

Slow and Stale Gradients Can Win the Race: Error-Runtime Trade-offs in Distributed SGD

Sanghamitra Dutta¹, Gauri Joshi¹, Soumyadip Ghosh².

Parijat Dube², and Priya Nagpurkar²

¹Carnegie Mellon University, ²IBM Research

May 11, 2018

Abstract

Distributed Stochastic Gradient Descent (SGD) when run in a synchronous manner, suffers from delays in waiting for the slowest learners (stragglers). Asynchronous methods can alleviate stragglers, but cause gradient staleness that can adversely affect convergence. In this work we present a novel theoretical characterization of the speed-up offered by asynchronous methods by analyzing the trade-off between the error in the trained model and the actual training runtime (wallclock time). The novelty in our work is that our runtime analysis considers random straggler delays, which helps us design and compare distributed SGD algorithms that strike a balance between stragglers and staleness. We also present a new convergence analysis of asynchronous SGD variants without bounded or exponential delay assumptions, and a novel learning rate schedule to compensate for gradient staleness.

1 INTRODUCTION

Stochastic gradient descent (SGD) is the backbone of most state-of-the-art machine learning algorithms. Thus, improving the stability and convergence rate of SGD algorithms is critical for making machine learning algorithms fast and efficient.

Traditionally SGD is run serially at a single node. However, for massive datasets, running SGD serially at a single server can be *prohibitively* slow. A solution that has proved successful in recent years is to parallelize the training across many learners (processing units). This method was first used at a large-scale in Google’s DistBelief [1] which used a central parameter server (PS) to aggregate gradients computed by learner nodes. While parallelism dramatically speeds up training, distributed machine learning frameworks face several challenges such as:

Straggling Learners. In synchronous SGD, the PS waits for all learners to push gradients before it updates the model parameters. Random delays in computation (referred to as straggling)

S. Dutta and G. Joshi are with the Department of Electrical and Computer Engineering, Carnegie Mellon University. S. Ghosh, P. Dube and P. Nagpurkar are with IBM Research. This work was done when G. Joshi was a research staff member at IBM Research and S. Dutta was an intern. Author Contacts: S. Dutta (sanghamd@andrew.cmu.edu), G. Joshi (gaurij@andrew.cmu.edu), S. Ghosh (ghoshs@us.ibm.com), P. Dube (pdube@us.ibm.com) and P. Nagpurkar (pnagpurkar@us.ibm.com).

Presented at the International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain.

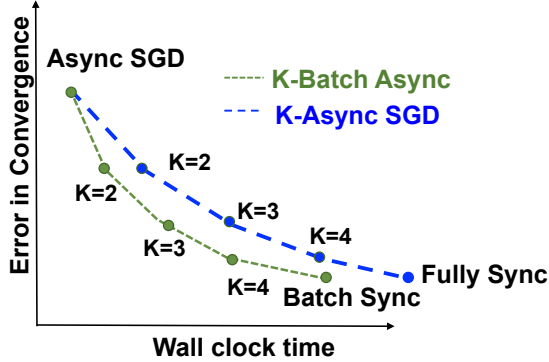


Figure 1: SGD variants span the error-runtime trade-off between fully Sync-SGD and fully Async-SGD. K is the number of learners or mini-batches the PS waits for before updating the model parameters, as we elaborate in Section 2.

are common in today’s distributed systems [2]. Waiting for slow and straggling learners can diminish the speed-up offered by parallelizing the training.

Gradient Staleness. To alleviate the problem of stragglers, SGD can be run in an asynchronous manner, where the central parameters are updated without waiting for all learners. However, learners may return *stale* gradients that were evaluated at an older version of the model, and this can make the algorithm unstable.

The key contributions of this work are:

1. Most SGD algorithms optimize the trade-off between training error, and the number of iterations or epochs. However, the wallclock time per iteration is a random variable that depends on the gradient aggregation algorithm. We present a rigorous analysis of the trade-off between error and the actual runtime (instead of iterations), *modelling runtimes as random variables with a general distribution*. This analysis is then used to compare different SGD variants such as K -sync SGD, K -async SGD and K -batch-async SGD, as illustrated in Figure 1.
2. We present a new convergence analysis of asynchronous SGD and some of its variants, where we relax several commonly made assumptions such as bounded delays and gradients, exponential service times, and independence of the staleness process.
3. We propose a novel learning rate schedule to compensate for gradient staleness, and improve the stability and convergence of asynchronous SGD, while preserving its fast runtime.

1.1 RELATED WORKS

Single Node SGD: Analysis of gradient descent dates back to classical works [3] in the optimization community. The problem of interest is the minimization of empirical risk of the form:

$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N f(\mathbf{w}, \xi_n) \right\}. \quad (1)$$

Here, ξ_n denotes the n -th data point and its label where $n = 1, 2, \dots, N$, and $f(\mathbf{w}, \xi_n)$ denotes the composite loss function. Gradient descent is a way to iteratively minimize this objective function by

updating the parameter \mathbf{w} in the opposite direction of the gradient of $F(\mathbf{w})$ at every iteration, as given by:

$$\mathbf{w}_{j+1} = \mathbf{w}_j - \eta \nabla F(\mathbf{w}_j) = \mathbf{w}_j - \frac{\eta}{N} \sum_{n=1}^N \nabla f(\mathbf{w}_j, \xi_n).$$

The computation of $\sum_{n=1}^N \nabla f(\mathbf{w}_j, \xi_n)$ over the entire dataset is expensive. Thus, stochastic gradient descent [4] with mini-batching is generally used in practice, where the gradient is evaluated over small, randomly chosen subsets of the data. Smaller mini-batches result in higher variance of the gradients, which affects convergence and error floor [5–7]. Algorithms such as AdaGrad [8] and Adam [9] gradually reduce learning rate to achieve a lower error floor. Another class of algorithms includes stochastic variation reduction techniques that include SVRG [10], SAGA [11] and their variants listed out in [12]. For a detailed survey of different SGD variants, refer to [13].

Synchronous SGD and Stragglers: To process large datasets, SGD is parallelized across multiple learners with a central PS. Each learner processes one mini-batch, and the PS aggregates all the gradients. The convergence of synchronous SGD is same as mini-batch SGD, with a P -fold larger mini-batch, where P is the number of learners. However, the time per iteration grows with the number of learners, because some straggling learners that slow down randomly [2]. Thus, it is important to juxtapose the error reduction per iteration with the runtime per iteration to understand the true convergence speed of distributed SGD.

To deal with stragglers and speed up machine learning, system designers have proposed several straggler mitigation techniques such as [14] that try to detect and avoid stragglers. An alternate direction of work is to use redundancy techniques, e.g., replication or erasure codes, as proposed in [15–35] to deal with the stragglers, as also discussed in Remark 1.

Asynchronous SGD and Staleness: A complementary approach to deal with the issue of straggling is to use asynchronous SGD. In asynchronous SGD, any learner can evaluate the gradient and update the central PS without waiting for the other learners. Asynchronous variants of existing SGD algorithms have also been proposed and implemented in systems [1, 36–39].

In general, analyzing the convergence of asynchronous SGD with the number of iterations is difficult in itself because of the randomness of gradient staleness. There are only a few pioneering works such as [40–52] in this direction. In [40], a fully decentralized analysis was proposed that considers no central PS. In [43], a new asynchronous algorithm called Hogwild was proposed and analyzed under bounded gradient and bounded delay assumptions. This direction of research has been followed upon by several interesting works such as [41] which proposed novel theoretical analysis under bounded delay assumption for other asynchronous SGD variants. In [48–51], the framework of ARock was proposed for parallel co-ordinate descent and analyzed using Lyapunov functions, relaxing several existing assumptions such as bounded delay assumption and the independence of the delays and the index of the blocks being updated. In algorithms such as Hogwild, ARock etc. every learner only updates a part of the central parameter vector \mathbf{w} at every iteration and are thus essentially different in spirit from conventional asynchronous SGD settings [41, 44] where every learner updates the entire \mathbf{w} . In an alternate direction of work [45], asynchrony is modelled as a perturbation.

1.2 OUR CONTRIBUTIONS

Existing machine learning algorithms mostly try to optimize the trade-off of error with the number of iterations, epochs or “work complexity” [7]. Time to complete a task has traditionally been

calculated in terms of work complexity measures [53], where the time taken to complete a task is a deterministic function of the size of the task (number of operations). However, due to straggling and synchronization bottle-necks in the system, the same task can often take different time to compute across different learners or iterations. We bring statistical perspective to the traditional work complexity analysis that incorporates the randomness introduced due to straggling. In this paper, we provide a systematic approach to analyze the expected error with runtime for both synchronous and asynchronous SGD, and some variants like K -sync, K -batch-sync, K -async and K -batch-async SGD by modelling the runtimes at each learner as i.i.d. random variables with a general distribution.

We also propose a new error convergence analysis for Async and K -async SGD that holds for strongly convex objectives and can also be extended to non-convex formulations. In this analysis we relax the bounded delay assumption in [41] and the bounded gradient assumption in [43]. We also remove the assumption of exponential computation time and the staleness process being independent of the parameter values [42] as we will elaborate in Section 3.2. Interestingly, our analysis also brings out the regimes where asynchrony can be better or worse than synchrony in terms of speed of convergence. Further, we propose a new learning rate schedule to compensate for staleness, and stabilize asynchronous SGD that is related but different from momentum tuning in [42, 54] as we clarify in Remark 2.

The rest of the paper is organized as follows. Section 2 describes our problem formulation introducing the system model and assumptions. Section 3 provides the main results of the paper – analytical characterization of expected runtime and new convergence analysis for Async and K -async SGD and the proposed learning rate schedule to compensate for staleness. The analysis of expected runtime is elaborated further in Section 4. Proofs and detailed discussions are presented in the Appendix.

2 PROBLEM FORMULATION

Our objective is to minimize the risk function of the parameter vector \mathbf{w} as mentioned in (1) given N training samples. Let S denote the total set of N training samples, *i.e.*, a collection of some data points with their corresponding labels or values. We use the notation ξ to denote a random seed $\in S$ which consists of either a single data and its label or a single mini-batch (m samples) of data and their labels.

2.1 SYSTEM MODEL

We assume that there is a central parameter server (PS) with P parallel learners as shown in Figure 2. The learners fetch the current parameter vector \mathbf{w}_j from the PS as and when instructed in the algorithm. Then they compute gradients using one mini-batch and push their gradients back to the PS as and when instructed in the algorithm. At each iteration, the PS aggregates the gradients computed by the learners and updates the parameter \mathbf{w} . Based on how these gradients are fetched and aggregated, we have different variants of synchronous or asynchronous SGD.

The time taken by a learner to compute gradient of one mini-batch is denoted by random variable X_i for $i = 1, 2, \dots, P$. We assume that the X_i s are i.i.d. across mini-batches and learners.

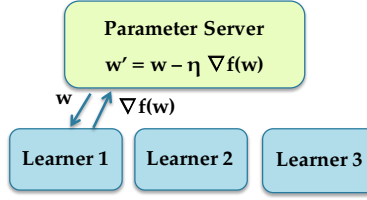


Figure 2: Parameter Server Model

2.2 PERFORMANCE METRICS

There are two metrics of interest: Expected Runtime and Error.

Definition 1 (Expected Runtime per iteration). *The expected runtime per iteration is the expected time (average time) taken to perform each iteration, i.e., the expected time between two consecutive updates of the parameter \mathbf{w} at the central PS.*

Definition 2 (Expected Error). *The expected error after j iterations is defined as $\mathbb{E}[F(\mathbf{w}_j) - F^*]$, the expected gap of the risk function from its optimal value.*

Our aim is to determine the trade-off between the expected error (measures the accuracy of the algorithm) and the expected runtime **after a total of J iterations** for the different SGD variants.

2.3 VARIANTS OF SGD

We now describe the SGD variants considered in this paper. Please refer to Figure 3 and Figure 4 for a pictorial illustration.

