
Communication trade-offs for Local-SGD with large step size

Kumar Kshitij PATEL

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

Aymeric DIEULEVEUT

MLO, EPFL, Lausanne, Switzerland
CMAP, Ecole Polytechnique, Palaiseau, France
aymeric.dieuleveut@polytechnique.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 $O\left(\frac{\sqrt{T}}{P^{3/2}}\right)$, 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 independent and identically distributed 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 $F : \mathbb{R}^d \rightarrow \mathbb{R}$, SGD iteratively updates an estimator $(\mathbf{v}_t)_{t \geq 0}$ for any $t \geq 1$

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

where $(\eta_t)_{t \geq 0}$ is a deterministic sequence of positive scalars, referred to as the *learning rate* and $\mathbf{g}_t(\mathbf{v}_{t-1})$ is an oracle on the gradient of the function F at \mathbf{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] 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

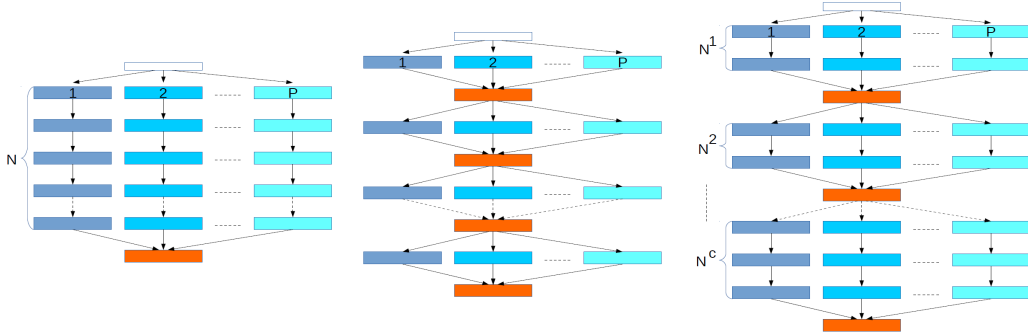


Figure 1: Schematic representation of one-shot averaging (left), mini-batch averaging (middle) and local-SGD (right). Vertical threads correspond to machines and orange boxes to communication rounds.

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 $F : \mathbb{R}^d \rightarrow \mathbb{R}$. We denote $\mathbf{w}^* := \operatorname{argmin}_{\mathbf{w}} F(\mathbf{w})$. 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 \preceq 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, it facilitates our proof technique and notation. At any $k \in [N^t]$, we denote by $\mathbf{w}_{p,k}^t$ the model proposed by worker p , at phase t , after k local iterations. All machines initially start from the same point \mathbf{w}_0 , that is for any $p \in [P]$, $\mathbf{w}_{p,0}^1 = \mathbf{w}_0$. The update rule is thus the following, for any $p \in [P]$, $t \in [C]$, $k \in [N^t]$:

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

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

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

with $\bar{\mathbf{w}}^t = \frac{1}{PN^t} \sum_{k=1}^{N^t} \sum_{p=1}^P \mathbf{w}_{p,k}^t$. We use the notation $\bar{\mathbf{w}}$ to underline the fact that iterates are averaged over one phase and $\bar{\bar{\mathbf{w}}}$ 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 \mathbf{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 $\mathbb{E}[\|\mathbf{v}_t - \mathbf{w}^*\|^2]$, which results in an upper bound on the mean excess of risk $\mathbb{E}[F(\bar{\mathbf{v}}_t) - F(\mathbf{w}^*)]$ [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 $(\bar{\mathbf{v}}_t - \mathbf{w}^*)_{t \geq 1}$: 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 A4 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 $\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] \leq C_1 \eta_t$ (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 A4 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 $F : \mathbb{R}^d \rightarrow \mathbb{R}$. 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: $\sup_{\mathbf{w} \in \mathbb{R}^d} \|F^{(2)}(\mathbf{w})\| < L$, and $\sup_{\mathbf{w} \in \mathbb{R}^d} \|F^{(3)}(\mathbf{w})\| < M$. 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 $\mathbf{w} \mapsto \|\Sigma^{1/2}(\mathbf{w} - \mathbf{w}^*)\|^2/2$.*

If Q1 is satisfied, then Assumptions A1, A2 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 $g_{p,k}^t(\mathbf{w})$ at a point \mathbf{w} . Formally, we make the following assumption:

A3 (Oracle on the gradient) *We observe unbiased estimators of the gradient $g_{p,k+1}^t(\mathbf{w})$: for any $(t, k) \in [C] \times [N^t]$ and $\mathbf{w} \in \mathbb{R}^d$, $\mathbb{E}[g_{p,k+1}^t(\mathbf{w}_{p,k}^t) | \mathbf{w}_{p,k}^t] = F'(\mathbf{w}_{p,k}^t)$. Moreover, for any fixed \mathbf{w} the functions $(g_{p,k}^t)_{(t,k)}(\mathbf{w})$ 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, $\mathbb{E}[\|g_{p,k}^t(\mathbf{w}_{p,k}^t) - F'(\mathbf{w}_{p,k}^t)\|^2]$ is uniformly upper bounded by σ_∞^2 , a constant which does not depend on the iteration.*

Assumption A4 is for example true if the sequence of random vectors $(g_{p,k+1}^t(\mathbf{w}_{p,k}^t) - F'(\mathbf{w}_{p,k}^t))_{t \in [C], k \in [N^t], p \in [P]}$ 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 $p \in [P]$, $t \in [C]$, $k \in [N^t]$, $g_{p,k}^t$ is almost surely L -co-coercive (with the same constant as in A2): that is, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$, $L \langle g_{p,k}^t(\mathbf{w}_1) - g_{p,k}^t(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle \geq \|g_{p,k}^t(\mathbf{w}_1) - g_{p,k}^t(\mathbf{w}_2)\|^2$.

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

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

It must be noted that A6 is a much weaker assumption than A4, 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 Q1, A3, A5 and A6. On the other hand, SGD for logistic regression satisfies A1, A2, A3 and A4. 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 $2\eta_k^t L \leq 1$. We consider two different types of learning rates:

1) in the *finite horizon* (FH) case, the step size $(\eta_k^t)_{(t,k) \in [C] \times [N^t]}$ 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: $\eta_k^t = \tilde{\eta}_{\sum_{t'=1}^t N^{t'} + k}$. 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: $\eta_k^t F''(\mathbf{w}^*)(\mathbf{w}_{p,k-1}^t - \mathbf{w}^*) = \mathbf{w}_{p,k-1}^t - \mathbf{w}_{p,k}^t - \eta_k^t [g_{p,k}^t(\mathbf{w}_{p,k-1}^t) - F'(\mathbf{w}_{p,k-1}^t)] - \eta_k^t [F'(\mathbf{w}_{p,k-1}^t) - F''(\mathbf{w}^*)(\mathbf{w}_{p,k-1}^t - \mathbf{w}^*)]$. We have added and subtracted a first order Taylor expansion around the optimal value \mathbf{w}^* of the gradient. Thus, using the definition of $\bar{\mathbf{w}}^C$:

$$F''(\mathbf{w}^*)(\bar{\mathbf{w}}^C - \mathbf{w}^*) = \frac{1}{P \sum_{t=1}^C N^t} \sum_{t=1}^C \sum_{p=1}^P \sum_{k=1}^{N^t} \left(\frac{\mathbf{w}_{p,k-1}^t - \mathbf{w}_{p,k}^t}{\eta_k^t} - [g_{p,k}^t(\mathbf{w}_{p,k-1}^t) - F'(\mathbf{w}_{p,k-1}^t)] - [F'(\mathbf{w}_{p,k-1}^t) - F''(\mathbf{w}^*)(\mathbf{w}_{p,k-1}^t - \mathbf{w}^*)] \right). \quad (3)$$

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 $\mathbb{E}[g_{p,k}^t(\mathbf{w}_{p,k-1}^t) - F'(\mathbf{w}_{p,k-1}^t)] = 0$), and the third is a *residual term* that accounts for the fact that the function is not quadratic (if F is quadratic, then $F'(\mathbf{w}_{p,k-1}^t) - F''(\mathbf{w}^*)(\mathbf{w}_{p,k-1}^t - \mathbf{w}^*) = 0$).

Controlling different terms in Equation (3). The variance of the noise $g_{p,k}^t(\mathbf{w}_{p,k-1}^t) - F'(\mathbf{w}_{p,k-1}^t)$ and the residual term both directly depend on the distance $\|\mathbf{w}_{p,k-1}^t - \mathbf{w}^*\|^2$. 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 $\mathbb{E}[\|\mathbf{w}_{p,k}^t - \mathbf{w}^*\|^2]$ ¹, (2) we provide the subsequent upper bound on $\mathbb{E}[\|F''(\mathbf{w}^*)(\bar{\mathbf{w}}^C - \mathbf{w}^*)\|^2]$.

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.

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

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: $Q_{bias} = 1 + \frac{M^2\eta}{\mu}\|\mathbf{w}^0 - \mathbf{w}^*\|^2 + \frac{L^2\eta}{\mu P}$, $Q_{1,var}(X) = \frac{L^2\eta}{\mu} + \frac{P}{X\eta\mu}$, $Q_{2,var}(X) = \frac{M^2XP\eta^2\sigma^2}{\mu^2}$.

