214/15-AOS1391. URL http://dx.doi.org/10.1214/15-AOS1391.

[46] J. C. Duchi, S. Chaturapruek, and C. Ré. Asynchronous stochastic convex optimization. *ArXiv e-prints*, 2015.

[47] D. Needell, R. Ward, and N. Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1017–1025. Curran Associates, Inc., 2014.

[48] A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

[49] M. Kamp, M. Boley, D. Keren, A. Schuster, and I. Sharfman. Communication-efficient distributed online prediction by dynamic model synchronization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 623–639. Springer, 2014.

[50] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004. ISBN 9781402075537. URL http://books.google.fr/books?id=VyYLem-l3CgC.

[51] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

[52] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. *ArXiv e-prints*, 2015.

[53] E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[54] J. Langford, A. Smola, and M. Zinkevich. Slow Learners are Fast. *ArXiv e-prints*, 2009.

[55] F. Niu, B. Recht, C. Re, and S. J. Wright. HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. *ArXiv e-prints*, 2011.

[56] A. Agarwal and J. C. Duchi. Distributed Delayed Stochastic Optimization. *ArXiv e-prints*, 2011.

[57] T. Paine, H. Jin, J. Yang, Z. Lin, and T. Huang. GPU Asynchronous Stochastic Gradient Descent to Speed Up Neural Network Training. *ArXiv e-prints*, 2013.

[58] M. Li, D. G. Andersen, A. Smola, and K. Yu. Communication efficient distributed machine learning with the parameter server. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 19–27, Cambridge, MA, USA, 2014. MIT Press. URL http://dl.acm.org/citation.cfm?id=2968826.2968829.

[59] S. Zhang, A. Choromanska, and Y. LeCun. Deep learning with Elastic Averaging SGD. *ArXiv e-prints*, 2014.

[60] J. Keuper and F.-J. Pfreundt. Asynchronous Parallel Stochastic Gradient Descent - A Numeric Core for Scalable Distributed Machine Learning Algorithms. *ArXiv e-prints*, 2015.

[61] S. De and T. Goldstein. Efficient Distributed SGD with Variance Reduction. *ArXiv e-prints*, 2015.

[62] H. R. Feyzmahdavian, A. Aytekin, and M. Johansson. An Asynchronous Mini-Batch Algorithm for Regularized Stochastic Optimization. *ArXiv e-prints*, 2015.

[63] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization. *ArXiv e-prints*, 2015.

[64] S.-Y. Zhao and W.-J. Li. Fast Asynchronous Parallel Stochastic Gradient Decent. *ArXiv e-prints*, 2015.

[65] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz. Revisiting Distributed Synchronous SGD. *ArXiv e-prints*, 2016.

[66] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. *ArXiv e-prints*, 2017.

[67] F. Pedregosa, R. Leblond, and S. Lacoste-Julien. Breaking the Nonsmooth Barrier: A Scalable Parallel Method for Composite Optimization. *ArXiv e-prints*, 2017.

[68] X. Lian, W. Zhang, C. Zhang, and J. Liu. Asynchronous Decentralized Parallel Stochastic Gradient Descent. *ArXiv e-prints*, 2017.

[69] R. Leblond, F. Pedregosa, and S. Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *ArXiv e-prints*, 2018.

[70] D. Alistarh, C. De Sa, and N. Konstantinov. The Convergence of Stochastic Gradient Descent in Asynchronous Shared Memory. *ArXiv e-prints*, 2018.

[71] J. Konecný, B. McMahan, and D. Ramage. Federated optimization: Distributed optimization beyond the datacenter. *CoRR*, abs/1511.03575, 2015. URL http://arxiv.org/abs/1511.03575.

[72] J. Konecný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492, 2016. URL http://arxiv.org/abs/1610.05492.

[73] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

[74] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning, 2017. URL http://proceedings.mlr.press/v70/zhang17e.html.

[75] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *CoRR*, abs/1705.07878, 2017. URL http://arxiv.org/abs/1705.07878.

[76] J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. *CoRR*, abs/1710.09854, 2017. URL http://arxiv.org/abs/1710.09854.

[77] C. D. Sa, C. Zhang, K. Olukotun, and C. Ré. Taming the wild: A unified analysis of hogwild!-style algorithms. *CoRR*, abs/1506.06438, 2015. URL http://arxiv.org/abs/1506.06438.

[78] T. Na, J. H. Ko, J. Kung, and S. Mukhopadhyay. On-chip training of recurrent neural networks with limited numerical precision. *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3716–3723, 2017.

[79] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep learning with limited numerical precision. *CoRR*, abs/1502.02551, 2015. URL http://arxiv.org/abs/1502.02551.

[80] D. Alistarh, J. Li, R. Tomioka, and M. Vojnovic. QSGD: randomized quantization for communication-optimal stochastic gradient descent. *CoRR*, abs/1610.02132, 2016. URL http://arxiv.org/abs/1610.02132.

[81] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson. Distributed learning with compressed gradients. *ArXiv e-prints*, 2018.

[82] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1): 1–122, 2011. ISSN 1935-8237. doi: 10.1561/2200000016. URL http://dx.doi.org/10.1561/2200000016.

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

[84] Y. Zhang and L. Xiao. Communication-Efficient Distributed Optimization of Self-Concordant Empirical Loss. *ArXiv e-prints*, 2015.

[85] S. J. Reddi, J. Konečný, P. Richtárik, B. Póczós, and A. Smola. AIDE: Fast and Communication Efficient Distributed Optimization. *ArXiv e-prints*, 2016.

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

[87] V. Smith, S. Forte, C. Ma, M. Takac, M. I. Jordan, and M. Jaggi. CoCoA: A General Framework for Communication-Efficient Distributed Optimization. *ArXiv e-prints*, 2016.

[88] C. Ma, V. Smith, M. Jaggi, M. I. Jordan, P. Richtárik, and M. Takáč. Adding vs. Averaging in Distributed Primal-Dual Optimization. *ArXiv e-prints*, 2015.

[89] K. Scaman, F. Bach, S. Bubeck, Y. Tat Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. *ArXiv e-prints*, 2017.

[90] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola. On Variance Reduction in Stochastic Gradient Descent and its Asynchronous Variants. *ArXiv e-prints*, 2015.

[91] S.-Y. Zhao and W.-J. Li. Fast asynchronous parallel stochastic gradient descent: A lock-free approach with convergence guarantee. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2379–2385. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3016100.3016231.

[92] J. D. Lee, Q. Lin, T. Ma, and T. Yang. Distributed Stochastic Variance Reduced Gradient Methods and A Lower Bound for Communication Complexity. *ArXiv e-prints*, 2015.

[93] M. M. Najafabadi, T. M. Khoshgoftaar, F. Villanustre, and J. Holt. Large-scale distributed l-bfgs. *Journal of Big Data*, 4(1):22, 2017.

[94] U. Şimşekli, Ç. Yıldız, T. H. Nguyen, G. Richard, and A. Taylan Cemgil. Asynchronous Stochastic Quasi-Newton MCMC for Non-Convex Optimization. *ArXiv e-prints*, 2018.

[95] Y. Arjevani and O. Shamir. Communication complexity of distributed convex learning and optimization. *CoRR*, abs/1506.01900, 2015. URL http://arxiv.org/abs/1506.01900.

[96] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Optimality guarantees for distributed statistical estimation. *ArXiv e-prints*, 2014.

[97] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. *CoRR*, abs/1506.07216, 2015. URL http://arxiv.org/abs/1506.07216.

[98] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2328–2336. Curran Associates, Inc., 2013.

[99] S. Lee, J. K. Kim, X. Zheng, Q. Ho, G. A. Gibson, and E. P. Xing. On model parallelization and scheduling strategies for distributed machine learning. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2834–2842. Curran Associates, Inc., 2014.

[100] C. Ma and M. Takáč. Partitioning Data on Features or Samples in Communication-Efficient Distributed Optimization? *ArXiv e-prints*, 2015.

[101] Z. Chen, L. Luo, and Z. Zhang. Communication Lower Bounds for Distributed Convex Optimization: Partition Data on Features. *ArXiv e-prints*, 2016.

[102] B. Fang and D. Klabjan. A Stochastic Large-scale Machine Learning Algorithm for Distributed Features and Observations. *ArXiv e-prints*, 2018.

[103] Z. Meng, A. Wiesel, and A. O. Hero. Distributed principal component analysis on networks via directed graphical models, 2012. ISSN 2379-190X.

[104] Y. Zhuang, W.-S. Chin, Y.-C. Juan, and C.-J. Lin. A fast parallel sgd for matrix factorization in shared memory systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 249–256, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0. doi: 10.1145/2507157.2507164. URL http://doi.acm.org/10.1145/2507157.2507164.

[105] F. Li, B. Wu, L. Xu, C. Shi, and J. Shi. A fast distributed stochastic gradient descent algorithm for matrix factorization, 2014. URL http://proceedings.mlr.press/v36/li14.html.

[106] W.-S. Chin, Y. Zhuang, Y.-C. Juan, and C.-J. Lin. A fast parallel stochastic gradient method for matrix factorization in shared memory systems. *ACM Trans. Intell. Syst. Technol.*, 6(1): 2:1–2:24, 2015. ISSN 2157-6904. doi: 10.1145/2668133. URL http://doi.acm.org/10.1145/2668133.

[107] J. Oh, W.-S. Han, H. Yu, and X. Jiang. Fast and robust parallel sgd matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 865–874, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783322. URL http://doi.acm.org/10.1145/2783258.2783322.

[108] B. A. Godichon and S. Saadane. On the rates of convergence of parallelized averaged stochastic gradient algorithms. *arXiv preprint arXiv:1710.07926*, 2017.

Communication trade-offs for synchronized distributed SGD with large step size

SUPPLEMENTARY MATERIAL

In this Appendix, we give the proofs of our main results, and auxiliary elements. In Section A.2, we provide results in the on-line setting where we consider the particular case of a decaying sequence $\eta_k^t = (\sum_{t'=1}^{t-1} N^{t'} + k)^{-\alpha}$, for some $\alpha \in (\frac{1}{2}, 1)$. In Appendix B, we describe the experimental evaluations that illustrate the behavior of the different processes. In Appendix C we provide some additional material (Tables, interpretations, etc.) which may help the reader navigate through our results. In Appendix D, we prove contraction Lemmas for $\mathbb{E}[\|\boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star\|^2]$. In Appendix E, we prove similar guarantees for moment of order 4. In Appendix G, we give the proof of the main results on $\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - ws)\|^2$ for mini-batch, one-shot averaging, and Local-SGD in the Finite Horizon setting. In Appendix H we give similar results in the online setting (for decaying step size). Finally, we provide a brief survey of distributed optimization techniques in Appendix I.

Contents

A Main results in the on-line Setting and tightness of Proposition 1

A.1 Most general assumption

Assumption 3 should be formally written as follows:

A3 (Oracle on the gradient) *There exists a filtration $(\mathcal{H}_k^t)_{(t,k)\in[C]\times[N^t]}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for any $(t,k) \in [C]\times[N^t]$ and $\boldsymbol{w} \in \mathbb{R}^d$, $g_{p,k+1}^t(\boldsymbol{w})$ is a $\mathcal{H}_{k+1}^t$-measurable random variable and $\mathbb{E}\left[g_{p,k+1}^t(\boldsymbol{w})|\mathcal{H}_k^t\right] = F'(\boldsymbol{w})$. In addition, we assume the functions $(g_{p,k}^t)_{(t,k)\in[C]\times[N^t]}$ to be independent and identically distributed (i.i.d.) random fields.*

A filtration is an increasing (i.e., for all $(t,k) \preccurlyeq (t',k')$, $\mathcal{H}_k^t \subset \mathcal{H}_{k'}^{t'}$), sequence of σ-algebras. **A**3 expresses that we have access to an i.i.d. sequence $(g_{p,k}^t)_{(t,k)\in[C]\times[N^t]}$ of unbiased estimators of F'. Remark that with such notations, for any $t \in [C], k \in [N^t], p \in [P]$, $\boldsymbol{w}_{p,k}^t$ is $\mathcal{H}_k^t$-measurable.

A.2 Main results: On-line Setting

In the on-line setting we consider the particular case of a decaying sequence $\eta_k^t = (\sum_{t'=1}^{t-1} N^{t'} + k)^{-\alpha}$, for some $\alpha \in (\frac{1}{2}, 1)$. The analysis is slightly more involved as Equation (3) results in more terms than in the finite horizon setting (sums do not directly telescope). While the decaying step-size case enables to improve some terms with respect to the finite horizon case (*e.g.* the speed at which one forgets the initial condition), the trade-offs concerning communication remain unchanged. We define the following constants to make the presentation clear, for $\alpha \in (1/2; 1)$:

$$R_{bias}(X) = 1 + X^{2\alpha}\exp\left(-\mu c_\eta X^{1-\alpha}\right) + \frac{1}{(\mu c_\eta)^{\frac{1}{1-\alpha}}} + \frac{M^2 c_\eta^2 \left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{(\mu c_\eta)^{\frac{2}{1-\alpha}}} + \frac{2L^2 c_\eta^2}{P(\mu c_\eta)^{\frac{1}{1-\alpha}}},$$

$$R_{1,var}(X) = \frac{X^{2\alpha-1}P}{2\alpha-1}\exp\left(-\frac{\mu X^{1-\alpha}}{2(1-\alpha)}\right) + \frac{P}{X^{1-\alpha}}\frac{1}{c_\eta \mu} + \frac{P}{X\mu^{\frac{2\alpha}{1-\alpha}} c_\eta^{\frac{2}{1-\alpha}}} + \frac{L^2 P c_\eta^2}{X^\alpha \mu^2}$$

$$R_{2,var}(X) = \frac{M^2\sigma^2 P c_\eta^2}{\mu^2 X^{2\alpha-1}}.$$

Now we present a result similar to Proposition 1 for mini-batch averaging and one shot averaging:

Proposition S7 (On-line Mini-batch Averaging and One-shot averaging) *Under the Assumptions **A**1, **A**2, **A**3, **A**5, **A**6 using $\eta_k^t = (\sum_{t'=1}^{t-1} N^{t'} + k)^{-\alpha}$ we have for respectively mini-batch averaging and one-shot averaging:*

$$\mathbb{E}\left[\left\|\nabla^2 F(\boldsymbol{w}^\star)(\boldsymbol{w} - \boldsymbol{w}^\star)\right\|^2\right] \precsim \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{X^2 c_\eta^2} R_{bias}(X) + \frac{2\sigma^2}{T}\left(1 + \frac{R_{1,var}(X)}{\kappa} + \frac{R_{2,var}(X)}{\kappa^2}\right),$$

with respectively $\kappa = 1$ and $X = N$ for one-shot averaging, and $\kappa = P$ and $X = C$ for mini-batch averaging.

Interpretation and comparison. This proposition is directly derived from Lemma S59 in Appendix H. This proposition is similar to Propositions 1 and 2, but the overall convergence rate is better as using decaying step size eventually performs better. For example, the bias term mainly decays as $1/X^2$ instead of $1/(\eta X)^2$. This underlines why in practice decaying step size is often preferable. Asymptotically, the variance term is now dominant, and as before, MBA and OSA have similar performance as $\sigma^2 T^{-1}$.

Optimal step size and asymptotic regimes for P, T For a fixed number of machine P, the bias is asymptotically vanishing, and if we ignore the linearly decaying terms and the dependence on μ, the resulting dominating term in $R_{1,2,var}$ is controlled by $X^{-\min\{(1-\alpha),\alpha,2\alpha-1\}}$, which would result in an optimal choice of $\alpha = 2/3$.

In the non asymptotic regime, where the total number of iterations and P grow simultaneously, the variance of OSA scales as T^{-1} as long as $PX^{-\min\{(1-\alpha),\alpha,2\alpha-1\}} = O(1)$. In other words, for $\alpha = 2/3$, we need $P \leq X^{1/3}$: the number of machines as to be smaller than the number of iterations to the power $1/3$, in other words, for 1000 iterations, one could only use 10 machines to reach the asymptotic regime where OSA performs similarly to MBA.

A.3 Tight bias term for finite horizon setting

For a simple 1-dimensional quadratic function $F(w) = h(w - \boldsymbol{w}^\star)^2$, with $h > 0$, without any noise (we observe $y_i = w_0 x_i + \varepsilon_i$, with $\varepsilon_i \equiv 0$, and $x_i \equiv \sqrt{h}$), we have for a step size η, for any $p \in \{1, \ldots, P\}, k \leq N$:

$$\boldsymbol{w}_{p,k}^1 - \boldsymbol{w}^\star = (1 - \eta h)^k (w_0 - \boldsymbol{w}^\star) \tag{S1}$$

$$h^2 \left(N^{-1} \sum_{k=0}^{N-1} \boldsymbol{w}_{p,k}^1 - \boldsymbol{w}^\star \right)^2 = \frac{1 - (1 - \eta h)^N}{(\eta N)^2} (w_0 - \boldsymbol{w}^\star)^2, \tag{S2}$$

$$\implies h^2 \left(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star \right)^2 = \frac{1 - (1 - \eta h)^N}{(\eta N)^2} (w_0 - \boldsymbol{w}^\star)^2, \tag{S3}$$

which exactly matches the Bias term in (5) (for a quadratic, $M = 0$) and $L^2 \eta / hP \leq 1/P$ is a small constant ($\eta L \leq 1$ and $L = \mu = h$).

B Experimental results

Table S1: Data-sets for experimentation.

Name of the Data-set	Task	Algorithm	Number of Samples	Number of Features
Epsilon	Classification	Logistic	400000	2000
Year Prediction MSD	Regression	Least-Squares	463715	90
CPU Stall	Regression	Least-Squares	8192	12

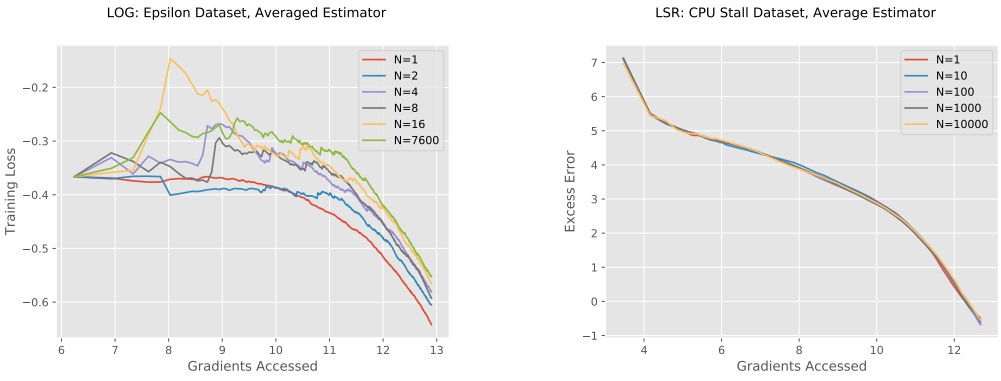

Figure S1: Performance of Local SGD

We perform experiments for three different data-sets[5], two for least-squares regression and one for logistic regression Table S1. For all the curves we use $\log(y)$ v/s $\log(x)$ plots unless explicitly mentioned. Moreover, to elucidate the theory we use the same learning rates for all the algorithms in an experiment. The number of workers is set to $P = 32$ every where, and plots are labeled w.r.t.