K -sync SGD: This is a generalized form of synchronous SGD, also suggested in [36, 55] to offer some resilience to straggling as the PS does not wait for all the learners to finish. The PS only waits for the first K out of P learners to push their gradients. Once it receives K gradients, it updates \mathbf{w}_j and cancels the remaining learners. The updated parameter vector \mathbf{w}_{j+1} is sent to all P learners for the next iteration. The update rule is given by:

$$\mathbf{w}_{j+1} = \mathbf{w}_j - \frac{\eta}{K} \sum_{l=1}^K g(\mathbf{w}_j, \xi_{l,j}). \quad (2)$$

Here $l = 1, 2, \dots, K$ denotes the index of the K learners that finish first, $\xi_{l,j}$ denotes the mini-batch of m samples used by the l -th learner at the j -th iteration and $g(\mathbf{w}_j, \xi_{l,j}) = \frac{1}{m} \sum_{\xi \in \xi_{l,j}} \nabla f(\mathbf{w}_j, \xi)$ denotes the average gradient of the loss function evaluated over the mini-batch $\xi_{l,j}$ of size m . For $K = P$, the algorithm is exactly equivalent to a fully synchronous SGD with P learners.

K -batch-sync: In K -batch-sync, all the P learners start computing gradients with the same \mathbf{w}_j . Whenever any learner finishes, it pushes its update to the PS and evaluates the gradient on the next mini-batch at the same \mathbf{w}_j . The PS updates using the first K mini-batches that finish and cancels the remaining learners. Theoretically, the update rule is still the same as (2) but here l now denotes the index of the mini-batch (out of the K mini-batches that finished first) instead of the learner. However K -batch-sync will offer advantages over K -sync in runtime per iteration as no learner is idle.

K -async SGD: This is a generalized version of asynchronous SGD, also suggested in [36]. In K -async SGD, all the P learners compute their respective gradients on a single mini-batch. The PS

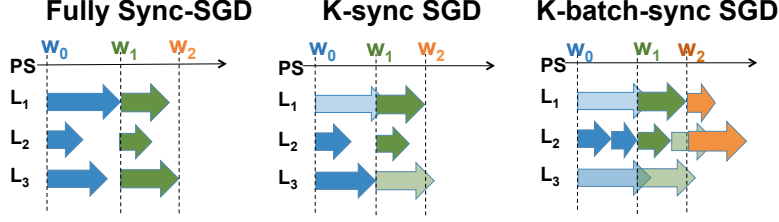


Figure 3: For $K = 2$ and $P = 3$, we illustrate the K -sync and K -batch-sync SGD in comparison with fully synchronous SGD. Lightly shaded arrows indicate straggling gradient computations that are cancelled.

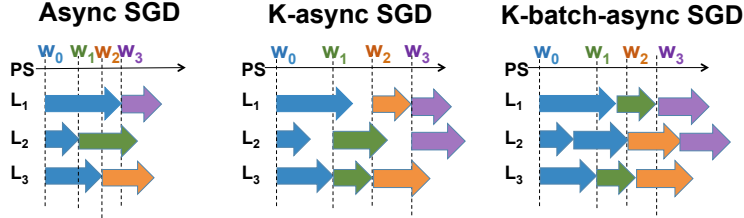


Figure 4: For $K = 2$ and $P = 3$, we illustrate the K -async and K -batch-async algorithms in comparison with fully asynchronous SGD.

waits for the first K out of P that finish first, but it does not cancel the remaining learners. As a result, for every update the gradients returned by each learner might be computed at a stale or older value of the parameter \mathbf{w} . The update rule is thus given by:

$$\mathbf{w}_{j+1} = \mathbf{w}_j - \frac{\eta}{K} \sum_{l=1}^K g(\mathbf{w}_{\tau(l,j)}, \xi_{l,j}). \quad (3)$$

Here $l = 1, 2, \dots, K$ denotes the index of the K learners that contribute to the update at the corresponding iteration, $\xi_{l,j}$ is one mini-batch of m samples used by the l -th learner at the j -th iteration and $\tau(l, j)$ denotes the iteration index when the l -th learner last read from the central PS where $\tau(l, j) \leq j$. Also, $g(\mathbf{w}_{\tau(l,j)}, \xi_{l,j}) = \frac{1}{m} \sum_{\xi \in \xi_{l,j}} \nabla f(\mathbf{w}_{\tau(l,j)}, \xi_{l,j})$ is the average gradient of the loss function evaluated over the mini-batch $\xi_{l,j}$ based on the stale value of the parameter $\mathbf{w}_{\tau(l,j)}$. For $K = 1$, the algorithm is exactly equivalent to fully asynchronous SGD, and the update rule can be simplified as:

$$\mathbf{w}_{j+1} = \mathbf{w}_j - \eta g(\mathbf{w}_{\tau(j)}, \xi_j). \quad (4)$$

Here ξ_j denotes the set of samples used by the learner that updates at the j -th iteration such that $|\xi_j| = m$ and $\tau(l, j)$ denotes the iteration index when that particular learner last read from the central PS. Note that $\tau(j) \leq j$.

K -batch-async: Observe in Figure 4 that K -async also suffers from some learners being idle while others are still working on their gradients until any K finish. In K -batch-async (proposed in [41]), the PS waits for K mini-batches before updating itself but irrespective of which learner they come from. So wherever any learner finishes, it pushes its gradient to the PS, fetches current parameter at PS and starts computing gradient on the next mini-batch based on the current value of the PS. Surprisingly, the update rule is again similar to (3) theoretically except that now l denotes

the indices of the K mini-batches that finish first instead of the learners and $\mathbf{w}_{\tau(l,j)}$ denotes the version of the parameter when the learner computing the l -th mini-batch last read from the PS. While the error convergence of K -batch-async is similar to K -async, it reduces the runtime per iteration as no learner is idle.

Remark 1. Recent works such as [20] propose erasure coding techniques to overcome straggling learners. Instead, the SGD variants considered in this paper such as K -sync and K -batch-sync SGD exploit the inherent redundancy in the data itself, and ignore the gradients returned by straggling learners. If the data is well-shuffled such that it can be assumed to be i.i.d. across learners, then for the same effective batch-size, ignoring straggling gradients will give equivalent error scaling as coded strategies, and at a lower computing cost. However, coding strategies may be useful in the non i.i.d. case, when the gradients supplied by each learner provide diverse information that is important to capture in the trained model.

2.4 ASSUMPTIONS

Closely following [7], we also make the following assumptions:

1. $F(\mathbf{w})$ is an L -smooth function. Thus,

$$\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\|_2 \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|_2. \quad (5)$$

2. $F(\mathbf{w})$ is strongly convex with parameter c . Thus,

$$2c(F(\mathbf{w}) - F^*) \leq \|\nabla F(\mathbf{w})\|_2^2 \quad \forall \mathbf{w}. \quad (6)$$

Refer to Appendix A for discussion on strong convexity. Our results also extend to non-convex objectives, as discussed in Section 3.

3. The stochastic gradient is an unbiased estimate of the true gradient:

$$\mathbb{E}_{\xi_j|\mathbf{w}_k} [g(\mathbf{w}_k, \xi_j)] = \nabla F(\mathbf{w}_k) \quad \forall k \leq j. \quad (7)$$

Observe that this is slightly different from the common assumption that says $\mathbb{E}_{\xi_j} [g(\mathbf{w}, \xi_j)] = \nabla F(\mathbf{w})$ for all \mathbf{w} . Observe that all \mathbf{w}_j for $j > k$ is actually not independent of the data ξ_j . We thus make the assumption more rigorous by conditioning on \mathbf{w}_k for $k \leq j$. Our requirement $k \leq j$ means that \mathbf{w}_k is the value of the parameter at the PS before the data ξ_j was accessed and can thus be assumed to be independent of the data ξ_j .

4. Similar to the previous assumption, we also assume that the variance of the stochastic update given \mathbf{w}_k at iteration k before the data point was accessed is also bounded as follows:

$$\mathbb{E}_{\xi_j|\mathbf{w}_k} [\|g(\mathbf{w}_k, \xi_j) - \nabla F(\mathbf{w}_k)\|_2^2] \leq \frac{\sigma^2}{m} + \frac{M_G}{m} \|\nabla F(\mathbf{w}_k)\|_2^2 \quad \forall k \leq j. \quad (8)$$

In the following Table 1, we provide a list of the notations used in this paper for referencing.

CONSTANTS		RANDOM VARIABLES	
Mini-batch Size	m	Runtime of a learner for one mini-batch	X_i
Total Iterations	J	Runtime per iteration	T
Number of learners (Processors)	P		
Number of learners to wait for	K		
Learning rate	η		
Lipschitz Constant	L		
Strong-convexity parameter	c		

Table 1: LIST OF NOTATIONS

3 MAIN RESULTS

3.1 RUNTIME ANALYSIS

We compare the theoretical wall clock runtime of the different SGD variants to illustrate the speed-up offered by different asynchronous and batch variants. A detailed discussion is provided in Section 4.

Theorem 1. *Let the wall clock time of each learner to process a single mini-batch be i.i.d. random variables X_1, X_2, \dots, X_P . Then the ratio of the expected runtimes per iteration for synchronous and asynchronous SGD is*

$$\frac{\mathbb{E}[T_{Sync}]}{\mathbb{E}[T_{Async}]} = P \frac{\mathbb{E}[X_{P:P}]}{\mathbb{E}[X]}$$

where $X_{(P:P)}$ is the P^{th} order statistic of P i.i.d. random variables X_1, X_2, \dots, X_P .

This result analytically characterizes the speed-up offered by asynchronous SGD for *any general distribution on the wall clock time of each learner*. To prove this result, we use ideas from renewal theory, as we discuss in Section 4. In the following corollary, we highlight this speed-up for the special case of exponential computation time.

Corollary 1. *Let the wall clock time of each learner to process a single mini-batch be i.i.d. exponential random variables $X_1, X_2, \dots, X_P \sim \exp(\mu)$. Then the ratio of the expected runtimes per iteration for synchronous and asynchronous SGD is approximately given by $P \log P$.*

Thus, the speed-up scales with P and can diverge to infinity for large P . We illustrate the speed-up for different distributions in Figure 5. It might be noted that a similar speed-up as Corollary 1 has also been obtained in a recent work [49] under exponential assumptions.

The next result illustrates the advantages offered by K -batch-sync and async over their corresponding counterparts K -sync and K -async respectively.

Theorem 2. *Let the wall clock time of each learner to process a single mini-batch be i.i.d. exponential random variables $X_1, X_2, \dots, X_P \sim \exp(\mu)$. Then the ratio of the expected runtimes per iteration for K -async (or sync) SGD and K -batch-async (or sync) SGD is*

$$\frac{\mathbb{E}[T_{K-async}]}{\mathbb{E}[T_{K-batch-async}]} = \frac{P\mathbb{E}[X_{K:P}]}{K\mathbb{E}[X]} \approx \frac{P \log \frac{P}{P-K}}{K}$$

where $X_{K:P}$ is the K^{th} order statistic of i.i.d. random variables X_1, X_2, \dots, X_P .

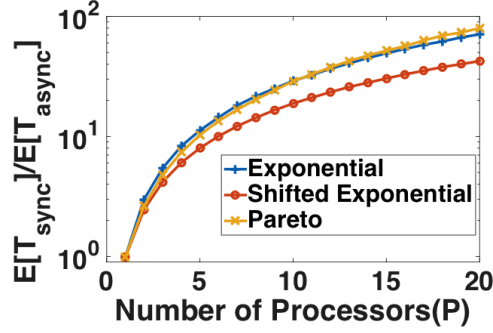


Figure 5: Plot of the speed-up using asynchronous over synchronous: $\log \frac{\mathbb{E}[T_{\text{sync}}]}{\mathbb{E}[T_{\text{async}}]}$ with P for different distributions - $\exp(1)$, $1 + \exp(1)$ and $\text{Pareto}(2, 1)$.

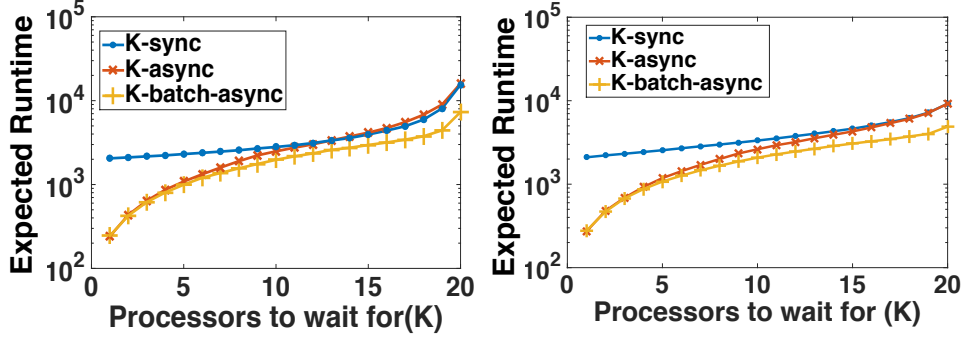


Figure 6: Plot of expected runtime for 2000 iterations: (Left) Pareto distribution $\text{Pareto}(2, 1)$ and (Right) Shifted exponential distribution $1 + \exp(1)$.

