Asynchronous Byzantine Machine Learning (the case of SGD) Supplementary Material

Abstract

We theoretically prove the Byzantine resilience (Section 1) and convergence (Section 2) of Kardam. Furthermore, we provide additional experimental results for the EMNIST dataset in Section 3.

1 Analysis of Byzantine Resilience

Definition 1 (Time)**.** *The global epoch (denoted by t) represents the global logical clock of the parameter server (or equivalently the number of model updates). The local timestamp (denoted by l_p) for a given worker p, represents the epoch of the model that the worker receives from the server and computes the gradient upon. The difference $t - l_p$ can be arbitrarily large due to the asynchrony of the network.*

We make the following assumptions about any honest worker p.

Assumption 1 (Unbiased gradient estimator)**.**

$$\mathbb{E}_{\xi_p} \boldsymbol{G}(\boldsymbol{x}_{l_p}, \xi_p) = \boldsymbol{\nabla} Q(\boldsymbol{x}_{l_p})$$

Assumption 2 (Bounded variance)**.**

$$\mathbb{E}_{\xi_p} \|\boldsymbol{G}(\boldsymbol{x}_{l_p}, \xi_p) - \boldsymbol{\nabla} Q(\boldsymbol{x}_{l_p})\|^2 \leq d\sigma^2$$

Assumptions 1 and 2 are common in the literature [2] and hold if the data used for computing the gradients is drawn uniformly and independently.

Assumption 3 (Linear growth of r-th moment)**.**

$$\mathbb{E}_{\xi_p} \|\boldsymbol{G}(\boldsymbol{x}, \xi_p)\|^r \leq A_r + B_r \|\boldsymbol{x}\|^r \quad \forall \boldsymbol{x} \in \mathbb{R}, \quad r = 2, 3, 4$$

Assumption 3 translates into "the r-th moment of the gradient estimator grows linearly with the r-th power of the norm of the model" as assumed in [2].

Assumption 4 (Lipschitz gradient)**.**

$$||\boldsymbol{\nabla} Q(\boldsymbol{x}_1) - \boldsymbol{\nabla} Q(\boldsymbol{x}_2)|| \leq K ||\boldsymbol{x}_1 - \boldsymbol{x}_2||$$

Assumption 5 (Convexity in the horizon)**.** *We require that beyond a certain horizon,* $\|\boldsymbol{x}\| \geq D$*, there exist* $\epsilon > 0$ *and* $0 \leq \beta < \pi/2$ *such that* $\|\boldsymbol{\nabla} Q(\boldsymbol{x})\| \geq \epsilon > 0$ *and* $\frac{\langle \boldsymbol{x}, \boldsymbol{\nabla} Q(\boldsymbol{x}) \rangle}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{\nabla} Q(\boldsymbol{x})\|} \geq \cos \beta$*.*

Assumption 5 is the same as in [1], which in turn is a slight refinement of a similar assumption in [2]. It essentially states that, beyond a certain horizon D in the parameter space, the opposite of the gradient points towards the origin.

Definition 2 (Byzantine resilience)**.** *Let* Q *be any cost function satisfying the aforementioned assumptions. Let* A *be any distributed SGD scheme. We say that* A *is Byzantine-resilient if the sequence* $\boldsymbol{\nabla} Q(\boldsymbol{x}_t) = 0$ *converges almost surely to zero, despite the presence of up to* f *Byzantine workers.*

Theorem 1 (Optimal Slowdown)**.** *We define the slowdown* SL *as the ratio between the number of updates from honest workers that pass the Lipschitz filter and the total number of updates delivered at the parameter server. We derive the upper and lower bounds of* SL *in the following.*

$$\frac{n-2f}{n-f} \leq SL \leq \frac{n-f}{n}$$

The upper and lower bounds are tight and hold when there are f *Byzantine workers and no Byzantine workers respectively. Therefore Kardam achieves the optimal bounds with respect to any Byzantine-resilient SGD scheme and* $n \approx 3f$ *workers.*

Proof. Any Byzantine-resilient SGD scheme assuming f Byzantine workers will at most use $\frac{n-f}{n}$ of the total available workers (upper bound). By definition, the Lipschitz filter accepts the gradients computed by $\frac{n-f}{n}$ of the total workers with empirical Lipschitzness below $\hat{K}_t$. If every worker is honest, then the filter accepts gradients from $\frac{n-f}{n}$ of the workers. We thus get the tightness of the upper bound for the slowdown of Kardam. For the lower bound, the Byzantine workers can know that putting a gradient proposition above $\hat{K}_t$ will get them filtered out and the parameter server will end up using only the honest workers available. The optimal attack would therefore be to slowdown the server by getting tiny-Lipschitz gradients accepted while preventing the model from actually changing. This way, the Byzantine workers will make the server filter gradients from a total of f out of the $n - f$ honest workers, leaving only $n - 2f$ useful workers for the server. □

Theorem 2 (Byzantine resilience in asynchrony)**.** *Let* A *be any distributed SGD scheme. If the maximum successive gradients that* A *accepts from a single worker and the maximum delay are both unbounded, then* A *cannot be Byzantine-resilient when* $f \geq 1$*.*

Proof. Without any restrictions, the parameter server could only accept successive gradients from the same Byzantine worker (without getting any update from any honest worker), for example, if the Byzantine worker is faster than any other

worker (which is true by the definition of a Byzantine worker and by the fact that delays on (honest) workers are unbounded). This way, the Byzantine worker can force the parameter server to follow arbitrarily bad directions and never converge. Hence, without any restriction on the number of gradients from the workers, we prove the impossibility of asynchronous Byzantine resilience. Readers familiar with distributed computing literature might note that if asynchrony was possible for Byzantine SGD without restricting the number of successive gradients from a single worker, this could be used as an abstraction to solve asynchronous Byzantine consensus (that is impossible to solve [3]). This provides another proof (by contradiction) for our theorem. □

Theorem 3 (Correct cone and bounded statistical moments)**.** *If $N > 3f + 1$ then for any $t \geq t_r$ (we show that $t_r \in \mathcal{O}(\frac{1}{K\sqrt{|\xi|}})$ where $|\xi|$ is the batch-size of honest workers):*

$$\mathbb{E}(\|\boldsymbol{Kar}_t\|^r) \leq A'_r + B'_r \|\boldsymbol{x}_t\|^r$$

for any $r = 2, 3, 4$ and

$$\langle \mathbb{E}(\boldsymbol{Kar}_t), \boldsymbol{\nabla} Q_t \rangle = \Omega(1 - \frac{\sqrt{d}\sigma}{\|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|}) \|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|^2$$

The expectation is on the random samples used for training.

Proof. First of all, it is important to note that a Byzantine worker can lie about its Lipschitz coefficient without being able to fool the parameter server. The median Lipschitz coefficient is always bounded between the Lipschitz coefficients of two correct worker, and it is against that the gradient of the Byzantine worker would be tested to be filtered out if harmful and accepted if useful.

Lemma 1 (Limit of successive gradients)**.** *The frequency filter ensures that any sequence of length $2f + 1$ consequently accepted gradients contains at least $f + 1$ gradients computed by honest workers.*

Proof. Given any sequence of $2f + 1$ consequently accepted gradients (L), we denote by S the set of workers that computed these gradients. The frequency filter guarantees that any f workers in S computed at most f gradients in L. At most f workers in S can be Byzantine, thus at least $f + 1$ gradients in L are from honest workers. □

We start the proof of Theorem 3 by proving that Kardam acts as self-stabilizing mechanism that guarantees the global confinement of the parameter vector using the following remark.

Lemma 2 (Global Confinement)**.** *Let $\boldsymbol{x}_t$ the sequence of parameter models visited by $\boldsymbol{Kar}$. There exist a constant $D > 0$ such that the sequence $\boldsymbol{x}_t$ almost surely verifies $\|\boldsymbol{x}_t\| \leq D$ when $t \mapsto \infty$.*

Proof. (Global Confinement) Lemma 2 can be proven by using Remark 1 and the proof of confinement in [2].

