




Adaptive Step-Size Methods for Compressed SGD With Memory Feedback

Adarsh M. Subramaniam , *Graduate Student Member, IEEE*,
Akshayaa Magesh , *Graduate Student Member, IEEE*, and Venugopal V. Veeravalli , *Fellow, IEEE*

I. INTRODUCTION

Abstract—Distributed and decentralized optimization problems, such as those encountered in federated learning, often suffer from a communication bottleneck at compute nodes when a large number of messages get aggregated. To address this communication bottleneck, compressed Stochastic Gradient Descent (SGD) algorithms have been proposed. Most existing compressed SGD algorithms use non-adaptive step-sizes (constant or diminishing) that depend on unknown system parameters to provide theoretical convergence guarantees, with the step-sizes being fine-tuned in practice to obtain good empirical performance for a given dataset and learning model. Such fine-tuning might be impractical in many scenarios. Therefore, it is of interest to study compressed SGD using adaptive step-sizes that do not depend on unknown system parameters. Motivated by prior work that employs adaptive step-sizes for uncompressed SGD, we develop an Armijo rule based step-size selection method for compressed SGD with feedback. In particular, we first motivate a novel scaling technique for Gradient Descent (GD) with Armijo step-size search, based on which we develop an Armijo step-search method with *scaling* for the descent step of the compressed SGD algorithm with memory feedback. We then establish that our algorithm has order-optimal convergence rates for convex-smooth and strong convex-smooth objectives under an *interpolation condition*, and for non-convex objectives under a *strong growth condition*. Experiments comparing our proposed algorithm with state-of-the-art compressed SGD methods with memory feedback demonstrate a notable improvement in terms of training loss across various levels of compression.

Index Terms—Adaptive step-sizes, compression, distributed optimization, memory feedback, stochastic gradient descent.

Manuscript received 10 January 2024; revised 30 March 2024; accepted 30 March 2024. Date of publication 5 April 2024; date of current version 17 May 2024. This work was supported in part by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196, through the University of Illinois at Urbana-Champaign. An earlier version of this paper was presented at the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [DOI: 10.1109/ICASSP49357.2023.10096611]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Laurent Condat. (*Corresponding author: Venugopal V. Veeravalli.*)

The authors are with the ECE Department and the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: adarshm2@ILLINOIS.EDU; amagesh2@ILLINOIS.EDU; vvv@ILLINOIS.EDU).

Digital Object Identifier 10.1109/TSP.2024.3385577

CONSIDER the following optimization problem, where the objective is to minimize the average of n functions:

$$\min_x f(x) = \min_x \frac{1}{n} \sum_{i=1}^n f_i(x). \quad (1)$$

This formulation is widely used in machine learning, where f_i is the loss function corresponding to datapoint i , the size of the training set is n , and x denotes the parameters of the learning model.

A practical approach to solving the optimization problem in (1) is Stochastic Gradient Descent (SGD), where one or a small batch of the component functions f_i 's is sampled at random at a given iteration step, and the iterate x is updated using the gradients of the sampled functions. In distributed and federated optimization [2], there is a *master node* and the component functions are split among multiple compute nodes, termed *worker nodes*. In the optimization process, the gradients are computed at the worker nodes and transmitted to the master node. In the scenario where multiple worker nodes transmit gradients to the master node at the same time, the communication of the gradients can create a communication bottleneck at the master node, and hence compressed versions of the gradients are used in the SGD iterations.

Compressed SGD algorithms have been studied in several recent works (see, e.g., [3], [4], [5], [6], [7]). In the earlier works [3], [4], only the top k (e.g., $k = 1\%$) components of the gradient (with feedback) are used in the iteration update. The convergence properties of such compressed SGD methods with feedback have also been discussed in [5], [8].

Existing works on compressed SGD algorithms use diminishing or constant step-sizes that depend on system parameters to obtain convergence guarantees. In machine learning applications, the step-sizes are fine-tuned to the dataset and the machine learning model to ensure good empirical performance. However, fine-tuning with data sampled from the real world during training is impractical. In order to address the limitations of fine-tuning, it is of interest to study adaptive step-size methods for compressed SGD that do not involve unknown system parameters. Recent literature has examined adaptive step-size methodologies for uncompressed SGD [9], [10], [11], [12], [13], [14]. In [9], an efficient adaptation of the classical Armijo step-search technique [15] for neural network training has been proposed. The proofs of convergence of these strategies assume

the existence of a point x^* that minimizes all functions f_i in the optimization problem (1). This condition is termed *interpolation* and theoretical justification for interpolation has been established in [16], [17].