To prove this, we derive an exact expression (see Lemma 5 in Section 4) for the expected runtime of K -batch-async SGD, for *any* given i.i.d. distribution of X_i s, not necessarily exponential. The expected runtime per iteration is obtained as $\frac{K\mathbb{E}[X]}{P}$, using ideas from renewal theory. The full proof of Theorem 2 is also provided in Section 4.

Theorem 2 shows that as $\frac{K}{P}$ increases, the speed-up using K -batch-async increases and can be upto $\log P$ times higher. For non-exponential distributions, we simulate the behaviour of expected runtime in Figure 6 for K -sync, K -async and K -batch-async respectively for Pareto and Shifted Exponential.

3.2 ERROR ANALYSIS UNDER FIXED LEARNING RATE

Theorem 3 below gives a convergence analysis of K -async SGD for fixed η , relaxing the following assumptions in existing literature.

- In several prior works such as [19, 21, 42, 49], it is often assumed, for the ease of analysis, that runtimes are exponentially distributed. In this paper, we extend our analysis for any general service time X_i .
- In [42], it is also assumed that the staleness process is independent of \mathbf{w} . While this assumption simplifies the analysis greatly, it is not true in practice. For instance, for a two learner case, the

parameter \mathbf{w}_2 after 2 iterations depends on whether the update from \mathbf{w}_1 to \mathbf{w}_2 was based on a stale gradient at \mathbf{w}_0 or the current gradient at \mathbf{w}_1 , depending on which learner finished first. In this work, we remove this independence assumption.

- Instead of the bounded delay assumption in [41], we use a general staleness bound

$$\mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \leq \gamma \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2]$$

which allows for large, but rare delays.

- In [43], the norm of the gradient is assumed to be bounded. However, if we assume that $\|\nabla F(\mathbf{w})\|_2^2 \leq M$ for some constant M , then using (6) we obtain $\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \frac{2}{c}(F(\mathbf{w}) - F^*) \leq \frac{M}{c^2}$ implying that \mathbf{w} itself is bounded which is a very strong and restrictive assumption, that we relax in this result.

Some of these assumptions have been addressed in the context of alternative asynchronous SGD variants in the recent works of [49–52].

Theorem 3. Suppose the objective $F(\mathbf{w})$ is c -strongly convex and the learning rate $\eta \leq \frac{1}{2L(\frac{M_G}{K^m} + \frac{1}{K})}$. Also assume that for some $\gamma \leq 1$,

$$\mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \leq \gamma \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2].$$

Then, the error of K -async SGD after J iterations is,

$$\mathbb{E} [F(\mathbf{w}_J)] - F^* \leq \frac{\eta L \sigma^2}{2c\gamma' K m} + (1 - \eta c \gamma')^J \left(\mathbb{E} [F(\mathbf{w}_0)] - F^* - \frac{\eta L \sigma^2}{2c\gamma' K m} \right) \quad (9)$$

where $\gamma' = 1 - \gamma + \frac{p_0}{2}$ and p_0 is a lower bound on the conditional probability that $\tau(l, j) = j$, given all the past delays and parameters.

Here, γ is a measure of staleness of the gradients returned by learners; smaller γ indicates less staleness.

The full proof is provided in Appendix C. We first prove the result for $K = 1$ in Appendix C.1 for ease of understanding, and then provide the more general proof for any K in Appendix C.2. We use Lemma 1 below to prove Theorem 3.

Lemma 1. Suppose that $p_0^{(l,j)}$ is the conditional probability that $\tau(l, j) = j$ given all the past delays and all the previous \mathbf{w} , and $p_0 \leq p_0^{(j)}$ for all j . Then,

$$\mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \geq p_0 \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2]. \quad (10)$$

Proof. By the law of total expectation,

$$\begin{aligned} \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] &= p_0^{(l,j)} \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2 | \tau(j) = j] + (1 - p_0^{(l,j)}) \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2 | \tau(j) \neq j] \\ &\geq p_0 \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2]. \end{aligned}$$

□

For the exponential distribution, p_0 is equal to $\frac{1}{P}$ as we discuss in Lemma 2. For non-exponential distributions, it is a constant in $[0, 1]$. For some special classes of distributions like new-longer-than-used (new-shorter-than-used) as defined in Definition 3, we can formally show that p_0 lies in $[0, \frac{1}{P}]$ ($[\frac{1}{P}, 1]$) respectively. The following Lemma 2 below provides bounds on p_0 .

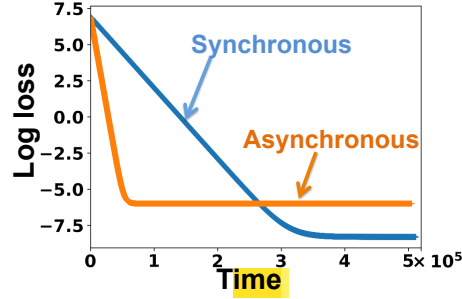


Figure 7: Theoretical error-runtime trade-off for Sync and Async-SGD with same η . Async-SGD has faster decay with time but a higher error floor.

Lemma 2 (Bounds on p_0). Define $p_0 = \inf_j p_0^{(j)}$, i.e. the largest constant such that $p_0 \leq p_0^{(j)} \forall j$.

- For exponential computation times, $p_0^{(j)} = \frac{1}{P}$ for all j and is thus invariant of j and $p_0 = \frac{1}{P}$.
- For new-longer-than-used (See Definition 3) computation times, $p_0^{(j)} \leq \frac{1}{P}$ and thus $p_0 \leq \frac{1}{P}$.
- For new-shorter-than-used computation times, $p_0^{(j)} \geq \frac{1}{P}$ and thus $p_0 \geq \frac{1}{P}$.

The proof is provided in Appendix C.1.1.

For K -batch-async, the update rule is same as K -async except that the index l denotes the index of the mini-batch. Thus, the error analysis will be exactly similar. Our analysis can also be extended to non-convex $F(\mathbf{w})$ as we show in Appendix C.2.1.

Now let us compare with K -sync SGD. We observe that the analysis of K -sync SGD is same as serial SGD with mini-batch size Km . Thus,

Lemma 3 (Error of K -sync). [7] Suppose that the objective $F(\mathbf{w})$ is c -strongly convex and learning rate $\eta \leq \frac{1}{2L(\frac{M_G}{Km} + 1)}$. Then, the error after J iterations of K -sync SGD is

$$\mathbb{E}[F(\mathbf{w}_J) - F^*] \leq \frac{\eta L \sigma^2}{2c(Km)} + (1 - \eta c)^J \left(F(\mathbf{w}_0) - F^* - \frac{\eta L \sigma^2}{2c(Km)} \right).$$

Can stale gradients win the race? For the same η , observe that the error given by Theorem 3 decays at the rate $(1 - \eta c(1 - \gamma + \frac{p_0}{2}))$ for K -async or K -batch-async SGD while for K -sync, the decay rate with number of iterations is $(1 - \eta c)$. Thus, depending on the values of γ and p_0 , the decay rate of K -async or K -batch-async SGD can be faster or slower than K -sync SGD. The decay rate of K -async or K -batch-async SGD is faster if $\frac{p_0}{2} > \gamma$. As an example, one might consider an exponential or new-shorter-than-used service time where $p_0 \geq \frac{1}{P}$ and γ can be made smaller by increasing K . It might be noted that asynchronous SGD can still be faster than synchronous SGD with respect to wall clock time even if its decay rate with respect to number of iterations is lower as every iteration is much faster in asynchronous SGD (Roughly $P \log P$ times faster for exponential service times).

The maximum allowable learning rate for synchronous SGD is $\max\{\frac{1}{c}, \frac{1}{2L(\frac{M_G}{Pm} + 1)}\}$ which can be much higher than that for asynchronous SGD, i.e., $\max\{\frac{1}{c(1 - \gamma + \frac{p_0}{2})}, \frac{1}{2L(\frac{M_G}{m} + 1)}\}$. Similarly the error-floor for synchronous is $\frac{\eta L \sigma^2}{2cPm}$ as compared to asynchronous whose error floor is $\frac{\eta L \sigma^2}{2c(1 - \gamma + \frac{p_0}{2})m}$.

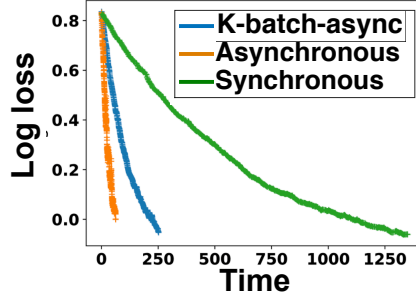


Figure 8: Error-runtime trade-off comparison of different SGD variants for logistic regression on MNIST, with $X_i \sim \exp(1)$, $P = 8$, $K = 4$, $\eta = 0.01$ and $m = 1$. K -batch-async gives intermediate performance, between Async and sync-SGD. (Details of setup provided in Appendix D.)

In Figure 7, we compare the theoretical trade-offs between synchronous ($K = P$ in Lemma 3) and asynchronous SGD ($K = 1$ in Theorem 3). Async-SGD converges very quickly, but to a higher floor. Figure 8 shows the same comparison on the MNIST dataset, along with K -batch-async SGD.

3.3 VARIABLE LEARNING RATE FOR STALENESS COMPENSATION

The staleness of the gradient is random, and can vary across iterations. Intuitively, if the gradient is less stale, we want to weigh it more while updating the parameter \mathbf{w} , and if it is more stale we want to scale down its contribution to the update. With this motivation, we propose the following condition on the learning rate at different iterations.

$$\eta_j \mathbb{E} [\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2^2] \leq C \quad (11)$$

for a constant C . This condition is also inspired from our error analysis in Theorem 3, because it helps remove the assumption $\mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \leq \gamma \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2]$. Using (11), we obtain the following convergence result.

Theorem 4. Suppose the learning rate in the j -th iteration $\eta_j \leq 1/2L(\frac{M_G}{m} + 1)$, and

$$\eta_j \mathbb{E} [\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2^2] \leq C$$

for some constant C . Then, we have

$$\mathbb{E} [F(\mathbf{w}_J)] - F^* \leq \Delta + (\mathbb{E} [F(\mathbf{w}_0)] - F^*) \prod_{j=1}^J (1 - \rho_j)$$

where $\rho_j = \eta_j(1 + \frac{p_0}{2})c$, and the error floor $\Delta = \Delta_J + (1 - \rho_J)\Delta_{J-1} + \dots + \prod_{j=1}^J (1 - \rho_j)\Delta_0$, where $\Delta_j = \frac{\eta_j^2 L \sigma^2}{2m} + \frac{CL^2}{2}$.

The proof is provided in Appendix C.3. In our analysis of Asynchronous SGD, we observe that the term $\frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2]$ is the most difficult to bound. For fixed learning rate, we had assumed that $\mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2]$ is bounded by $\gamma \|\nabla F(\mathbf{w}_j)\|_2^2$. However, if we

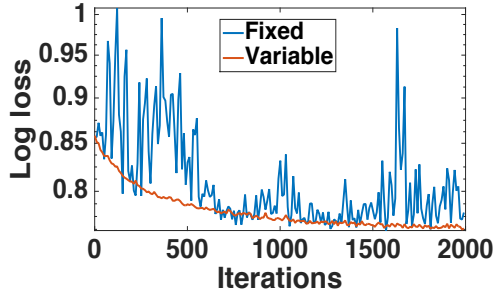


Figure 9: Async-SGD on CIFAR10 dataset, with $X \sim \exp 20$, mini-batch size $m = 250$ and $P = 40$ learners. We compare fixed $\eta = 0.01$, and the variable schedule given in (13) for $\eta_{max} = 0.01$ and $C = 0.005\eta_{max}$. Observe that the proposed schedule can give fast convergence, and also maintain stability, while the fixed η algorithm becomes unstable.

impose the condition (11) on η , we do not require this assumption. Our proposed condition actually provides a bound for the staleness term as follows:

$$\frac{\eta_j}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \leq \frac{\eta_j L^2}{2} \mathbb{E} [\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2^2] \leq \frac{CL^2}{2}. \quad (12)$$

Proposed Algorithmic Modification Inspired by this analysis, we propose the learning rate schedule,

$$\eta_j = \min \left\{ \frac{C}{\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2^2}, \eta_{max} \right\} \quad (13)$$

where η_{max} is a suitably large ceiling on learning rate. It ensures stability when the first term in (13) becomes large due to the staleness $\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2$ being small. The C is chosen of the same order as the desired error floor. To implement this schedule, the PS needs to store the last read model parameters for every learner. In Figure 9 we illustrate how this schedule can stabilize asynchronous SGD. We also show simulation results that characterize the performance of this algorithm in comparison with naive asynchronous SGD with fixed learning rate.