Remark 1. *Let $r = 2, 3, 4$. There exist $A'_r \geq 0$ and $B'_r \geq 0$ such that:* $(\forall t \geq 0)\mathbb{E}\|\boldsymbol{Kar}_t(\boldsymbol{x}_t, \xi)\|^r \leq A'_r + B'_r\|\boldsymbol{x}_t\|^r$

Proof. (Remark 1). Note that if $\boldsymbol{Kar}_t(\boldsymbol{x}_t)$ comes from a honest worker, we have $\boldsymbol{Kar}_t(\boldsymbol{x}_t, \xi) = \boldsymbol{G}(\boldsymbol{x}_t, \xi)$ therefore, $(\forall t \geq 0)\mathbb{E}\|\boldsymbol{Kar}_t(\boldsymbol{x}_t, \xi)\|^r \leq A_r + B_r\|\boldsymbol{x}_t\|^r$ since by assumption on the estimator $\boldsymbol{G}$ used by honest workers, we have $(\forall \boldsymbol{x} \in \mathbb{R})\mathbb{E}\|\boldsymbol{G}(\boldsymbol{x}, \xi)\|^r \leq A_r + B_r\|\boldsymbol{x}\|^r$.

Let $t > 2f+1$ be any epoch at the parameter server. Because of the Lipschitz filter (passed by $\boldsymbol{Kar}_t$), there exists $i \leq f$ such that $\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})$ comes from an honest worker. Therefore, $\|\boldsymbol{x}_{t-i}\| \leq \|\boldsymbol{x}_t\| + \sum_{l=1}^{i} \gamma'_{t-l}\boldsymbol{Kar}_{t-l}(\boldsymbol{x}_{t-l}) \leq \|\boldsymbol{x}_t\| + \sum_{l=1}^{i} \gamma_{t-l} \cdot \frac{\min(\boldsymbol{Kar}_0, \|\boldsymbol{Kar}_{t-l}(\boldsymbol{x}_{t-l})\|)}{\|\boldsymbol{Kar}_{t-l}(\boldsymbol{x}_{t-l})\|} \cdot \boldsymbol{Kar}_{t-l}(\boldsymbol{x}_{t-l}) \leq f \cdot \boldsymbol{Kar}_0 + \|\boldsymbol{x}_t\|$.

So, for $r = 2, 3, 4$ there exists C_r such that $\|\boldsymbol{x}_{t-i}\|^r \leq (f \cdot \boldsymbol{Kar}_0)^r + C_r\|\boldsymbol{x}_t\|^r$.

According to the Lipschitz criteria:

$$
\begin{aligned}
&\|\boldsymbol{Kar}_t(\boldsymbol{x}_t)\| \\
&\leq K_t(\|\boldsymbol{x}_t\| + \|\boldsymbol{x}_{t-1}\|) + \|\boldsymbol{Kar}_{t-1}(\boldsymbol{x}_{t-1})\| \\
&\leq \sum_{l=1}^{i} K_{t-l+1}(\|\boldsymbol{x}_{t-l+1}\| + \|\boldsymbol{x}_{t-l}\|) + \|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\| \\
&\leq 2K\sum_{l=0}^{i} \|\boldsymbol{x}_{t-l}\| + \|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\| \\
&\leq 2K\sum_{l=0}^{i}\sum_{s=l}^{i-1} [\gamma'_{t-s} \cdot \|\boldsymbol{Kar}_{t-s}(\boldsymbol{x}_{t-s})\| + \|\boldsymbol{x}_{t-i}\|] + \|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\| \\
&\leq 2K\sum_{l=0}^{i}\sum_{s=l}^{i-1} \gamma_{t-s}\|\boldsymbol{Kar}_{t-s}(\boldsymbol{x}_{t-s})\| \cdot \frac{\min(\boldsymbol{Kar}_0, \|\boldsymbol{Kar}_{t-s}(\boldsymbol{x}_{t-s})\|)}{\|\boldsymbol{Kar}_{t-s}(\boldsymbol{x}_{t-s})\|} \\
&\quad + 2fK\|\boldsymbol{x}_{t-i}\| + \|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\| \\
&\leq Kf(f-1)\boldsymbol{Kar}_0 + 2fK\|\boldsymbol{x}_{t-i}\| + \|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\| \\
&= D + E\|\boldsymbol{x}_{t-i}\| + F\|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\|
\end{aligned}
$$

Where K is the global Lipschitz. (We do not need to know the value of K to implement $\boldsymbol{Kar}$ but we use it for the proofs.) Taking both side of the above inequality to the power r, we have the following for $r = 2 \dots 4$ for constants D_r, E_r and F_r:

$\|\boldsymbol{Kar}_t(\boldsymbol{x}_t)\|^r \leq D_r + E_r \cdot \|\boldsymbol{x}_{t-i}\|^r + F_r \cdot E\|\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})\|^r$

As $\boldsymbol{Kar}_{t-i}(\boldsymbol{x}_{t-i})$ comes from an honest worker, using the Jensen inequality and the assumption on honest workers. We can take the expected value on ξ.

$$
\begin{aligned}
&\mathbb{E}\|\boldsymbol{Kar}_t(\boldsymbol{x}_t)\|^r \\
&\leq D_r + E_r \cdot \|\boldsymbol{x}_{t-i}\|^r + F_r[A_r + B_r\|\boldsymbol{x}_{t-i}\|^r] \\
&= D_r + F_rA_r + \|\boldsymbol{x}_{t-i}\|^r[E_r + F_rB_r] \\
&\leq D_r + F_rA_r + [(f \cdot \boldsymbol{Kar}_0)^r + C_r\|\boldsymbol{x}_t\|^r] \cdot [E_r + F_rB_r] \\
&= D_r + F_rA_r + f^r\boldsymbol{Kar}_0^r[E_r + F_rB_r] + [E_r + F_rB_r] \cdot \|\boldsymbol{x}_t\|^r
\end{aligned}
$$

We denote by $A'_r = D_r + F_r A_r + (f \cdot \boldsymbol{Kar}_0)^r \cdot [E_r + F_r B_r]$ and $B'_r = E_r + F_r B_r$, we obtain:

$\mathbb{E}\|\boldsymbol{Kar}_t(\boldsymbol{x}_t)\|^r \leq A'_r + B'_r \|\boldsymbol{x}_t\|^r$ □

Remark 1 shows that with $\boldsymbol{Kar}$, all the assumptions of Bottou [2] (Section 5.2) are holding even in the presence of Byzantine workers, and thus, the global confinement of $\boldsymbol{x}_t$ stated in Lemma 2. □

Remark 1 have proved the first part of Theorem 3 To continue the proof of this Theorem, the goal is to find a lower bound on the scalar product between Kardam and the real gradient of the cost. This is achieved via an upper bound on: $\|\mathbb{E}\boldsymbol{Kar}_t - \boldsymbol{\nabla} Q_t(\boldsymbol{x}_t)\|$. Let p the worker whose gradient estimation $\boldsymbol{g_p}$ was selected by Kardam to be the update for epoch t at the parameter server. According to Lemma 1, considering the latest $2f+1$ timestamps, at least $f+1$ of updates came from honest workers. Hence, there exists $i < f$ such that, $\boldsymbol{Kar}_{t-i}$ came from an honest worker. Hence, $\mathbb{E}\boldsymbol{Kar}_{t-i} = \boldsymbol{\nabla} Q_{t-i}$. By applying the triangle inequality twice, we have:

$$\begin{aligned}\|\boldsymbol{Kar}_t - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| \leq \|&\boldsymbol{Kar}_t - \boldsymbol{Kar}_{t-i}\| \\ &+ \|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\| \\ &+ \|\boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i}) - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|\end{aligned}$$