Motivated by adaptive step-size methodologies for uncompressed SGD, we introduce an adaptive step-size method for compressed SGD with feedback based on Armijo step-size search. To this end, we propose a *scaling technique*, which is crucial to the convergence of the proposed adaptive step-size compressed SGD algorithm. To provide some intuition for the *scaling* technique, we study classical GD with Armijo step-size search and scaling, and illustrate through examples that scaling can improve convergence in asymmetric objective functions. We then draw parallels between gradients in asymmetric functions and compressed gradients with memory feedback to motivate the use of scaling for compressed SGD algorithms with feedback.

In this work, we build on our results in [1] and establish that the convergence rate for GD on convex functions with Armijo step-size search is $O(1/T)$ over a horizon of length T , when the step-size returned by Armijo step-size search is scaled by a constant and applied to the descent step. Further, we illustrate the impact of scaling through examples. The lemmas presented rigorously characterize the properties of the involved constants and experiments in addition to those presented in [1], reiterate that the proposed algorithm demonstrates accelerated convergence as compared to prevailing non-adaptive step-size methods with feedback at equivalent compression levels. Results presented in this work also illustrate that compressed SGD with memory feedback and Armijo step-size search may not converge without scaling.

To the best of our knowledge, the algorithm proposed in this paper is the first adaptive step-size method for compressed SGD with feedback with theoretical guarantees.

Our Contributions

- 1) We propose a *scaling* technique for GD with Armijo step-size search, and present examples that illustrate that such scaling results in faster convergence of GD in asymmetric and convex problems.
- 2) We establish for a time horizon T that GD on convex functions with Armijo step-size search and scaling converges at the rate of $O(\frac{1}{T})$ for all values of $\sigma \in (0, 1)$ (see (4) for a definition).
- 3) We propose a computationally feasible and efficient adaptive step-size search algorithm for compressed SGD with feedback. Our proposed approach incorporates a scaling technique developed in this work and employs a biased compression operator.
- 4) Under the *interpolation condition* (Definition 2), we establish that our algorithm has a convergence rate of $O(\frac{1}{T})$ for convex and smooth objective functions, and $O(\beta^T)$ ($\beta < 1$) for strongly-convex and smooth objective functions. Additionally, we prove that under a *strong growth condition* (Definition 3), the algorithm has a convergence rate of $O(\frac{1}{T})$ in the non-convex setting.
- 5) We present a comprehensive analysis of the dependence of the constants in the convergence proofs on the scaling parameter.
- 6) We present in our experiments that the proposed scaling technique is more than just a proof technicality. We also illustrate that our algorithm for compressed SGD outperforms existing non-adaptive compressed SGD methods with feedback on neural network training tasks at various levels of compression.

II. PRELIMINARIES

In this section, we study the application of the classic Armijo step-size search within the context of compressed stochastic gradients. Consider the optimization problem in (1). At any given time t , let x_t denote the iterate of the optimization algorithm, and i_t represent the selected datapoint or data batch. Let the loss function corresponding to the batch i_t at the iterate x_t be denoted by $f_{i_t}(x_t)$, and the step-size utilized at time t be denoted by η_t .

A. Compression Operator top_k

In the experimental evaluation of our proposed algorithm, the gradient in each iteration is compressed using the top_k compression operator (see, e.g., [5], [6]). Let $x \in \mathbb{R}^d$. Let \mathbb{T}_k be the set of indices of the k elements of x with the highest magnitudes. The top_k operator compresses a vector x such that

$$\text{top}_k(x)_i = \begin{cases} (x)_i & \text{If } i \in \mathbb{T}_k \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

The top_k compression operator is an element of the class of biased compression operators \mathcal{C} [18] with the following property.