In the following, we use the \lesssim 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 $t \in [C]$,*

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

$$\mathbb{E} \left[\|F''(\mathbf{w}^*)(\overline{\mathbf{w}}^C - \mathbf{w}^*)\|^2 \right] \lesssim \frac{\|\mathbf{w}^0 - \mathbf{w}^*\|^2}{\eta^2 C^2} Q_{bias} + \frac{\sigma^2}{T} \left(1 + \frac{Q_{1,var}(C)}{P} + \frac{Q_{2,var}(C)}{P^2} \right). \quad (5)$$

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

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

$$\mathbb{E} \left[\|F''(\mathbf{w}^*)(\overline{\mathbf{w}}^C - \mathbf{w}^*)\|^2 \right] \lesssim \frac{\|\mathbf{w}^0 - \mathbf{w}^*\|^2}{\eta^2 N^2} Q_{bias} + \frac{\sigma^2}{T} (1 + Q_{1,var}(N) + Q_{2,var}(N)). \quad (7)$$

Interpretation, fixed P . Using mini-batch naturally reduces the variance of the process $(\mathbf{w}_{p,k}^t)_{p \in [P], t \in [C], k \in [N^t]}$. 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 $\sigma^2 T^{-1}$ as $T \rightarrow \infty$, 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 Appendix A.2.

Interpretation, $P, T \rightarrow \infty$. 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 $\|F''(\mathbf{w}^*)(\bar{\mathbf{w}}^C - \mathbf{w}^*)\|^2$: it is the natural quantity in such a setting [10, 15, 27]. These results could be translated into function value convergence $F(\bar{\mathbf{w}}^C) - F(\mathbf{w}^*)$, using the inequality $F(\bar{\mathbf{w}}^C) - F(\mathbf{w}^*) \leq L\mu^{-2} \|F''(\mathbf{w}^*)(\bar{\mathbf{w}}^C - \mathbf{w}^*)\|^2$ 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 setting³, but bounds are much harder to read see 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.

²A 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 $1/N^2 + 1/PN$).

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

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 $\hat{\mathbf{w}}^t$ (i.e., the restart point of the next phase), and on the second order moment of any iterate. For $t \in [C]$, we denote $N_1^t := \sum_{t'=1}^t N^{t'}$.

Proposition 3 (Local-SGD: Quadratic Functions with Bounded Noise) *Under Assumptions Q 1, A3, A4, we have the following bound for local-SGD: for any $p \in [P], t \in [C], k \in [N^t]$,*

$$\begin{aligned} \mathbb{E} \left[\|\hat{\mathbf{w}}^{t-1} - \mathbf{w}^*\|^2 \right] &\leq (1 - \eta\mu)^{N_1^{t-1}} \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{\sigma_\infty^2 \eta}{P} \frac{1 - (1 - \eta\mu)^{N_1^{t-1}}}{\mu} \\ \mathbb{E} \left[\|\mathbf{w}_{p,k}^t - \mathbf{w}^*\|^2 \right] &\leq (1 - \eta\mu)^{N_1^{t-1} + k} \|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \sigma_\infty^2 \eta \left(\underbrace{\frac{1 - (1 - \eta\mu)^{N_1^{t-1}}}{P\mu}}_{\text{long term reduced variance}} + \underbrace{\frac{1 - (1 - \eta\mu)^k}{\mu}}_{\text{local iteration variance}} \right). \end{aligned}$$

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

Interpretation. The variance bound for the iterates “just after” communication, $\hat{\mathbf{w}}^t$ 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 $\mathbf{w}_{p,k}^t$ 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 $\mathbf{w}_{p,k}^t$ admits the same upper bound as the mini-batch iterate $\hat{\mathbf{w}}_{MB}^{N_1^{t-1} + k}$ (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 $\bar{\mathbf{w}}^C$ 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 variance⁴ rate, local-SGD communicates $O(\sqrt{NC}/(P\mu))$ times less as compared to mini-batch averaging.*

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

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 **A4** and the quadratic assumption **Q1**, 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{w}^C is as good as in the mini-batch case, up to a constant:*

- (i) Assume **Q1**, **A3**, **A5**, **A6**, and for any $t \in [C]$, $N^t \leq (\mu\eta P)^{-1}$ and $\mu\eta^2 N_1^t = O(1)$.
- (ii) Assume **A1**, **A2**, **A3**, **A4**, and for any $t \in [C]$, $N^t \leq \inf((\eta P M \mathbb{E}[\|\hat{w}^t - w^*\|])^{-1}, (\mu\eta P)^{-1})$.

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 $\exp(\mu\eta^2 N_1^t)$, 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 $\mathbb{E}[\|\hat{w}^t - w^*\|]$ 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 $\mathbb{E}[\|\hat{w}^t - w^*\|]$ 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.

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 approximation 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 30th 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 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.