We know:

$$\begin{aligned}\|\boldsymbol{Kar}_{t-i} - \boldsymbol{Kar}_t\| &\leq \sum_{k=i}^{1} \|\boldsymbol{Kar}_{t-k} - \boldsymbol{Kar}_{t-k+1}\| \leq K \sum_{k=1}^{i} \|\boldsymbol{x}_{t-k+1} - \boldsymbol{x}_{t-k}\| \\ &\leq K \sum_{k=1}^{i} \gamma_{t-k} \|\boldsymbol{Kar}_{t-k}\| \leq i \cdot K \cdot \gamma_{t-i} \cdot \|\boldsymbol{Kar}\|_{max(t,i)}\end{aligned}$$

where, $\|\boldsymbol{Kar}\|_{max(t,i)}$ is the upper-bound on the norm of $\boldsymbol{Kar}$ in the list from $t-i$ to $t-1$. Since $i < f$, we have $\|\boldsymbol{Kar}_{t-i} - \boldsymbol{Kar}_t\| \leq fK\gamma_{t-i}\|\boldsymbol{Kar}\|_{max(t,i)}$. Since $\boldsymbol{x}_t$ is globally confined (Lemma 2), by continuous differentiability of Q, so will be $\|\boldsymbol{\nabla} Q(\boldsymbol{x}_{t,i})\|$, therefore $fK\|\boldsymbol{Kar}\|_{max(t,i)}$ is bounded, and multiplies γ_{t-i} in the right hand side of the last inequality, and we know from the hypothesis on the learning rate that $\lim_{t\to\infty} \gamma_t = 0$ (sequence of summable squares, therefore goes to zero). Since $i < f$ (and obviously, f, as a global variable, is independent of t), then we also have $\lim_{t\to\infty} \gamma_{t-i} = 0$. This means that for every $\epsilon > 0$, eventually, the left hand-side of the above inequality is bounded by $\epsilon\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\|$, more precisely, since γ_t is typically $\mathcal{O}(\frac{1}{t})$, this will hold after t_r such that $t_r = \Omega(\frac{1}{\epsilon K})$.

By replacing in Formula 1, we get:

$$\begin{aligned}\|\boldsymbol{Kar}_t - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| &\leq (1+\epsilon)\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\| + \|\boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i}) - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| \\ &\leq (1+\epsilon)\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\| + \sum_{s=1}^{i} K_{t-s}\gamma_{t-s}\|\boldsymbol{\nabla} Q(\boldsymbol{x}_{t-s})\| \\ &\leq (1+\epsilon)\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\| + f \cdot K \cdot \gamma_{t-i} \cdot \|\boldsymbol{\nabla} Q\|_{max(t,i)}.\end{aligned}$$

Where K_{t-s} is the real local Lipschitz coefficient of the loss function at epoch $t-s$. Let $j = \min(\frac{\sqrt{d}\sigma}{2}, \|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| - \sqrt{d}\sigma)$, $C = \frac{j}{2\epsilon\sqrt{d}\sigma}$, $\epsilon' = \frac{j}{2.C}$. As $\lim_{t\to\infty} \gamma_t = 0$ and $\|\boldsymbol{\nabla} Q\|_{max(t,i)}$ is bounded, there exist a time after which, the above quantity can be made bounded as

$$\|\boldsymbol{Kar}_t - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| \leq (1+\epsilon)\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\| + \epsilon'.$$

And hence:

$$\begin{aligned}\|\mathbb{E}(\boldsymbol{Kar}_t) - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| &\leq \mathbb{E}(\|\boldsymbol{Kar}_t - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|) \\ &\leq (1+\epsilon)\mathbb{E}(\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\|) + \epsilon'.\end{aligned}$$

since $\boldsymbol{Kar}_{t-i}$ comes from a correct worker, we have:

$$\mathbb{E}(\|\boldsymbol{Kar}_{t-i} - \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-i})\|) \leq \sqrt{d}\sigma$$

Therefore, $\|\mathbb{E}(\boldsymbol{Kar}_t) - \boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| \leq (1+\epsilon)\sqrt{d}\sigma + \epsilon'$. Consequently, Kardam only selects vectors that live on average in the cone of radius α around the true gradient, where α is given by:

$\sin(\alpha) = \frac{(1+\epsilon)\sqrt{d}\sigma + \epsilon'}{\|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|}$. (as long as $\|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| > (1+\epsilon)\sqrt{d}\sigma + \epsilon'$, this has a sense)

Note:

- The $\sqrt{d}$ in $\|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\| > \sqrt{d}\sigma$ is not a harsh requirement, we are using the conventional notation where $\sqrt{d}\sigma$ is the upper bound on the variance, σ should be seen as the "component-wise" standard deviation, therefore, the norm of a non-trivial gradient is naturally larger than the vector-wise standard deviation of its estimator, which is typically $\sqrt{d}\sigma$.

- As long as the true gradient has *a nontrivial meaning* (it is larger than the standard deviation of its correct estimators), α is strictly bounded between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, which means that as long as there is no convergence to null gradients, Kardam is selecting vectors in the correct cone around the true gradient. Most importantly, this angle shrinks to zero when the variance is too small compared to the norm of the gradient, i.e., with large batch-sizes, Kardam boils down to be an unbiased gradient estimator. However, we only require the "component-wise" condition.

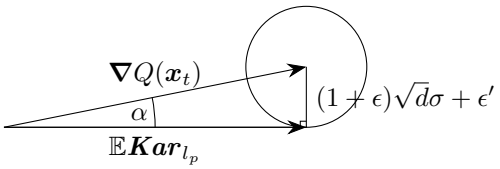


Figure 1: If $\left\|\mathbb{E}\boldsymbol{Kar}_{l_p} - \nabla Q(\boldsymbol{x}_t)\right\| \leq (1+\epsilon)\sqrt{d}\sigma + \epsilon'$ then $\langle \mathbb{E}\boldsymbol{Kar}_{l_p}, \nabla Q(\boldsymbol{x}_t)\rangle$ is upper bounded by $(1-\sin\alpha)\|\nabla Q(\boldsymbol{x}_t)\|^2$ where $\sin\alpha = \frac{(1+\epsilon)\sqrt{d}\sigma+\epsilon'}{\|\nabla Q(\boldsymbol{x}_t)\|}$.

In fact, as long as $\|\nabla Q(\boldsymbol{x}_t)\| > \sqrt{d}.\sigma$, we can consider small enough ϵ and ϵ' such that $D_1 = (1+\frac{3}{4C})\frac{\sqrt{d}\sigma}{\|\nabla Q(\boldsymbol{x}_t)\|}$, $D_2 = \frac{1}{C} + \frac{C-1}{C}\frac{\sqrt{d}\sigma}{\|\nabla Q(\boldsymbol{x}_t)\|}$, and $\sin(\alpha) = \min(D_1, D_2) < 1$. This indeed guarantees that $\alpha < \frac{\pi}{2}$, moreover, it is enough to take $C >> \frac{\nabla Q(\boldsymbol{x}_t)\|}{\sqrt{d}\sigma}$ and α would satisfy $\sin(\alpha) \approx \frac{\sqrt{d}\sigma}{\|\nabla Q(\boldsymbol{x}_t)\|}$.

Actually, in a list of L previous selected vectors, more than half of the vectors are from correct workers. (progress is made: liveness)

Consider a sublist of L from L_i to L_j. At the time of adding a worker in L_j, the frequency criteria was checked for the new addition to L. The active table at that time assure that in any new sublist of L, especially L_i^j), any f workers appear at most $\frac{j-i}{2}$ times. As the number of Byzantine workers is maximum f. in sublist L_i^j, the Byzantine workers did less than half of the updates. In other words, at least half of the updates come from honest workers. This proves the safety of Kardam.