Definition 1: A compressor $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is an element of the class of biased compression operators \mathcal{C} iff $\forall v \in \mathbb{R}^d$ there exists $\gamma \in [0, 1)$ such that

$$\|v - \phi(v)\|^2 \leq (1 - \gamma)\|v\|^2. \quad (3)$$

For a compression operator ϕ , we define γ as the compression factor. The top_k compression operator satisfies the conditions in Definition 1 with $\gamma = \frac{k}{d}$.

B. Armijo Step-Size Search in Gradient Descent

The adaptive step-size compressed SGD algorithm proposed in this work efficiently adapts the classical Armijo step-size search method [15]. At each iteration t , the Armijo step-size search algorithm initiates the step-size search at α_{\max} (a parameter of the algorithm) for a given function f . Subsequently, it systematically reduces the step-size by a factor of $\rho < 1$ until the condition in Equation (4), termed Armijo condition, is satisfied.

$$f(x_t - \alpha_t \nabla f(x_t)) \leq f(x_t) - \sigma \alpha_t \|\nabla f(x_t)\|^2. \quad (4)$$

Here, σ is a parameter of the Armijo step-size search method. The first step-size α_t that satisfies the Armijo condition is then utilized in the t -th iteration of the Gradient Descent (GD) algorithm. Algorithm 1 presents the pseudocode for the Armijo step-size search method.

Algorithm 1 Armijo Step-Size Search

```

1: procedure ARMIMO STEP-SIZE SEARCH( $f, \alpha_{\max}, x_t, t, \rho$ )
2:    $\alpha_t = \alpha_{\max}$ 
3:   repeat
4:      $\alpha_t \leftarrow \alpha_t \rho$ 
5:      $\tilde{x}_{t+1} \leftarrow x_t - \alpha_t \nabla f(x_t)$ 
6:   until  $f(\tilde{x}_{t+1}) \leq f(x_t) - \sigma \alpha_t \|\nabla f(x_t)\|^2$ 
7:   return  $\alpha_t$ 
8: end procedure

```

C. Definitions

To establish the convergence of our algorithm, we require the following interpolation condition in the convex and strong-convex setting, and the strong growth condition in the non-convex setting.

Definition 2 (Interpolation [13]): A function $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ satisfies the interpolation condition if

$$\exists x^* \text{ s.t. } \nabla f_i(x^*) = 0, \forall i \in \{1, 2, \dots, n\}. \quad (5)$$

Interpolation is particularly applicable to neural networks characterized by a substantially larger number of parameters compared to the number of data points on which they are trained. Rigorous mathematical characterizations of interpolation are presented in [16], [17].

Definition 3 (Strong Growth Condition [9]): A function $f = \frac{1}{n} \sum_{i=1}^n f_i(x)$ satisfies the strong growth condition with constant ν if

$$\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x)\|^2 \leq \nu \|\nabla f(x)\|^2. \quad (6)$$

The Strong Growth Condition (SGC) has been widely used to show accelerated convergence rates for SGD on convex and strong-convex objectives [19]. Under SGC, accelerated convergence rates for incremental gradient methods [20], [21] and superlinear convergence for Gauss-Newton methods can be established.

III. GRADIENT DESCENT WITH ARMIMO STEP-SIZE SEARCH AND SCALING

In this section, we motivate the idea of *scaling* for GD with Armijo step-size search.

Optimization problems in machine learning applications often exhibit asymmetry, with gradients that do not align with the direction of a minimizer. For instance, consider GD applied to minimize the function $f(x) = \frac{x_1^2}{2} + \frac{x_2^2}{5}$. At a given point x , the negative gradient $-\nabla f(x)$ is orthogonal to the tangent and may not necessarily align towards the minimizer $(0, 0)$. Consequently, GD with Armijo step-size search can experience slower convergence, particularly when the maximum step-size α_{\max} is large. To address this challenge, we propose the idea of scaling. Here, the step-size α_t determined by Armijo's rule (4) is scaled by a factor $a < 2\sigma$, and the scaled step-size $\eta_t \triangleq a\alpha_t$ is utilized in the descent step, i.e.,

$$x_{t+1} = x_t - \eta_t \nabla f(x_t).$$

TABLE I
LOGARITHM OF FUNCTION VALUES ($\log_{10}(f(x_t))$) AFTER ≈ 2000
ITERATIONS OF GD WITH ARMIMO STEP-SIZE SEARCH;
COMPARING RESULTS WITH AND WITHOUT SCALING