Remark 2. The idea of variable learning rate is related to the idea of momentum tuning in [42, 54] and may have a similar effect of stabilizing the convergence of asynchronous SGD. However, learning rate tuning is arguably more general since asynchrony results in a momentum term in the gradient update (as shown in [42, 54]) only under the assumption that the staleness process is geometric and independent of \mathbf{w} .

4 RUNTIME ANALYSIS

In this section, we provide our analysis of the expected runtime of different variants of SGD. These lemmas are then used in the proofs of Theorem 1 and Theorem 2.

4.1 RUNTIME OF K -SYNC SGD

Lemma 4 (Runtime of K -sync SGD). *The expected runtime per iteration for K -sync SGD is,*

$$\mathbb{E}[T] = \mathbb{E}[X_{K:P}] \quad (14)$$

where $X_{K:P}$ is the K^{th} order statistic of P i.i.d. random variables X_1, X_2, \dots, X_P .

Proof of Lemma 4. We assume that the P learners have an i.i.d. computation times. When all the learners start together, and we wait for the first K out of P i.i.d. random variables to finish, the expected computation time for that iteration is $\mathbb{E}[X_{K:P}]$, where $X_{K:P}$ denotes the K -th statistic of P i.i.d. random variables X_1, X_2, \dots, X_P . \square

Thus, for a total of J iterations, the expected runtime is given by $J\mathbb{E}[X_{K:P}]$.

Remark 3. For $X_i \sim \exp(\mu)$, the expected runtime per iteration is given by,

$$\mathbb{E}[T] = \frac{1}{\mu} \sum_{i=P-K+1}^P \frac{1}{i} \approx \frac{1}{\mu} \left(\frac{\log \frac{P}{P-K}}{\mu} \right)$$

where the last step uses an approximation from [56]. For justification, the reader is referred to Appendix B.1.

4.2 RUNTIME OF K -BATCH-SYNC SGD

The expected runtime of K -batch-sync SGD is not analytically tractable in general, but for $X_i \sim \exp(\mu)$, the runtime per iteration is distributed as $\text{Erlang}(K, P\mu)$. Refer to Appendix B.2 for explanation. Thus, for K -batch-sync SGD, the expected time per iteration is given by,

$$\mathbb{E}[T] = \frac{K}{P\mu}.$$

4.3 RUNTIME OF K -BATCH-ASYNC SGD

Lemma 5 (Runtime of K -batch-async SGD). The expected runtime per iteration for K -batch-async SGD in the limit of large number of iterations is given by:

$$\mathbb{E}[T] = \frac{K\mathbb{E}[X]}{P}. \quad (15)$$

Unlike the results for the synchronous variants, this result on average runtime per iteration holds only in the limit of large number of iterations. To prove the result we use ideas from renewal theory. For a brief background on renewal theory, the reader is referred to Appendix B.3.

Proof of Lemma 5. For the i -th learner, let $\{N_i(t), t > 0\}$ be the number of times the i -th learner pushes its gradient to the PS over in time t . The time between two pushes is an independent realization of X_i . Thus, the inter-arrival times $X_i^{(1)}, X_i^{(2)}, \dots$ are i.i.d. with mean inter-arrival time $\mathbb{E}[X_i]$. Using the elementary renewal theorem [57, Chapter 5] we have,

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}[N_i(t)]}{t} = \frac{1}{\mathbb{E}[X_i]}. \quad (16)$$

Thus, the rate of gradient pushes by the i -th learner is $1/\mathbb{E}[X_i]$. As there are P learners, we have a superposition of P renewal processes and thus the average rate of gradient pushes to the PS is

$$\lim_{t \rightarrow \infty} \sum_{i=1}^P \frac{\mathbb{E}[N_i(t)]}{t} = \sum_{i=1}^P \frac{1}{\mathbb{E}[X_i]} = \frac{P}{\mathbb{E}[X]}. \quad (17)$$

Every K pushes are one iteration. Thus, the expected runtime per iteration or effectively the expected time for K pushes is given by $\mathbb{E}[T] = \frac{K\mathbb{E}[X]}{P}$. \square

Thus, for a total of J iterations, the average runtime can be approximated as $\frac{JK\mathbb{E}[X]}{P}$ when J is large. Note that Fully-Synchronous SGD is actually K -sync SGD with $K = P$, i.e., waiting for all the P learners to finish. On the other hand, Fully-Asynchronous SGD is actually K -batch-async with $K = 1$. Now, we provide the proofs of Theorem 1 and Corollary 1 respectively, that provide a comparison between these two variants.

Proof of Theorem 1. By taking the ratio of the expected runtimes per iteration in Lemma 4 with $K = P$ and Lemma 5 with $K = 1$, we get the result in Theorem 1. \square

Proof of Corollary 1. The expectation of the maximum of P i.i.d. $X_i \sim \exp(\mu)$ is $\mathbb{E}[X_{P:P}] = \sum_{i=1}^P \frac{1}{i\mu} \approx \frac{\log P}{\mu}$ [56]. This can be substituted in Theorem 1 to get Corollary 1. \square

4.4 RUNTIME OF K -ASYNC SGD

The expected runtime per iteration of K -async SGD is not analytically tractable for non-exponential X_i , but we obtain an upper bound on it for a class of distributions called the “new-longer-than-used” distributions, as defined below.

Definition 3 (New-longer-than-used). *A random variable is said to have a new-longer-than-used distribution if the following holds for all $t, u \geq 0$:*

$$\Pr(U > u + t | U > t) \leq \Pr(U > u).$$

Most of the continuous distributions we encounter like normal, exponential, gamma, beta are new-longer-than-used. Alternately, the hyper exponential distribution is new-shorter-than-used and it satisfies $\Pr(U > u + t | U > t) \geq \Pr(U > u)$ for all $t, u \geq 0$.

Lemma 6 (Runtime of K -async SGD). *Suppose that each X_i has a new-longer-than-used distribution. Then, the expected runtime per iteration for K -async is upper-bounded as*

$$\mathbb{E}[T] \leq \mathbb{E}[X_{K:P}] \tag{18}$$

where $X_{K:P}$ is the K^{th} order statistic of P i.i.d. random variables X_1, X_2, \dots, X_P .

The proof of this lemma is provided in Appendix B.4.

We provided a comparison of the expected runtimes of K -async and K -batch-async SGD variants in Theorem 2, for the special case of exponential computation times. Here, we provide the proof of Theorem 2.

Proof of Theorem 2. For the exponential X_i , equality holds in (18) in Lemma 6, as we justify in Appendix B.4.1. The expectation can be derived as $\mathbb{E}[X_{K:P}] = \sum_{i=P-K+1}^P \frac{1}{i\mu} \approx \frac{\log(P/P-K)}{\mu}$. For exponential X_i , the expected runtime per iteration for K -batch-async is given by $\mathbb{E}[T] = \frac{\mu \mathbb{E}[X]}{P} = \frac{K}{\mu P}$ from Lemma 5. \square

In Figure 10, we pictorially illustrate the expected error-runtime trade-offs of K -async with K -batch-async SGD.

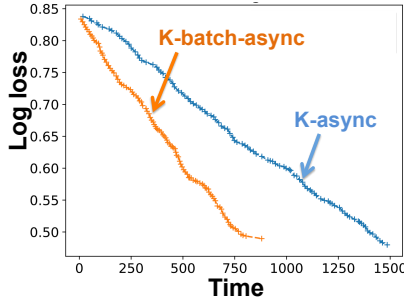


Figure 10: Error-Runtime Trade-off on MNIST Dataset: Comparison of K -async with K -batch-async under exponential computation time with $X_i \sim \exp(1)$. As derived theoretically, the K -batch-async has a sharper fall with time as compared to K -async even though the error floor attained is similar. (Details in Appendix D.)

5 CONCLUSIONS

The speed of distributed SGD depends on the error reduction per iteration, as well as the runtime per iteration. This paper presents a novel runtime analysis of synchronous and asynchronous SGD, and their variants for any general distribution on the wall-clock time of each learner. When juxtaposed with the error analysis, we get error-runtime trade-offs that can be used to compare different SGD algorithms. We also give a new analysis of asynchronous SGD by relaxing some commonly made assumptions and also propose a novel learning rate schedule to compensate for gradient staleness.

In the future we plan to explore methods to gradually increase synchrony, so that we can achieve fast convergence as well as low error floor. We are also looking into the use of local updates to minimize the frequency of communication between the PS and learners, that is closely related to [58–61].

Acknowledgements

The authors thank Mark Wegman, Pulkit Grover and Jianyu Wang for their suggestions and feedback.

References

- [1] Jeffrey Dean et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.
- [2] Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.
- [3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [4] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

- [5] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1):165–202, 2012.
- [6] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. Efficient mini-batch training for stochastic optimization. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 661–670, 2014.
- [7] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.
- [8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2, July 2011.
- [9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.
- [10] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
- [11] Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.
- [12] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.
- [13] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [14] Aaron Harlap, Henggang Cui, Wei Dai, Jinliang Wei, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, and Eric P. Xing. Addressing the straggler problem for iterative convergent parallel ml. In *ACM Symposium on Cloud Computing (SoCC)*, pages 98–111, 2016.
- [15] Gauri Joshi, Yanpei Liu, and Emina Soljanin. On the delay-storage trade-off in content download from coded distributed storage systems. *IEEE Journal on Selected Areas in Communications*, 32(5):989–997, 2014.
- [16] Da Wang, Gauri Joshi, and Gregory Wornell. Using straggler replication to reduce latency in large-scale parallel computing. *ACM SIGMETRICS Performance Evaluation Review*, 43(3):7–11, 2015.
- [17] Gauri Joshi, Emina Soljanin, and Gregory Wornell. Queues with redundancy: Latency-cost analysis. *ACM SIGMETRICS Performance Evaluation Review*, 43(2):54–56, 2015.
- [18] Gauri Joshi, Emina Soljanin, and Gregory Wornell. Efficient redundancy techniques for latency reduction in cloud systems. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 2(12), may 2017.
- [19] Kangwook Lee, Maximilian Lam, Ramtin Pedarsani, Dimitris Papailiopoulos, and Kannan Ramchandran. Speeding up distributed machine learning using codes. *IEEE Transactions on Information Theory*, 64(3):1514–1529, 2018.

- [20] Rashish Tandon, Qi Lei, Alexandros G. Dimakis, and Nikos Karampatziakis. Gradient coding: Avoiding stragglers in distributed learning. In *International Conference on Machine Learning*, pages 3368–3376, 2017.
- [21] Sanghamitra Dutta, Viveck Cadambe, and Pulkit Grover. Short-dot: Computing large linear transforms distributedly using coded short dot products. In *Advances In Neural Information Processing Systems*, pages 2100–2108, 2016.
- [22] Wael Halbawi, Navid Azizan-Ruhi, Fariborz Salehi, and Babak Hassibi. Improving distributed gradient descent using reed-solomon codes. *arXiv preprint arXiv:1706.05436*, 2017.
- [23] Yaoqing Yang, Pulkit Grover, and Soummya Kar. Coded distributed computing for inverse problems. In *Advances in Neural Information Processing Systems*, pages 709–719, 2017.
- [24] Yaoqing Yang, Pulkit Grover, and Soummya Kar. Fault-tolerant distributed logistic regression using unreliable components. In *Communication, Control, and Computing (Allerton)*, pages 940–947. IEEE, 2016.
- [25] Qian Yu, Mohammad Maddah-Ali, and Salman Avestimehr. Polynomial codes: an optimal design for high-dimensional coded matrix multiplication. In *Advances in Neural Information Processing Systems*, pages 4406–4416, 2017.
- [26] Can Karakus, Yifan Sun, and Suhas Diggavi. Encoded distributed optimization. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2890–2894, 2017.
- [27] Can Karakus, Yifan Sun, Suhas Diggavi, and Wotao Yin. Straggler mitigation in distributed optimization through data encoding. In *Advances in Neural Information Processing Systems*, pages 5440–5448, 2017.
- [28] Zachary Charles, Dimitris Papailiopoulos, and Jordan Ellenberg. Approximate gradient coding via sparse random graphs. *arXiv preprint arXiv:1711.06771*, 2017.
- [29] Songze Li, Sucha Supittayapornpong, Mohammad Ali Maddah-Ali, and Salman Avestimehr. Coded terasort. In *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 389–398, 2017.
- [30] Mohammad Fahim, Haewon Jeong, Farzin Haddadpour, Sanghamitra Dutta, Viveck Cadambe, and Pulkit Grover. On the optimal recovery threshold of coded matrix multiplication. In *Communication, Control, and Computing (Allerton)*, pages 1264–1270. IEEE, 2017.
- [31] Min Ye and Emmanuel Abbe. Communication-computation efficient gradient coding. *arXiv preprint arXiv:1802.03475*, 2018.
- [32] Songze Li, Mohammad Ali Maddah-Ali, Qian Yu, and A Salman Avestimehr. A fundamental tradeoff between computation and communication in distributed computing. *IEEE Transactions on Information Theory*, 64(1):109–128, 2018.
- [33] Viveck Cadambe and Pulkit Grover. Codes for Distributed Computing: A Tutorial. *IEEE Information Theory Society Newsletter*, 67(4):3–15, December 2017.
- [34] Sanghamitra Dutta, Ziqian Bai, Haewon Jeong, Tze Meng Low, and Pulkit Grover. A Unified Coded Deep Neural Network Training Strategy based on Generalized PolyDot codes. In *IEEE International Symposium on Information Theory (ISIT) - To Appear*, 2018.