The Byzantine workers may stop sending updates or send incorrect updates. In the case where the Byzantine workers stop sending updates, Kardam still guarantees liveness. The reason is that there are at least $2f+1$ honest workers who update the model. □

2 Convergence Analysis

Definition 3 (Dampening function)**.** *We employ a bijective and strictly decreasing dampening function $\tau \mapsto \Lambda(\tau)$ with $\Lambda(0) = 1$.*[1] *Note that every bijective function is also invertible, i.e., $\Lambda^{-1}(\nu)$ exists for every ν in the range of the Λ function.*

Let $\boldsymbol{\Lambda}_t$ be the set of Λ values associated with the gradients at epoch t.

$$\boldsymbol{\Lambda}_t = \{\Lambda(\tau_{tl}) \mid [g,l] \in \mathcal{G}_t\}$$

We partition the set $\mathcal{G}_t$ of gradients at epoch t according to their Λ-value as follows.

$$\mathcal{G}_t = \bigsqcup_{\lambda \in \boldsymbol{\Lambda}_t} \mathcal{G}_{t\lambda}$$
$$\mathcal{G}_{t\lambda} = \{[g,l] \in \mathcal{G}_t \mid \Lambda(\tau_{tl}) = \lambda\}$$

[1] If $\Lambda(0) = 1$, then there is no decay for gradients computed on the latest version of the model, i.e., $\tau_{tl} = 0$.

Therefore, the update equation can be reformulated as follows.

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma_t \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \cdot \sum_{[\boldsymbol{G}(\boldsymbol{x}_l;\xi),l] \in \mathcal{G}_{t\lambda}} \boldsymbol{G}(\boldsymbol{x}_l;\xi)$$

Definition 4 (Adaptive learning rate)**.** *Given the Lipschitz constant K, the total number of epochs T, and the total number of gradients in each epoch as M, we define γ_t as follows.*

$$\gamma_t = \underbrace{\sqrt{\frac{Q(\boldsymbol{x}_1) - Q(\boldsymbol{x}^*)}{K \cdot T \cdot M \cdot d \cdot \sigma^2}}}_{\gamma} \cdot \underbrace{\frac{M}{\sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \cdot |\mathcal{G}_{t\lambda}|}}_{\mu_t} \tag{1}$$

where γ is the baseline component of the learning rate and μ_t is the adaptive component that depends on the amount of stale updates that the server receives at epoch t. Moreover, μ_t incorporates the total staleness at any epoch t based on the staleness coefficients (λ) associated with all the gradients received in epoch t.

Comments on $Q(\boldsymbol{x}^*)$. $\boldsymbol{x}^*$ refers to the (not necessarily global) optimum we are heading to, and on which our adaptive learning rate depends. Assuming this value is known was made just for the sake of a proof, as is usually done in proofs for the speed of convergence of SGD (e.g., the references provided by the reviewer). In practice, one does not need to know $Q(\boldsymbol{x}^*)$ and can assume it to be lower bounded (Bottou1998). This will produce overshooting (large steps) in the early phases of Kardam, but will get to small enough step sizes: The baseline part of our adaptive learning rate contains a term $1/T$, where T is the total number of iterations (also unknown before we run SGD). In practice, it is replaced by $1/t$ (t: epoch at the server). This part of our learning rate decreases with t and will compensate for the overshooting described above (overcoming the overshooting in at most $O(1/Q(x_1))$ steps).

Remark 2 (Correct cone)**.** *As a consequence of passing the filter and of Theorem 3, $\boldsymbol{G}$ satisfies the following.*

$$\langle \mathbb{E}_\xi \boldsymbol{G}(\boldsymbol{x};\xi), \boldsymbol{\nabla} Q(\boldsymbol{x}) \rangle > \Omega\left((\|\boldsymbol{\nabla} Q(x_t)\| - \sqrt{d}\sigma) \|\boldsymbol{\nabla} Q(x_t)\| \right)$$

The theoretical guarantee for the convergence rate of Kardam depends on Assumptions 2,4 and Remark 2. These assumptions are weaker than the convergence guarantees in [7, 4]. In particular, due to unbounded delays and the potential presence of Byzantine workers, we only assume the unbiased gradient estimator $\boldsymbol{G}(\cdot)$ for honest workers (Assumption 1). We instead employ (Remark 2) the fact that $\boldsymbol{G}(\cdot)$ and $\boldsymbol{\nabla} Q(\boldsymbol{x})$ make a lower bounded angle together (and subsequently a lower bounded scalar product) for all the workers. The classical unbiased assumption is more restrictive as it requires this angle to be exactly equal to 0, and the scalar product to be equal to $\|\boldsymbol{\nabla} Q(\boldsymbol{x})\| \cdot \|\boldsymbol{G}(\boldsymbol{x})\|$. Most importantly, we highlight the fact that those assumption are satisfied by Kardam, since every gradient used in this section to compute the ***Kar*** update has passed the Lipschitz filter of the previous section.

Theorem 4 (Convergence guarantee)**.** *We provide the convergence guarantee in terms of the ergodic convergence that is the weighted average of the $\mathcal{L}_2$ norm of all gradients ($||\boldsymbol{\nabla} Q(\boldsymbol{x}_t)||^2$). Using the above-mentioned assumptions, and the maximum adaptive rate $\mu_{\max} = \max\{\mu_1, \ldots, \mu_t\}$, we have the following bound on the ergodic convergence rate.*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|^2 \leq (2+\mu_{\max}+\gamma K M \chi \mu_{\max})\cdot\gamma K\cdot d\sigma^2 + d\cdot\sigma^2 + 2DK\sigma\sqrt{d} + K^2D^2 \tag{2}$$

under the prerequisite that

$$\sum_{\lambda\in\boldsymbol{\Lambda}_t}\left\{K\gamma_t^2|\boldsymbol{\Lambda}_t| + \sum_{s=1}^{\infty}\sum_{\nu\in\boldsymbol{\Lambda}_{t+s}}\gamma_{t+s}K^2\nu|\mathcal{G}_{t+s,\nu}|\Lambda^{-1}(\nu)\mathbb{I}_{(s\leq\Lambda^{-1}(\nu))}\gamma_t^2|\boldsymbol{\Lambda}_t|\right\}\lambda^2 \leq \sum_{\lambda\in\boldsymbol{\Lambda}_t}\frac{\gamma_t\lambda}{|\mathcal{G}_{t\lambda}|} \tag{3}$$

where the Iverson indicator function is defined as follows.

$$\mathbb{I}_{(s\leq\Delta)} = \begin{cases} 1 & \textit{if } s\leq\Delta \\ 0 & \textit{otherwise.}\end{cases}$$

It is important to note that the prerequisite (Inequality 3) holds for any decay function Λ (since $\lambda < 1$ holds by definition) and for any standard learning rate schedule such that $\gamma_t < 1$. Various GD approaches [7, 5, 6, 4] provide convergence guarantees with similar prerequisites.

Proof. We provide the convergence guarantee in terms of *ergodic convergence*—the weighted average of the $\mathcal{L}_2$ norm of all gradients ($||\boldsymbol{\nabla} Q(\boldsymbol{x}_t)||^2$). For the sake of clarity in the proofs, if X is a set, we also denote its cardinality by X.

Lemma 3. *1 Assume that, for all epochs $1 \leq t \leq T$*

$$\sum_{\lambda\in\boldsymbol{\Lambda}_t}\left\{K\gamma_t^2|\boldsymbol{\Lambda}_t| + \sum_{s=1}^{\infty}\sum_{\nu\in\boldsymbol{\Lambda}_{t+s}}\gamma_{t+s}K^2\nu|\mathcal{G}_{t+s,\nu}|\Lambda^{-1}(\nu)\mathbb{I}_{(s\leq\Lambda^{-1}(\nu))}\gamma_t^2|\boldsymbol{\Lambda}_t|\right\}\lambda^2 \leq \sum_{\lambda\in\boldsymbol{\Lambda}_t}\frac{\gamma_t\lambda}{|\mathcal{G}_{t\lambda}|}$$

Then, the ergodic convergence rate is bounded as follows.

$$\frac{\sum_{t=1}^{T}\left(\gamma_t \sum_{\lambda\in\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\right)\mathbb{E}||\boldsymbol{\nabla}Q(\boldsymbol{x}_t)||^2}{\sum_{t=1}^{T}\gamma_t\sum_{\lambda\in\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}} \leq \frac{2(Q(\boldsymbol{x}_1)-Q(\boldsymbol{x}^*))}{\sum_{t=1}^{T}\gamma_t\sum_{\lambda\in\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}}$$
$$+\frac{\left(\sum_{t=1}^{T}K\gamma_t^2\sum_{\lambda\in\mathbf{\Lambda}_t}\lambda^2\mathcal{G}_{t\lambda}+\gamma_t K^2\sum_{\lambda\in\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1}\gamma_j^2\sum_{\lambda'\in\mathbf{\Lambda}_j}\lambda'^2\mathcal{G}_{j\lambda'}\right)\cdot d\cdot\sigma^2}{\sum_{t=1}^{T}\gamma_t\sum_{\lambda\in\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}}$$