[5]Data available at https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/.

the number of local steps N which we don't change along the different phases. We do the following experiments:

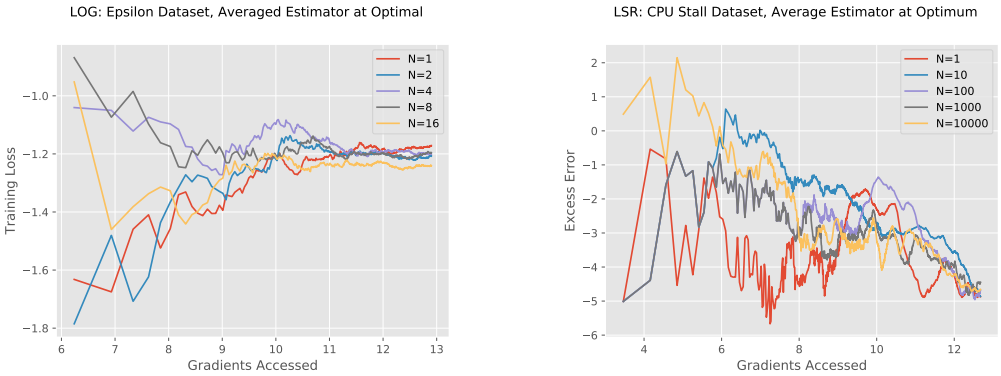

Figure S2: Performance of Local SGD at the optimal

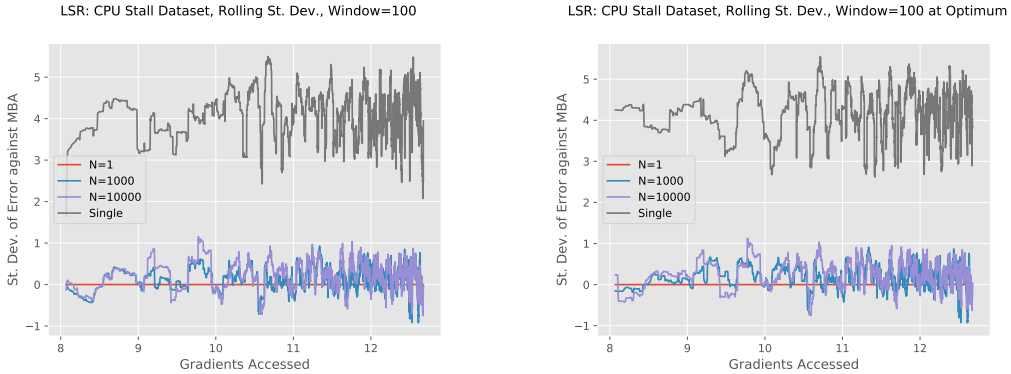

Figure S3: Variance of the loss function compared to MBA

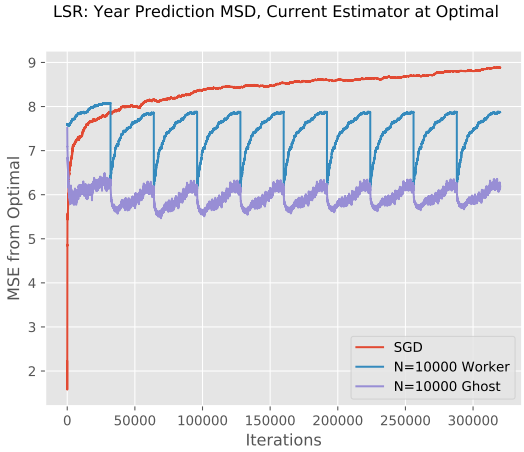

Figure S4: Iterate Convergence of a single process against SGD and the ghost process.

1. Performance of local SGD with different number of local steps spanning OSA to MBA (Figure S1). We globally find MBA to perform the best. Besides, as we increase the number of local steps N the performance gets closer to OSA. This observation aligns with our theoretical guarantees. We use the averaged iterate (i.e., $\overline{\overline{w}}$, the average over all the iterates

till that point) for reporting the performance. The current iterate (i.e., $\breve{\boldsymbol{w}}_k^t$, the ghost iterate for the current iteration) is omitted as the graphs are too noisy to be interpreted, and a variance of the loss is used instead.

2. Performance of local SGD with different number of local steps when started at the optimal point (Figure S2). We expect that if we start at $\boldsymbol{w}^\star$ then the bias term goes to zero and the difference between the algorithms becomes sharper. This is because our results predict that for constant learning rate, the initial conditions are forgotten at the same rate. We see that mini-batch outperforms OSA no the first iterations, but not asymptotically.
3. Variance of the estimators, for loss (Figure S3) and iterate values (Figure S4). We expect that a larger mini-batch size predicts a lower variance for these cases, and we observe the same through our experiments. In fact, the mean squared error of the parameters at the optimal is observed to be following a periodic curve. The value on an individual worker rises until it communicates, but always remains lower than a single SGD process run for the same number of iterations. This, verifies our theory and results for iterate convergence. Moreover, the variance at the loss function follows a similar pattern which elucidates the fact the intuitions developed in the paper also hold for functional convergence.

C Some Additional Material

C.1 Pseudo codes

Pseudo codes of both algorithms are given in Figure S5.

procedure SGD
  Input: $F : \mathbb{R}^d \to \mathbb{R}$
  $\boldsymbol{v}_0 \leftarrow$ **Initialize**
  for t = 0,1,2,...T **do**
    $\boldsymbol{g}_t(\boldsymbol{v}_{t-1}) \leftarrow$ **SFO**$(F, \boldsymbol{v}_{t-1})$
    $\boldsymbol{v}_t \leftarrow \boldsymbol{v}_{t-1} - \eta_t \boldsymbol{g}_t(\boldsymbol{v}_{t-1})$
  Output:
$S(\boldsymbol{v}_0, \boldsymbol{v}_1, .., \boldsymbol{v}_{T-1}, \boldsymbol{v}_T) \in \mathbb{R}^d$

Algorithm 1: Vanilla-SGD

procedure LOCAL-SGD
  Input: $F : \mathbb{R}^d \to \mathbb{R}$
  $\hat{\boldsymbol{w}}^0 = \boldsymbol{w}^0 \leftarrow$ **Initialize**
  for $t = 1, 2, ...C$ **do**
    parfor i=1,2,...P **do**
      $\boldsymbol{w}_{i,0}^t \leftarrow \hat{\boldsymbol{w}}^{t-1}$
      for k=0,1,2,...N^t **do**
        $\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t) \leftarrow$ **SFO**$(F, \boldsymbol{w}_{i,k-1}^t)$
        $\boldsymbol{w}_{i,k}^t \leftarrow \boldsymbol{w}_{i,k-1}^t - \eta_k^t \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)$
      $\overline{\boldsymbol{w}}_i^t \leftarrow \frac{1}{N_t} \sum_{k=1}^{N_t} \boldsymbol{w}_{i,k}^t$
    end parfor
    $\overline{\boldsymbol{w}}^t \leftarrow \frac{1}{P} \sum_{i=1}^{P} \overline{\boldsymbol{w}}_i^t$
    $\hat{\boldsymbol{w}}^t \leftarrow \frac{1}{p} \sum_{i=1}^{P} \boldsymbol{w}_{i,N_t}^t$
  Output: $\overline{\overline{\boldsymbol{w}}}^T = \frac{1}{C} \sum_{t=1}^{C} \overline{\boldsymbol{w}}^t \in \mathbb{R}^d$

Algorithm 2: Local-SGD

Figure S5: Serial and parallel SGD algorithms. **SFO** stands for the stochastic first order oracle. Note that every node has access to the full function i.e., the data is not distributed across nodes.

C.2 Summary of Results

In the table below, we specify for which algorithm our results apply (mini batch, one shot, or local SGD), under which assumptions they are proved and if they apply to the on-line setting(OL) or just the finite horizon(FH) case.

C.3 Example: Learning from i.i.d. observations

Our main motivation comes from machine learning; consider two sets $\mathcal{X}, \mathcal{Y}$ and a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$. The generalization error is defined as $F_\ell(\boldsymbol{w}) = \mathbb{E}_{X,Y}[\ell(X, Y, \boldsymbol{w})]$, where

		Assumptions							Setting	
Proposition	Algorithm	**A**1	**A**2	**Q**1	**A**3	**A**4	**A**5	**A**6	FH	OL
Proposition 1	Mini-Batch	✓	✓		✓		✓	✓	✓	
Proposition 2	One-shot averaging	✓	✓		✓		✓	✓	✓	
Proposition S7	Mini-Batch &OS	✓	✓		✓		✓	✓		✓
Proposition 3	Local SGD			✓	✓	✓			✓	
Corollary S17	Local SGD			✓	✓		✓	✓	✓	✓
Corollary S21	Local SGD	✓	✓		✓	✓			✓	✓
Theorem 6 1.	Local SGD			✓	✓		✓	✓	✓	
Theorem 6 2.	Local SGD	✓	✓		✓	✓			✓	

Table S2: Summary of results

(X, Y) are some random variables. Given i.i.d. observations $(X_k, Y_k)_{k\in\mathbb{N}^*}$ with the same distribution as (X, Y), for any $k \in \mathbb{N}^*$, we define $f_k(\cdot) = \ell(X_k, Y_k, \cdot)$ the loss with respect to observation k. SGD can be used in two contexts:

1. *Stochastic Approximation*: We use *independent* observations at each iteration. The total number of iterations is thus at most the number of observations we access. SGD then corresponds to following the gradient of the loss f_k on a single independent observation (X_k, Y_k). As the gradients we use are then unbiased gradients of the generalization error, this means that SGD directly minimizes this (unknown) function.
2. *Empirical Risk Minimization*: We define the empirical risk as $\hat{F}_\ell(\boldsymbol{w}) = N^{-1}\sum_{k=1}^{N}[\ell(X_k, Y_k, \boldsymbol{w})]$. At each step t, we sample an index i_t *uniformly on* $[N]$, and use the gradient of the loss f_{i_t}. Here the number of iterations is not limited, but the algorithm will converge to the minimum of the empirical risk.

In practice, this means that in the first situation, we want to optimize the precision of the algorithm for a limited number of oracle calls, while in the second situation one would rather optimize the number of outer iterations of the algorithm (*i.e.* its running time). In both these assumptions, Assumption **A**3 is satisfied for the filtration generated by all the observations before time (t, k) (respectively all the indices sampled before time (t, k)).

Two typical situations regarding loss functions are worth mentioning. On the first hand, in *least-squares regression*, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the loss function is $\ell(X, Y, \boldsymbol{w}) = (\langle X, \boldsymbol{w}\rangle - Y)^2$. Then F_Σ is the quadratic function $\boldsymbol{w} \mapsto \left\|\Sigma^{1/2}(\boldsymbol{w} - \boldsymbol{w}^\star)\right\|^2/2$, with $\Sigma = \mathbb{E}[XX^\top]$, which satisfies Assumption **Q**1. For any $\boldsymbol{w} \in \mathbb{R}^d$,

$$f_k'(\boldsymbol{w}) - F_\Sigma'(\boldsymbol{w}) = (X_k X_k^\top - \Sigma)(\boldsymbol{w} - \boldsymbol{w}^\star) - (X_k^\top \boldsymbol{w}^\star - Y_k)X_k \tag{S4}$$

Then, Assumption **A**5 and **A**6 are satisfied, if X is bounded and Y has finite variance.

On the other hand, in *logistic regression*, where $\ell(X, Y, \boldsymbol{w}) = \log(1 + \exp(-Y\langle X, \boldsymbol{w}\rangle))$. Assumptions **A**2 and **A**4 are then satisfied [44], as is Assumption **A**1 under an additional restriction to a compact set or if an extra regularization is added.

SGD for least squares regression typically satisfies **Q**1, **A**3, **A**5 and **A**6. On the other hand, SGD for logistic regression satisfies **A**1, **A**2, **A**3 and **A**4.

D Convergence guaranties for the second order moment.

In this section, we prove several Lemmas that allow to control the second order moment for the iterate. We first recall a few useful inequalities that will be used in the following. See for example [50].

If F is convex and smooth (*e.g.* satisfies **A**2), the gradient of F is cocoercive, thus for any $\boldsymbol{w} \in \mathbb{R}^d$:

$$L \left\langle F'(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \right\rangle \geq \left\| F'(\boldsymbol{w}) \right\|. \tag{S5}$$

If the function is strongly-convex (Assumption **A**1), then for any $\boldsymbol{w} \in \mathbb{R}^d$:

$$\left\langle F'(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^\star \right\rangle \geq \mu \left\| \boldsymbol{w} - \boldsymbol{w}^\star \right\|^2. \tag{S6}$$

D.1 Inner iteration Lemma

We first recall the proof of the convergence for inner iterates. This proof corresponds to what happens on one machine, and can be found in the literature [10, 25] for example.

For any $p \in [P], t \in [C], k \in [N^t]$, under Assumptions **A**1, **A**2, **A**3, **A**5, **A**6, we have

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star \right\|^2\right] \leq \mathbb{E}\left[\left\| \boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star \right\|^2\right] - \eta_k^t \left\langle F'(\boldsymbol{w}_{p,k-1}^t), \boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star \right\rangle + 2(\eta_k^t)^2 \sigma^2 \tag{S7}$$

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star \right\|^2\right] \leq (1 - \eta_k^t \mu) \mathbb{E}\left[\left\| \boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star \right\|^2\right] + 2\eta_k^t \sigma^2.$$

Using the second equation recursively results in:

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star \right\|^2\right] \leq \prod_{m=1}^{k} (1 - \eta_m^t \mu) \mathbb{E}\left[\left\| \boldsymbol{w}_{p,0}^t - \boldsymbol{w}^\star \right\|^2\right] + 2\sigma^2 \sum_{m=1}^{k} (\eta_m^t)^2 \prod_{l=m+1}^{k} (1 - \eta_l^t \mu). \tag{S8}$$

More precisely, for precise reference in the following proofs, we referenced this inequality with the following specific cases:

Lemma S8 *Under Assumptions **A**1, **A**2, **A**3, **A**5, **A**6, for mini-batch SGD with batch-size P and step-size η we have,*

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{MB}^t - \boldsymbol{w}^\star \right\|^2\right] \leq \prod_{m=1}^{t} (1 - \mu\eta) \mathbb{E}\left[\left\| \boldsymbol{w}^0 - \boldsymbol{w}^\star \right\|^2\right] + \frac{2\sigma^2\eta^2}{P} \sum_{m=1}^{t} \prod_{l=m+1}^{t} (1 - \mu\eta).$$

Such a result on reduced variance for mini-batch SGD ($\frac{\sigma^2}{P}$) can be found in many previous works like [51]. Since mini-batch SGD is trivial to parallelize, this result also holds for the averaged iterate for outer iteration t while using mini-batch averaging. Similarly, for decaying step sizes,

Lemma S9 *Under Assumptions **A**1, **A**2, **A**3, **A**5, **A**6, and $\tilde{\eta}_t = \frac{c_\eta}{t^\alpha}$ for mini-batch SGD, for any $t \in [C]$ we have,*

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{MB}^t - \boldsymbol{w}^\star \right\|^2\right] \leq \prod_{m=1}^{t} (1 - \mu\tilde{\eta}_m) \left\| \boldsymbol{w}^0 - \boldsymbol{w}^\star \right\|^2 + 2\sigma^2 \frac{1}{P} \sum_{m=1}^{t} (\tilde{\eta}_m)^2 \prod_{l=m+1}^{t} (1 - \mu\tilde{\eta}_l).$$

Similarly, in the case of one-shot averaging,

Lemma S10 *Under Assumptions **A**1, **A**2, **A**3, **A**5, **A**6 and a constant step-size η using one-shot averaging, for any $K \in [N^1]$ and $i \in [P]$ we have,*

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{i,K}^1 - \boldsymbol{w}^\star \right\|^2\right] \leq \prod_{m=1}^{K} (1 - \mu\eta) \left\| \boldsymbol{w}^0 - \boldsymbol{w}^\star \right\|^2 + 2\sigma^2\eta^2 \sum_{m=1}^{K} \prod_{l=m+1}^{K} (1 - \mu\eta_l^1).$$

Lemma S11 *Under Assumptions **A**1, **A**2, **A**3, **A**5, **A**6, and $\eta_k^1 = \tilde{\eta}_k = \frac{c_\eta}{k^\alpha}$ using one-shot averaging for any $K \in [N^1]$ and $i \in [P]$ we have,*

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{i,K}^1 - \boldsymbol{w}^\star \right\|^2\right] \leq \prod_{m=1}^{K} (1 - \mu\eta_m^1) \left\| \boldsymbol{w}^0 - \boldsymbol{w}^\star \right\|^2 + 2\sigma^2 \sum_{m=1}^{K} (\eta_m^1)^2 \prod_{l=m+1}^{K} (1 - \mu\eta_l^1).$$

D.2 Proof of Proposition 3

In this Section we prove Proposition 3. In order to provide a bound on the mean squared distance to the optimum of the outer iterates, we introduce a *ghost* sequence [52], *i.e.*, a sequence of iterates which is not actually computed. For any $t \in [C], k \in [N^t]$, we define

$$\breve{\boldsymbol{w}}_k^t := \frac{1}{P}\sum_{i=1}^{P} \boldsymbol{w}_{i,k}^t. \tag{S9}$$

We prove the following Lemma:

Lemma S12 *Under Assumptions **Q**1, **A**3 and **A**4, for any $t \in [C], K \in [N^t]$, we have:*

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_K^t - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{m=1}^{K}(1-\mu\eta_m^t)\left\|\breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2}{P}\sum_{m=1}^{K}(\eta_m^t)^2\prod_{l=m+1}^{K}(1-\mu\eta_l^t)\,. \tag{S10}$$

Remarking that for any $t \in [C]$, $\breve{\boldsymbol{w}}_{N^t}^t = \hat{\boldsymbol{w}}^t$ this implies the *first inequality* of Proposition 3. Note that this Lemma is valid for both decaying steps and and a constant learning rate. Especially, for a constant step size η, and $K = N^t$:

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] \leq (1-\mu\eta)^{N^t}\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2}{P}\eta\frac{1-(1-\mu\eta)^{N^t}}{\mu}\,.$$

More generally, we also have the following corollary, if we denote $(\tilde{\eta}_k)_{k\geq 0}$ the sequence such that $\eta_k^t = \tilde{\eta}_{\{\sum_{t'=1}^{t-1}N^{t'}+k\}}$ (this just corresponds to re-indexing the sequence):

Corollary S13 *Under Assumptions **Q**1, **A**3 and **A**4, for any $T \in [C]$, we have:*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^T - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{k=1}^{\sum_{t=1}^{T}N^t}(1-\mu\tilde{\eta}_k)\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2}{P}\sum_{t=1}^{\sum_{t=1}^{T}N^t}\tilde{\eta}_k^2\prod_{j=k+1}^{\sum_{t=1}^{T}N^t}(1-\mu\tilde{\eta}_j)\,. \tag{S11}$$

Proof 14 (Proof of Corollary S13) *By induction, Lemma S12 implies that for any $T \in [C]$*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^T - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{t=1}^{T}\prod_{k=1}^{N^t}(1-\mu\eta_k^t)\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2}{P}\sum_{t=1}^{T}\prod_{t'=t+1}^{T}\prod_{k=1}^{N^{t'}}(1-\mu\eta_k^t)\sum_{k=1}^{N^t}(\eta_k^t)^2\prod_{j=k+1}^{N^t}(1-\mu\eta_j^t)\,. \tag{S12}$$

Then using $\eta_k^t = \tilde{\eta}_{\{\sum_{t'=1}^{t-1}N^{t'}+k\}}$, the corollary is just re-writing of Equation (S12).

To prove the second inequality of Proposition 3, we combine Lemma S12 and Equation (S8), using the fact that $\boldsymbol{w}_{p,0}^t = \hat{\boldsymbol{w}}^{t-1}$.

This results means that for a quadratic function with gradients having uniformly bounded variance, the outer iteration decay is the same as for mini-batch iterations (but for mini-batch, it is true under the weaker set of Assumptions **A**1, **A**2, **A**3, **A**5, **A**6).

D.2.1 Proof

Proof 15 (Proof of Lemma S12) *By definition of $\breve{\boldsymbol{w}}_k^t$, we have for any $t \in [C], k \in [N^t]$, using the linearity of F' (Assumption **Q**1):*

$$\frac{1}{P}\sum_{i=1}^{P}\boldsymbol{w}_{i,k+1}^t = \frac{1}{P}\sum_{i=1}^{P}\boldsymbol{w}_{i,k}^t - \frac{1}{P}\sum_{i=1}^{P}\eta_{k+1}^t g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)$$

$$\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star = \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star - \frac{1}{P}\sum_{i=1}^{P}\eta_{k+1}^t g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)$$

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_{k,t}\right] \leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 - 2\eta_{k+1}^t\left\langle\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\right\rangle$$

$$+ (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 |\mathcal{H}_{k,t}\right] . \tag{S13}$$

Now analyzing just the last term,

$$\begin{aligned}
&(\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&= (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P \left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t)\right)\right\|^2 |\mathcal{H}_{k,t}\right] + (\eta_{k+1}^t)^2 \left\|F'(\breve{\boldsymbol{w}}_k^t)\right\|^2 . \qquad \text{(S14)}
\end{aligned}$$

*Under the independence of the noises (Assumption **A**3), then the uniform upper bound on the variance (Assumption **A**4), we have the following upper bound :*

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P \left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t)\right)\right\|^2 |\mathcal{H}_{k,t}\right] &= \frac{1}{P^2}\sum_{i=1}^P \mathbb{E}\left[\left\|\left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t)\right)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&\leq \frac{1}{P}\sigma_\infty^2 .
\end{aligned}$$

*Under Assumption **Q**1, F' is co-coercive, thus using Equation (S5), we have the following upper bound:*

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_{k,t}\right] \leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 - 2\eta_{k+1}^t(1-\eta_{k+1}^t L)\left\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\right\rangle + \frac{(\eta_{k+1}^t)^2\sigma_\infty^2}{P} .$$

And using strong convexity (esp. Equation (S6)), and the fact that $\eta_{k+1}^t L \leq \frac{1}{2}$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_{k,t}\right] &\leq (1 - 2\mu\eta_{k+1}^t(1-\eta_{k+1}^t L))\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 + \frac{(\eta_{k+1}^t)^2\sigma_\infty^2}{P} \\
&\leq (1 - \mu\eta_{k+1}^t)\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 + \frac{(\eta_{k+1}^t)^2\sigma_\infty^2}{P} . \qquad \text{(S15)}
\end{aligned}$$

By recursion, we then have, for any $K \in [N^t]$:

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_K^t - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{k=1}^K (1-\mu\eta_k^t)\left\|\breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2}{P}\sum_{k=1}^K (\eta_k^t)^2 \prod_{j=k}^K (1-\mu\eta_j^t) .$$

This concludes the proof.