- [35] Ankur Mallick, Malhar Chaudhari, and Gauri Joshi. Rateless codes for near-perfect load balancing in distributed matrix-vector multiplication. *arXiv preprint arXiv:1804.10331*, 2018.
- [36] Suyog Gupta, Wei Zhang, and Fei Wang. Model accuracy and runtime tradeoff in distributed deep learning: A systematic study. In *International Conference on Data Mining*, pages 171–180, 2016.
- [37] James Cipar, Qirong Ho, Jin Kyu Kim, Seunghak Lee, Gregory R. Ganger, Garth Gibson, Kimberly Keeton, and Eric Xing. Solving the straggler problem with bounded staleness. In *Workshop on Hot Topics in Operating Systems*, 2013.
- [38] Henggang Cui, James Cipar, Qirong Ho, Jin Kyu Kim, Seunghak Lee, Abhimanu Kumar, Jinliang Wei, Wei Dai, Gregory R Ganger, Phillip B Gibbons, et al. Exploiting bounded staleness to speed up big data analytics. In *USENIX Annual Technical Conference (ATC)*, pages 37–48, 2014.
- [39] Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B Gibbons, Garth A Gibson, Greg Ganger, and Eric P Xing. More effective distributed ml via a stale synchronous parallel parameter server. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2013.
- [40] John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.
- [41] Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.
- [42] Ioannis Mitliagkas, Ce Zhang, Stefan Hadjis, and Christopher Ré. Asynchrony begets momentum, with an application to deep learning. In *Allerton Conference on Communication, Control, and Computing*, pages 997–1004. IEEE, 2016.
- [43] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.
- [44] Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.
- [45] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.
- [46] Sorathan Chaturapruek, John C Duchi, and Christopher Ré. Asynchronous stochastic convex optimization: the noise is in the noise and sgd don’t care. In *Advances in Neural Information Processing Systems*, pages 1531–1539, 2015.
- [47] Wei Zhang, Suyog Gupta, Xiangru Lian, and Ji Liu. Staleness-aware async-sgd for distributed deep learning. In *International Joint Conference on Artificial Intelligence*, pages 2350–2356. AAAI Press, 2016.

- [48] Zhimin Peng, Yangyang Xu, Ming Yan, and Wotao Yin. Arock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879, 2016.
- [49] Robert Hannah and Wotao Yin. More iterations per second, same quality—why asynchronous algorithms may drastically outperform traditional ones. *arXiv preprint arXiv:1708.05136*, 2017.
- [50] Robert Hannah and Wotao Yin. On unbounded delays in asynchronous parallel fixed-point algorithms. *Journal of Scientific Computing*, pages 1–28, 2016.
- [51] Tao Sun, Robert Hannah, and Wotao Yin. Asynchronous coordinate descent under more realistic assumptions. In *Advances in Neural Information Processing Systems*, pages 6183–6191, 2017.
- [52] Rémi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Asaga: Asynchronous parallel saga. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- [53] Robert Sedgewick and Kevin Wayne. *Algorithms*. Addison-Wesley Professional, 2011.
- [54] Jian Zhang, Ioannis Mitliagkas, and Christopher Ré. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.
- [55] Jianmin Chen, Rajat Monga, Samy Bengio, and Rafal Józefowicz. Revisiting distributed synchronous SGD. *CoRR*, abs/1604.00981, 2016.
- [56] Ross Sheldon. *A first course in probability*. Pearson Education India, 2002.
- [57] Robert G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 1st edition, 2013.
- [58] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.
- [59] Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity empowers distributed learning. *arXiv preprint arXiv:1706.05699*, 2017.
- [60] Fan Zhou and Guojing Cong. On the convergence properties of a k -step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012*, 2017.
- [61] Sixin Zhang, Anna E Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. In *Advances in Neural Information Processing Systems*, pages 685–693, 2015.
- [62] David M Kreps. *A course in microeconomic theory*, volume 41. JSTOR, 1990.
- [63] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [64] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.
- [65] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

A STRONG CONVEXITY DISCUSSION

Definition 4 (Strong-Convexity). *A function $h(\mathbf{u})$ is defined to be c -strongly convex, if the following holds for all \mathbf{u}_1 and \mathbf{u}_2 in the domain:*

$$h(\mathbf{u}_2) \geq h(\mathbf{u}_1) + [\nabla h(\mathbf{u}_1)]^T (\mathbf{u}_2 - \mathbf{u}_1) + \frac{c}{2} \|\mathbf{u}_2 - \mathbf{u}_1\|_2^2$$

For strongly convex functions, the following result holds for all \mathbf{u} in the domain of $h(\cdot)$.

$$2c(h(\mathbf{u}) - h^*) \leq \|\nabla h(\mathbf{u})\|_2^2 \quad (19)$$

The proof is derived in [7]. For completeness, we give the sketch here.

Proof. Given a particular \mathbf{u} , let us define the quadratic function as follows:

$$q(\mathbf{u}') = h(\mathbf{u}) + \nabla h(\mathbf{u})^T (\mathbf{u}' - \mathbf{u}) + \frac{c}{2} \|\mathbf{u}' - \mathbf{u}\|_2^2$$

Now, $q(\mathbf{u}')$ is minimized at $\mathbf{u}' = \mathbf{u} - \frac{1}{c} \nabla h(\mathbf{u})$ and the value is $h(\mathbf{u}) - \frac{1}{2c} \|\nabla h(\mathbf{u})\|_2^2$. Thus, from the definition of strong convexity we now have,

$$\begin{aligned} h^* &\geq h(\mathbf{u}) + \nabla h(\mathbf{u})^T (\mathbf{u}' - \mathbf{u}) + \frac{c}{2} \|\mathbf{u}' - \mathbf{u}\|_2^2 \\ &\geq h(\mathbf{u}) - \frac{1}{2c} \|\nabla h(\mathbf{u})\|_2^2 \quad [\text{minimum value of } q(\mathbf{u}')]. \end{aligned}$$

□

B RUNTIME ANALYSIS PROOFS

Here we provide all the remaining proofs and supplementary information for the results in Section 4.

B.1 Runtime of K -sync SGD

K -th statistic of exponential distributions: Here we give a sketch of why the K -th order statistic of P exponentials scales as $\log(P/P - K)$. A detailed derivation can be obtained in [56]. Consider P i.i.d. exponential distributions with parameter μ . The minimum $X_{1:P}$ of P independent exponential random variables with parameter μ is exponential with parameter $P\mu$. Conditional on $X_{1:P}$, the second smallest value $X_{2:P}$ is distributed like the sum of $X_{1:P}$ and an independent exponential random variable with parameter $(P - 1)\mu$. And so on, until the K -th smallest value $X_{K:P}$ which is distributed like the sum of $X_{(K-1):P}$ and an independent exponential random variable with parameter $(P - K + 1)\mu$. Thus,

$$X_{K:P} = Y_P + Y_{P-1} + \cdots + Y_{P-K+1}$$

where the random variables Y_i s are independent and exponential with parameter $i\mu$. Thus,

$$\mathbb{E}[X_{K:P}] = \sum_{i=P-K+1}^P \frac{1}{i\mu} = \frac{H_P - H_{P-K}}{\mu} \approx \frac{\log \frac{P}{P-K}}{\mu}.$$

Here H_P and H_{P-K} denote the P -th and $(P - K)$ -th harmonic numbers respectively.

For the case where $K = P$, the expectation is given by,

$$\mathbb{E}[X_{P:P}] = \frac{1}{\mu} \sum_{i=1}^P \frac{1}{i} = \frac{1}{\mu} H_P \approx \frac{1}{\mu} \log P.$$

B.2 Runtime of K -batch-sync SGD

In general, the expected runtime per iteration of K -batch-sync SGD is not tractable but for the special case of exponentials it follows the distribution $\text{Erlang}(K, P\mu)$. This is obtained from the memoryless property of exponentials.

All the learners start their computation together. The expected time taken by the first mini-batch to be completed is the minimum of P i.i.d. exponential random variables $X_1, X_2, \dots, X_P \sim \exp(\mu)$ is another exponential random variable distributed as $\exp(P\mu)$. At the time when the first mini-batch is complete, from the memoryless property of exponentials, it may be viewed as P i.i.d. exponential random variables $X_1, X_2, \dots, X_P \sim \exp(\mu)$ starting afresh again. Thus, the time to complete each mini-batch is distributed as $\exp(P\mu)$, and an iteration being the sum of the time to complete K such mini-batches, has the distribution $\text{Erlang}(K, P\mu)$.

B.3 Runtime of K -batch-async SGD

Here we include a discussion on renewal processes for completeness, to provide a background for the proof of Lemma 5, which gives the expected runtime of K -batch-async SGD. The familiar reader can merely skim through this and refer to the proof provided in the main section of the paper in Section 4.

Definition 5 (Renewal Process). *A renewal process is an arrival process where the inter-arrival intervals are positive, independent and identically distributed random variables.*

Lemma 7 (Elementary Renewal Theorem). *[57, Chapter 5] Let $\{N(t), t > 0\}$ be a renewal counting process denoting the number of renewals in time t . Let $\mathbb{E}[Z]$ be the mean inter-arrival time. Then,*

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}[N(t)]}{t} = \frac{1}{\mathbb{E}[Z]}. \quad (20)$$

Observe that for asynchronous SGD or K -batch-async SGD, every gradient push by a learner to the PS can be thought of as an arrival process. The time between two consecutive pushes by a learner follows the distribution of X_i and is independent as computation time has been assumed to be independent across learners and mini-batches. Thus the inter-arrival intervals are positive, independent and identically distributed and hence, the gradient pushes are a renewal process.

B.4 Runtime of K -async SGD

Proof of Lemma 6. For new-longer-than-used distributions observe that the following holds:

$$\Pr(X_i > u + t | X_i > t) \leq \Pr(X_i > u). \quad (21)$$

Thus the random variable $X_i - t | X_i > t$ is thus stochastically dominated by X_i . Now let us assume we want to compute the expected computation time of one iteration of K -async starting at time instant t_0 . Let us also assume that the learners last read their parameter values at time instants t_1, t_2, \dots, t_P respectively where any K of these t_1, t_2, \dots, t_P are equal to t_0 as K out of P learners were updated at time t_0 and the remaining $(P - K)$ of these t_1, t_2, \dots, t_P are $< t_0$. Let Y_1, Y_2, \dots, Y_P be the random variables denoting the computation time of the P learners starting from time t_0 . Thus,

$$Y_i = X_i - (t_0 - t_i) | X_i > (t_0 - t_i) \quad \forall i = 1, 2, \dots, P. \quad (22)$$

Now each of the Y_i s are independent and are stochastically dominated by X_i s.

$$\Pr(Y_i > u) \leq \Pr(X_i > u) \quad \forall i, j = 1, 2, \dots, P. \quad (23)$$

The expectation of the K -th statistic of $\{Y_1, Y_2, \dots, Y_P\}$ is the expected runtime of the iteration. Let us denote $h_K(x_1, x_2, \dots, x_P)$ as the K -th statistic of P numbers (x_1, x_2, \dots, x_P) . And let us denote $g_{K,s}(x)$ as the K -th statistic of P numbers where $P-1$ of them are given as $\mathbf{s}_{1 \times (P-1)}$ and x is the P -th number. Thus

$$g_{K,s}(x) = h_K(x, s(1), s(2), \dots, s(P-1))$$

First observe that $g_{K,s}(x)$ is an increasing function of x since given the other $P-1$ values, the K -th order statistic will either stay the same or increase with x . Now we use the property that if Y_i is stochastically dominated by X_i , then for any increasing function $g(\cdot)$, we have

$$\mathbb{E}_{Y_1} [g(Y_1)] \leq \mathbb{E}_{X_1} [g(X_1)].$$

This result is derived in [62].