Remark 3. *Given a list of vectors $u_1,\ldots,u_N$, we implicitly use the following inequality in our proof.*

$$\left\|\sum_{i=1}^{N}u_i\right\|^2 \leq N\cdot\sum_{i=1}^{N}\|u_i\|^2 \tag{3}$$

Proof. For the sake of concision, for every $m=[g,l]\in\mathcal{G}_{t\lambda}$, we denote by $\xi_{[t]}$ the set of ξ values that the server sends during epoch t. Let $\xi_{[t,*\neq m]}$ denote the set $\xi_{[t]}$ minus the variable ξ corresponding to message m. Additionally, $G[tm]\triangleq\boldsymbol{G}(\boldsymbol{x}_{t-\tau_{tl}};\xi)$ and $\boldsymbol{\nabla}Q[tm]\triangleq\boldsymbol{\nabla}Q(\boldsymbol{x}_{t-\tau_{tl}})$.

A second order expansion of Q, followed by the application of the Lipschitz inequality to $\boldsymbol{\nabla}Q$ yields the following.

$$\begin{aligned}Q(\boldsymbol{x}_{t+1})-Q(\boldsymbol{x}_t) &\leq \langle\boldsymbol{\nabla}Q(\boldsymbol{x}_t),\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle+\frac{K}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2\\ &\leq -\gamma_t\sum_{\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\langle\boldsymbol{\nabla}Q(\boldsymbol{x}_t),\frac{1}{\mathcal{G}_{t\lambda}}\sum_{\mathcal{G}_{t\lambda}}G[tm]\rangle+\frac{K}{2}\gamma_t^2\left\|\sum_{\mathbf{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}G[tm]\right\|^2\end{aligned}$$

Taking the expectation and using the correct cone property, we have:

$$\begin{aligned}\mathbb{E}_{\xi_{[t]}}Q(\boldsymbol{x}^{(t+1)})-Q(\boldsymbol{x}_t) &\leq -\gamma_t\sum_{\mathbf{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\langle\boldsymbol{\nabla}Q(\boldsymbol{x}_t),\frac{1}{\mathcal{G}_{t\lambda}}\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\rangle\\ &+\frac{K}{2}\gamma_t^2\mathbb{E}_{\xi_{[t]}}\left\|\sum_{\mathbf{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}G[tm]\right\|^2\end{aligned}$$

Using $\langle a, b\rangle = \frac{||a||^2+||b||^2-||a-b||^2}{2}$, we obtain the following inequality.

$$\mathbb{E}_{\xi_{[t]}} Q(\boldsymbol{x}_{t+1}) - Q(\boldsymbol{x}_t) \leq -\frac{\gamma_t}{2}\sum_{\boldsymbol{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\left\|\boldsymbol{\nabla}Q(\boldsymbol{x}_t)\right\|^2$$
$$-\frac{\gamma_t}{2}\sum_{\boldsymbol{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\left\|\frac{1}{\mathcal{G}_{t\lambda}}\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\right\|^2 + \frac{K\gamma_t^2}{2}\underbrace{\mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}G[tm]\right\|^2}_{S_1}$$
$$+\frac{\gamma_t}{2}\sum_{\boldsymbol{\Lambda}_t}\lambda\mathcal{G}_{t\lambda}\underbrace{\left\|\boldsymbol{\nabla}Q(\boldsymbol{x}_t) - \frac{1}{\mathcal{G}_{t\lambda}}\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\right\|^2}_{S_2}$$

We now define two terms S_1 and S_2 as follows.

$$S_1 = \mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}(G[tm]-\boldsymbol{\nabla}Q[tm]) + \sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\right\|^2$$
$$= \mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}(G[tm]-\boldsymbol{\nabla}Q[tm])\right\|^2 + \mathbb{E}_{\xi_{[m]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\right\|^2$$
$$+ 2\mathbb{E}_{\xi_{[t]}}\langle\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}(G[tm]-\boldsymbol{\nabla}Q[tm]), \sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\rangle$$
$$= \mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}(G[tm]-\boldsymbol{\nabla}Q[tm])\right\|^2 + \mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\right\|^2$$
$$+ 2\langle\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}(\boldsymbol{\nabla}Q[tm]-\boldsymbol{\nabla}Q[tm]), \sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\rangle$$
$$= \underbrace{\mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}(G[tm]-\boldsymbol{\nabla}Q[tm])\right\|^2}_{A_1} + \underbrace{\mathbb{E}_{\xi_{[t]}}\left\|\sum_{\boldsymbol{\Lambda}_t}\lambda\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\right\|^2}_{A_2}$$

Regarding A_2, applying Equation 3 yields the following inequality.

$$A_2 \leq \mathbb{E}_{\xi_{[t]}}\boldsymbol{\Lambda}_t\cdot\sum_{\boldsymbol{\Lambda}_t}\lambda^2\|\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\|^2 \leq \boldsymbol{\Lambda}_t\cdot\sum_{\boldsymbol{\Lambda}_t}\lambda^2\mathbb{E}_{\xi_{[t]}}\|\sum_{\mathcal{G}_{t\lambda}}\boldsymbol{\nabla}Q[tm]\|^2$$

Regarding A_1, the term $\|\ldots\|^2$ is expressed as a scalar product and expanded

as follows.

$$
\begin{aligned}
A_1 &= \mathbb{E}_{\xi_{[t]}} \sum_{\lambda,\lambda' \in \boldsymbol{\Lambda}_t} \Big(\sum_{\substack{m \in \mathcal{G}_{t\lambda}, \\ m' \in \mathcal{G}_{t\lambda'}}} \lambda\lambda' \cdot \langle G[tm] - \boldsymbol{\nabla} Q[tm], G[tm'] - \boldsymbol{\nabla} Q[tm'] \rangle \Big) \\
&= \text{diagonal } + \text{off-diagonal} \\
&= \sum_{\lambda \in \boldsymbol{\Lambda}_t} \sum_{m \in \mathcal{G}_{t\lambda}} \lambda^2 \cdot \mathbb{E}_{\xi_{[t]}} \|G[tm] - \boldsymbol{\nabla} Q[tm]\|^2 + \mathbb{E}_{\xi_{[t,m' \neq m]}} \left(\mathbb{E}_{\xi} \langle G[tm] - \boldsymbol{\nabla} Q[tm], G[tm'] - \boldsymbol{\nabla} Q[tm'] \rangle \right) \\
&\leq \sum_{\boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda} \cdot d \cdot \sigma^2 + d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2 D^2
\end{aligned}
$$

The sum over the off-diagonal terms (i.e., $(\lambda, m) \neq (\lambda', m')$) is bounded by $d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2 D^2$. Moreover, if $\lambda \neq \lambda'$, then $m \neq m'$ because $\mathcal{G}_{t\lambda}$ and $\mathcal{G}_{t\lambda'}$ are disjoint sets and thus for any off-diagonal pair $(\lambda, m), (\lambda, m')$ we have $m \neq m'$.