D.3 Proof of Proposition S16

In this Section we prove Proposition S16.

D.3.1 Statement of Proposition S16

Proposition S16 (Local-SGD: Quadratic Functions) *Under Assumptions **Q**1,**A**3,**A**5,**A**6, we have the following bound for one shot averaging: $p \in [P], t \in [C], k \in [N^t]$,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] \leq \kappa_2^t \prod_{k=1}^{\sum_{t'=1}^t N^{t'}} (1-\mu\tilde{\eta}_k)\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + 2\kappa_1^t\kappa_2^t\frac{\sigma^2}{P}\sum_{t=1}^{\sum_{k=1}^t N^t} \tilde{\eta}_k^2 \prod_{j=k+1}^{\sum_{t'=1}^t N^{t'}} (1-\mu\tilde{\eta}_j) \tag{S16}$$

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star\right\|^2\right] \leq \kappa_2^t \prod_{k=1}^{\sum_{t'=1}^t N^{t'}+k} (1-\mu\tilde{\eta}_k)\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + 2\kappa_1^t\kappa_2^t\frac{\sigma^2}{P}\sum_{u=1}^{\sum_{t'=1}^t N^{t'}} \tilde{\eta}_u^2 \prod_{j=k+1}^{\sum_{t'=1}^t N^{t'}+k} (1-\mu\tilde{\eta}_j)$$

$$+ 2\frac{\sigma^2}{P} \sum_{u=\sum_{t'=1}^{t} N^{t'}}^{\sum_{t'=1}^{t} N^{t'}+k} \tilde{\eta}_u^2 \prod_{j=u+1}^{\sum_{t'=1}^{t} N^{t'}+k} (1-\mu\tilde{\eta}_j), \tag{S17}$$

with, for $t \in [C]$, $\kappa_1^t = \left(4 + \mu \sum_{k=1}^{N^t} (\eta_k^t)^2\right)$, *and* $\kappa_2^t := \exp\left(\mu \sum_{t'=0}^{t} \sum_{k=1}^{N^t} (\eta_k^t)^2\right)$.

When considering a constant step size η, we have the following corollary.

Corollary S17 (Local-SGD: Quadratic Functions) *Under Assumptions* ***Q****1,****A****3,****A****5,****A****6, we have the following bound for one shot averaging:* $p \in [P], t \in [C], k \in [N^t]$, *constant learning rate* η,

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right] \leq \tau_2^t (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}} \left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + 2\tau_1^t \tau_2^t \frac{\sigma^2 \eta}{P} \frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{\mu} \tag{S18}$$

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star\right\|^2\right] &\leq \tau_2^t (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}+k} \left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 \\ &\quad + 2\sigma^2\eta \left(\sup_{t'=1\dots t} (\tau_1^t)\tau_2^t \frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{P\mu} + \frac{1-(1-\eta\mu)^k}{\mu} \right).\end{aligned} \tag{S19}$$

Where we have $\tau_1^t = 4 + \mu \boldsymbol{N}^t \eta^2$ and $\tau_2^t = \exp\left(\mu \boldsymbol{N}_1^t \eta^2\right)$. Under the latter requirement (for optimality) that for any t, $\boldsymbol{N}^t \mu P \eta \leq 1$, we have $\mu \boldsymbol{N}_1^t \eta^2 \leq C\eta P^{-1}$, thus this is generally a small constant. This result is a consequence of Lemma S18.

Interpretation. As before, the first bound shows that the variance of the iterates *after communication* is reduced by a factor of P *w.r.t.* the serial case, thus almost as good as mini-batch averaging. However, the constants involved are worse than in the additive noise setting (Proposition 3). Consequently, and similarly to Proposition 3, the bound for the current iterates is composed of two terms for the variance: a "reduced variance" coming from the communication step, and a "inner loop" variance, that does not benefit from the number of machines.

Finally, we provide a convergence result in the most general case, removing the quadratic assumption. For the sake of concision, we skip the bound for the averaged iterate after a communication round, and directly give the result for the inner process.

D.3.2 Proof

This result is a consequence of Lemma S18, which implies Equation (S18). Indeed, using it recursively, and using $(1+x) \leq \exp(x)$, we get:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^T - \boldsymbol{w}^\star\right\|^2\right] &\leq \exp\left(\mu \sum_{t'=0}^{T} \sum_{k=1}^{N^t} (\eta_k^t)^2\right) \prod_{t'=1}^{T} \prod_{k=1}^{N^{t'}} (1-\mu\eta_k^t) \mathbb{E}\left[\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2\right] \\ &\quad + 2\kappa_1 \exp\left(\mu \sum_{t'=0}^{t} \sum_{k=1}^{N^t} (\eta_k^t)^2\right) \frac{\sigma^2}{P} \sum_{t=1}^{T} \prod_{t'=t+1}^{T} \prod_{k=1}^{N^{t'}} (1-\mu\eta_k^t) \sum_{k=1}^{N^t} (\eta_k^t)^2 \prod_{j=k+1}^{N^t} (1-\mu\eta_j^t)\end{aligned}$$

With, for $t \in [C]$, $\kappa_1^t = \left(4 + \mu \sum_{k=1}^{N^t} (\eta_k^t)^2\right)$, and $\kappa_2^t := \exp\left(\mu \sum_{t'=0}^{t} \sum_{k=1}^{N^t} (\eta_k^t)^2\right)$, and re-writing everything in terms of the sequence $\tilde{\eta}_k$, it gives Equation (S16). The second inequality naturally follows.

Lemma S18 *Under Assumptions* ***Q****1, A****3****, A****5****, A****6****, for any* $t \in [C], K \in [N^t]$, *we have:*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] \leq \left(1 + \mu \sum_{k=1}^{N^t} (\eta_k^t)^2\right) \prod_{k=1}^{N^t} (1-\mu\eta_k) \mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right] \tag{S20}$$

$$+ 2\left(4 + \mu \sum_{k=1}^{N^t} (\eta_k^t)^2\right) \frac{\sigma^2}{P} \sum_{k=0}^{N^t} (\eta_k^t)^2 \prod_{j=k+1}^{N^t} (1-\mu\eta_j^t). \tag{S21}$$

The proof is a bit technical, so we summarize here the 2 main steps:

1. We prove an inequality (namely Equation (S23)) that is comparable to Equation (S15), but with an extra term.
2. We use the control on the inner process (Appendix D.1) to control the extra term.

Proof 19 *We consider again the ghost process defined at Equation (S9). Equations (S13) and (S14) are still valid. We now use the following decomposition*[6]*:*

$$\begin{aligned}
\square &:= (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&= (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P \left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t)\right)\right\|^2 |\mathcal{H}_{k,t}\right] + (\eta_{k+1}^t)^2 \left\|F'(\breve{\boldsymbol{w}}_k^t)\right\|^2 \\
&\leq 2(\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P \left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t) - g_{i,k+1}^t(\boldsymbol{w}^\star)\right)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&+ 2(\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}^\star)\right\|^2 |\mathcal{H}_{k,t}\right] + (\eta_{k+1}^t)^2 \left\|F'(\breve{\boldsymbol{w}}_k^t)\right\|^2 .
\end{aligned}$$

*Using the independence of the noises (Assumption **A**3) we have,*

$$\begin{aligned}
\square &\leq \frac{2(\eta_{k+1}^t)^2}{P^2}\sum_{i=1}^P \mathbb{E}\left[\left\|\left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t) - g_{i,k+1}^t(\boldsymbol{w}^\star)\right)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&+ \frac{2(\eta_{k+1}^t)^2}{P}\mathbb{E}\left[\left\|g_{i,k+1}^t(\boldsymbol{w}^\star)\right\|^2 |\mathcal{H}_{k,t}\right] + (\eta_{k+1}^t)^2 \left\|F'(\breve{\boldsymbol{w}}_k^t)\right\|^2 \\
&\leq \frac{4(\eta_{k+1}^t)^2}{P^2}\sum_{i=1}^P \left(\mathbb{E}\left[\left\|\left(g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - g_{i,k+1}^t(\boldsymbol{w}^\star)\right)\right\|^2 |\mathcal{H}_{k,t}\right] + \mathbb{E}\left[\left\|\left(F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star)\right)\right\|^2 |\mathcal{H}_{k,t}\right]\right) \\
&+ \frac{2(\eta_{k+1}^t)^2}{P}\mathbb{E}\left[\left\|g_{i,k+1}^t(\boldsymbol{w}^\star)\right\|^2 |\mathcal{H}_{k,t}\right] + (\eta_{k+1}^t)^2 \left\|F'(\breve{\boldsymbol{w}}_k^t)\right\|^2 .
\end{aligned}$$

*Using Assumption **A**5 (co-coercivity for $(g_{i,k}^t)$-s and F) we obtain,*

$$\begin{aligned}
\square &\leq \frac{8L(\eta_{k+1}^t)^2}{P^2}\sum_{i=1}^P \left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle + \frac{2(\eta_{k+1}^t)^2}{P}\mathbb{E}\left[\left\|g_{i,k+1}^t(\boldsymbol{w}^\star)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&+ (\eta_{k+1}^t)^2 L\left\langle F'(\breve{\boldsymbol{w}}_k^t), \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\rangle . \qquad \text{(S22)}
\end{aligned}$$

*This leads to, combining Equations (S13) and (S22), and the upper bound on the variance of the noise at the optimum (Assumption **A**6)*

$$\begin{aligned}
\diamond &:= \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_{k,t}\right] \\
&\leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 - 2\eta_{k+1}^t\left\langle\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\right\rangle + \frac{2(\eta_{k+1}^t)^2}{P}\mathbb{E}\left[\left\|g_{i,k+1}^t(\boldsymbol{w}^\star)\right\|^2 |\mathcal{H}_{k,t}\right] \\
&+ \frac{8L(\eta_{k+1}^t)^2}{P^2}\sum_{i=1}^P \left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle + (\eta_{k+1}^t)^2 L\left\langle F'(\breve{\boldsymbol{w}}_k^t), \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\rangle \\
&\leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 - 2\eta_{k+1}^t(1 - \eta_{k+1}^t L)\left\langle\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\right\rangle + 2\frac{(\eta_{k+1}^t)^2}{P}\sigma^2 \\
&+ \frac{8L(\eta_{k+1}^t)^2}{P^2}\sum_{i=1}^P \left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle .
\end{aligned}$$

[6] In the following, $\square, \diamond, \clubsuit$, etc. are used as symbolic notations to ease presentation.

Using $L\eta_{k+1}^t \leq \frac{1}{2}$, and strong-convexity (Assumption A1)

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_{k,t}\right] \leq (1-\mu\eta_{k+1}^t)\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 + \frac{2(\eta_{k+1}^t)^2\sigma^2}{P} + \frac{8L(\eta_{k+1}^t)^2}{P^2}\sum_{i=1}^P \left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle . \tag{S23}$$

This inequality should be compared to Equation (S15). It is interesting to remark that the last term is not an artifact of the proof: this is easy to check for least-squares regression.

Using recursively the above inequality and using the definition of $\breve{\boldsymbol{w}}^t$, and taking expectation on the historical randomness we have, for any $N \in [N^t - 1]$

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{N+1}^t - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{k=0}^N (1-\mu\eta_{k+1}^t)\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star\right\|^2\right] + 2\frac{\sigma^2}{P}\sum_{k=0}^N(\eta_{k+1}^t)^2\prod_{j=k+1}^N(1-\mu\eta_{j+1}^t) + \frac{8L}{P^2}\mathbb{E}\left[\sum_{k=0}^N(\eta_{k+1}^t)^2\sum_{i=1}^P\left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle\prod_{j=k+1}^N(1-\mu\eta_{j+1}^t)\right].$$

Especially, for $N = N^t - 1$, $\breve{\boldsymbol{w}}_{N^t}^t = \hat{\boldsymbol{w}}^t$, and moreover $\breve{\boldsymbol{w}}_0^t = \hat{\boldsymbol{w}}^{t-1}$:

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{k=0}^{N^t-1} (1-\mu\eta_{k+1}^t)\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right] + 2\frac{\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t) + \frac{8L}{P^2}\mathbb{E}\left[\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\sum_{i=1}^P\left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)\right]. \tag{S24}$$

To upper bound the last term in the above equation, we use Equation (S7),

$$\begin{aligned}
\clubsuit &:= \frac{8L}{P^2}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\sum_{i=1}^P\left\langle F'(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}^\star), \boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\rangle\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t) \\
&\leq \frac{8L}{P^2}\sum_{k=0}^{N^t-1}\eta_{k+1}^t\sum_{i=1}^P\left(\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\|^2\right] - \mathbb{E}\left[\left\|\boldsymbol{w}_{i,k+1}^t - \boldsymbol{w}^\star\right\|^2\right]\right. \\
&\left. + 2(\eta_{k+1}^t)^2\sigma^2\right)\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t) \\
&\leq \frac{8L}{P^2}\sum_{k=0}^{N^t-1}\eta_{k+1}^t\sum_{i=1}^P\left(\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\|^2\right] - \mathbb{E}\left[\left\|\boldsymbol{w}_{i,k+1}^t - \boldsymbol{w}^\star\right\|^2\right]\right)\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t) \\
&+ \frac{16L\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^3\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t) .
\end{aligned}$$

Note that since the mean squared distance doesn't depend on the machine, we can assume to be working on machine 1. This leads to, using an Abel transform:

$$\begin{aligned}
\clubsuit &\leq \frac{8L}{P}\sum_{k=0}^{N^t-1}\left(\mathbb{E}\left[\left\|\boldsymbol{w}_{1,k}^t - \boldsymbol{w}^\star\right\|^2\right] - \mathbb{E}\left[\left\|\boldsymbol{w}_{1,k+1}^t - \boldsymbol{w}^\star\right\|^2\right]\right)\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)\eta_{k+1}^t \\
&+ \frac{16L\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^3\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)
\end{aligned}$$

$$\leq \frac{8L}{P}\Bigg(\sum_{k=0}^{N^t-1} \mathbb{E}\left[\left\|\boldsymbol{w}_{1,k}^t - \boldsymbol{w}^\star\right\|^2\right] \left(\eta_{k+1}^t \prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t) - \eta_k^t \prod_{j=k}^{N^t-1}(1-\mu\eta_{j+1}^t)\right)$$
$$+ \mathbb{E}\left[\left\|\boldsymbol{w}_{1,0}^t - \boldsymbol{w}^\star\right\|^2\right] \prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1}^t)\eta_0^t - \mathbb{E}\left[\left\|\boldsymbol{w}_{1,N^t}^t - \boldsymbol{w}^\star\right\|^2\right]\eta_{N^t}^t\Bigg)$$
$$+ \frac{16L\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^3 \prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)\,.$$

Finally, using convexity, we have that

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{N^t}^t - \boldsymbol{w}^\star\right\|^2\right] \leq \frac{1}{P}\sum_{p=1}^{P}\mathbb{E}\left[\left\|\boldsymbol{w}_{p,N^t}^t - \boldsymbol{w}^\star\right\|^2\right] = \mathbb{E}\left[\left\|\boldsymbol{w}_{1,N^t}^t - \boldsymbol{w}^\star\right\|^2\right].$$

Thus:

$$\clubsuit \leq \frac{8L}{P}\sum_{k=0}^{N^t-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{1,k}^t - \boldsymbol{w}^\star\right\|^2\right]\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1})\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$+ \frac{8L}{P}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right]\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1}^t)\eta_0^t - \frac{8L}{P}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|\right]\eta_{N^t}^t$$
$$+ \frac{16L\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^3\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)\,. \tag{S25}$$

We now use Equation (S8). It leads to the following,

$$\frac{8L}{P}\sum_{k=0}^{N^t-1}\mathbb{E}\left[\left\|\boldsymbol{w}_{1,k}^t - \boldsymbol{w}^\star\right\|^2\right]\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1})\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$\leq \frac{8L}{P}\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right]\sum_{k=0}^{N^t-1}\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$+ \frac{8L}{P}\sum_{k=0}^{N^t-1}\big(2\sigma^2\sum_{l=1}^{k}(\eta_l^t)^2\prod_{m=l+1}^{k}(1-\mu\eta_m^t)\big)\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1})\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$\leq \frac{8L}{P}\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right]\sum_{k=0}^{N^t-1}\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$+ \frac{16\sigma^2L}{P}\sum_{k=0}^{N^t-1}\sum_{l=1}^{k}(\eta_l^t)^2\prod_{j=l+1}^{N^t-1}(1-\mu\eta_{j+1})\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$\leq \frac{8L}{P}\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right]\big(\eta_{N^t-1}^t - \eta_0^t + \sum_{k=0}^{N^t-1}\mu(\eta_k^t)^2\big)$$
$$+ \frac{16\sigma^2L}{P}\sum_{l=1}^{N^t-1}\sum_{k=l}^{N^t-1}(\eta_l^t)^2\prod_{j=l+1}^{N^t-1}(1-\mu\eta_{j+1})\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$
$$\leq \frac{8L}{P}\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right]\big(\eta_{N^t-1}^t - \eta_0^t + \sum_{k=0}^{N^t-1}\mu(\eta_k^t)^2\big)$$
$$+ \frac{16\sigma^2L}{P}\sum_{l=1}^{N^t-1}(\eta_l^t)^2\prod_{j=l+1}^{N^t-1}(1-\mu\eta_{j+1})\sum_{k=0}^{N^t-1}\big(\eta_{k+1}^t - \eta_k^t(1-\mu\eta_{k+1}^t)\big)$$

$$\leq \frac{8L}{P}\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^2\right]\left(\eta_{N^t}^t-\eta_0^t+\sum_{k=0}^{N^t-1}\mu(\eta_{k+1}^t)^2\right)$$
$$+\frac{16\sigma^2 L}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1})\left(\eta_{N^t}^t-\eta_0^t+\sum_{k=0}^{N^t-1}\mu(\eta_{k+1}^t)^2\right). \tag{S26}$$

Combining Equations (S24) to (S26), we get, denoting $C_{N^t}=\eta_{N^t}^t+\sum_{k=0}^{N^t-1}\mu(\eta_{k+1}^t)^2$*:*

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^2\right] &\leq \prod_{k=0}^{N^t-1}(1-\mu\eta_{k+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^2\right]+2\frac{\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)\\
&+\frac{8L}{P}\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^2\right](C_{N^t}-\eta_0^t)-\frac{8L}{P}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|\right]\eta_{N^t}^t\\
&+\frac{16\sigma^2 L}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1})(C_{N^t}-\eta_0^t)\\
&+\frac{8L}{P}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^2\right]\prod_{j=0}^{N^t-1}(1-\mu\eta_{j+1}^t)\eta_0^t+\frac{16L\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^3\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t)\,.\end{aligned}$$

Thus, simplifying:

$$\begin{aligned}&\left(1+\frac{8L}{P}\eta_{N^t}^t\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^2\right]\\
&\leq\left(1+\frac{8L}{P}\eta_{N^t}^t+\sum_{k=0}^{N^t-1}\mu(\eta_{k+1}^t)^2\right)\prod_{k=0}^{N^t-1}(1-\mu\eta_{k+1})\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^2\right]\\
&+2\frac{\sigma^2}{P}\sum_{k=0}^{N^t-1}(\eta_{k+1}^t)^2\left(1+\frac{8L}{P}\eta_{N^t}^t+\sum_{k=0}^{N^t-1}\mu(\eta_{k+1}^t)^2+L\eta_{k+1}^t\right)\prod_{j=k+1}^{N^t-1}(1-\mu\eta_{j+1}^t).\end{aligned}$$

This concludes the proof of the Lemma, using $L\eta_k^t\leq 1/2$*:*

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^2\right] &\leq \left(1+\mu\sum_{k=1}^{N^t}(\eta_k^t)^2\right)\prod_{k=1}^{N^t}(1-\mu\eta_k)\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^2\right]\\
&+2\left(4+\mu\sum_{k=1}^{N^t}(\eta_k^t)^2\right)\frac{\sigma^2}{P}\sum_{k=0}^{N^t}(\eta_k^t)^2\prod_{j=k+1}^{N^t}(1-\mu\eta_j^t).\end{aligned}$$

This result can be used recursively. It implies that if $\mu\sum_{t=1}^{C}\sum_{k=1}^{N^t}(\eta_k^t)^2\leq K$, then the upper bound on the outer iterates is as good as the one for mini-batch, up to a constant.

D.4 Proof of Proposition S20

In this Section we prove the first upper bound of Corollary S21.

D.4.1 Statement of Proposition S20

Finally, we provide a convergence result in the most general case, removing the quadratic assumption.