This implies that for a given \mathbf{s} ,

$$\mathbb{E}_{Y_1} [g_{K,s}(Y_1)] \leq \mathbb{E}_{X_1} [g_{K,s}(X_1)].$$

This leads to,

$$\begin{aligned} \mathbb{E}_{Y_1|Y_2=s(1), Y_3=s(2), \dots, Y_P=s(P-1)} [h_K(Y_1, Y_2, \dots, Y_P)] \\ \leq \mathbb{E}_{X_1|Y_2=s(1), Y_3=s(2), \dots, Y_P=s(P-1)} [h_K(X_1, Y_2, \dots, Y_P)]. \end{aligned} \quad (24)$$

From this,

$$\begin{aligned} \mathbb{E} [h_K(Y_1, Y_2, \dots, Y_P)] &= \mathbb{E}_{Y_2, \dots, Y_P} [\mathbb{E}_{Y_1|Y_2, Y_3, \dots, Y_P} [h_K(Y_1, Y_2, \dots, Y_P)]] \\ &\leq \mathbb{E}_{Y_2, \dots, Y_P} [\mathbb{E}_{X_1|Y_2, Y_3, \dots, Y_P} [h_K(X_1, Y_2, \dots, Y_P)]] \\ &= \mathbb{E} [h_K(X_1, Y_2, \dots, Y_P)]. \end{aligned} \quad (25)$$

This step proceeds inductively. Thus, similarly

$$\begin{aligned} \mathbb{E} [h_K(X_1, Y_2, \dots, Y_P)] &= \mathbb{E}_{X_1, Y_3, \dots, Y_P} [\mathbb{E}_{Y_2|X_1, Y_3, \dots, Y_P} [h_K(X_1, Y_2, \dots, Y_P)]] \\ &\leq \mathbb{E}_{X_1, Y_3, \dots, Y_P} [\mathbb{E}_{X_2|X_1, Y_3, \dots, Y_P} [h_K(X_1, X_2, Y_3, \dots, Y_P)]] \\ &= \mathbb{E} [h_K(X_1, X_2, Y_3, \dots, Y_P)]. \end{aligned} \quad (26)$$

Thus, finally combining, we have,

$$\begin{aligned} \mathbb{E} [h_K(Y_1, Y_2, \dots, Y_P)] &\leq \mathbb{E} [h_K(X_1, Y_2, \dots, Y_P)] \\ &\leq \mathbb{E} [h_K(X_1, X_2, Y_3, \dots, Y_P)] \\ &\leq \dots \\ &\leq \mathbb{E} [h_K(X_1, X_2, X_3, \dots, X_P)]. \end{aligned} \quad (27)$$

□

B.4.1 Exponential Computation time

For exponential distributions, the inequality in Lemma 6 holds with equality. This follows from the memoryless property of exponentials. Let us consider the scenario of the proof of Lemma 6 where we similarly define $Y_i = X_i - (t_0 - t_i) | X_i > (t_0 - t_i)$. From the memoryless property of exponentials [56], if $X_i \sim \exp(\mu)$, then $Y_i \sim \exp(\mu)$. Thus, the expectation of the K -th statistic of Y_i s can be easily derived as all the Y_i s are now i.i.d. with distribution $\exp(\mu)$. Thus, the expected runtime per iteration is given by,

$$\mathbb{E}[T] = \mathbb{E}[Y_{K:P}] = \frac{1}{\mu} \sum_{i=P-K+1}^P \frac{1}{i} \approx \frac{1}{\mu} \log \frac{P}{P-K}.$$

C ASYNC-SGD ANALYSIS PROOFS

In this section, we provide a proof of the error convergence of asynchronous SGD.

C.1 Async-SGD with fixed learning rate

First we prove a simplified version of Theorem 3 for the case $K = 1$. While this is actually a corollary of the more general Theorem 3, we prove this first for ease of understanding and simplicity. The proof of the more general Theorem 3 is then provided in Appendix C.2.

The corollary is as follows:

Corollary 2. *Suppose that the objective function $F(\mathbf{w})$ is strongly convex with parameter c and the learning rate $\eta \leq \frac{1}{2L(\frac{M_G}{m} + 1)}$. Also assume that $\mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \leq \gamma \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2]$ for some constant $\gamma \leq 1$. Then, the error after J iterations of Async SGD is given by,*

$$\mathbb{E}[F(\mathbf{w}_J)] - F^* \leq \frac{\eta L \sigma^2}{2c\gamma'm} + (1 - \eta c \gamma')^J \left(\mathbb{E}[F(\mathbf{w}_0)] - F^* - \frac{\eta L \sigma^2}{2c\gamma'm} \right),$$

where $\gamma' = 1 - \gamma + \frac{p_0}{2}$ and p_0 is a non-negative lower bound on the conditional probability that $\tau(j) = j$ given all the past delays and parameters.

To prove the result, we will use the following lemma.

Lemma 8. *Let us denote $\mathbf{v}_j = g(\mathbf{w}_{\tau(j)}, \xi_j)$, and assume that $\mathbb{E}_{\xi_j|\mathbf{w}}[g(\mathbf{w}, \xi_j)] = \nabla F(\mathbf{w})$. Then,*

$$\begin{aligned} \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \mathbf{v}_j\|_2^2] &\leq \mathbb{E}[\|\mathbf{v}_j\|_2^2] - \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \\ &\quad + \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2]. \end{aligned}$$

Proof of Lemma 8. Observe that,

$$\begin{aligned} \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \mathbf{v}_j\|_2^2] &= \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)}) + \nabla F(\mathbf{w}_{\tau(j)}) - \mathbf{v}_j\|_2^2] \\ &= \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \mathbb{E}[\|\mathbf{v}_j - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2]. \end{aligned} \quad (28)$$

The last line holds since the cross term is 0 as derived below.

$$\begin{aligned}
& \mathbb{E} [(\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)}))^T (\mathbf{v}_j - \nabla F(\mathbf{w}_{\tau(j)}))] \\
&= \mathbb{E}_{\mathbf{w}_{\tau(j)}, \mathbf{w}_j} [(\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)}))^T \mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}, \mathbf{w}_j} [\mathbf{v}_j - \nabla F(\mathbf{w}_{\tau(j)})]] \\
&= \mathbb{E}_{\mathbf{w}_{\tau(j)}, \mathbf{w}_j} [(\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)}))^T (\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}} [\mathbf{v}_j] - \nabla F(\mathbf{w}_{\tau(j)}))] \\
&= 0.
\end{aligned}$$

Here again the last line follows from Assumption 2 in Section 2 which states that

$$\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}} [\mathbf{v}_j] = \nabla F(\mathbf{w}_{\tau(j)}).$$

Returning to (28), observe that the second term can be further decomposed as,

$$\begin{aligned}
\mathbb{E} [\|\mathbf{v}_j - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] &= \mathbb{E}_{\mathbf{w}_{\tau(j)}} \left[\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}} [\|\mathbf{v}_j - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \right] \\
&= \mathbb{E}_{\mathbf{w}_{\tau(j)}} \left[\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}} [\|\mathbf{v}_j\|_2^2] - 2\mathbb{E}_{\mathbf{w}_{\tau(j)}} \left[\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}} [\mathbf{v}_j^T \nabla F(\mathbf{w}_{\tau(j)})] \right] \right. \\
&\quad \left. + \mathbb{E}_{\mathbf{w}_{\tau(j)}} \left[\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(j)}} [\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \right] \right] \\
&= \mathbb{E} [\|\mathbf{v}_j\|_2^2] - 2\mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \\
&= \mathbb{E} [\|\mathbf{v}_j\|_2^2] - \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2].
\end{aligned}$$

□

We will also be proving a K -learner version of this lemma Appendix C.2 to prove Theorem 3. Now we proceed to provide the proof of Corollary 2.

Proof of Corollary 2.

$$\begin{aligned}
F(\mathbf{w}_{j+1}) &\leq F(\mathbf{w}_j) + (\mathbf{w}_{j+1} - \mathbf{w}_j)^T \nabla F(\mathbf{w}_j) + \frac{L}{2} \|\mathbf{w}_{j+1} - \mathbf{w}_j\|_2^2 \\
&= F(\mathbf{w}_j) + (-\eta \mathbf{v}_j)^T \nabla F(\mathbf{w}_j) + \frac{L\eta^2}{2} \|\mathbf{v}_j\|_2^2 \\
&= F(\mathbf{w}_j) - \frac{\eta}{2} \|\nabla F(\mathbf{w}_j)\|_2^2 - \frac{\eta}{2} \|\mathbf{v}_j\|_2^2 + \frac{\eta}{2} \|\nabla F(\mathbf{w}_j) - \mathbf{v}_j\|_2^2 + \frac{L\eta^2}{2} \|\mathbf{v}_j\|_2^2. \tag{29}
\end{aligned}$$

Here the last line follows from $2\mathbf{a}^T \mathbf{b} = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - \|\mathbf{a} - \mathbf{b}\|_2^2$. Taking expectation,

$$\begin{aligned}
\mathbb{E} [F(\mathbf{w}_{j+1})] &\leq \mathbb{E} [F(\mathbf{w}_j)] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta}{2} \mathbb{E} [\|\mathbf{v}_j\|_2^2] + \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j) - \mathbf{v}_j\|_2^2] + \frac{L\eta^2}{2} \mathbb{E} [\|\mathbf{v}_j\|_2^2] \\
&\stackrel{(a)}{\leq} \mathbb{E} [F(\mathbf{w}_j)] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta}{2} \mathbb{E} [\|\mathbf{v}_j\|_2^2] + \frac{\eta}{2} \mathbb{E} [\|\mathbf{v}_j\|_2^2] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \\
&\quad + \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \frac{L\eta^2}{2} \mathbb{E} [\|\mathbf{v}_j\|_2^2]. \tag{30}
\end{aligned}$$

Here, (a) follows from Lemma 8 that we just derived. Now, again bounding from (30), we have

$$\begin{aligned}
\mathbb{E} [F(\mathbf{w}_{j+1})] &\stackrel{(b)}{\leq} \mathbb{E} [F(\mathbf{w}_j)] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_j)\|_2^2] \\
&\quad + \frac{L\eta^2}{2} \mathbb{E} [\|\mathbf{v}_j\|_2^2]
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(c)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2}(1-\gamma)\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{L\eta^2\sigma^2}{2m} \\
&\quad - \frac{\eta}{2}\left(1 - L\eta\left(\frac{M_G}{m} + 1\right)\right)\mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \\
&\stackrel{(d)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2}(1-\gamma)\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{L\eta^2\sigma^2}{2m} - \frac{\eta}{4}\mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \\
&\stackrel{(e)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2}(1-\gamma)\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{L\eta^2\sigma^2}{2m} - \frac{\eta}{4}p_0\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2]. \quad (31)
\end{aligned}$$

Here (b) follows from the statement of the theorem that

$$\mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \leq \gamma\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2]$$

for some constant $\gamma \leq 1$. The next step (c) follows from Assumption 4 in Section 2 which lead to

$$\mathbb{E}[\|\mathbf{v}_j\|_2^2] \leq \frac{\sigma^2}{m} + \left(\frac{M_G}{m} + 1\right)\mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2].$$

Step (d) follows from choosing $\eta < \frac{1}{2L(\frac{M_G}{m} + 1)}$ and finally (e) follows from Lemma 1.