$$
\begin{aligned}
&\mathbb{E}_{\xi_{[t]}} \langle G[tm] - \boldsymbol{\nabla} Q[tm], G[tm'] - \boldsymbol{\nabla} Q[tm'] \rangle \\
&\quad = \mathbb{E}_{\xi_{[t,m' \neq m]}} \left(\mathbb{E}_{\xi} \langle G[tm] - \boldsymbol{\nabla} Q[tm], G[tm'] - \boldsymbol{\nabla} Q[tm'] \rangle \right) \\
&\quad = \mathbb{E}_{\xi_{[t,m' \neq m]}} \langle \mathbb{E}_{\xi} G[tm] - \boldsymbol{\nabla} Q[tm], G[tm'] - \boldsymbol{\nabla} Q[tm'] \rangle \\
&\quad = \mathbb{E}_{\xi_{[t,m' \neq m]}} (\langle \mathbb{E}_{\xi} G[tm], G[tm'] \rangle - \langle \boldsymbol{\nabla} Q[tm], G[tm'] \rangle - \langle \mathbb{E}_{\xi} G[tm], \boldsymbol{\nabla} Q[tm'] \rangle + \langle \boldsymbol{\nabla} Q[tm], \boldsymbol{\nabla} Q[tm'] \rangle) \\
&\quad \leq \mathbb{E}_{\xi_{[t,m' \neq m]}} (\|\mathbb{E}_{\xi} G[tm]\| \cdot \|G[tm']\| + \|\boldsymbol{\nabla} Q[tm]\| \cdot \|G[tm']\| \\
&\qquad + \|\mathbb{E}_{\xi} G[tm]\| \cdot \|\boldsymbol{\nabla} Q[tm']\| + \|\boldsymbol{\nabla} Q[tm]\| \cdot \|\boldsymbol{\nabla} Q[tm']\|) \\
&\quad \leq d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2 D^2
\end{aligned}
$$

Hence, we obtain the following inequalities for S_1 and S_2.

$$
S_1 \leq \sum_{\boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda} \cdot d \cdot \sigma^2 + \boldsymbol{\Lambda}_t \cdot \sum_{\boldsymbol{\Lambda}_t} \lambda^2 \mathbb{E}_{\xi_{[t]}} \| \sum_{\mathcal{G}_{t\lambda}} \boldsymbol{\nabla} Q[tm] \|^2 + d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2 D^2
$$

$$
S_2 \leq \left\| \frac{1}{\mathcal{G}_{t\lambda}} \sum_{\mathcal{G}_{t\lambda}} \boldsymbol{\nabla} Q(\boldsymbol{x}_t) - \boldsymbol{\nabla} Q[tm] \right\|^2
$$

Recall that, since $m = [g, l] \in \mathcal{G}_{t\lambda}$, we have $\boldsymbol{\nabla} Q[tm] = \boldsymbol{\nabla} Q(\boldsymbol{x}_{t-\tau_{tl}})$. By

applying the Lipschitz inequality, we get:

$$
\begin{aligned}
S_2 &\le K^2 \|\boldsymbol{x}_t - \boldsymbol{x}_{t-\Lambda^{-1}(\lambda)}\|^2 \\
&\le K^2 \left\| \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \boldsymbol{x}_{j+1} - \boldsymbol{x}_j \right\|^2 \le K^2 \left\| \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j \sum_{\nu\in\boldsymbol{\Lambda}_j} \nu \sum_{\mathcal{G}_{j\nu}} G[jm] \right\|^2 \\
&\le \underbrace{K^2 \left\| \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j \sum_{\nu\in\boldsymbol{\Lambda}_j} \nu \sum_{\mathcal{G}_{j\nu}} (G[jm] - \boldsymbol{\nabla} Q[jm]) \right\|^2}_{S_3 = \|a\|^2} \\
&+ \underbrace{K^2 \left\| \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j \sum_{\nu\in\boldsymbol{\Lambda}_j} \nu \sum_{\mathcal{G}_{j\nu}} \boldsymbol{\nabla} Q[jm] \right\|^2}_{S_4 = \|b\|^2} + 2K^2 \langle a, b \rangle
\end{aligned}
$$

Hence, we obtain the following inequalities for S_3 and S_4.

$$
\mathbb{E}_{\xi_{[j]},\ldots} S_3 \le \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathcal{G}_{j\nu} \cdot d \cdot \sigma^2 \quad \text{(cross-products vanish)}
$$

$$
\mathbb{E}_{\xi_{[j]},\ldots} S_4 \le \Lambda^{-1}(\lambda) \sum_{j=k-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \boldsymbol{\Lambda}_j \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathbb{E} \left\| \sum_{\mathcal{G}_{j\nu}} \boldsymbol{\nabla} Q[jm'] \right\|^2 \quad \text{(by Eq. 3).}
$$

Moreover, we have $\mathbb{E}_* \langle a, b \rangle = \langle \mathbb{E}_* a, b \rangle = 0$.

$$
\begin{aligned}
\mathbb{E} S_2 \le\ & K^2 \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathcal{G}_{j\nu} \cdot d \cdot \sigma^2 \\
& + K^2 \Lambda^{-1}(\lambda) \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \boldsymbol{\Lambda}_j \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathbb{E} \left\| \sum_{\mathcal{G}_{j\nu}} \boldsymbol{\nabla} Q[jm'] \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\xi_{[t]}} Q(\boldsymbol{x}_{t+1}) - Q(\boldsymbol{x}_t) \le & -\frac{\gamma_t}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|^2 \\
& + \sum_{\boldsymbol{\Lambda}_t} \left(\frac{K \gamma_t^2 \boldsymbol{\Lambda}_t \lambda^2}{2} - \frac{\gamma_t \lambda}{2 \mathcal{G}_{t\lambda}} \right) \mathbb{E} \left\| \sum_{\mathcal{G}_{t\lambda}} \boldsymbol{\nabla} Q[tm] \right\|^2 \\
& + \left(\frac{K \gamma_t^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda} + \frac{\gamma_t K^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathcal{G}_{j\nu} \right) \cdot d \cdot \sigma^2 \\
& + \frac{\gamma_t K^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \Lambda^{-1}(\lambda) \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \boldsymbol{\Lambda}_j \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathbb{E} \left\| \sum_{j\nu} \boldsymbol{\nabla} Q[jm'] \right\|^2
\end{aligned}
$$

Summing for $t = 1, \ldots, T$, we arrive at the following inequality.

$$\mathbb{E}Q(\boldsymbol{x}_{t+1}) - Q(\boldsymbol{x}_1) \leq -\sum_t \frac{1}{2}\left(\gamma_t \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda}\right) \|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|^2$$
$$+ \sum_t \left(\frac{K\gamma_t^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda} + \frac{\gamma_t K^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathcal{G}_{j\nu} \right) \cdot d \cdot \sigma^2$$
$$+ \sum_t \sum_{\boldsymbol{\Lambda}_t} \left(\frac{K\gamma_t^2 \boldsymbol{\Lambda}_t \lambda^2}{2} - \frac{\gamma_t \lambda}{2\mathcal{G}_{t\lambda}} \right) \mathbb{E} \left\| \sum_{\mathcal{G}_{t\lambda}} \boldsymbol{\nabla} Q[tm] \right\|^2$$
$$+ \sum_t \left(\sum_{s=1}^{\infty} \sum_{\boldsymbol{\Lambda}_{t+s}} \gamma_{t+s} K^2 \nu \mathcal{G}_{t+s,\nu} \Lambda^{-1}(\nu) \mathbb{I}(s \leq \Lambda^{-1}(\nu)) \right) \frac{\gamma_t \boldsymbol{\Lambda}_t \lambda^2}{2} \mathbb{E} \left\| \sum_{\mathcal{G}_{t\lambda}} \boldsymbol{\nabla} Q[tm] \right\|^2$$

The last term comes from the following observation.

$$\sum_{t=1}^{T} \sum_{\boldsymbol{\Lambda}_t} \sum_{s=1}^{\infty} Q_t^{\lambda} Z_{t-s} \mathbb{I}(s \leq \Lambda^{-1}(\lambda)) = \sum_{s=1}^{\infty} \sum_{t=1}^{T} \sum_{\boldsymbol{\Lambda}_t} Q_t^{\lambda} Z_{t-s} \mathbb{I}(s \leq \Lambda^{-1}(\lambda))$$
$$= \sum_{s=1}^{\infty} \sum_{l=1-s}^{T-s} \sum_{\boldsymbol{\Lambda}_{l+s}} Q_{l+s}^{\lambda} Z_l \mathbb{I}(s \leq \Lambda^{-1}(\lambda)) = \sum_{s=1}^{\infty} \sum_{t=1}^{T} \sum_{\boldsymbol{\Lambda}_{t+s}} Q_{t+s}^{\lambda} Z_t \mathbb{I}(s \leq \Lambda^{-1}(\lambda))$$
$$= \sum_{t=1}^{T} \left(\sum_{s=1}^{\infty} \sum_{\boldsymbol{\Lambda}_{t+s}} Q_{t+s}^{\lambda} \mathbb{I}(s \leq \Lambda^{-1}(\lambda)) \right) Z_t$$