Proposition S20 (Local-SGD: General Functions) *Under Assumptions A1, A2, A3, A4 we have:*

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t-\boldsymbol{w}^\star\right\|^2\right]\leq\kappa_2\prod_{k=1}^{\sum_{t'=1}^{t}N^{t'}+k}(1-\mu\tilde{\eta}_k)\left\|\boldsymbol{w}_0-\boldsymbol{w}^\star\right\|^2+2\frac{\sigma^2}{P}\sum_{u=\sum_{t'=1}^{t}N^{t'}}^{\sum_{t'=1}^{t}N^{t'}+k}\tilde{\eta}_u^2\prod_{j=u+1}^{\sum_{t'=1}^{t}N^{t'}+k}(1-\mu\tilde{\eta}_j)$$

$$+ \left(\sup_{t'=1\dots t} C_{P,M,K,t'} \right) \frac{\sigma^2}{P} \sum_{u=1}^{\sum_{t'=1}^{t} N^{t'}} \tilde{\eta}_u^2 \prod_{j=k+1}^{\sum_{t'=1}^{t} N^{t'}+k} (1-\mu\tilde{\eta}_j),$$

with $C_{P,M,K,t} = 1 + MP \sum_{k=1}^{K} \eta_k^t \left\| \breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star \right\|$.

Interpretation: if $(\sup_{t'=1\dots t} C_{P,M,K,t})$ is uniformly bounded, we perform as well as minibatch SGD for the outer iterations (up to a constant).

For a constant step size η, the proposition has the following corollary:

Corollary S21 (Local-SGD: General Functions) *Under Assumptions A1, A2, A3, A4 we have:*

$$\mathbb{E}\left[\left\| \boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star \right\|^2 \right] \leq \tau_2^t (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}+k} \left\| \boldsymbol{w}_0 - \boldsymbol{w}^\star \right\|^2 + \sigma_\infty^2 \left(\left(\sup_{t'=1\dots t} C_{P,M,t'} \right) \frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{P\mu} + 2\frac{1-(1-\eta\mu)^k}{\mu} \right).$$

Where $C_{P,M,t} = 1 + MP\eta \sum_{k=1}^{N^t} \mathbb{E}\left[\left\| \breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star \right\| \right]$. We prove the on-line case of the result using Lemma S22 in supplementary material.

Interpretation. When communication occurs, averaging the different models over the machines results in a variance reduction, but at each phase, the variance accumulated within the phase is degraded with respect to the simplest setting by at most $C_{P,M,t}$. This constant increases with the number of machines and the step size, and also depends on the mean distance $\sum_{k=1}^{N^t} \mathbb{E}\left[\left\| \breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star \right\| \right]$ during phase t. As a consequence if $C_{P,M,t}$ is uniformly bounded, we perform as well as mini-batch SGD. If $\mathbb{E}\left[\left\| \breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star \right\| \right]$ is assumed to be decaying, this is true if for any $t \in [T]$, $N^t \eta MP \mathbb{E}\left[\left\| \hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star \right\| \right] \leq O(1)$.

In the following, we alternatively relax the bounded variance assumption **A**4 and the quadratic assumption **Q**1, and show similar results for local SGD. This allows us to successively cover the cases of least squares regression (LSR) and logistic regression (LR).

D.4.2 Proof

Proposition S20 follows from Lemma S22. We have for any $t \in [C], K \in [N^t]$,

$$\mathbb{E}\left\| \breve{\boldsymbol{w}}_K^t - \boldsymbol{w}^\star \right\|^2 \leq \prod_{k=1}^{K} (1-\mu\eta_k^t) \mathbb{E}\left\| \breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star \right\|^2 + C_{P,M,K,t} \frac{\sigma_\infty^2}{P} \sum_{k=1}^{K} (\eta_k^t)^2 \prod_{j=k+1}^{K} (1-\mu\eta_j^t),$$

with $C_{P,M,K,t} = 1 + MP \sum_{k=1}^{K} \eta_k^t \left\| \breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star \right\|$.

As in the two previous sections, we first focus on upper bounding $\mathbb{E}\left[\left\| \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star \right\|^2 \right]$. We prove the following Lemma:

Lemma S22 *For any $t \in [C], K \in [N^t]$, under Assumptions A1, A2, A3, A4 we have:*

$$\mathbb{E}\left\| \breve{\boldsymbol{w}}_K^t - \boldsymbol{w}^\star \right\|^2 \leq \prod_{k=1}^{K} (1-\mu\eta_k^t) \mathbb{E}\left\| \breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star \right\|^2 + C_{P,M,K,t} \frac{\sigma_\infty^2}{P} \sum_{k=1}^{K} (\eta_k^t)^2 \prod_{j=k+1}^{K} (1-\mu\eta_j^t),$$

with $C_{P,M,K,t} = 1 + MP \sum_{k=1}^{K} \eta_k^t \mathbb{E}\left[\left\| \breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star \right\| \right]$.

This means, if we have consider an weak upper bound on $\mathbb{E}\left[\left\| \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star \right\| \right] \leq R$ that the inner loops keeps the same variance as the mini-batch case if $MP \sum_{k=1}^{K} \eta_k^t = O(1)$. For example, for a constant step size η, it results in $PN^t\eta \leq 1$, *i.e.* $N^t \leq \frac{1}{P\eta}$. Note that the number of inner steps one can make increases with the phases, as $\mathbb{E}\left[\|\hat{w}^t - \boldsymbol{w}^\star\| \right]$ decreases.

D.4.3 Proof of Lemma S22

We rely on the following decomposition. Almost surely, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_t^k\right] &\leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 - 2\eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\rangle \\ &\quad + (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 |\mathcal{H}_{k,t}\right] \\ &\quad + 2\eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle. \end{aligned} \tag{S27}$$

The first two lines correspond to the quadratic case (Equation (S13)), that has been analyzed in Lemma S18. The third term accounts for the difference between the mean gradient and the gradient at the mean point. We use Assumption **A**2 to control this term.

We then use the following Lemma, which control how the inner iterates $\boldsymbol{w}_{p,k}^t$ deviate from their average $\breve{\boldsymbol{w}}_k^t$:

Lemma S23 *For any $t \in [C], k \in [N^t]$, under Assumptions **A**1, **A**2, **A**3, **A**4 we have a.s.:*

$$\frac{1}{P}\sum_{p=1}^P \mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2\right] \leq \sigma_\infty^2 \sum_{j=1}^k (\eta_j^t)^2 \prod_{s=j+1}^k (1 - \eta_s^t \mu).$$

The proof of this Lemma is postponed to Appendix D.4.4.

Using Cauchy-Schwarz inequality and the bound on the third order derivative of F, we have:

$$2\eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle \leq 2\eta_{k+1}^t \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\| \left\|F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\|, \tag{S28}$$

and, using a second order expansion of the gradient at $\breve{\boldsymbol{w}}_k^t$ together with Assumption **A**2 we have:

$$\left\|F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\| \leq \frac{M}{P}\sum_{p=1}^P \left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2. \tag{S29}$$

Using the proof of Equation (S15), and combining Equations (S27) to (S29) and Lemma S23, we have, for any $t \in [C], k \in [N^t]$:

$$\begin{aligned}\triangle &:= \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 |\mathcal{H}_t^k\right] \\ \triangle &\leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 - 2\eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\rangle + (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 |\mathcal{H}_{k,t}\right] \\ &\quad + 2\eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle \\ \mathbb{E}[\triangle] &\leq (1 - \mu\eta_{k+1}^t)\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2\right] + (\eta_{k+1}^t)^2 \frac{1}{P}\sigma_\infty^2 \\ &\quad + 2\eta_{k+1}^t \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|\right] M \sum_{j=1}^k (\eta_j^t)^2 \sigma_\infty^2 \prod_{s=j+1}^k (1 - \eta_s^t \mu). \end{aligned} \tag{S30}$$

Thus by induction, for any $t \in [C], K \in [N^t]$:

$$\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_K^t - \boldsymbol{w}^\star\right\|^2\right] \leq \prod_{k=1}^K (1 - \mu\eta_k^t)\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star\right\|^2\right] + \frac{1}{P}\sigma_\infty^2 \sum_{k=1}^K (\eta_k^t)^2 \prod_{j=k+1}^K (1 - \mu\eta_j^t)$$

$$
\begin{aligned}
&+ 2\sigma_\infty^2 M \sum_{k=1}^{K} \eta_k^t \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star\right\|\right] \sum_{j=1}^{k} (\eta_j^t)^2 \prod_{s=j+1}^{k} (1-\eta_s^t\mu) \prod_{j=k+1}^{K} (1-\mu\eta_j^t) \\
&= \prod_{k=1}^{K} (1-\mu\eta_k^t) \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star\right\|^2\right] + \frac{1}{P}\sigma_\infty^2 \sum_{k=1}^{K} (\eta_k^t)^2 \prod_{j=k+1}^{K} (1-\mu\eta_j^t) \\
&+ 2M\sigma_\infty^2 \sum_{j=1}^{K} (\eta_j^t)^2 \prod_{s=j+1}^{K} (1-\mu\eta_j^t) \sum_{k=j}^{K} \eta_k^t \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star\right\|\right] \\
&= \prod_{k=1}^{K} (1-\mu\eta_k^t) \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_0^t - \boldsymbol{w}^\star\right\|^2\right] + C_{P,M,K,t} \frac{\sigma_\infty^2}{P} \sum_{k=1}^{K} (\eta_k^t)^2 \prod_{j=k+1}^{K} (1-\mu\eta_j^t),
\end{aligned}
$$

with $C_{P,M,K,t} = 1 + MP\sum_{k=1}^{K} \eta_k^t \mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star\right\|\right]$. This concludes the proof.
In the following section, we proved the auxiliary Lemma that was used in the proof.

D.4.4 Proof of Lemma S23

We now study $\frac{1}{P}\sum_{p=1}^{P} \left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2$ as k increases. Note that initially ($k=0$), this quantity is 0. For any $k \in [N^t], p \in [P]$:

$$
\begin{aligned}
\left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2 &= \left\|\boldsymbol{w}_{p,k-1}^t - \eta_k^t g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - \breve{\boldsymbol{w}}_{k-1}^t + \eta_k^t \frac{1}{P}\sum_{i=1}^{P} g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2 \\
&= \left\|\boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t\right\|^2 - 2\eta_k^t \left\langle \boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t, g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\rangle \\
&+ (\eta_k^t)^2 \left\|g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2.
\end{aligned}
$$

Thus, expanding and using cocoercivity Assumption:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2 |\mathcal{H}_{k-1}^t\right] &= \left\|\boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t\right\|^2 \\
&- 2\eta_k^t \left\langle \boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t, F'(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} F'(\boldsymbol{w}_{i,k-1}^t)\right\rangle \\
&+ \mathbb{E}\left[(\eta_k^t)^2 \left\|g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2 |\mathcal{H}_{k-1}^t\right] \\
&= \left\|\boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t\right\|^2 - 2\eta_k^t \left\langle \boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t, F'(\boldsymbol{w}_{p,k-1}^t) - F'(\breve{\boldsymbol{w}}_{k-1}^t)\right\rangle \\
&+ 2\eta_k^t \left\langle \boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t, F'(\breve{\boldsymbol{w}}_{k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} F'(\boldsymbol{w}_{i,k-1}^t)\right\rangle \\
&+ \mathbb{E}\left[(\eta_k^t)^2 \left\|g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2 |\mathcal{H}_{k-1}^t\right] \\
&\leq (1 - 2\eta_k^t\mu(1-\eta_k^t L)) \left\|\boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t\right\|^2 \\
&+ 2\eta_k^t \left\langle \boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t, F'(\breve{\boldsymbol{w}}_{k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} F'(\boldsymbol{w}_{i,k-1}^t)\right\rangle \\
&+ \mathbb{E}\left[(\eta_k^t)^2 \left\|(g_{p,k}^t - F')(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} (g_{i,k}^t - F')(\boldsymbol{w}_{i,k-1}^t)\right\|^2 |\mathcal{H}_{k-1}^t\right].
\end{aligned}
$$

Summing over $p \in [P]$:

$$\sum_{p=1}^{P} \mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2 |\mathcal{H}_{k-1}^t\right] \leq (1-\eta_k^t\mu)\sum_{p=1}^{P}\left\|\boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t\right\|^2$$
$$+ 2\eta_k^t \Bigg\langle \underbrace{\sum_{p=1}^{P}(\boldsymbol{w}_{p,k-1}^t - \breve{\boldsymbol{w}}_{k-1}^t)}_{=0}, F'(\breve{\boldsymbol{w}}_{k-1}^t) - \frac{1}{P}\sum_{i=1}^{P} F'(\boldsymbol{w}_{i,k-1}^t)\Bigg\rangle$$
$$+ \sum_{p=1}^{P}\mathbb{E}\left[(\eta_k^t)^2\left\|(g_{p,k}^t - F')(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P}(g_{i,k}^t - F')(\boldsymbol{w}_{i,k-1}^t)\right\|^2 |\mathcal{H}_{k-1}^t\right].$$

If we denote $\delta_k^t = \frac{1}{P}\sum_{p=1}^{P}\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \breve{\boldsymbol{w}}_k^t\right\|^2\right]$, we thus have $\delta_0 = 0$ and

$$\delta_k^t \leq (1-\eta_k^t\mu)\delta_{k-1}^t + \frac{1}{P}\sum_{p=1}^{P}\mathbb{E}\left[(\eta_k^t)^2\left\|(g_{p,k}^t - F')(\boldsymbol{w}_{p,k-1}^t) - \frac{1}{P}\sum_{i=1}^{P}(g_{i,k}^t - F')(\boldsymbol{w}_{i,k-1}^t)\right\|^2 |\mathcal{H}_{k-1}^t\right].$$
$$\leq \frac{1}{P}\sum_{p=1}^{P}\sum_{j=1}^{k}\mathbb{E}\left[(\eta_j^t)^2\left\|(g_{p,j}^t - F')(\boldsymbol{w}_{p,j-1}^t) - \frac{1}{P}\sum_{i=1}^{P}(g_{i,j}^t - F')(\boldsymbol{w}_{i,j-1}^t)\right\|^2\right]\prod_{s=j+1}^{k}(1-\eta_s^t\mu)$$
$$\leq \sum_{j=1}^{k}\mathbb{E}\left[(\eta_j^t)^2\left\|(g_{1,j}^t - F')(\boldsymbol{w}_{1,j-1}^t) - \frac{1}{P}\sum_{i=1}^{P}(g_{i,j}^t - F')(\boldsymbol{w}_{i,j-1}^t)\right\|^2\right]\prod_{s=j+1}^{k}(1-\eta_s^t\mu)$$
$$\leq \sum_{j=1}^{k}\mathbb{E}\left[(\eta_j^t)^2\left\|(g_{1,j}^t - F')(\boldsymbol{w}_{1,j-1}^t)\right\|^2\right]\prod_{s=j+1}^{k}(1-\eta_s^t\mu).$$

Note that everything is tight until the last line for $P = 1$ (then for all k, $\delta_k^t = 0$). Under Assumption **A**4, we thus have:

$$\delta_k^t \leq \sum_{j=1}^{k}(\eta_j^t)^2\sigma_\infty^2\prod_{s=j+1}^{k}(1-\eta_s^t\mu).$$

This concludes the proof.

E Convergence guaranties for the fourth order moment.

In this section, we prove several Lemmas that allow to control the fourth order moment of the iterate. While controlling the second order moment is sufficient for quadratic functions as no "residual" term appears in Equation (3) (the "residual" corresponds to the rest of a linear expansion of the gradient, which is thus exact for a quadratic function), in the general case, we also need to control the 4th order moment.

We first give guarantees for the inner iterates (within a phase) in Appendix E.1, then in the local SGD framework in Appendix E.2.

E.1 Inner Iteration Lemmas

Here, we can use the following Lemma from [25], that gives a recursion for the 4th order moment.

Lemma S24 *Under the Assumptions* ***A****1,* ***A****2,* ***A****3,* ***A****5 for th* 4^{th}*-order moment, assuming* $\eta \leq \frac{1}{18L}$ *we have,*

$$\mathbb{E}\left[(\left\|\boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\|)^4\right]^{1/2} \leq (1-\eta\mu)\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\|^4\right]^{1/2} + 20\eta^2\sigma^2$$

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq(1-\eta\mu)^k\mathbb{E}\left[\left\|\boldsymbol{w}_{i,0}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}+\frac{20\eta\sigma^2}{\mu}.$$

In the mini-batch setting, we have of course the same result with a variance reduction:

Lemma S25 *Under the Assumptions* ***A1****,* ***A2****,* ***A3****,* ***A5*** *for th 4^{th}-order moment for mini-batch averaging we have, assuming $\eta\leq\frac{1}{18L}$ we have,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq\big(1-\eta\mu\big)\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^4\right]^{1/2}+\frac{20\eta^2}{P}\sigma^2$$
$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq\big(1-\eta\mu\big)^t\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\eta}{P\mu}\sigma^2.$$

Analogous to Lemma S24 we have the following result for fourth order moments,

Lemma S26 *Under the Assumptions* ***A1****,* ***A2****,* ***A3****,* ***A5*** *for th 4^{th}-order moment, assuming $\eta\leq\frac{1}{18L}$ we have,*

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq\big(1-\eta_k^t\mu\big)\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}+20\eta^2\sigma^2$$
$$\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq\prod_{j=1}^k(1-\eta_j^t\mu)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+20\sigma^2\sum_{j=1}^k\prod_{l=j+1}^k(1-\mu\eta_l^t)(\eta_j^t)^2.$$

Similarly for mini-batch analogous to Lemma S25,

Lemma S27 *Under the Assumptions* ***A1****,* ***A2****,* ***A3****,* ***A5*** *for th 4^{th}-order moment for mini-batch averaging and decreasing step size we have, assuming $\eta\leq\frac{1}{18L}$ we have,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq\big(1-\eta^t\mu\big)\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star\right\|^4\right]^{1/2}+\frac{20\eta^2}{P}\sigma^2$$
$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\leq\prod_{j=1}^t\left(1-\eta^j\mu\right)\left\|\hat{\boldsymbol{w}}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\sigma^2}{P}\sum_{j=1}^t\prod_{l=j+1}^t(1-\mu\eta^l)(\eta^j)^2.$$

The proof is included for completeness and because the same proof technique is used afterwards in Appendix E.2.