Now one might recall that the function $F(w)$ was defined to be strongly convex with parameter c . Using the standard result of strong-convexity (6) in (31), we obtain the following result:

$$\mathbb{E}[F(\mathbf{w}_{j+1})] - F^* \leq \frac{\eta^2 L \sigma^2}{2m} + (1 - \eta c(1 - \gamma + \frac{p_0}{2}))(\mathbb{E}[F(\mathbf{w}_j)] - F^*).$$

Let us denote $\gamma' = (1 - \gamma + \frac{p_0}{2})$. Then, using the above recursion, we thus have,

$$\mathbb{E}[F(\mathbf{w}_J)] - F^* \leq \frac{\eta L \sigma^2}{2c\gamma' m} + (1 - \eta\gamma' c)^J(\mathbb{E}[F(\mathbf{w}_0)] - F^* - \frac{\eta L \sigma^2}{2c\gamma' m}).$$

□

C.1.1 Discussion on range of p_0

Let us denote the conditional probability of $\tau(j) = j$ given all the past delays and parameters as $p_0^{(j)}$. Now $p_0 \leq p_0^{(j)} \forall j$. Clearly the value of $p_0^{(j)}$ will differ for different distributions and accordingly the value of p_0 will differ. Here we include a brief discussion on the possible values of p_0 for different distributions. These also hold for K -async and K -batch-async SGD.

Proof of Lemma 2. Let t_0 be the time when the j -th iteration occurs, and suppose that learner i' pushed its gradient in the j -th iteration. Now similar to the proof of Lemma 6, let us also assume that the learners last read their parameter values at time instants t_1, t_2, \dots, t_P respectively where $t'_i = t_0$ and the remaining $(P - 1)$ of these t_i s are $< t_0$. Let Y_1, Y_2, \dots, Y_P be the random variables denoting the computation time of the P learners starting from time t_0 . Thus, $Y_i = X_i - (t_0 - t_i) | X_i > (t_0 - t_i)$. For exponentials, from the memoryless property, all these Y_i s become i.i.d. and thus from symmetry the probability of i' finishing before all the others is equal, i.e. $\frac{1}{P}$. Thus, $p_0^{(j)} = p_0 = \frac{1}{P}$. For new-longer-than-used distributions, as we have discussed before all the Y_i s with $i \neq i'$ will be stochastically dominated by $Y_{i'} = X_{i'}$. Thus, probability of i with $i \neq i'$ finishing first is higher than i' . Thus, $p_0^{(j)} \leq \frac{1}{P}$ and so is p_0 . Similarly, for new-shorter-than-used distributions, $Y_{i'}$ is stochastically dominated by all the Y_i s and thus probability of i' finishing first is more. So, $p_0^{(j)} \geq \frac{1}{P}$ and so is p_0 . □

C.2 K-async SGD under fixed learning rate

In this subsection, we provide a proof of Theorem 3.

Before we proceed to the proof of this theorem, we first extend our Assumption 4 from the variance of a single stochastic gradient to sum of stochastic gradients in the following Lemma.

Lemma 9. *If the variance of the stochastic updates is bounded as*

$$\mathbb{E}_{\xi_j | \mathbf{w}_{\tau(l,j)}} [\|g(\mathbf{w}_{\tau(l,j)}, \xi_{l,j}) - \nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \leq \frac{\sigma^2}{m} + \frac{M_G}{m} \|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2 \quad \forall \tau(l,j) \leq j,$$

then for K -async, the variance of the sum of stochastic updates given all the parameter values $\mathbf{w}_{\tau(l,j)}$ is also bounded as follows:

$$\begin{aligned} & \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \mathbf{w}_{\tau(1,j)} \dots \mathbf{w}_{\tau(K,j)}} \left[\left\| \sum_{l=1}^K g(\mathbf{w}_{l,j}, \xi_{l,j}) \right\|_2^2 \right] \\ & \leq \frac{K\sigma^2}{m} + \left(\frac{M_G}{m} + K \right) \left\| \sum_{l=1}^K \nabla F(\mathbf{w}_{\tau(l,j)}) \right\|_2^2. \end{aligned} \quad (32)$$

Proof. First let us consider the expectation of any cross term such that $l \neq l'$. For the ease of writing, let $\Omega = \{\mathbf{w}_{\tau(1,j)} \dots \mathbf{w}_{\tau(K,j)}\}$.

Now observe the conditional expectation of the cross term as follows:

$$\begin{aligned} & \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \Omega} [(g(\mathbf{w}_{l,j}, \xi_{l,j}) - \nabla F(\mathbf{w}_{\tau(l,j)}))^T ((g(\mathbf{w}_{l',j}, \xi_{l',j}) - \nabla F(\mathbf{w}_{\tau(l',j)})))] \\ & = \mathbb{E}_{\xi_{l,j}, \xi_{l',j} | \Omega} [(g(\mathbf{w}_{l,j}, \xi_{l,j}) - \nabla F(\mathbf{w}_{\tau(l,j)}))^T ((g(\mathbf{w}_{l',j}, \xi_{l',j}) - \nabla F(\mathbf{w}_{\tau(l',j)})))] \\ & = \mathbb{E}_{\xi_{l',j} | \Omega} [\mathbb{E}_{\xi_{l,j} | \xi_{l',j}, \Omega} [(g(\mathbf{w}_{l,j}, \xi_{l,j}) - \nabla F(\mathbf{w}_{\tau(l,j)}))^T] (g(\mathbf{w}_{l',j}, \xi_{l',j}) - \nabla F(\mathbf{w}_{\tau(l',j)}))] \\ & = \mathbb{E}_{\xi_{l',j} | \Omega} [0^T (g(\mathbf{w}_{l',j}, \xi_{l',j}) - \nabla F(\mathbf{w}_{\tau(l',j)}))] = 0. \end{aligned} \quad (33)$$

Thus the cross terms are all 0. So the expression simplifies as,

$$\begin{aligned} \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \Omega} \left[\left\| \sum_{l=1}^K g(\mathbf{w}_{l,j}, \xi_{l,j}) - F(\mathbf{w}_{\tau(l,j)}) \right\|_2^2 \right] & \stackrel{(a)}{=} \sum_{l=1}^K \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \Omega} [\|g(\mathbf{w}_{l,j}, \xi_{l,j}) - F(\mathbf{w}_{\tau(l,j)})\|_2^2] \\ & \leq \sum_{l=1}^K \frac{\sigma^2}{m} + \frac{M_G}{m} \|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2. \end{aligned} \quad (34)$$

Thus,

$$\begin{aligned} & \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \Omega} \left[\left\| \sum_{l=1}^K g(\mathbf{w}_{l,j}, \xi_{l,j}) \right\|_2^2 \right] \\ & = \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \Omega} \left[\left\| \sum_{l=1}^K g(\mathbf{w}_{l,j}, \xi_{l,j}) - F(\mathbf{w}_{\tau(l,j)}) \right\|_2^2 \right] + \mathbb{E}_{\xi_{1,j}, \dots, \xi_{K,j} | \Omega} \left[\left\| \sum_{l=1}^K F(\mathbf{w}_{\tau(l,j)}) \right\|_2^2 \right] \\ & \leq \frac{K\sigma^2}{m} + \sum_{l=1}^K \frac{M_G}{m} \|F(\mathbf{w}_{\tau(l,j)})\|_2^2 + \left\| \sum_{l=1}^K F(\mathbf{w}_{\tau(l,j)}) \right\|_2^2 \\ & \leq \frac{K\sigma^2}{m} + \sum_{l=1}^K \frac{M_G}{m} \|F(\mathbf{w}_{\tau(l,j)})\|_2^2 + \sum_{l=1}^K K \|F(\mathbf{w}_{\tau(l,j)})\|_2^2. \end{aligned} \quad (35)$$

□

Now we return to the proof of the theorem.

Proof of Theorem 3. Let $\mathbf{v}_j = \frac{1}{K} \sum_{l=1}^K g(\mathbf{w}_{l,j}, \xi_{l,j})$. Following steps similar to the Async-SGD proof, from Lipschitz continuity we have the following.

$$\begin{aligned}
F(\mathbf{w}_{j+1}) &\leq F(\mathbf{w}_j) + (\mathbf{w}_{j+1} - \mathbf{w}_j)^T \nabla F(\mathbf{w}_j) + \frac{L}{2} \|\mathbf{w}_{j+1} - \mathbf{w}_j\|_2^2 \\
&= F(\mathbf{w}_j) - \frac{\eta}{K} \sum_{l=1}^K g(\mathbf{w}_{l,j}, \xi_{l,j})^T \nabla F(\mathbf{w}_j) + \frac{L}{2} \|\eta \mathbf{v}_j\|_2^2 \\
&\stackrel{(a)}{=} F(\mathbf{w}_j) - \frac{\eta}{2K} \sum_{l=1}^K \|\nabla F(\mathbf{w}_j)\|_2^2 - \frac{\eta}{2K} \sum_{l=1}^K \|g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2 \\
&\quad + \frac{\eta}{2K} \sum_{l=1}^K \|g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2 - \frac{\eta}{2K} \sum_{l=1}^K \|\nabla F(\mathbf{w}_j)\|_2^2 + \frac{L\eta^2}{2} \|\mathbf{v}_j\|_2^2 \\
&= F(\mathbf{w}_j) - \frac{\eta}{2} \|\nabla F(\mathbf{w}_j)\|_2^2 - \frac{\eta}{2K} \sum_{l=1}^K \|g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2 \\
&\quad + \frac{\eta}{2K} \sum_{l=1}^K \|g(\mathbf{w}_{l,j}, \xi_{l,j}) - \nabla F(\mathbf{w}_j)\|_2^2 + \frac{L\eta^2}{2} \|\mathbf{v}_j\|_2^2. \tag{36}
\end{aligned}$$

Here (a) follows from $2\mathbf{a}^T \mathbf{b} = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - \|\mathbf{a} - \mathbf{b}\|_2^2$. Taking expectation,

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}_{j+1})] &\leq \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2] \\
&\quad + \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|\nabla F(\mathbf{w}_j) - g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2] + \frac{L\eta^2}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] \\
&\stackrel{(a)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2] \\
&\quad + \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|g(\mathbf{w}_{l,j}, \xi_{l,j})\|_2^2] - \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \\
&\quad + \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] + \frac{L\eta^2}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] \tag{37} \\
&\stackrel{(b)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta}{2K} \sum_{l=1}^K \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] + \frac{\eta}{2} \gamma \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] \\
&\quad + \frac{L\eta^2}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] \\
&\stackrel{(c)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2} (1 - \gamma) \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{L\eta^2 \sigma^2}{2Km} \\
&\quad - \frac{\eta}{2K} \sum_{l=1}^K \left(1 - L\eta \left(\frac{M_G}{Km} + \frac{1}{K}\right)\right) \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \\
&\stackrel{(d)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2} (1 - \gamma) \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{L\eta^2 \sigma^2}{2Km} - \frac{\eta}{4K} \sum_{l=1}^K \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2]
\end{aligned}$$

$$\stackrel{(e)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta}{2}(1-\gamma)\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{L\eta^2\sigma^2}{2Km} - \frac{\eta}{4}p_0\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2]. \quad (38)$$

Here step (a) follows from Lemma 8 and step (b) follows from the assumption that

$$\mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \leq \gamma\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2]$$

for some constant $\gamma \leq 1$. The next step (c) follows from the Lemma 9 that bounds the variance of the sum of stochastic gradients. Step (d) follows from choosing $\eta < \frac{1}{2L(\frac{M_G}{K_m} + \frac{1}{K})}$ and finally (e) follows from Lemma 1 in Section 3 that says $\mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(l,j)})\|_2^2] \geq p_0\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2]$ for some non-negative constant p_0 which is a lower bound on the conditional probability that $\tau(l, j) = j$ given all past delays and parameter values.

Finally, since $F(\mathbf{w})$ is strongly convex, using the inequality $2c(F(\mathbf{w}) - F^*) \leq \|\nabla F(\mathbf{w})\|_2^2$ in (38), we finally obtain the desired result. \square

C.2.1 Extension to Non-Convex case

The analysis can be extended to provide weaker guarantees for non-convex objectives. Let $\gamma' = 1 - \gamma + \frac{p_0}{2}$

For non-convex objectives, we have the following result.

Theorem 5. *For non-convex objective function, we have the following ergodic convergence result given by:*

$$\frac{1}{J+1} \sum_{j=0}^J \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] \leq \frac{2(F(\mathbf{w}_0) - F^*)}{(J+1)\eta\gamma'} + \frac{L\eta\sigma^2}{Km\gamma'}$$

where $F^* = \min_{\mathbf{w}} F(\mathbf{w})$.