Since the two last terms sum to a non-positive value, we arrive at the following inequality.

$$\sum_t \frac{1}{2}\left(\gamma_t \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda}\right) \|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|^2 \leq Q(\boldsymbol{x}_1) - Q(\boldsymbol{x}^*)$$
$$+ \sum_t \left(\frac{K\gamma_t^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda} + \frac{\gamma_t K^2}{2} \sum_{\boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \sum_{\boldsymbol{\Lambda}_j} \nu^2 \mathcal{G}_{j\nu} \right) \cdot d \cdot \sigma^2 + \mathcal{O}\left(\frac{1}{K \cdot \sqrt{|\xi|}}\right)$$

□

We first recall Definition 4, which introduces the adaptive learning rate schedule, before we prove Theorem 4 via employing Lemma 3. Due to the choice of the learning rate (Definition 4), the inequality in Theorem 4 reduces to the following inequality.

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|Q(\boldsymbol{x}_t)\|^2 \leq S_5 + S_6 + S_7$$

First, we obtain the following equality for S_5.

$$S_5 = \frac{2(Q(\boldsymbol{x}_1) - Q(\boldsymbol{x}^*))}{\sum_{t=1}^{T} \gamma_t \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda}} = \frac{2\gamma^2 KTM \cdot d \cdot \sigma^2}{\gamma TM} = 2\gamma K \cdot d \cdot \sigma^2$$

Regarding S_6, we obtain the following inequality.

$$S_6 = \frac{\sum_{t=1}^{T} K\gamma_t^2 \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda}}{\sum_{t=1}^{T} \gamma_t \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda}} \cdot d \cdot \sigma^2 = \frac{K\gamma^2 \sum_{t=1}^{T} \mu_t^2 \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda^2 \mathcal{G}_{t\lambda}}{\gamma TM} \cdot d \cdot \sigma^2$$

$$\leq \frac{K\gamma^2 \sum_{t=1}^{T} \mu_t^2 \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda}}{\gamma TM} \cdot d \cdot \sigma^2 \text{ (since } \lambda^2 \leq \lambda \leq 1)$$

$$\leq \frac{K\gamma^2 TM\mu_{\max}}{\gamma TM} \cdot d \cdot \sigma^2 = \mu_{\max}\gamma K \cdot d \cdot \sigma^2$$

Finally, we obtain the following inequality for S_7.

$$S_7 = \frac{\sum_{t=1}^{T} \gamma_t K^2 \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \gamma_j^2 \sum_{\lambda' \in \boldsymbol{\Lambda}_j} \lambda'^2 M_{j\lambda'}}{\gamma TM} \cdot d \cdot \sigma^2$$

$$\leq \frac{K^2\gamma^3 \sum_{t=1}^{T} \mu_t \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} \sum_{j=t-\Lambda^{-1}(\lambda)}^{t-1} \mu_j^2 \sum_{\lambda' \in \boldsymbol{\Lambda}_j} \lambda'^2 M_{j\lambda'}}{\gamma TM} \cdot d \cdot \sigma^2$$

$$\leq \frac{K^2\gamma^3 \sum_{t=1}^{T} \sum_{\lambda \in \boldsymbol{\Lambda}_t} \lambda \mathcal{G}_{t\lambda} M\Lambda^{-1}(\lambda)\mu_{\max}}{\gamma TM} \cdot d \cdot \sigma^2$$

$$\leq \frac{K^2\gamma^3 \sum_{t=1}^{T} \sum_{\lambda \in \boldsymbol{\Lambda}_t} \mathcal{G}_{t\lambda} M\chi\mu_{\max}}{\gamma TM} \cdot d \cdot \sigma^2 \leq \frac{K^2\gamma^3 TM^2\chi\mu_{\max}}{\gamma TM} \cdot d \cdot \sigma^2$$

$$\leq \gamma^2 K^2 M\chi\mu_{\max} \cdot d \cdot \sigma^2$$

Hence, we prove the ergodic convergence rate.

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\boldsymbol{\nabla} Q(\boldsymbol{x}_t)\|^2 \leq (2 + \mu_{\max} + \gamma KM\chi\mu_{\max}) \cdot \gamma K \cdot d \cdot \sigma^2 + d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2D^2$$

□

Theorem 5 (Convergence time complexity)**.** *Given a mini-batch size $|\xi|$, the number of gradients M the server waits for before updating the model, and the total number of epochs T, the time complexity for convergence of Kardam is:*

$$\mathcal{O}\left(\frac{\mu_{\max}}{\sqrt{T \cdot |\xi| \cdot M}} + \frac{\chi \cdot \mu_{\max}}{T} + d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2D^2\right)$$

where χ denotes a constant such that for all τ_{tl}, the following inequality holds:

$$\tau_{tl} \cdot \Lambda(\tau_{tl}) \leq \chi \tag{4}$$

Theorem 5 highlights the relation between the staleness and the convergence time complexity. Furthermore, this time complexity is linearly dependent on the decay bound (χ) and the maximum adaptive rate (μ_{max}).

We now prove Theorem 5 by employing Theorem 4 along with Definition 4.

Proof. Substituting the value of γ from Definition 4 in RHS of Theorem 4, we get the following.

$$\begin{aligned}
&(2 + \mu_{\max} + \gamma K M \chi \mu_{\max}) \cdot \gamma K \cdot d \cdot \sigma^2 + d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2D^2 \\
&= \mathcal{O}\left(\frac{\mu_{\max}}{\sqrt{T \cdot |\xi| \cdot M}} + \frac{\chi \cdot \mu_{\max}}{T} + d \cdot \sigma^2 + 2DK\sigma\sqrt{d} + K^2D^2 \right)
\end{aligned}$$

Note that $\sigma = \mathcal{O}(1/\sqrt{|\xi|})$ (Assumption 2) and therefore the bound is also dependent on n.

□

Remark 4 (Dampening functions comparison)**.** *Given two dampening functions* $\Lambda_1(\tau) = \frac{1}{1+\tau}$ *and* $\Lambda_2(\tau) = exp(-\alpha \sqrt[\beta]{\tau})$*, and the convergence time complexity from Theorem 5,* $\Lambda_2(\tau)$ *converges faster than* $\Lambda_1(\tau)$ *when* $\frac{\beta}{e} < \alpha \leq \frac{ln(\tau+1)}{\sqrt[\beta]{\tau}}$.

We also empirically highlight this remark by comparing these two functions in our main paper where DynSGD [4] employs Λ_1 and Kardam employs Λ_2.