Proof 28 *For $i\in[P]$, $k\in[N_t]$ and $t\in[C]$ we define the notation $\phi_{i,k}^t=\left\|\boldsymbol{w}_{i,k}^t-\boldsymbol{w}^\star\right\|$. We have that,*

$$\begin{aligned}(\phi_{i,k}^t)^4&=\left(\left\|\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\right\|^2-2\eta\langle\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t),\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\rangle+\eta^2\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2\right)^2\\&=\left((\phi_{i,k-1}^t)^2-2\eta\langle\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t),\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\rangle+\eta^2\left\|g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2\right)^2\\&=(\phi_{i,k-1}^t)^4-4\eta(\phi_{i,k-1}^t)^2\langle\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t),\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\rangle\\&+4\eta^2\langle g_{i,k}^t(\boldsymbol{w}_{i,k-1}^t),\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\rangle^2+2\eta^2(\phi_{i,k-1}^t)^2\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2\\&-4\eta^3\langle\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t),\boldsymbol{w}_{i,k-1}^t-\boldsymbol{w}^\star\rangle\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2+\eta^4\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^4.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^p|\mathbb{H}_{k-1}^t\right]&\leq2^{p-1}\big(\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)-\boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^p|\mathbb{H}_{k-1}^t\right]+\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^p|\mathbb{H}_{k-1}^t\right]\big)\\&\leq2^{p-1}\big(\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)-\boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^p\right]+\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^p|\mathbb{H}_{k-1}^t\right]\big)\\&\leq2^{p-1}\big(\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)-\boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^p+\sigma^p\big),\end{aligned}\tag{S31}$$

Where we have used at the first line Minkowski's inequality and the fact that $x\mapsto x^p$ is convex on $\mathbb{R}^+$ for $p=1,\ldots,4$ thus $(x+y)^p\leq2^{p-1}(x^p+y^p)$, and at the last line the Assumption ***A5*** *on the noise : $\mathbb{E}\left[\left\|f_{i,k}^t(\boldsymbol{w}^\star)\right\|^p|\mathbb{H}_{k-1}^t\right]\leq\sigma^p$.*

This leads to

$$
\begin{aligned}
\blacktriangle &:= \mathbb{E}\left[(\phi_{i,k}^t)^4|\mathbb{H}_{k-1}^t\right] \\
&\leq (\phi_{i,k-1}^t)^4 - 4\eta(\phi_{i,k-1}^t)^2\mathbb{E}\left[\left\langle \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle|\mathbb{H}_{k-1}^t\right] \\
&+ 4\eta^2\mathbb{E}\left[\left\langle \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle^2|\mathbb{H}_{k-1}^t\right] + 2\eta^2(\phi_{i,k-1}^t)^2\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2|\mathbb{H}_{k-1}^t\right] \\
&- 4\eta^3\mathbb{E}\left[\left\langle \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2|\mathbb{H}_{k-1}^t\right] + \eta^4\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^4|\mathbb{H}_{k-1}^t\right] \\
&\leq (\phi_{i,k-1}^t)^4 - 4\eta(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle + 4\eta^2\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2(\phi_{i,k-1}^t)^2|\mathbb{H}_{k-1}^t\right] \\
&+ 2\eta^2(\phi_{i,k-1}^t)^2\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2|\mathbb{H}_{k-1}^t\right] + 4\eta^3\phi_{i,k-1}^t\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^3|\mathbb{H}_{k-1}^t\right] \\
&+ \eta^4\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^4|\mathbb{H}_{k-1}^t\right] \\
&\leq (\phi_{i,k-1}^t)^4 - 4\eta(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle + 12\eta^2\sigma^2(\phi_{i,k-1}^t)^2 + 16\eta^3\phi_{i,k-1}^t\sigma^3 + 8\eta^4\sigma^4 \\
&+ 12\eta^2(\phi_{i,k-1}^t)^2\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^2|\mathbb{H}_{k-1}^t\right] \\
&+ 16\eta^3\phi_{i,k-1}^t\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^3|\mathbb{H}_{k-1}^t\right] + 8\eta^4\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^4|\mathbb{H}_{k-1}^t\right].
\end{aligned}
$$

Above we have used Cauchy Schwartz inequality several times for the second inequality and equation (S31) for the third one.

$$
\begin{aligned}
\bigstar &:= \mathbb{E}\left[(\phi_{i,k}^t)^4|\mathbb{H}_{k-1}^t\right] \\
&\leq (\phi_{i,k-1}^t)^4 - 4\eta(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle + 12\eta^2L(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle \\
&+ 16\eta^3L^2(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle + 8\eta^4L^3(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle \\
&+ 12\eta\sigma^2(\phi_{i,k-1}^t)^2 + 8\eta^2\sigma^2(\phi_{i,k-1}^t)^2 + 8\eta^4\sigma^4 + 8\eta^4\sigma^4 \\
&= (\phi_{i,k-1}^t)^4 + (-4\eta + 12\eta^2L + 16\eta^3L^2 + 8\eta^4L^3)(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle \\
&+ (12\eta^2\sigma^2 + 8\eta^2\sigma^2)(\phi_{i,k-1}^t)^2 + 16\eta^4\sigma^4 \\
&\leq (\phi_{i,k-1}^t)^4 - 4\eta(1-9\eta L)(\phi_{i,k-1}^t)^2\left\langle F'(\boldsymbol{w}_{i,k-1}^t), \boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\rangle + 20\eta^2\sigma^2(\phi_{i,k-1}^t)^2 + 16\eta^4\sigma^4.
\end{aligned}
$$

Above we used $\eta L \leq 1$ in the last line. Finally, using strong convexity, we have:

$$
\mathbb{E}\left[(\phi_{i,k}^t)^4|\mathbb{H}_{k-1}^t\right] \leq \big(1 - 4\eta\mu(1-9\eta L)\big)(\phi_{i,k-1}^t)^4 + 20\eta^2\sigma^2(\phi_{i,k-1}^t)^2 + 16\eta^4\sigma^4,
$$

Now $\mathbb{E}\left[(\phi_{i,k-1}^t)^2\right] \leq \mathbb{E}\left[(\phi_{i,k-1}^t)^4\right]^{1/2}$ by Jensen's inequality. Also since we assume $\eta \leq \frac{1}{9L}$ and $\frac{\mu}{L} \leq 1$ we can obtain $(1-4\eta\mu(1-9\eta L))^{1/2} \geq (1-4\eta\mu)^{1/2} \geq (1-\frac{4\mu}{9L})^{1/2} \geq (1-\frac{4}{9})^{1/2} \geq 1/2$. This finally leads to $20\eta^2\sigma^2\mathbb{E}\left[(\phi_{i,k-1}^t)^2\right] \leq (1-4\eta\mu(1-9\eta L))^{1/2}\mathbb{E}\left[(\phi_{i,k-1}^t)^4\right]^{1/2} 40\eta^2\sigma^2$, which can be used below to obtain

$$
\begin{aligned}
\mathbb{E}\left[(\phi_{i,k}^t)^4|\mathbb{H}_{k-1}^t\right] &\leq \big(1 - 4\eta\mu(1-9\eta L)\big)\mathbb{E}\left[(\phi_{i,k-1}^t)^4\right] + 20\eta^2\sigma^2\mathbb{E}\left[(\phi_{i,k-1}^t)^2\right] + 16\eta^4\sigma^4 \\
&\leq \left(\big(1 - 4\eta\mu(1-9\eta L)\big)^{1/2}\mathbb{E}\left[(\phi_{i,k-1}^t)^4\right]^{1/2} + 20\eta^2\sigma^2\right)^2 \\
\mathbb{E}\left[(\phi_{i,k}^t)^4\right]^{1/2} &\leq \big(1 - 2\eta\mu(1-9\eta L)\big)\mathbb{E}\left[(\phi_{i,k-1}^t)^4\right]^{1/2} + 20\eta^2\sigma^2.
\end{aligned}
$$

This Concludes the proof.

E.2 Proof of Lemma S29

In this section, we prove the following Lemma, which is necessary to conclude the proof for the second set of Assumptions in Theorem 6. Indeed, we need to control the moment of order 4 to be able to control the residual term that arises from linear expansion of the gradient around $\boldsymbol{w}^\star$.

Lemma S29 *There exist absolute constants C_4, D_4, E_4, such that if $\eta_k^t L \leq \frac{1}{C_4}$:*

$$
\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^4\right]^{1/2} \leq (1-\eta_k^t\mu)\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^4\right]^{1/2} + D_4(\eta_k^t)^2\frac{\sigma_\infty^2}{P}
$$

$$+ E_4 \eta_{k+1}^t \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\| \left\| F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\|. \tag{S32}$$

In other words, $\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^4\right]^{1/2}$ satisfies the same recursion as $\mathbb{E}\left[\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2\right]$, as this equation is the same as Equation (S30) (up to absolute constants).

Proof 30 *This proof combines element from the classical bound for the fourth order moment, and from the proof of Lemma S22, which addresses the similar setting but only for the second order moment. We start from the definition of* $\breve{\boldsymbol{w}}_{k+1}^t$*:*

$$\begin{aligned}\left\|\breve{\boldsymbol{w}}_{k+1}^t - \boldsymbol{w}^\star\right\|^2 \leq \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^2 &- 2\eta_{k+1}^t \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t)\big\rangle \\ &+ (\eta_{k+1}^t)^2 \left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 \\ &+ 2\eta_{k+1}^t \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\big\rangle. \end{aligned} \tag{S33}$$

Thus, squaring this equation we get, denoting $\breve{\phi}_k^t = \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|$*:*

$$\begin{aligned}(\breve{\phi}_{k+1}^t)^4 \leq (\breve{\phi}_k^t)^4 &- 4(\breve{\phi}_k^t)^2 \eta_{k+1}^t \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t)\big\rangle \\ &+ 2(\breve{\phi}_k^t)^2 (\eta_{k+1}^t)^2 \left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^2 \\ &+ 4(\breve{\phi}_k^t)^2 \eta_{k+1}^t \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\big\rangle \\ &+ 3(\eta_{k+1}^t)^2 \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t)\big\rangle^2 \\ &+ 3(\eta_{k+1}^t)^4 \left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t)\right\|^4 \\ &+ 3(2\eta_{k+1}^t)^2 \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\big\rangle^2, \end{aligned}$$

formally, we have used $(a+b+c+d)^2 \leq a^2 + 2ab + 2ac + 2ad + 3b^2 + 3c^2 + 3d^2$.

That is, conditioning on the past, and using Assumption A5 (cocoercivity and the fact that g_k^t *is a.s. L-Lipshitz):*

$$\begin{aligned}\mathbb{E}\left[(\breve{\phi}_{k+1}^t)^4 | \mathcal{H}_k^t\right] \leq (\breve{\phi}_k^t)^4 &- 4(\breve{\phi}_k^t)^2 \eta_{k+1}^t (1 - \eta_k^t L) \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\big\rangle \\ &+ 2(\breve{\phi}_k^t)^2 (\eta_{k+1}^t)^2 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t)\right\|^2 | \mathcal{H}_k^t\right] \\ &+ 4(\breve{\phi}_k^t)^2 \eta_{k+1}^t \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\big\rangle \\ &+ 3(\eta_{k+1}^t)^2 \big\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P F'(\breve{\boldsymbol{w}}_k^t)\big\rangle L(\breve{\phi}_k^t)^2 \end{aligned}$$

$$
\begin{aligned}
&+ 6(\eta_{k+1}^t)^4 \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\boldsymbol{w}_{i,k}^t) - F'(\boldsymbol{w}_{i,k}^t)\right\|^4 |\mathcal{H}_k^t\right] \\
&+ 6(\eta_{k+1}^t)^4 L^2 (\breve{\phi}_k^t)^2 \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\rangle \\
&+ 3(2\eta_{k+1}^t)^2 \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle^2.
\end{aligned}
$$

*Rearranging terms and using the uniform upper bound on the 4-th moment of the noise **A**6, we have:*

$$
\begin{aligned}
\mathbb{E}\left[(\breve{\phi}_{k+1}^t)^4 | \mathcal{H}_k^t\right] &\leq (\breve{\phi}_k^t)^4 - 4(\breve{\phi}_k^t)^2 \eta_{k+1}^t (1 - \eta_k^t L - 3\eta_k^t L - 6(\eta_{k+1}^t)^4 L^2) \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t)\rangle \\
&\quad + 2(\breve{\phi}_k^t)^2 (\eta_{k+1}^t)^2 \frac{\sigma_\infty^2}{P} + 6(\eta_{k+1}^t)^4 \frac{\sigma_\infty^4}{P^2} \\
&\quad + 4(\breve{\phi}_k^t)^2 \eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle \\
&\quad + 3(2\eta_{k+1}^t)^2 \mathbb{E}\left[\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle^2 | \mathcal{H}_k^t\right].
\end{aligned} \tag{S34}
$$

The first 2 lines of Equation (S34) correspond to the expansion in Equation (S32) (the constants are slightly different because we use a uniform bound on the gradient instead of co-coercivity). The last two lines correspond to the residual term, for which we will use Lemma S23.

We have:

$$
\begin{aligned}
&4(\breve{\phi}_k^t)^2 \eta_{k+1}^t \langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle \\
&+ 6(2\eta_{k+1}^t)^2 \mathbb{E}\left[\langle \breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star, \frac{1}{P}\sum_{i=1}^P g_{i,k+1}^t(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\rangle^2 | \mathcal{H}_k^t\right] \\
&\qquad \leq 4(\breve{\phi}_k^t)^3 \eta_{k+1}^t \left\|F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\| \\
&\qquad + 6(2\eta_{k+1}^t)^2 L \left\|\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right\|^3 \left\|\frac{1}{P}\sum_{i=1}^P F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\| \\
&\qquad = (\breve{\phi}_k^t)^3 \eta_{k+1}^t (4 + 24\eta_k^t L) \left\|F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\|.
\end{aligned}
$$

As a result, there exist absolute constants ("numbers") C_4, D_4, E_4, such that if $\eta_k^t L \leq \frac{1}{C_4}$:

$$
\begin{aligned}
\mathbb{E}\left[(\breve{\phi}_{k+1}^t)^4\right]^{1/2} &\leq (1 - \eta_k^t \mu)\mathbb{E}\left[(\breve{\phi}_k^t)^4\right]^{1/2} + D_4 (\eta_k^t)^2 \frac{\sigma_\infty^2}{P} \\
&\quad + E_4 \eta_{k+1}^t \mathbb{E}\left[(\breve{\phi}_k^t)\left\|F'(\breve{\boldsymbol{w}}_k^t) - \frac{1}{P}\sum_{p=1}^P F'(\boldsymbol{w}_{p,k}^t)\right\|\right].
\end{aligned} \tag{S35}
$$

This is the result of the Lemma.

F Main error decomposition

F.1 General decomposition

In this section, we prove the following decomposition for the on-line setting.

Lemma S31 *Under the differentiability of **A**2 we have*[7]*,*

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{P\left(\boldsymbol{w}^0 - \boldsymbol{w}^\star\right)}{T\eta_1^1} - \frac{P\left(\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star\right)}{T\eta_{N^C+1}^C} - \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^t} - \frac{1}{\eta_{k+1}^t}\right)$$
$$+ \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\xi_{i,k}^t,$$

where $\xi_{i,k}^t = F'(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)$ *and* $\delta_{i,k}^t = F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star) - F'(\boldsymbol{w}_{i,k-1}^t)$*.*

Proof 32 *Below, we have* $\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)$ *as the stochastic gradient at step* k *on machine* i *for communication phase* t*. After adding and subtracting few quantities and rearranging we have,*

$$\boldsymbol{w}_{i,k}^t = \boldsymbol{w}_{i,k-1}^t - \eta_k^t \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)$$
$$\boldsymbol{w}_{i,k}^t = \boldsymbol{w}_{i,k-1}^t - \eta_k^t F'(\boldsymbol{w}_{i,k-1}^t) + \eta_k^t\left(F'(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right)$$
$$\boldsymbol{w}_{i,k}^t = \boldsymbol{w}_{i,k-1}^t - \eta_k^t F'(\boldsymbol{w}_{i,k-1}^t) + \eta_k^t\delta_{i,k}^t + \eta F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star) - \eta_k^t F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star)$$
$$\boldsymbol{w}_{i,k}^t = \boldsymbol{w}_{i,k-1}^t + \eta_k^t\xi_{i,k}^t + \eta_k^t\delta_{i,k}^t - \eta_k^t F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star).$$

where $\xi_{i,k}^t$ *and* $\delta_{i,k}^t$ *are respectively terms related to stochastic noise and quadratic residual. Obtaining the horizontal average over all the machines and recalling the definition of the ghost process* $\breve{\boldsymbol{w}}_k^t$ *as defined above we have,*

$$\frac{1}{P}\sum_{i=1}^{P}F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star) = \frac{1}{P}\sum_{i=1}^{P}\frac{1}{\eta_k^t}\left(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}_{i,k}^t\right) + \frac{1}{P}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{P}\sum_{i=1}^{P}\xi_{i,k}^t$$
$$F''(\boldsymbol{w}^\star)(\breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star) = \frac{\breve{\boldsymbol{w}}_{k-1}^t - \breve{\boldsymbol{w}}_k^t}{\eta_k^t} + \frac{1}{P}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{P}\sum_{i=1}^{P}\xi_{i,k}^t.$$

Obtaining the vertical average over all the machines first within a communication phase and then among different phases we have,

$$\frac{1}{N^t}\sum_{k=1}^{N^t}F''(\boldsymbol{w}^\star)(\breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star) = \frac{1}{N^t}\sum_{k=1}^{N^t}\frac{\breve{\boldsymbol{w}}_{k-1}^t - \breve{\boldsymbol{w}}_k^t}{\eta_k^t} + \frac{1}{N^t P}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{N^t P}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\xi_{i,k}^t$$
$$\frac{1}{\sum_{t=1}^{C}N^t}\sum_{t=1}^{C}\sum_{k=1}^{N^t}F''(\boldsymbol{w}^\star)(\breve{\boldsymbol{w}}_{k-1}^t - \boldsymbol{w}^\star) = \frac{1}{\sum_{t=1}^{C}N^t}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\frac{\breve{\boldsymbol{w}}_{k-1}^t - \breve{\boldsymbol{w}}_k^t}{\eta_k^t} + \frac{1}{P\sum_{t=1}^{C}N^t}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\delta_{i,k}^t$$
$$+ \frac{1}{P\sum_{t=1}^{C}N^t}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\xi_{i,k}^t.$$

Now recalling the definitions for the overall iterate $\overline{\overline{\boldsymbol{w}}}^C = \frac{1}{\sum_{t=1}^{C}N^t}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\breve{\boldsymbol{w}}_k^t$*,* $\hat{\boldsymbol{w}}^t = \breve{\boldsymbol{w}}_{N^t}^t$*, the initial point* $\hat{\boldsymbol{w}}^0 = \boldsymbol{w}^0$*, and the total number of gradients* $T = P\sum_{t=1}^{C}N^t$ *as we have defined above. After making these changes and on rearranging we obtain,*

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{P}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\frac{\breve{\boldsymbol{w}}_{k-1}^t - \breve{\boldsymbol{w}}_k^t}{\eta_k^t} + \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\xi_{i,k}^t$$
$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{P\left(\boldsymbol{w}^0 - \boldsymbol{w}^\star\right)}{T\eta_1^1} - \frac{P\left(\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star\right)}{T\eta_{N^C+1}^C} - \frac{P}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\left(\breve{\boldsymbol{w}}_k^t - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^t} - \frac{1}{\eta_{k+1}^t}\right)$$
$$+ \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\xi_{i,k}^t.$$

[7] Note that after the final iteration of the phase the learning rate (which the algorithm uses nowhere) corresponds to the first learning rate for the next phase. This anomaly in notation is a direct result of us considering the ghost process, which runs continuously till the end.

Thus we have obtained the required result as,

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{P\left(\boldsymbol{w}^0 - \boldsymbol{w}^\star\right)}{T\eta_1^1} - \frac{P\left(\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star\right)}{T\eta_{N^C+1}^C} - \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^t} - \frac{1}{\eta_{k+1}^t}\right)$$
$$+ \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\delta_{i,k}^t + \frac{1}{T}\sum_{t=1}^{C}\sum_{k=1}^{N^t}\sum_{i=1}^{P}\xi_{i,k}^t.$$

F.2 Bounding the noise term

The stochastic noise term which appears above can be bounded using the following lemma,

Lemma S33 *Under the Assumptions A3, A5, A6 we have*

$$\mathbb{E}\left[\left\|\xi_{i,k}^t\right\|^2\right] \leq 2L^2\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^t\boldsymbol{w}^\star\right\|^2\right] + 2\sigma^2.$$

Proof 34 *Using Assumptions A3, A5, A6 respectively we prove the result*

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_{i,k}^t\right\|^2\right] &= \mathbb{E}\left[\left\|F'(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2\right] \leq \mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t)\right\|^2\right] - \left\|F'(\boldsymbol{w}_{i,k-1}^t)\right\|^2 \\ &\leq 2\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}_{i,k-1}^t) - \boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^2\right] + 2\mathbb{E}\left[\left\|\boldsymbol{g}_{i,k}^t(\boldsymbol{w}^\star)\right\|^2\right] \\ &\leq 2L^2\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star\right\|^2\right] + 2\sigma^2.\end{aligned}$$

G Proofs for OSA, MBA and Local-SGD in the finite horizon setting

In this Section and Appendix H we prove convergence results for $\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\right\|\right]$. The proof technique is the one proposed by Polyak and Judisky in the original article on averaging [1]. This proof technique has also been used in [10, 15]. We notice here the following differences, that justify including the proofs:

1. Polyak and Judisky were mainly interested in the asymptotic analysis, and the set of assumptions considered was different.
2. In [10], the authors prove comparable bounds in the case of bounded gradients. However, their analysis in the smooth and strongly convex setting is not optimal. Precisely, they use a sub-optimal upper bound when controlling the second order moments, that significantly worsens the subsequent proof. This point was underlined in [25, 47]. The result they provide under our set of assumptions is eventually 1) not optimal, 2) uselessly complex, and 3) only for serial-SGD.
3. In [15], authors prove a result close to us, using a similar approach for one-shot averaging. Their bounds only apply to decaying step size. Moreover, they rely on the following asymptotic upper bound: $\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k}^t - \boldsymbol{w}^\star\right\|^2\right] \leq C_1\eta_k^t$: this bound is correct but the constant C_1 is "asymptotic" (see for e.g., [34]). On contrary, we use non-asymptotic upper bounds on the second order moment involved. As a consequence, our bounds are both simpler and tighter.

G.1 Technical Lemmas

Lemma S35 (Jensen's Inequality) *For $a_i \in \mathbb{R}^d$, $\left\|\frac{1}{P}\sum_{i=1}^P a_i\right\|^2 \leq \frac{1}{P}\sum_{i=1}^P \|a_i\|^2$.*

Proof 36 *The result is an application of Jensen's inequality with the convex function $f(.) = \|.\|^2$.*

Lemma S37 (Minkowski's Inequality) *For* $a_i \in \mathbb{R}^d$, $\mathbb{E}\left[\left\|\sum_{i=1}^{P} a_i\right\|^2\right] \leq \left(\sum_{i=1}^{P} \mathbb{E}\left[\|a_i\|^2\right]^{\frac{1}{2}}\right)^2$

Proof 38 *The inequality is an application of Minkowski's inequality (or simply triangle's inequality) with the norm* $\|.\|_E = \mathbb{E}\left[\|.\|^2\right]^{\frac{1}{2}}$.