Proof. Recall the recursion derived in the last proof in (38). After re-arrangement, we obtain the following:

$$\mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] \leq \frac{2(\mathbb{E}[F(\mathbf{w}_j)] - \mathbb{E}[F(\mathbf{w}_{j+1})])}{\eta\gamma'} + \frac{L\eta\sigma^2}{Km\gamma'}. \quad (39)$$

Taking summation from $j = 0$ to $j = J$, we get,

$$\begin{aligned} \frac{1}{J+1} \sum_{j=0}^J \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] &\leq \frac{2(\mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}_J)])}{(J+1)\eta\gamma'} + \frac{L\eta\sigma^2}{Km\gamma'} \\ &\stackrel{(a)}{\leq} \frac{2(F(\mathbf{w}_0) - F^*)}{(J+1)\eta\gamma'} + \frac{L\eta\sigma^2}{Km\gamma'}. \end{aligned} \quad (40)$$

Here (a) follows since we assume \mathbf{w}_0 to be known and also from $\mathbb{E}[F(\mathbf{w}_J)] \geq F^*$. \square

C.3 Variable Learning Rate Schedule

We propose a new heuristic for learning rate schedule that is more stable than fixed learning rate for asynchronous SGD. Our learning rate schedule is $\eta_j = \min \left\{ \frac{C}{\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2^2}, \eta_{max} \right\}$, where η_{max} is a suitably large value of learning rate beyond which the convergence diverges. This heuristic is inspired from the assumption in Theorem 4 given by $\eta_j \mathbb{E}[\|\mathbf{w}_j - \mathbf{w}_{\tau(j)}\|_2^2] \leq C$. In this section, we derive the accuracy trade-off mentioned in Theorem 4 based on this assumption.

Proof of Theorem 4. Following steps similar to (29), we first obtain the following:

$$\begin{aligned} F(\mathbf{w}_{j+1}) &\leq F(\mathbf{w}_j) - \frac{\eta_j}{2} \|\nabla F(\mathbf{w}_j)\|_2^2 - \frac{\eta_j}{2} \|\mathbf{v}_j\|_2^2 \\ &\quad + \frac{\eta_j}{2} \|\nabla F(\mathbf{w}_j) - \mathbf{v}_j\|_2^2 + \frac{L\eta_j^2}{2} \|\mathbf{v}_j\|_2^2. \end{aligned} \quad (41)$$

Now taking expectation, we obtain the following result.

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_{j+1})] &\stackrel{(a)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta_j}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] + \frac{\eta_j}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] \\ &\quad - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \frac{L\eta_j^2}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] \\ &\stackrel{(b)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \frac{CL^2}{2} + \frac{L\eta_j^2}{2} \mathbb{E}[\|\mathbf{v}_j\|_2^2] \\ &\stackrel{(c)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{CL^2}{2} + \frac{L\eta_j^2\sigma^2}{2m} \\ &\quad - \frac{\eta_j}{2} \left(1 - L\eta_j\left(\frac{M_G}{m} + 1\right)\right) \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] \\ &\stackrel{(e)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \frac{CL^2}{2} + \frac{\eta_j^2 L\sigma^2}{2m} - \frac{\eta_j}{4} \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2]. \end{aligned} \quad (42)$$

Here (a) follows from (30), (b) follows from (12), (c) follows from Assumption 4 and (d) follows as $\eta_j \leq \frac{1}{2L(\frac{M_G}{m} + 1)}$. Let us define $\Delta_j = \frac{CL^2}{2} + \frac{\eta_j^2 L\sigma^2}{2m}$. Thus, the recursion can be written as,

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_{j+1})] &\leq \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta_j}{2} \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] - \frac{\eta_j}{4} \mathbb{E}[\|\nabla F(\mathbf{w}_{\tau(j)})\|_2^2] + \Delta_j \\ &\stackrel{(e)}{\leq} \mathbb{E}[F(\mathbf{w}_j)] - \frac{\eta_j}{2} \left(1 + \frac{p_0}{2}\right) \mathbb{E}[\|\nabla F(\mathbf{w}_j)\|_2^2] + \Delta_j. \end{aligned} \quad (43)$$

Here (e) follows from Lemma 1. If the loss function $F(\mathbf{w})$ is strongly convex with parameter c , then for all \mathbf{w} , we have $2c(F(\mathbf{w}) - F^*) \leq \|\nabla F(\mathbf{w})\|_2^2$. Using this result, we obtain

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_{j+1})] - F^* &\leq (1 - \eta_j(1 + \frac{p_0}{2})c)(\mathbb{E}[F(\mathbf{w}_j)] - F^*) + \Delta_j \\ &\leq (1 - \eta_j(1 + \frac{p_0}{2})c)(1 - \eta_{j-1}(1 + \frac{p_0}{2})c)(\mathbb{E}[F(\mathbf{w}_{j-1})] - F^*) \\ &\quad + (1 - \eta_j(1 + \frac{p_0}{2})c)\Delta_{j-1} + \Delta_j \\ &\leq (1 - \rho_j)(1 - \rho_{j-1}) \dots (1 - \rho_0)(\mathbb{E}[F(\mathbf{w}_0)] - F^*) + \Delta, \end{aligned} \quad (44)$$

where $\rho_j = \eta_j(1 + \frac{p_0}{2})c$ and $\Delta = \Delta_j + (1 - \rho_j)\Delta_{j-1} + \dots + (1 - \rho_j)(1 - \rho_{j-1}) \dots (1 - \rho_1)\Delta_0$. \square

D SIMULATION SETUP DETAILS

MNIST [63]: For the simulations on MNIST dataset, we first convert the 28×28 images into single vectors of length 784. We use a single layer of neurons followed by soft-max cross entropy with logits loss function. Thus effectively the parameters consist of a weight matrix \mathbf{W} of size 784×10 and a bias vector \mathbf{b} of size 1×10 . We use a regularizer of value 0.01, mini-batch size $m = 1$, and learning rate $\eta = 0.01$. For implementation we used Tensorflow with Python3. Thus, the model is as follows:

```

X=tf.placeholder(tf.float32,[None,784])
Y=tf.placeholder(tf.float32,[None,10])
W=tf.Variable(tf.random_normal(shape=[784,10],
                               stddev=0.01), name="weights")
b=tf.Variable(tf.random_normal(shape=[1,10],
                               stddev=0.01), name="bias")

logits=tf.matmul(X,W) + b
entropy=tf.nn.softmax_cross_entropy_with
        _logits(logits=logits,labels=Y) +
        lamda*tf.square(tf.norm(W))

loss=tf.reduce_mean( entropy)

```

For the run-time simulations, we generate random variables from the respective distributions in python to represent the computation times.

CIFAR10 [64]: For the CIFAR10 simulations, similar to MNIST, we convert the images into vectors of length 1024. We combine the three colour variants in the ratio $[0.2989, 0.5870, 0.114]$ to generate a single vector of length 1024 for every image. We use a single layer of neurons again followed by soft-max cross entropy with logits in tensorflow. Thus, the parameters consist of a weight matrix \mathbf{W} of size 1024×10 and a bias vector \mathbf{b} of size 1×10 . We use a mini-batch size of 250, regularizer of 0.05.

We use a similar model as follows:

```

X=tf.placeholder(tf.float32,[None,1024])
Y=tf.placeholder(tf.float32,[None,10])
W=tf.Variable(tf.random_normal(shape=[1024,10],
                               stddev= 0.01),name="weights")
b=tf.Variable(tf.random_normal(shape=[1,10],
                               stddev = 0.01),name="bias")

logits=tf.matmul(X,W) + b
entropy=tf.nn.softmax_cross_entropy_with
        _logits(logits=logits,labels=Y) +
        lamda*tf.square(tf.norm(W))

loss=tf.reduce_mean(entropy)

```

The computation time as each learner is generated from exponential distribution.

E CHOICE OF HYPERPARAMETERS

Our analysis techniques can also inform the choice of hyperparameters for synchronous and K -sync SGD.

E.1 Varying K in K -sync

We first perform some simulations of K -sync SGD applied on the MNIST dataset. For the simulation set-up, we consider 8 parallel learners with fixed mini-batch size $m = 1$ and fixed learning rate

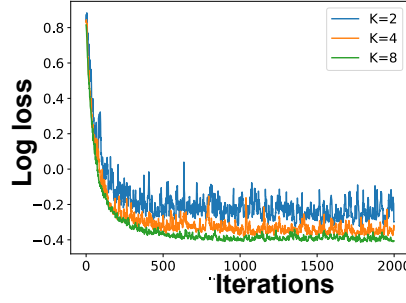


Figure 11: Error-Iterations tradeoff on MNIST dataset: Simulation of K -sync SGD for different values of K . Observe that accuracy improves with increasing K which means increasing effective batch size ($\eta = 0.05$).

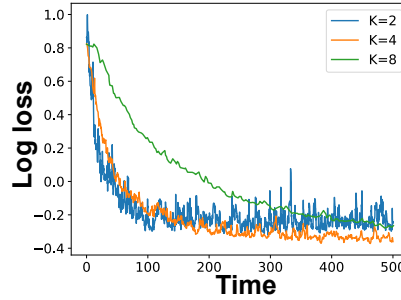


Figure 12: Error-Runtime tradeoff on MNIST dataset: Simulation of K -sync SGD for different values of K ($\eta = 0.05$).

0.05. The number of learners to wait for in K -sync, *i.e.* K is varied and the error-runtime trade-off is observed. The runtimes are generated from a shifted exponential distribution given by $X_i \sim m + \exp \mu$.

Observe that in the plot of error with the number of iterations in Figure 11, the error improves with increasing K , which means increasing the effective mini-batch and reducing the variability in the gradient. However, if we look at the same error plotted against runtime (See Figure 12) instead of the number of iterations, observe that increasing K naively does not always lead to a better trade-off. As K increases, the central PS has to wait for more learners to finish at every iteration, thus suffering from increased straggler effect. The best error-runtime trade-off is obtained at an intermediate $K = 4$. Thus, the current analysis informs the optimal choice of K to achieve a good error-runtime trade-off.

E.2 Varying mini-batch m

We consider the training of Alexnet on ImageNet dataset [65] using $P = 4$ learners. For this simulation, we perform fully synchronous SGD, *i.e.* K -sync with $K = P = 4$. We fix the learning rate and vary the mini-batch used for training. The runtimes are generated from a shifted exponential distribution given by $X_i \sim m + \exp \mu$, that depends on the mini-batch size. Intuitively, this distribution makes sense since to compute one mini-batch, a processor would atleast need a time m (Work Complexity). However, due to delays, it has the additional exponential tail. The error-runtime trade-offs are observed in Figure 13 and Figure 14.

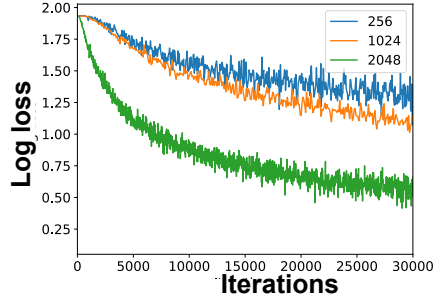


Figure 13: Error-Iterations tradeoff on IMAGENET dataset: Simulation of fully synchronous SGD ($K = P = 4$) for different values of mini-batch m . Observe that accuracy improves with increasing m which means increasing effective batch size.

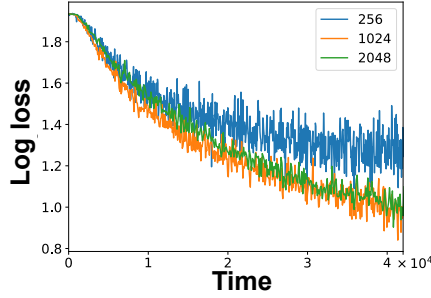


Figure 14: Error-Runtime tradeoff on IMAGENET dataset: Same simulation of fully synchronous SGD ($K = P = 4$) for different values of mini-batch m plotted against time. Observe that higher m does not necessarily mean the best trade-off with runtime as higher mini-batch also has longer time.

Again, observe that the plot of error with the number of iterations improves with the mini-batch size, as also expected from theory. However, increasing the mini-batch also changes the runtime distribution. Thus, when we plot the same error against expected runtime, we again observe that increasing the mini-batch size naively does not necessarily lead to the best trade-off. Instead, the best error-runtime trade-off is observed with an intermediate mini-batch value of 1024. Thus, our analysis informs the choice of the optimal mini-batch.