Proof. From Inequality 4, we have the following for Λ_1 and Λ_2.

$$\begin{aligned}
\chi_1 &= \max_{\tau} \left\{ \frac{\tau}{\tau + 1} \right\} \\
\chi_2 &= \max_{\tau} \left\{ \tau \cdot exp(-\alpha \sqrt[\beta]{\tau}) \right\}
\end{aligned}$$

The maximum value of $\{\tau \cdot exp(-\alpha \sqrt[\beta]{\tau})\}$ is $\left(\frac{\beta}{e\alpha}\right)^{\beta}$ when $\tau = \left(\frac{\beta}{\alpha}\right)^{\beta}$. We get that $\chi_1 \geq \chi_2$ when the following holds.

$$\frac{\tau}{\tau + 1} \geq \left(\frac{\beta}{e\alpha}\right)^{\beta}$$

Hence, from the above inequality, we get the following.

$$\tau \geq \frac{1}{\left(\frac{e\alpha}{\beta}\right)^{\beta} - 1}$$

Note that since $\tau > 0$, we get $\left(\frac{e\alpha}{\beta}\right)^{\beta} > 1$ which leads to the following lower bound on α.

$$\alpha > \frac{\beta}{e} \tag{5}$$

Furthermore, for the μ_{max} terms, we compare the values between the two dampening functions.

$$\mu_1 = \max_{\tau}\left\{\frac{M}{\sum_{\mathbf{\Lambda}_t}\lambda\cdot|\mathcal{G}_{t\lambda}|}\right\} = \max_{\tau}\left\{\frac{M}{\sum_{\tau}\frac{1}{\tau+1}\cdot|\mathcal{G}_{t\lambda}|}\right\}$$
$$\mu_2 = \max_{\tau}\left\{\frac{M}{\sum_{\mathbf{\Lambda}_t}\lambda\cdot|\mathcal{G}_{t\lambda}|}\right\} = \max_{\tau}\left\{\frac{M}{\sum_{\tau}exp(-\alpha\sqrt[\beta]{\tau})\cdot|\mathcal{G}_{t\lambda}|}\right\}$$

Hence, for $\mu_1 \geq \mu_2$, we need to show that $\frac{1}{\tau+1} \leq exp(-\alpha\sqrt[\beta]{\tau})$, i.e., $\tau+1 \geq exp(\alpha\sqrt[\beta]{\tau})$. The relation holds for any α with the upper bound as follows.

$$\alpha \leq \frac{ln(\tau+1)}{\sqrt[\beta]{\tau}} \tag{6}$$

From Inequalities 5 and 6, we get the following.

$$\frac{\beta}{e} < \alpha \leq \frac{ln(\tau+1)}{\sqrt[\beta]{\tau}}$$

One possible setting is $\beta \approx 1.85$ when $1 \leq \tau \leq 10$, $\beta \approx 3.1$ when $11 \leq \tau \leq 33$, and $\beta \approx 4$ when $34 \leq \tau \leq 75$. Given these values of β and τ, $\Lambda_2(\tau)$ has a smaller convergence time complexity (Theorem 5) than $\Lambda_1(\tau)$. Hence, $\Lambda_2(\tau)$ converges faster than $\Lambda_1(\tau)$.

□

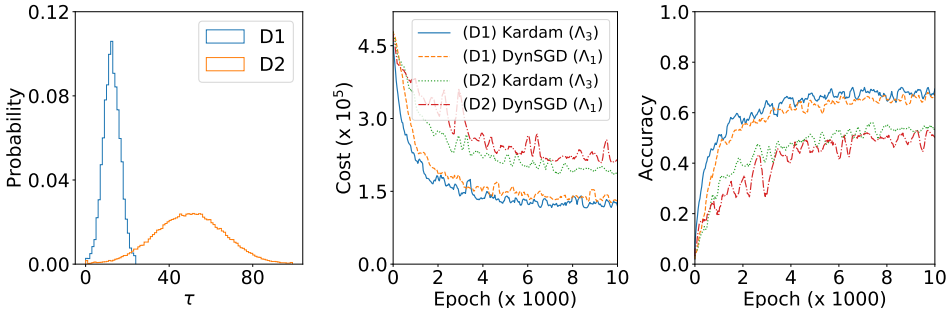
3 Additional Experimental Results.

We also evaluate the performance of Kardam for image classification on the EMNIST dataset[2] consisting 814,255 examples of handwritten characters and digits (62 classes). We perform min-max scaling normalization as a pre-processing step resulting in 784 normalized input. We split the dataset into 697,932 training and 116,323 test examples and employ a base learning rate of $8 * 10^{-4}$ alongside a mini-batch of 100 examples if not stated otherwise.

***M*-soft-asynchronous.** Recall that parameter M in M-soft-async denotes the number of required responses that Kardam aggregates before performing one model update. We evaluate Kardam with different values for M and analyze the effect on convergence. Figure 3 depicts that a larger M leads to a faster convergence and less noise in the learning curve w.r.t cost. The explanation for this observation is that by increasing M, Kardam increases the robustness of each update.

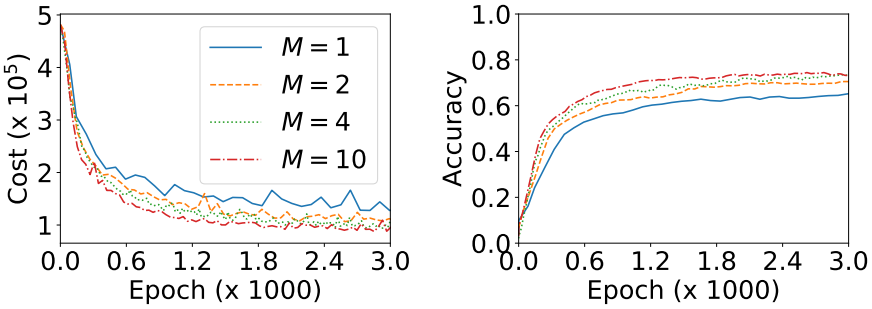
Choosing the optimal value for the parameter M involves balancing the trade-off between the robustness of each update and the update latency. Increasing the value of M triggers a decrease in the model update throughput. Therefore, the model converges faster in terms of epochs but slower in terms of time. The staleness in the accumulated gradients diminishes this increase in the robustness.

[2] `https://www.nist.gov/itl/iad/image-group/emnist-dataset`



(a) Staleness distributions (b) Cost (c) Accuracy

Figure 2: Impact of staleness for EMNIST.

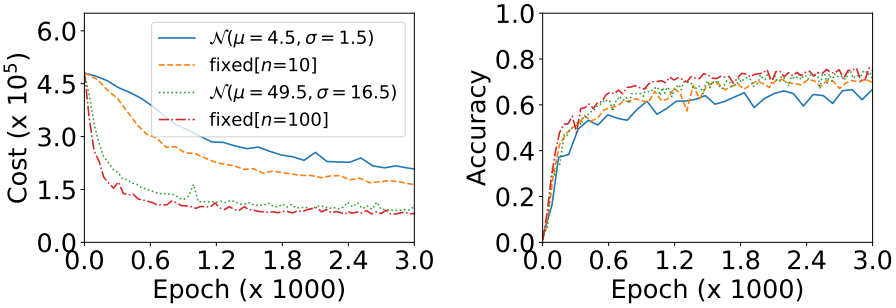


(a) Cost (b) Accuracy

Figure 3: Effect of M-soft-async for EMNIST.

Mini-batch size. We empirically show the trade-off between the mini-batch size and the convergence speed. To exclude the effect of staleness, we perform synchronous updates (SSGD) for this experiment. We have two settings to compare the effect of a variable mini-batch size. We fix the mini-batch size for the first one to $|\xi|$, and sample the size from a Gaussian distribution $\mathcal{N}(\mu = \frac{|\xi|-1}{2}, \sigma = \frac{|\xi|-1}{6})$ for the second.

Figure 4 depicts that the fixed scenario results in faster convergence as expected. The main reason is that a higher mini-batch size leads to a more robust gradient estimation with less noise. Theorem 5 also justifies this observation as the convergence time is inversely proportional to $\sqrt{|\xi|}$.



(a) Cost (b) Accuracy

Figure 4: Effect of mini-batch size ($|\xi|$) for EMNIST.

References

[1] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS*, pages 118–128, 2017.

[2] L. Bottou. Online learning and stochastic approximations. *Online learning in neural networks*, 17(9):142, 1998.

[3] M. J. Fischer, N. A. Lynch, and M. S. Paterson. Impossibility of distributed consensus with one faulty process. *JACM*, 32(2):374–382, 1985.

[4] J. Jiang, B. Cui, C. Zhang, and L. Yu. Heterogeneity-aware distributed parameter servers. In *SIGMOD*, pages 463–478, 2017.

[5] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *NIPS*, pages 2737–2745, 2015.

[6] R. Zhang, S. Zheng, and J. T. Kwok. Asynchronous distributed semi-stochastic gradient optimization. In *AAAI*, pages 2323–2329, 2016.

[7] W. Zhang, S. Gupta, X. Lian, and J. Liu. Staleness-aware async-sgd for distributed deep learning. In *IJCAI*, pages 2350–2356, 2016.