G.2 Proof of Proposition 1 (Mini-batch case)

Lemma S8 proves the first part of the proposition. We prove the second part of the proposition here following the approach by [1]. Using Lemma S31, Lemma S24 we can obtain an upper bound on $\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\right\|^2\right]$, which is in-fact a tighter quantity when compared to $\mathbb{E}\left[\left\|\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star\right\|^2\right]$. We prove the following lemma,

Lemma S39 *Under the Assumptions* ***A1****,* ***A2****,* ***A3****,* ***A5****,* ***A6*** *we have,*

$$\mathbb{E}\left[\left\|\nabla^2 F(\boldsymbol{w}^\star)(\boldsymbol{w} - \boldsymbol{w}^\star)\right\|^2\right] \leq 4\sum_{i=1}^{5} A_{i,P,C}^2,$$

where the terms are respectively,

$$A_{1,P,C}^2 = \frac{P^2}{T^2\eta^2}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2, A_{2,P,C}^2 = \frac{P^2}{T^2\eta^2}\left((1-\mu\eta)^C\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{\eta}{\mu P}\right),$$
$$A_{3,P,C}^2 = \frac{P^2M^2}{T^2\mu^2\eta^2}\left(\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{C20\eta^2}{P}\sigma^2\right)^2, A_{4,P,C}^2 = \frac{2\sigma^2}{T},$$
$$A_{3,P,C}^2 = \frac{2L^2P}{T^2}\left(\frac{1}{\mu\eta}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{\left(C\mu\eta - 1 + (1-\mu\eta)^C\right)}{\mu^2 P}\right).$$

Proof 40 *In order to upper bound the expectation we need to separately upper bound all the terms that appear in the result for Lemma S31. But before that we can actually simplify the result with constant step size and using* $N^t = 1\ \forall t \in [C]$ *as follows,*

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{\boldsymbol{w}^0 - \boldsymbol{w}^\star}{C\eta} - \frac{\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star}{C\eta} + \frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^t + \frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^t.$$

Now we bound each of the terms in the above decomposition one by one. For the first term,

$$\mathbb{E}\left[\left\|\frac{1}{C\eta}\left(\boldsymbol{w}^0 - \boldsymbol{w}^\star\right)\right\|^2\right] = \frac{P^2}{T^2\eta^2}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 = A_{1,P,C}^2.$$

For the second term using Lemma S8,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{C\eta}\left(\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star\right)\right\|^2\right] &= \frac{P^2}{T^2\eta^2}\mathbb{E}\left[\left\|\boldsymbol{w}_{MB}^C - \boldsymbol{w}^\star\right\|^2\right] \\
&\leq \frac{P^2}{T^2\eta^2}\left(\prod_{k=1}^{C}(1-\mu\eta)\mathbb{E}\left[\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2\right] + 2\sigma^2\frac{1}{P}\sum_{k=1}^{C}\prod_{l=k+1}^{C}(1-\mu\eta)\eta^2\right) \\
&\leq \frac{P^2}{T^2\eta^2}\left((1-\mu\eta)^C\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{1}{P}\left(\frac{1-(1-\mu\eta)^C}{\mu\eta}\right)\eta^2\right) \\
&\leq \frac{P^2}{T^2\eta^2}\left((1-\mu\eta)^C\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{\eta}{\mu P}\right) = A_{2,P,C}^2.
\end{aligned}$$

For the third term using Lemma S35 and Lemma S37 we get,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^{t}\right\|^2\right] = \frac{1}{T^2}\mathbb{E}\left[\left\|\sum_{t=1}^{C}\sum_{i=1}^{P}\left(F'(\boldsymbol{w}_{i,0}^{t}) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,0}^{t} - \boldsymbol{w}^\star)\right)\right\|^2\right]$$
$$\leq \frac{P}{T^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\sum_{t=1}^{C}\left(F'(\hat{\boldsymbol{w}}^{t-1}) - F''(\boldsymbol{w}^\star)(\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star)\right)\right\|^2\right]$$
$$\leq \frac{P^2}{T^2}\left(\sum_{t=1}^{C}\sqrt{\mathbb{E}\left[\left\|\left(F'(\hat{\boldsymbol{w}}^{t-1}) - F''(\boldsymbol{w}^\star)(\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star)\right)\right\|^2\right]}\right)^2.$$

*Now using the upper bound from **A**2 followed by Lemma S25 we get,*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^{t}\right\|^2\right] \leq \frac{P^2M^2}{T^2}\left(\sum_{t=1}^{C}\sqrt{\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^4\right]}\right)^2$$
$$\leq \frac{P^2M^2}{T^2}\left(\sum_{t=1}^{C}\left((1-\eta\mu)^{t-1}\mathbb{E}\left[(\hat{\boldsymbol{w}}^0 - \boldsymbol{w}^\star)^4\right]^{1/2} + \frac{20\eta}{P\mu}\sigma^2\right)\right)^2$$
$$\leq \frac{P^2M^2}{T^2}\left(\frac{1-(1-\eta\mu)^C}{\eta\mu}\mathbb{E}\left[(\hat{\boldsymbol{w}}^0 - \boldsymbol{w}^\star)^4\right]^{1/2} + \frac{20C\eta}{P\mu}\sigma^2\right)^2$$
$$\leq \frac{P^2M^2}{T^2\mu^2\eta^2}\left(\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{20C\eta^2}{P}\sigma^2\right)^2 = A_{3,P,C}^2.$$

*For the fourth term, note that we are sampling i.i.d observations and thus the stochastic noise across all machines and iterations is independent and equal to zero in expectation (see **A**3). This implies the first equation below while the second inequality is obtained using Lemma S33,*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^{t}\right\|^2\right] = \frac{1}{T^2}\sum_{t=1}^{C}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\xi_{i,1}^{t}\right\|^2\right] \leq \frac{1}{T^2}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(2L^2\mathbb{E}\left[\left\|\boldsymbol{w}_{i,0}^{t} - \boldsymbol{w}^\star\right\|^2\right] + 2\sigma^2\right)$$
$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\sum_{t=1}^{C}\mathbb{E}\left[\left\|\boldsymbol{w}_{1,0}^{t} - \boldsymbol{w}^\star\right\|^2\right].$$

Now using Lemma S8 we have,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^{t}\right\|^2\right] \leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\sum_{t=1}^{C}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}_{MB}^{t-1} - \boldsymbol{w}^\star\right\|^2\right]$$
$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\sum_{t=1}^{C}\left((1-\mu\eta)^{t-1}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{\eta\left(1-(1-\mu\eta)^C\right)}{\mu P}\right)$$
$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\left(\frac{1-(1-\mu\eta)^C}{\mu\eta}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{\left(C\mu\eta - (1-(1-\mu\eta)^C)\right)}{\mu^2 P}\right)$$
$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\left(\frac{1}{\mu\eta}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{C\eta}{\mu P}\right)$$
$$= A_{4,P,C}^2 + A_{5,P,C}^2.$$

Now using Lemma S35, we have proved the lemma.

It can be seen in the above lemma that there are two kinds of terms: one that depend on the history or initialization and second the ones that depend on the variance bound. This implies that it would be possible to restate Lemma S39 as follows,

Lemma S41 *Under the assumptions A1, A2, A3, A5, A6 we have,*

$$\mathbb{E}\left[\left\|\nabla^2 F(\boldsymbol{w}^\star)(\boldsymbol{w}-\boldsymbol{w}^\star)\right\|^2\right] \leq 4(\hat{A}^2_{1,P,C}+\hat{A}^2_{2,P,C})$$

Where the terms are respectively,

$$\hat{A}^2_{1,P,C} = \frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{\eta^2 C^2}\left(1+(1-\mu\eta)^C+\frac{2M^2}{\mu^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2L^2\eta}{\mu P}\right),$$
$$\hat{A}^2_{2,P,C} = \frac{2\sigma^2}{T}\left(1+\frac{P}{T\eta\mu}+\frac{400M^2C^2\eta^2\sigma^2}{T\mu^2}+\frac{2L^2C\eta}{T\mu}\right).$$

Ignoring constants the above constants can be upper bounded as follows,

$$\begin{aligned}
\hat{A}^2_{1,P,C} &\leq \frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{\eta^2 C^2}\left(1+1+\frac{2M^2}{\mu^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2L^2\eta}{\mu P}\right)\\
&\leq 2\frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{\eta^2 C^2}\left(1+\frac{M^2}{\mu^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{L^2\eta}{\mu P}\right)\\
&\precsim \frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{\eta^2 C^2}\left(1+\frac{M^2}{\mu^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{L^2\eta}{\mu P}\right),\\
\hat{A}^2_{2,P,C} &\leq 800\frac{\sigma^2}{T}\left(1+\frac{P}{T\eta\mu}+\frac{M^2C^2\eta^2\sigma^2}{T\mu^2}+\frac{L^2C\eta}{T\mu}\right)\\
&\precsim \frac{\sigma^2}{T}\left(1+\frac{P}{T\eta\mu}+\frac{M^2C^2\eta^2\sigma^2}{T\mu^2}+\frac{L^2C\eta}{T\mu}\right).
\end{aligned}$$

Thus, we recover Proposition 1.

G.3 Proof Proposition 2 (One-shot averaging case)

To prove the proposition we need to prove a bound on second moment of the inner iterations followed by a bound on the final average outer iteration. For inner iterations we follow the result from [53] as the process on a single worker is completely independent of any other worker. We have the following lemma,

Lemma S42 *Under the Assumptions A1, A2, A3, A5, A6 for constant step size for one shot averaging we have,*

$$\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\boldsymbol{w}^1_{i,k}-\boldsymbol{w}^\star)\right\|^2\right] \leq 4\sum_{i=1}^{5} B^2_{i,P,N^1}$$

where the terms are respectively,

$$B^2_{1,P,N^1} = \frac{P^2}{T^2\eta^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2, B^2_{2,P,N^1} = \frac{P^2}{T^2\eta^2}\left((1-\mu\eta)^{N^1}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2\eta}{\mu}\right),$$
$$B^2_{3,P,N^1} = \frac{P^2M^2}{T^2\mu\eta}\left(\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+20\eta^2N^1\sigma^2\right)^2, B^2_{4,P,N^1} = \frac{2\sigma^2}{T},$$
$$B^2_{5,P,N^1} = \frac{2L^2P}{T^2}\left(\frac{1}{\mu\eta}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2N^1\eta}{\mu}\right).$$

Proof 43 *We follow the same line of proof as before. We can use the decomposition from Lemma S31 with constant step size and $C=1$, which results in the following simpler decomposition,*

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C-\boldsymbol{w}^\star) = \frac{\boldsymbol{w}^0-\boldsymbol{w}^\star}{N^\eta}-\frac{\hat{\boldsymbol{w}}^1-\boldsymbol{w}^\star}{N^1\eta}+\frac{1}{T}\sum_{k=1}^{N^1}\sum_{i=1}^{P}\delta^1_{i,k}+\frac{1}{T}\sum_{k=1}^{N^1}\sum_{i=1}^{P}\xi^1_{i,k}$$

For the first term,

$$\mathbb{E}\left[\left\|\frac{\boldsymbol{w}^0-\boldsymbol{w}^\star}{N^1\eta}\right\|^2\right] \leq \frac{P^2}{T^2\eta^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2 = B^2_{1,P,N^1}.$$

For the second term using Lemma S10 and rearranging we have,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{\hat{\boldsymbol{w}}^1-\boldsymbol{w}^\star}{N^1\eta}\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{PN^1\eta}\sum_{i=1}^{P}\boldsymbol{w}^1_{i,N^1}-\boldsymbol{w}^\star\right\|^2\right] \leq \frac{P}{T^2\eta^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\boldsymbol{w}^1_{i,N^1}-\boldsymbol{w}^\star\right\|^2\right] \\
&\leq \frac{P^2}{T^2\eta^2}\left(\prod_{l=1}^{N^1}(1-\mu\eta)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2 + 2\sigma^2\sum_{l=1}^{N^1}\prod_{m=l+1}^{N^1}(1-\mu\eta)\eta^2\right) \\
&\leq \frac{P^2}{T^2\eta^2}\left((1-\mu\eta)^{N^1}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{1-(1-\mu\eta)^{N^1}}{\mu\eta}\eta^2\right) \\
&\leq \frac{P^2}{T^2\eta^2}\left((1-\mu\eta)^{N^1}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2\eta}{\mu}\right) = B^2_{2,P,N^1}.
\end{aligned}$$

For the third term using Lemma S35 and Lemma S37 we obtain,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\delta^1_{i,k}\right\|^2\right] &= \frac{1}{T^2}\mathbb{E}\left[\left\|\sum_{i=1}^{P}\sum_{k=1}^{N^1}F'(\boldsymbol{w}^t_{i,k-1})-F''(\boldsymbol{w}^\star)(\boldsymbol{w}^t_{i,k-1}-\boldsymbol{w}^\star)\right\|^2\right] \\
&\leq \frac{P}{T^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\sum_{k=1}^{N^1}F'(\boldsymbol{w}^t_{i,k-1})-F''(\boldsymbol{w}^\star)(\boldsymbol{w}^t_{i,k-1}-\boldsymbol{w}^\star)\right\|^2\right] \\
&\leq \frac{P}{T^2}\sum_{i=1}^{P}\left(\sum_{k=1}^{N^1}\sqrt{\mathbb{E}\left[\left\|F'(\boldsymbol{w}^1_{i,k-1})-F''(\boldsymbol{w}^\star)(\boldsymbol{w}^1_{i,k-1}-\boldsymbol{w}^\star)\right\|^2\right]}\right)^2
\end{aligned}$$

*Now first using the upper bound of **A**2, followed by Lemma S24 and some rearranging we can obtain the following,*

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\delta^1_{i,k}\right\|^2\right] &\leq \frac{PM^2}{T^2}\sum_{i=1}^{P}\left(\sum_{k=1}^{N^1}\mathbb{E}\left[\left\|\boldsymbol{w}^1_{i,k-1}-\boldsymbol{w}^\star\right\|^4\right]^{1/2}\right)^2 \\
&\leq \frac{PM^2}{T^2}\sum_{i=1}^{P}\left(\sum_{k=1}^{N^1}\left((1-\mu\eta)^{k-1}\mathbb{E}\left[\left\|\boldsymbol{w}^1_{i,0}-\boldsymbol{w}^\star\right\|^4\right]^{1/2}+\frac{20\eta\sigma^2}{\mu}\right)\right)^2 \\
&\leq \frac{P^2M^2}{T^2}\left(\sum_{k=1}^{N^1}\left((1-\mu\eta)^{k-1}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\eta\sigma^2}{\mu}\right)\right)^2 \\
&\leq \frac{P^2M^2}{T^2}\left(\frac{1-(1-\mu\eta)^{N^1}}{\mu\eta}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\eta N^1\sigma^2}{\mu}\right)^2 \\
&\leq \frac{P^2M^2}{T^2\mu^2\eta^2}\left(\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+20\eta^2N^1\sigma^2\right)^2 = B^2_{3,P,N^1}.
\end{aligned}$$

*For the fourth term, using the fact that on different machines noise of the gradient is i.i.d. over different iterations and zero in expectation (**A**3) we obtain,*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\xi^1_{i,k}\right\|^2\right] = \frac{1}{T^2}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\mathbb{E}\left[\left\|\xi^1_{i,k}\right\|^2\right].$$

Now using Lemma S33 we have,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\xi_{i,k}^1\right\|^2\right] \leq \frac{1}{T^2}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\left(2L^2\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^1-\boldsymbol{w}^\star\right\|^2\right]+2\sigma^2\right)$$
$$\leq \frac{2\sigma^2}{T}+\frac{2L^2}{T^2}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^1-\boldsymbol{w}^\star\right\|^2\right].$$

Now using Lemma S10 we have,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\xi_{i,k}^1\right\|^2\right] \leq \frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\sum_{k=1}^{N^1}\left(\prod_{l=1}^{k-1}(1-\mu\eta)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+2\sigma^2\sum_{l=1}^{k-1}\prod_{m=l+1}^{k-1}(1-\mu\eta)\eta^2\right)$$
$$\leq \frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\sum_{k=1}^{N^1}\left((1-\mu\eta)^{k-1}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2\eta}{\mu}\right)$$
$$\leq \frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\left(\frac{1}{\mu\eta}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{N^1 2\sigma^2\eta}{\mu}\right)$$
$$= B_{4,P,N^1}^2+B_{5,P,N^1}^2.$$

Finally using Lemma S37, concludes the proof.

Similar to the mini-batch case, there are two kinds of terms one that depend on the history or initialization and second that depend on the variance bound of the functions. This implies that it would be possible to restate Lemma S42 as follows,

Lemma S44 *Under the Assumptions* ***A3****,* ***A2****,* ***A1****,* ***A5****,* ***A6*** *we have,*

$$\mathbb{E}\left[\left\|\nabla^2F(\boldsymbol{w}^\star)(\boldsymbol{w}-\boldsymbol{w}^\star)\right\|^2\right] \leq 4(\hat{B}_{1,P,N^1}^2+\hat{B}_{2,P,N^1}^2)$$

Where the terms are respectively,

$$\hat{B}_{1,P,N^1}^2 = \frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{(N^1)^2\eta^2}\left(1+(1-\mu\eta)^{N^1}+\frac{2M^2\eta}{\mu}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2L^2\eta}{P\mu}\right),$$
$$\hat{B}_{2,P,N^1}^2 = \frac{2\sigma^2}{T}\left(1+\frac{2L^2\eta}{\mu}+\frac{P^2}{T\mu\eta}+\frac{400M^2\sigma^2\eta^2T}{\mu^2}\right).$$

On upper-bounding the above two terms while ignoring the constants,

$$\hat{B}_{1,P,N^1}^2 \leq \frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{(N^1)^2\eta^2}\left(1+1+\frac{2M^2\eta}{\mu}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2L^2\eta}{P\mu}\right)$$
$$\leq 2\frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{(N^1)^2\eta^2}\left(1+\frac{M^2\eta}{\mu}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{L^2\eta}{P\mu}\right)$$
$$\precsim \frac{\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2}{(N^1)^2\eta^2}\left(1+\frac{M^2\eta}{\mu}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{L^2\eta}{P\mu}\right),$$
$$\hat{B}_{2,P,N^1}^2 \leq 800\frac{\sigma^2}{T}\left(1+\frac{L^2\eta}{\mu}+\frac{P^2}{T\mu\eta}+\frac{M^2\sigma^2\eta^2T}{\mu^2}\right)$$
$$\hat{B}_{2,P,N^1}^2 \precsim \frac{\sigma^2}{T}\left(1+\frac{L^2\eta}{\mu}+\frac{P^2}{T\mu\eta}+\frac{M^2\sigma^2\eta^2T}{\mu^2}\right).$$

Thus we have recovered Proposition 2.

H Proofs for OSA, MBA and Local-SGD in the online setting

Recall that the step size at iteration $(t,k), \in [C] \times [N^t]$ is defined as $\eta_k^t = \frac{c_\eta}{\left(\sum_{t'=1}^{t-1} N^t + k\right)^\alpha}$ where $\alpha \in (0,1)$. Though our results can be extended for the entire range of learning rates, we prove results only for $\alpha \in (\frac{1}{2}, 1)$.

H.1 Technical Lemmas

We first state a few technical results which are helpful in the following proofs.

Lemma S45 *For $\tilde{\eta}_m = \frac{c_\eta}{m^\alpha}$, $\alpha \in (0,1)$ we have $\prod_{m=1}^t (1 - \mu\tilde{\eta}_m) \leq \exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\right)$.*

Proof 46 *The proof simply follows from applying the inequality $1 + x \leq \exp(x)$, followed by an integral bound over the series as $\sum_{m=1}^t \frac{1}{m^\alpha} \geq \frac{1}{2}\int_0^t \frac{1}{m^\alpha} dm = \frac{t^{1-\alpha}}{1-\alpha}$. Note that it is possible to consider $\alpha = 1$ but the integral bound changes. For brevity we don't include it here.*

Lemma S47 *For $\tilde{\eta}_m = \frac{c_\eta}{m^\alpha}$, $\alpha \in (0,1)$ we have*

$$\sum_{m=1}^t (\tilde{\eta}_m)^2 \prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) \leq \exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1 - \frac{1}{2^{1-\alpha}}\right)\right) c_\eta^2 \left(1 + \frac{t^{1-2\alpha} - 1}{1 - 2\alpha}\right) + \frac{2c_\eta}{t^\alpha \mu}.$$

Further if $\alpha \in (\frac{1}{2}, 1)$, then for large t, $\sum_{m=1}^t (\tilde{\eta}_m)^2 \prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) \leq \exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1 - \frac{1}{2^{1-\alpha}}\right)\right) \frac{2\alpha c_\eta^2}{2\alpha - 1} + \frac{2c_\eta}{t^\alpha \mu}$.

Proof 48 *First we decompose the term, then use $1 + x \leq \exp(x)$, followed by a series of integral bounds like Lemma S45,*

$$\begin{aligned}
\sum_{m=1}^t \tilde{\eta}_m^2 \prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) &\leq \sum_{m=1}^{\frac{t}{2}} (\tilde{\eta}_m)^2 \prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) + \sum_{m=\frac{t}{2}}^t (\tilde{\eta}_m)^2 \prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) \\
&\leq \prod_{l=\frac{t}{2}+1}^t (1 - \mu\tilde{\eta}_l) \sum_{m=1}^{\frac{t}{2}} (\tilde{\eta}_m)^2 + \sum_{m=\frac{t}{2}}^t \frac{\tilde{\eta}_m}{\mu}\left(\prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) - \prod_{l=m}^t (1 - \mu\tilde{\eta}_l)\right) \\
&\leq \exp\left(-\mu \sum_{l=\frac{t}{2}+1}^t \tilde{\eta}_l\right) \sum_{m=1}^t (\tilde{\eta}_m)^2 + \frac{\tilde{\eta}_{\frac{t}{2}}}{\mu} \sum_{m=\frac{t}{2}}^t \left(\prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) - \prod_{l=m}^t (1 - \mu\tilde{\eta}_l)\right) \\
&\leq \exp\left(-\mu c_\eta \frac{t^{1-\alpha} - \left(\frac{t}{2}\right)^{1-\alpha}}{2(1-\alpha)}\right) \sum_{m=1}^t \frac{c_\eta^2}{m^{2\alpha}} + \frac{\tilde{\eta}_{\frac{t}{2}}}{\mu}\left(1 - \prod_{l=\frac{t}{2}+1}^t (1 - \mu\tilde{\eta}_l)\right) \\
&\leq \exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1 - \frac{1}{2^{1-\alpha}}\right)\right) c_\eta^2 \left(1 + \frac{t^{1-2\alpha} - 1}{1 - 2\alpha}\right) + \frac{2c_\eta}{t^\alpha \mu}.
\end{aligned}$$

The additional condition on α is obtained by simply taking the limiting case for $t \to \infty$. Also note that this upper bound is tight up to constants (for both terms), especially one could easily show $\sum_{m=1}^t (\tilde{\eta}_m)^2 \prod_{l=m+1}^t (1 - \mu\tilde{\eta}_l) \geq \frac{c_\eta}{2t^\alpha \mu}$.

Lemma S49 *For the gamma function $\Gamma(s) = \int_0^\infty y^{s-1} \exp(-y) dy$ we have, $\sum_{t=1}^C \exp\left(-at^b\right) \leq \frac{1}{ba^{1/b}}\Gamma(\frac{1}{b})$.*

Proof 50 *First we use an integral bound as $\sum_{t=1}^C \exp\left(-at^b\right) \leq \int_0^\infty \exp\left(-az^b\right) dz$, followed by the integral substitution $u = az^b$ after which the proof follows from the definition of the gamma function.*

Lemma S51 *For the gamma function* $\Gamma(s) = \int_0^\infty y^{s-1}\exp(-y)dy$ *we have,* $\sum_{t=1}^{C}\frac{\exp\left(-at^b\right)}{t^c} \leq \frac{1}{ba^{(1-c)/b}}\Gamma(\frac{1-c}{b})$.

Proof 52 *First we use an integral bound as* $\sum_{t=1}^{C}\frac{\exp\left(-at^b\right)}{t^c} \leq \int_0^\infty \frac{\exp\left(-az^b\right)}{z^c}dz$*, followed by the integral substitution* $u = az^b$ *after which the proof follows from the definition of the gamma function.*

Lemma S53 *For* $a \in (0,1)$, $\sum_{t=1}^{C}\frac{1}{t^{1-a}} \leq \frac{C^a}{a}$.

Proof 54 *It is a simple application of the integral bound on a decreasing function,* $\sum_{t=1}^{C}\frac{1}{t^{1-a}} \leq \int_0^C x^{a-1}dx = \frac{C^a}{a}$.

Lemma S55 (Weighted Minkowski) *For* $b_i \in \mathbb{R}$ *and* $a_i \in \mathbb{R}^d$*, we have* $\mathbb{E}\left[\left\|\sum_{i=1}^{P} a_i b_i\right\|^2\right] \leq \left(\sum_{i=1}^{P} b_i \sqrt{\mathbb{E}\left[\|a_i\|^2\right]}\right)^2$.

Proof 56 *We consider again the norm* $\|.\|_E = \mathbb{E}\left[\|.\|^2\right]^{\frac{1}{2}}$*. Now the above result follows by first applying triangle inequality as* $\left\|\sum_{i=1}^{P} a_i b_i\right\|_E \leq \sum_{i=1}^{P}\|a_i b_i\|_E$*, followed by Holder's inequality to give* $\sum_{i=1}^{P} b_i \|a_i\|_E$.

H.2 Proof of Proposition S7 (Mini-batch Averaging Case)

We have the following lemma for mini-batch averaging for the decreasing step-size case,

Lemma S57 *Under the Assumptions* ***A1****,* ***A2****,* ***A3****,* ***A5****,* ***A6*** *we have for mini-batch averaging,*

$$\mathbb{E}\left[\left\|\nabla^2 F(\boldsymbol{w}^\star)(\boldsymbol{w}-\boldsymbol{w}^\star)\right\|^2\right] \leq 5\sum_{i=1}^{6} C_{i,P,C}^2.$$

Where the terms are,

$$\begin{aligned}
C_{1,P,C}^2 =& \frac{1}{C^2 c_\eta^2}\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2,\\
C_{2,P,C}^2 =& \frac{4}{C^{2-2\alpha}c_\eta^2}\left(\exp\left(-\frac{\mu c_\eta C^{1-\alpha}}{2(1-\alpha)}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2\right.\\
&\left.+\frac{2\sigma^2}{P}\left(\exp\left(-\frac{\mu C^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha-1}+\frac{2c_\eta}{C^\alpha\mu}\right)\right),\\
C_{3,P,C}^2 =& \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(\beta_1\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\beta_2\frac{\sigma^2}{P}+\beta_3\frac{\sigma^2 C^\alpha}{P}\right),\\
C_{4,P,C}^2 =& \frac{P^2M^2}{T^2}\left(2\beta_1^2\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^4+2\frac{400\sigma^4}{P^2}\left(\beta_2^2+\beta_3^2C^{2-2\alpha}\right)\right),\\
C_{5,P,C}^2 =& \frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\left(\beta_1\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\beta_2\frac{\sigma^2}{P}+\beta_3\frac{\sigma^2C^{1-\alpha}}{P}\right).
\end{aligned}$$

And the constants are,

$$\beta_1 = \frac{2^{\frac{1+3\alpha}{1-\alpha}}(1-\alpha)^{\frac{4\alpha-2}{1-\alpha}}}{(\mu c_\eta)^{\frac{2\alpha}{1-\alpha}}}\Gamma(\frac{\alpha}{1-\alpha})^2, \beta_2 = \frac{4^{\frac{1+2\alpha-\alpha^2}{(1-\alpha)}}(1-\alpha)^{\frac{2\alpha-1}{(1-\alpha)}}c_\eta^2}{(2\alpha-1)\left(\mu c_\eta(2^{1-\alpha}-1)\right)^{\frac{2\alpha}{(1-\alpha)}}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)^2, \beta_3 = \frac{32c_\eta}{\alpha^2\mu},$$

$$\beta_4 = \frac{2^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}}{(\mu c_\eta)^{\frac{1}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right), \beta_5 = \frac{2^{\frac{3-2\alpha}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}\alpha c_\eta^2}{(2\alpha-1)\left(\mu c_\eta(2^{1-\alpha}-1)\right)^{\frac{1}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right), \beta_6 = \frac{2c_\eta}{(1-\alpha)\mu}.$$

Proof 58 *Using again the decomposition in Lemma S31, we can obtain the following simpler version for mini-batch averaging,*

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{\boldsymbol{w}^0 - \boldsymbol{w}^\star}{C\eta_1^1} - \frac{\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star}{C\eta_2^C} - \frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_1^t} - \frac{1}{\eta_2^t}\right) + \frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^t + \frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^t.$$

Note again that we assume $\alpha \in (\frac{1}{2}, 1)$, just for the sake of brevity. For the first term,

$$\mathbb{E}\left[\left\|\frac{\boldsymbol{w}^0 - \boldsymbol{w}^\star}{C\eta_1^1}\right\|^2\right] = \frac{1}{C^2 c_\eta^2}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 = C_{1,P,C}^2.$$

For the second term using Lemma S9, followed by Lemma S45 and Lemma S47 we obtain,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{\hat{\boldsymbol{w}}^C - \boldsymbol{w}^\star}{C\eta_2^C}\right\|^2\right] &= \frac{(C+1)^{2\alpha}}{C^2 c_\eta^2}\mathbb{E}\left[\left\|\boldsymbol{w}_{MB}^C - \boldsymbol{w}^\star\right\|^2\right] \\
&\leq \frac{2^{2\alpha}}{C^{2-2\alpha}c_\eta^2}\left(\prod_{m=1}^{C}(1-\mu\tilde{\eta}_m)\mathbb{E}\left[\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2\right] + 2\sigma^2\frac{1}{P}\sum_{m=1}^{C}(\tilde{\eta}_m)^2\prod_{l=m+1}^{C}(1-\mu\tilde{\eta}_l)\right) \\
&\leq \frac{4}{C^{2-2\alpha}c_\eta^2}\left(\exp\left(-\frac{\mu c_\eta C^{1-\alpha}}{2(1-\alpha)}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2\right. \\
&\left. + \frac{2\sigma^2}{P}\left(\exp\left(-\frac{\mu C^{1-\alpha}}{2(1-\alpha)}\left(1 - \frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha - 1} + \frac{2c_\eta}{C^\alpha \mu}\right)\right) = C_{2,P,C}^2
\end{aligned}$$

For the third term using Lemma S55 and $(t+1)^\alpha - t^\alpha \leq \alpha t^{\alpha-1}$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_1^t} - \frac{1}{\eta_2^t}\right)\right\|^2\right] \\
&\leq \frac{1}{T^2 c_\eta^2}\mathbb{E}\left[\left\|\sum_{t=1}^{C}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t - \boldsymbol{w}^\star\right)\left((t+1)^\alpha - t^\alpha\right)\right\|^2\right] \\
&\leq \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(\sum_{t=1}^{C}\left((t+1)^\alpha - t^\alpha\right)\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t - \boldsymbol{w}^\star\right)\right\|^2\right]}\right)^2 \\
&\leq \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(\sum_{t=1}^{C}t^{\alpha-1}\sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{MB}^t - \boldsymbol{w}^\star\right\|^2\right]}\right)^2.
\end{aligned}$$

Now using Lemma S9, Lemma S45, Lemma S47 and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we get,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_1^t} - \frac{1}{\eta_2^t}\right)\right\|^2\right] \\
&\leq \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(\sum_{t=1}^{C}t^{\alpha-1}\sqrt{\prod_{m=1}^{t}(1-\mu\tilde{\eta}_m)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\frac{1}{P}\sum_{m=1}^{t}(\tilde{\eta}_m)^2\prod_{l=m+1}^{t}(1-\mu\tilde{\eta}_l)}\right)^2 \\
&\leq \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(\sum_{t=1}^{C}t^{\alpha-1}\sqrt{\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2}{P}\left(\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1 - \frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha - 1} + \frac{2c_\eta}{t^\alpha \mu}\right)}\right)^2
\end{aligned}$$

$$\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\Bigg(\sum_{t=1}^{C} t^{\alpha-1}\Bigg(\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{4(1-\alpha)}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\| + \sqrt{\frac{2\sigma^2}{P}\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha-1}}$$
$$+\sqrt{\frac{4c_\eta\sigma^2}{Pt^\alpha\mu}}\Bigg)\Bigg)^2$$
$$\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\Bigg(\sum_{t=1}^{C} t^{\alpha-1}\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{4(1-\alpha)}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\| + \sum_{t=1}^{C} t^{\alpha-1}\sqrt{\frac{2\sigma^2c_\eta^2}{P(2\alpha-1)}\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)}$$
$$+\sum_{t=1}^{C} t^{\frac{\alpha}{2}-1}\sqrt{\frac{4c_\eta\sigma^2}{P\mu}}\Bigg)^2$$
$$\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\Bigg(\sum_{t=1}^{C} t^{\alpha-1}\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{4(1-\alpha)}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\| + \sqrt{\frac{2\sigma^2c_\eta^2}{P(2\alpha-1)}}\sum_{t=1}^{C} t^{\alpha-1}\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{4(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)$$
$$+\sqrt{\frac{4c_\eta\sigma^2}{P\mu}}\sum_{t=1}^{C}\frac{1}{t^{1-\frac{\alpha}{2}}}\Bigg)^2.$$

Now using Lemma S51 (with $b=1-\alpha$, $c=1-\alpha$ and $a=\frac{\mu c_\eta}{4(1-\alpha)}$), followed by using Lemma S51 again (with $a=\frac{\mu c_\eta}{4(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)$, $b=1-\alpha$ and $c=1-\alpha$) and Lemma S53 (with $a=\frac{\alpha}{2}$) we get,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t-\boldsymbol{w}^\star\right)\left(\frac{1}{\eta_1^t}-\frac{1}{\eta_2^t}\right)\right\|^2\right]$$
$$\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\Bigg(\frac{4^{\frac{\alpha}{1-\alpha}}(1-\alpha)^{\frac{2\alpha-1}{1-\alpha}}}{(\mu c_\eta)^{\frac{\alpha}{1-\alpha}}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\| + \sqrt{\frac{2\sigma^2c_\eta^2}{P(2\alpha-1)}}\frac{2^{\frac{\alpha(3-\alpha)}{1-\alpha}}(1-\alpha)^{\frac{2\alpha-1}{1-\alpha}}}{(\mu c_\eta(2^{1-\alpha}-1))^{\frac{\alpha}{1-\alpha}}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)$$
$$+\sqrt{\frac{4c_\eta\sigma^2}{P\mu}}\frac{2C^{\frac{\alpha}{2}}}{\alpha}\Bigg)^2.$$

Finally using Lemma S35 and re-organizing with constants defined as above,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,1}^t-\boldsymbol{w}^\star\right)\left(\frac{1}{\eta_1^t}-\frac{1}{\eta_2^t}\right)\right\|^2\right]$$
$$\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\Bigg(2\frac{4^{\frac{2\alpha}{1-\alpha}}(1-\alpha)^{\frac{4\alpha-2}{1-\alpha}}}{(\mu c_\eta)^{\frac{2\alpha}{1-\alpha}}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)^2\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2 + 2\frac{2\sigma^2c_\eta^2}{P(2\alpha-1)}\frac{4^{\frac{\alpha(3-\alpha)}{(1-\alpha)}}(1-\alpha)^{\frac{2\alpha-1}{(1-\alpha)}}}{(\mu c_\eta(2^{1-\alpha}-1))^{\frac{2\alpha}{(1-\alpha)}}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)^2$$
$$+2\frac{4c_\eta\sigma^2}{P\mu}\frac{4C^\alpha}{\alpha^2}\Bigg)$$
$$\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(\beta_1\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\beta_2\frac{\sigma^2}{P}+\beta_3\frac{\sigma^2C^\alpha}{P}\right)=C_{3,P,C}^2.$$

For the fourth term first proceeding as in Lemma S39 with Lemma S35 and Lemma S37 we can obtain,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^t\right\|^2\right] = \frac{1}{T^2}\mathbb{E}\left[\left\|\sum_{t=1}^{C}\sum_{i=1}^{P}\left(F'(\boldsymbol{w}_{i,0}^t)-F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,0}^t-\boldsymbol{w}^\star)\right)\right\|^2\right]$$
$$\leq \frac{P}{T^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\sum_{t=1}^{C}\left(F'(\hat{\boldsymbol{w}}^{t-1})-F''(\boldsymbol{w}^\star)(\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star)\right)\right\|^2\right]$$
$$\leq \frac{P}{T^2}\sum_{i=1}^{P}\left(\sum_{t=1}^{C}\sqrt{\mathbb{E}\left[\left\|\left(F'(\hat{\boldsymbol{w}}^{t-1})-F''(\boldsymbol{w}^\star)(\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star)\right)\right\|^2\right]}\right)^2$$

$$\leq \frac{PM^2}{T^2}\sum_{i=1}^{P}\left(\sum_{t=1}^{C}\sqrt{\mathbb{E}\left[(\hat{\boldsymbol{w}}^{t-1}-\boldsymbol{w}^\star)^4\right]}\right)^2$$
$$\leq \frac{P^2M^2}{T^2}\left(\sum_{t=1}^{C}\sqrt{\mathbb{E}\left[(\boldsymbol{w}_{MB}^{t-1}-\boldsymbol{w}^\star)^4\right]}\right)^2.$$

Now using Lemma S26, followed by Lemma S45 and Lemma S47 we get[8],

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^t\right\|^2\right] \leq \frac{P^2M^2}{T^2}\left(\sum_{t=1}^{C}\left(\prod_{j=1}^{t-1}(1-\tilde{\eta}_j\mu)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\sigma^2}{P}\sum_{j=1}^{t-1}(\tilde{\eta}_j)^2\prod_{l=j+1}^{t-1}(1-\mu\tilde{\eta}_l)\right)\right)^2$$
$$\leq \frac{P^2M^2}{T^2}\left(\sum_{t=1}^{C}\exp\left(-\frac{\mu c_\eta(t-1)^{1-\alpha}}{2(1-\alpha)}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2\right.$$
$$\left.+\sum_{t=2}^{C}\frac{20\sigma^2}{P}\left(\exp\left(-\frac{\mu c_\eta(t-1)^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha-1}+\frac{2c_\eta}{(t-1)^\alpha\mu}\right)\right)^2$$
$$\leq \frac{P^2M^2}{T^2}\left(\sum_{t=1}^{C}\exp\left(-\frac{\mu c_\eta(t-1)^{1-\alpha}}{2(1-\alpha)}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2\right.$$
$$\left.+\sum_{t=1}^{C}\frac{20\sigma^2}{P}\left(\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha-1}+\sum_{t=1}^{C}\frac{2c_\eta}{t^\alpha\mu}\right)\right)^2.$$

Now using Lemma S49 (with $b=1-\alpha$ and $a=\frac{\mu c_\eta}{2(1-\alpha)}$), followed by Lemma S49 again (with $a=\frac{\mu c_\eta}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)$ and $b=1-\alpha$), followed by Lemma S53 (with $a=1-\alpha$) and Lemma S35 we get,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^t\right\|^2\right]$$
$$\leq \frac{P^2M^2}{T^2}\left(\frac{2^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}}{(\mu c_\eta)^{\frac{1}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2\right.$$
$$\left.+\frac{20\sigma^2}{P}\left(\frac{2^{\frac{2-\alpha}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}}{(\mu c_\eta(2^{1-\alpha}-1))^{\frac{1}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right)\frac{2\alpha c_\eta^2}{2\alpha-1}+\frac{2c_\eta C^{1-\alpha}}{(1-\alpha)\mu}\right)\right)^2$$
$$\leq \frac{P^2M^2}{T^2}\left(2\frac{2^{\frac{2}{1-\alpha}}(1-\alpha)^{\frac{2\alpha}{1-\alpha}}}{(\mu c_\eta)^{\frac{2}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right)^2\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^4\right.$$
$$\left.+2\frac{400\sigma^4}{P^2}\left(\frac{2^{\frac{4-2\alpha}{1-\alpha}}(1-\alpha)^{\frac{2\alpha}{1-\alpha}}}{(\mu c_\eta(2^{1-\alpha}-1))^{\frac{2}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right)^2\frac{4\alpha^2c_\eta^4}{(2\alpha-1)^2}+\frac{4c_\eta^2C^{2-2\alpha}}{(1-\alpha)^2\mu^2}\right)\right)$$

Bounding again with the constants defined above,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\delta_{i,1}^t\right\|^2\right] \leq \frac{P^2M^2}{T^2}\left(2\beta_4^2\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^4+2\frac{400\sigma^4}{P^2}\left(\beta_5^2+\beta_6^2C^{2-2\alpha}\right)\right)=C_{4,P,C}^2.$$

For the fifth term, proceeding as in Lemma S39,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^t\right\|^2\right]=\frac{1}{T^2}\sum_{t=1}^{C}\sum_{i=1}^{P}\left(2L^2\mathbb{E}\left[\left\|\boldsymbol{w}_{i,0}^t-\boldsymbol{w}^\star\right\|^2\right]+2\sigma^2\right)$$

[8]Note that we ignore t=1 in second inequality for second term as we have already incorporated it in the first term

$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\sum_{t=1}^{C}\mathbb{E}\left[\left\|\boldsymbol{w}_{1,0}^t - \boldsymbol{w}^\star\right\|^2\right]$$
$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\sum_{t=1}^{C}\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}_{MB}^{t-1} - \boldsymbol{w}^\star\right\|^2\right].$$

Now using Lemma S9, Lemma S45 and Lemma S47 like before,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^t\right\|^2\right] \leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\sum_{t=1}^{C}\left(\exp\left(-\frac{\mu c_\eta}{2(1-\alpha)}t^{1-\alpha}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2\right.$$
$$\left. + \frac{2\sigma^2}{P}\exp\left(-\frac{\mu c_\eta t^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right)\frac{2\alpha c_\eta^2}{2\alpha-1} + \frac{4\sigma^2 c_\eta}{Pt^\alpha\mu}\right).$$

Further using Lemma S49 (with $b = 1-\alpha$ and $a = \frac{\mu c_\eta}{2(1-\alpha)}$), followed by Lemma S49 again (with $a = \frac{\mu c_\eta}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)$ and $b = 1-\alpha$), followed by Lemma S53 (with $a = 1-\alpha$) and the constants as used above we get,

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{C}\sum_{i=1}^{P}\xi_{i,1}^t\right\|^2\right] \leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\left(\frac{2^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}}{(\mu c_\eta)^{\frac{1}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2\right.$$
$$+ \frac{2^{\frac{2-\alpha}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}}{(\mu c_\eta(2^{1-\alpha}-1))^{\frac{1}{1-\alpha}}}\Gamma\left(\frac{1}{1-\alpha}\right)\frac{2\alpha c_\eta^2}{2\alpha-1} + \left.\frac{2c_\eta C^{1-\alpha}}{(1-\alpha)\mu}\right)$$
$$\leq \frac{2\sigma^2}{T} + \frac{2L^2P}{T^2}\left(\beta_4\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \beta_5\frac{\sigma^2}{P} + \beta_6\frac{\sigma^2 C^{1-\alpha}}{P}\right) = C_{5,P,C}^2.$$

Finally using Lemma S35 we have proved the lemma.

The following lemma separates the terms above into bias and variance terms, following which we can easily prove Proposition S7,

Lemma S59 *Under the Assumptions **A1**, **A2**, **A3**, **A5**, **A6** we have for mini-batch averaging,*

$$\mathbb{E}\left[\left\|\nabla^2 F(\boldsymbol{w}^\star)(\boldsymbol{w} - \boldsymbol{w}^\star)\right\|^2\right] \leq 5\left(\hat{C}_{1,P,C}^2 + \hat{C}_{2,P,C}^2\right)$$

Where for constants defined as above the terms are,

$$\hat{C}_{1,P,C}^2 = \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{C^2 c_\eta^2}\left(1 + 4C^{2\alpha}\exp\left(-\frac{\mu c_\eta C^{1-\alpha}}{2(1-\alpha)}\right) + \alpha^2\beta_1 + 2M^2 c_\eta^2\beta_1^2\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2L^2\beta_1 c_\eta^2}{P}\right),$$

$$\hat{C}_{2,P,C}^2 = \frac{2\sigma^2}{T}\left(1 + \frac{8\alpha C^{2\alpha-1}}{2\alpha-1}\exp\left(-\frac{\mu C^{1-\alpha}}{2(1-\alpha)}\left(1-\frac{1}{2^{1-\alpha}}\right)\right) + \frac{8}{C^{1-\alpha}c_\eta\mu} + \frac{\alpha^2\beta_2}{2Cc_\eta^2} + \frac{\alpha^2\beta_3}{2C^{1-\alpha}c_\eta^2}\right.$$
$$\left. + \frac{400M^2\sigma^2}{T}\left(\beta_2^2 + \beta_3^2 C^{2-2\alpha}\right) + \frac{L^2}{T}\left(\beta_2 + \beta_3 C^{1-\alpha}\right)\right).$$

To get Proposition S7, we upper bound every term up to constants depending only on α. Specifically, we use $\beta_1 \precsim (\mu c_\eta)^{-\frac{1}{1-\alpha}}$, $\beta_2 \precsim (\mu c_\eta)^{-\frac{\alpha}{1-\alpha}}$, and $\beta_3 \precsim \frac{c_\eta}{\mu}$.

H.3 Proof of Proposition S7 (One-shot Averaging case)

The analysis for the one-shot case is very similar to the mini-batch case, just like the constant step-size case. In fact at many place the communications C of MBA get replaced by N^1 and the form of the bound remains the same. This intuitive conversion strengthens our analysis, which smoothly extends to both the extreme cases.

Lemma S60 *Under the Assumptions A1, A2, A3, A5, A6 for decreasing step size, for one shot averaging we have,*

$$\mathbb{E}\left[\left\|\nabla^2 F(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k}^1 - \boldsymbol{w}^\star)\right\|^2\right] \leq 5\sum_{i=1}^{6} D_{i,P,C}^2$$

where the terms are,

$$D_{1,P,N^1}^2 = \frac{P^2}{T^2 c_\eta^2}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2, D_{2,P,N^1}^2 = \frac{4}{(N^1)^{2-2\alpha}c_\eta^2}\left(\exp\left(-\frac{\mu c_\eta (N^1)^{1-\alpha}}{1-\alpha}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2 c_\eta}{\mu}\right),$$

$$D_{3,P,N^1}^2 = \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(4\beta^2\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2 (N^1)^{2\alpha} c_\eta}{\mu\alpha^2}\right), D_{4,P,N^1}^2 = \frac{P^2 M^2}{T^2}\left(\beta\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{20\sigma^2 N^1 c_\eta}{\mu}\right)^2,$$

$$D_{5,P,N^1}^2 = \frac{2\sigma^2}{T}, D_{6,P,N^1}^2 = \frac{2L^2 P}{T^2}\left(\beta\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2 N^1 c_\eta}{\mu}\right).$$

And the constants are $\beta_1 = 1 + \left(\frac{(1-\alpha)^\alpha}{\mu c_\eta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)$ *and* $\beta_2 = \left(2^\alpha \frac{(1-\alpha)^{2\alpha-1}}{(\mu c_\eta)^\alpha}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)$.

Proof 61 *We follow an analysis similar to [15]. We can simplify the decomposition from Lemma S31 for one outer phase as follows,*

$$F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star) = \frac{\boldsymbol{w}^0 - \boldsymbol{w}^\star}{N^1\eta_1^1} - \frac{\hat{\boldsymbol{w}}^1 - \boldsymbol{w}^\star}{N^1\eta_{N^1+1}^1} - \frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\left(\boldsymbol{w}_{i,k}^1 - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^1} - \frac{1}{\eta_{k+1}^1}\right)$$
$$+ \frac{1}{T}\sum_{k=1}^{N^1}\sum_{i=1}^{P}\delta_{i,k}^1 + \frac{1}{T}\sum_{k=1}^{N^1}\sum_{i=1}^{P}\xi_{i,k}^1.$$

For the first term,

$$\mathbb{E}\left[\left\|\frac{\boldsymbol{w}^0 - \boldsymbol{w}^\star}{N^1\eta_1^1}\right\|^2\right] \leq \frac{P^2}{T^2 c_\eta^2}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 = D_{1,P,N^1}^2.$$

For the second term note that the inner iterate bound is independent for different machines using Lemma S11 for say machine 1*, followed by Lemma S45 and Lemma S47 we get,*

$$\mathbb{E}\left[\left\|\frac{\hat{\boldsymbol{w}}^1 - \boldsymbol{w}^\star}{N^1\eta_{N^1+1}^1}\right\|^2\right] \leq \frac{(N^1+1)^{2\alpha}}{(N^1)^2 c_\eta^2}\mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,N^1}^1 - \boldsymbol{w}^\star\right)\right\|^2\right]$$
$$\leq \frac{2^{2\alpha}}{(N^1)^{2-2\alpha}c_\eta^2}\mathbb{E}\left[\left\|\boldsymbol{w}_{1,N^1}^1 - \boldsymbol{w}^\star\right\|^2\right]$$
$$\leq \frac{4}{(N^1)^{2-2\alpha}c_\eta^2}\left(\prod_{m=1}^{N^1}(1-\mu\eta_m^1)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\sum_{m=1}^{N^1}(\eta_m^1)^2\prod_{l=m+1}^{N^1}(1-\mu\eta_l^1)\right)$$
$$\leq \frac{4}{(N^1)^{2-2\alpha}c_\eta^2}\left(\exp\left(-\frac{\mu c_\eta (N^1)^{1-\alpha}}{1-\alpha}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2 c_\eta}{\mu}\right) = D_{2,P,N^1}^2.$$

For the third term using $(k+1)^\alpha - k^\alpha \leq \alpha k^{\alpha-1}$*, Lemma S55, and noting that the individual bounds on inner iterates for different machines are the same, thus using machine* 1 *for brevity we can obtain,*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\left(\boldsymbol{w}_{i,k}^1 - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^1} - \frac{1}{\eta_{k+1}^1}\right)\right\|^2\right] \leq \frac{P^2\alpha^2}{T^2 c_\eta^2}\mathbb{E}\left[\left\|\sum_{k=1}^{N^1}k^{\alpha-1}\left(\boldsymbol{w}_{1,k}^1 - \boldsymbol{w}^\star\right)\right\|^2\right]$$
$$\leq \frac{P^2\alpha^2}{T^2 c_\eta^2}\left(\sum_{k=1}^{N^1}k^{\alpha-1}\sqrt{\mathbb{E}\left[\left\|\boldsymbol{w}_{1,k}^1 - \boldsymbol{w}^\star\right\|^2\right]}\right)^2.$$

Now using Lemma S11, Lemma S45, Lemma S47 and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we get,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{T}\sum_{k=1}^{N^1}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,k}^1 - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^1} - \frac{1}{\eta_{k+1}^1}\right)\right\|^2\right] \\
&\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(\sum_{k=1}^{N^1} k^{\alpha-1}\sqrt{\mathbb{E}\left[\prod_{m=1}^{k}(1-\mu\tilde{\eta}_m)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\sum_{m=1}^{k}(\tilde{\eta}_m)^2\prod_{l=m+1}^{k}(1-\mu\tilde{\eta}_l)\right]}\right)^2 \\
&\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(\sum_{k=1}^{N^1} k^{\alpha-1}\sqrt{\exp\left(-\frac{\mu c_\eta k^{1-\alpha}}{1-\alpha}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2 c_\eta}{\mu}}\right)^2 \\
&\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(\sum_{k=1}^{N^1} k^{\alpha-1}\left(\exp\left(-\frac{\mu c_\eta k^{1-\alpha}}{2(1-\alpha)}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\| + \sqrt{\frac{2\sigma^2 c_\eta}{\mu}}\right)\right)^2.
\end{aligned}$$

Now using Lemma S51 again with $b = 1-\alpha$ and $a = \frac{\mu c_\eta}{2(1-\alpha)}$ with β_2 defined as above and Lemma S53 we get,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{T}\sum_{k=1}^{N^1}\sum_{i=1}^{P}\left(\boldsymbol{w}_{i,k}^1 - \boldsymbol{w}^\star\right)\left(\frac{1}{\eta_k^1} - \frac{1}{\eta_{k+1}^1}\right)\right\|^2\right] \\
&\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(\left(2^\alpha\frac{(1-\alpha)^{2\alpha-1}}{(\mu c_\eta)^\alpha}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{\alpha}{1-\alpha}\right)\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\| + \sqrt{\frac{2\sigma^2(N^1)^{2\alpha}c_\eta}{\mu\alpha^2}}\right)^2 \\
&\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(\beta_2\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\| + \sqrt{\frac{2\sigma^2(N^1)^{2\alpha}c_\eta}{P\mu\alpha^2}}\right)^2 \\
&\leq \frac{P^2\alpha^2}{T^2c_\eta^2}\left(2\beta_2^2\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{4\sigma^2(N^1)^{2\alpha}c_\eta}{\mu\alpha^2}\right) = D_{3,P,N^1}^2.
\end{aligned}$$

Now for the fourth term proceeding as in Lemma S42 with Lemma S35 and Lemma S37 we can obtain ,

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\delta_{i,k}^1\right\|^2\right] &= \frac{1}{T^2}\mathbb{E}\left[\left\|\sum_{i=1}^{P}\sum_{k=1}^{N^1}F'(\boldsymbol{w}_{i,k-1}^t) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star)\right\|^2\right] \\
&\leq \frac{P}{T^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\sum_{k=1}^{N^1}F'(\boldsymbol{w}_{i,k-1}^t) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^t - \boldsymbol{w}^\star)\right\|^2\right] \\
&\leq \frac{P}{T^2}\sum_{i=1}^{P}\left(\sum_{k=1}^{N^1}\sqrt{\mathbb{E}\left[\left\|F'(\boldsymbol{w}_{i,k-1}^1) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{i,k-1}^1 - \boldsymbol{w}^\star)\right\|^2\right]}\right)^2
\end{aligned}$$

*Now first using the upper bound of **A**2, followed by Lemma S26, Lemma S45, Lemma S47 and Lemma S49 we can obtain the following,*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\delta_{i,k}^1\right\|^2\right] \leq \frac{PM^2}{T^2}\sum_{i=1}^{P}\left(\sum_{k=1}^{N^1}\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^1 - \boldsymbol{w}^\star\right\|^4\right]^{1/2}\right)^2$$

$$
\begin{aligned}
&\leq \frac{P^2M^2}{T^2}\left(\sum_{k=1}^{N^1}\left(\prod_{j=1}^{k-1}(1-\eta_j^1\mu)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+20\sigma^2\sum_{j=1}^{k-1}\prod_{l=j+1}^{k-1}(1-\mu\eta_l^1)(\eta_j^1)^2\right)\right)^2\\
&\leq \frac{P^2M^2}{T^2}\left(\sum_{k=1}^{N^1}\left(\exp\left(-\frac{\mu c_\eta(k-1)^{1-\alpha}}{1-\alpha}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\sigma^2c_\eta}{\mu}\right)\right)^2\\
&\leq \frac{P^2M^2}{T^2}\left(\left(1+\left(\frac{(1-\alpha)^\alpha}{\mu c_\eta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\sigma^2N^1c_\eta}{\mu}\right)^2\\
&\leq \frac{P^2M^2}{T^2}\left(\beta_1\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{20\sigma^2N^1c_\eta}{\mu}\right)^2=D_{4,P,N^1}^2.
\end{aligned}
$$

*For the fifth term, using the fact that for different machines noise is independent, zero in expectation (**A**3) we obtain,*

$$
\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\xi_{i,k}^1\right\|^2\right]=\frac{1}{T^2}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\mathbb{E}\left[\left\|\xi_{i,k}^1\right\|^2\right].
$$

Now using Lemma S33 we have,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\xi_{i,k}^1\right\|^2\right]&\leq\frac{1}{T^2}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\left(2L^2\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^1-\boldsymbol{w}^\star\right\|^2\right]+2\sigma^2\right)\\
&\leq\frac{2\sigma^2}{T}+\frac{2L^2}{T^2}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\mathbb{E}\left[\left\|\boldsymbol{w}_{i,k-1}^1-\boldsymbol{w}^\star\right\|^2\right].
\end{aligned}
$$

Now using Lemma S11, followed by Lemma S45, Lemma S47 and Lemma S49 with definition of β as before, and we have,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{T}\sum_{i=1}^{P}\sum_{k=1}^{N^1}\xi_{i,k}^1\right\|^2\right]&\leq\frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\sum_{k=1}^{N^1}\left(\prod_{m=1}^{k-1}(1-\mu\eta_m^1)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+2\sigma^2\sum_{m=1}^{k-1}(\eta_m^1)^2\prod_{l=m+1}^{k-1}(1-\mu\eta_l^1)\right)\\
&\leq\frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\sum_{k=1}^{N^1}\left(\exp\left(-\frac{\mu c_\eta(k-1)^{1-\alpha}}{1-\alpha}\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2c_\eta}{\mu}\right)\\
&\leq\frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\left(\left(1+\left(\frac{(1-\alpha)^\alpha}{\mu c_\eta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2N^1c_\eta}{\mu}\right)\\
&\leq\frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\left(\left(1+\left(\frac{(1-\alpha)^\alpha}{\mu c_\eta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)\right)\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2N^1c_\eta}{\mu}\right)\\
&\leq\frac{2\sigma^2}{T}+\frac{2L^2P}{T^2}\left(\beta\left\|\boldsymbol{w}^0-\boldsymbol{w}^\star\right\|^2+\frac{2\sigma^2Nc_\eta}{\mu}\right)=D_{5,P,N^1}^2+D_{6,P,N^1}^2.
\end{aligned}
$$

Thus using Lemma S35 we have proved the lemma.

We can get the following lemma combining the bias and variance terms separately,

Lemma S62 *Under the Assumptions **A**1, **A**2, **A**3, **A**5, **A**6 for decreasing step size, for one shot averaging we have,*

$$
\mathbb{E}\left[\left\|\nabla^2F(\boldsymbol{w}^\star)(\boldsymbol{w}-\boldsymbol{w}^\star)\right\|^2\right]\leq 5\left(\hat{D}_{1,P,N^1}^2+\hat{D}_{2,P,N^1}^2\right)
$$

Where for constants defined as above the terms are,

$$\hat{D}^2_{1,P,N^1} = \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{(N^1)^2 c_\eta^2}\left(1 + 4(N^1)^{2\alpha}\exp\left(-\frac{\mu c_\eta (N^1)^{1-\alpha}}{2(1-\alpha)}\right) + \alpha^2\beta_1 + 2M^2 c_\eta^2 \beta_1^2\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{2L^2\beta_1 c_\eta^2}{P}\right),$$

$$\hat{D}^2_{2,P,N^1} = \frac{2\sigma^2}{T}\left(1 + \frac{8\alpha P (N^1)^{2\alpha-1}}{2\alpha - 1}\exp\left(-\frac{\mu (N^1)^{1-\alpha}}{2(1-\alpha)}\left(1 - \frac{1}{2^{1-\alpha}}\right)\right) + \frac{8P}{(N^1)^{1-\alpha} c_\eta \mu} + \frac{\alpha^2 P\beta_2}{2N^1 c_\eta^2} + \frac{\alpha^2 P\beta_3}{2(N^1)^{1-\alpha} c_\eta^2}\right.$$
$$\left. + \frac{400M^2 P\sigma^2}{N^1}\left(\beta_2^2 + \beta_3^2 (N^1)^{2-2\alpha}\right) + \frac{L^2}{N^1}\left(\beta_2 + \beta_3 (N^1)^{1-\alpha}\right)\right).$$

I Brief overview of distributed optimization

The above three schemes (OSA, MBA, Local-SGD) are the most studied synchronous parallel schemes. However, communication latencies often make it difficult to use these algorithms for large-scale problems. Thus many alternative parallelization schemes which minimize communication or perform better have been studied. The major problem with some of these variants is that they are often difficult to tune, are not as stable and don't scale well to non-convex optimization problems. Result-wise, most of the machine learning packages use centralized mini-batch synchronous SGD.

Asynchronous SGD: These techniques are characterized by avoiding a centralized synchronization, using delayed updates, maintaining parameter server estimates and being fault tolerant. Some of the notable references in a chronological order are [46, 52, 54–70].

Federated optimization: This setting is characterized by a huge number of mobile user devices, which run their local model in a decentralized manner with often unbalanced data, but aim to train jointly. Many research questions still remain open but the direction is very relevant for distributed AI. Some references are [71–73].

Compressed Communication: A common strategy to combat the communication overhead is to introduce lossless or lossy compression of exchanged information, often the gradients. Some of the work in this direction can be found in [74–81].

Non-SGD methods: Many other optimization algorithms (coordinate descent, quasi newton, etc.) have also been studied in the parallel setting, owing to their better distributivity or convergence for some applications compared to the SGD algorithm. Some of them are [82] (ADMM), [83] (DANE), [84] (DiSCO), [85] (AIDE), [86–88] (COCOA) and some of the references therein. Recently [89] gave provably optimal algorithms for the strongly convex and smooth functions for both synchronous and asynchronous cases. More broadly speaking, variance reduction methods are often the methods of choice in better understood, convex optimization problems [add reference]. Yet, their usage in the deep learning community has been relatively scarce, and often they are more difficult to parallelize [add reference]. Some of the works for instance are [61, 90–92]. Among second order methods, quasi newton methods like distributed L-BFGS [93, 94] are also widely popular among the machine learning community.

Communication Lower Bounds: On a broader level our work is related to communication lower bounds which arise from information and learning-theoretic considerations. Unfortunately, these bounds are difficult to match for convex optimization as they are provided in [95]. Similar bounds have also been provided for the generally easier statistical estimation setting in [96–98].

Feature Distribution: As clearly evident training data is not the only element of our optimization scheme which can be parallelized. Often in many problems in natural language processing and linear estimation, the features number in hundreds of thousands, and it might be of some interest to distribute the features alongside or beside training data. Some relevant references are [87, 99–102].

There has also been work in parallelizing stochastic optimization algorithms for specific problems (like PCA) in the past, for e.g., [31, 32, 103–107].

Reference	Setting	Limitations
Zhang et. al. [33]	OSA	Small learning rates $\frac{c}{\mu t}$; μ often unknown; Non-asymptotic bound on single worker convergence rate is used ([34]);
Jain et. al. [20]	OSA, MBA	Results for least square regression (LSR) in finite horizon setting only;
Godichon et. al. [108]	OSA	Uses uniform gradient bound **A**4 and thus not usable for LSR; Non-asymptotic result ([34]) is used;
Stich [40]	Local SGD	Small learning rates $\frac{c}{\mu t}$; μ often unknown; Uses uniform gradient bound **A**4 and thus not usable for LSR; Doesn't capture the need for an adaptive communication frequency [21]; Doesn't extend to one-shot averaging, implying it is not tight enough;

Table S3: Limitations of the previously existing results.

We also provide a brief overview of some other techniques in distributed optimization in Appendix I.