Function $f(x)$	Method	$\alpha_{\max} = 10$	$\alpha_{\max} = 100$	$\alpha_{\max} = 500$
$\sum_{i=1}^{10} \frac{x_i^2}{2^i}$	Scaled	-18	-20	-227
	Non-scaled	-284	-246	-263
$\sum_{i=1}^{10} \frac{x_i^2}{3^i}$	Scaled	-37	-12	-234
	Non-scaled	-205	-285	-290
$\sum_{i=1}^{10} \frac{x_i^2}{2^i}$	Scaled	-7	-15	-30
	Non-scaled	-9	-8	-9
$\sum_{i=1}^{10} \frac{x_i^2}{i^2}$	Scaled	-65	-87	-108
	Non-scaled	-47	-48	-50
$\sum_{i=1}^{10} \frac{x_i^2}{i^3}$	Scaled	-5	-11	-12
	Non-scaled	-4	-3	-3

To illustrate the effect of scaling on asymmetric problems, we tabulate in Table I, the performance of GD with scaled ($a = 1.5\sigma$) and non-scaled ($a = 1$) Armijo step-size search. On symmetric problems with objectives $\sum_{i=1}^{10} \frac{x_i^2}{2^i}$ and $\sum_{i=1}^{10} \frac{x_i^2}{3^i}$, non-scaled GD outperforms scaled GD. However, on asymmetric problems with objectives $\sum_{i=1}^{10} \frac{x_i^2}{2^i}$, $\sum_{i=1}^{10} \frac{x_i^2}{i^2}$ and $\sum_{i=1}^{10} \frac{x_i^2}{i^3}$, scaled GD outperforms non-scaled GD. The difference is more pronounced when α_{\max} is large. The convergence of GD with Armijo step-size search and scaling is presented in the following theorem.

Theorem 1 (Deterministic-Uncompressed-Convex): Let f be convex and L smooth with a minimum x^* . The Armijo step-size search gradient descent for $\sigma \in (0, 1)$ and scale factor $a < 2\sigma$ satisfies

$$f(\bar{x}_T) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{\frac{2(1-\sigma)}{L} \rho (2a - \frac{a^2}{\sigma}) T}, \quad (7)$$

where $\bar{x}_T = \frac{1}{T} \sum_{i=0}^{T-1} x_i$.

The proof is presented in the Appendix. Note that, while the GD algorithm with Armijo step-size search converges for all values of $\sigma \in (0, 1)$, a convergence rate of $O(\frac{1}{T})$ for convex objectives has been proven only for values of $\sigma \geq 0.5$ [15]. However, typical values of σ used in practice are in the range $[10^{-5}, 10^{-1}]$ [22]. For our modified Armijo step-size search rule with scaling, we prove a convergence rate of $O(\frac{1}{T})$ for GD on convex objectives for all values of $\sigma \in (0, 1)$. To the best of our knowledge, this is the first such proof.

IV. COMPRESSED SGD WITH ADAPTIVE STEP-SIZES

In this section, drawing inspiration from the scaling technique employed in Gradient Descent (GD) with Armijo step-size search, we introduce an adaptive step-size algorithm for compressed SGD with feedback.

A. Compressed SGD With Armijo Step-Size Search and Scaling (CSGD-ASSS)

SGD methods with compressed gradients (see, e.g., [3], [4], [5], [6], [7]) utilize memory feedback to compensate for the error due to compression. The update rule for the iterate at time t is given by

$$x_{t+1} = x_t - \phi(m_t + \eta_t \nabla f_{i_t}(x_t)), \quad (8)$$

Algorithm 2 Compressed SGD + Armijo Step-Size Search and Scaling (CSGD-ASSS)

```

1:  $\alpha_{\max} = \alpha_0$ 
2: for  $t = 1, \dots, T$  do
3:   Sample batch  $i_t$  of data
4:    $\alpha_{\max} = \omega\alpha_{t-1}$ 
5:    $\alpha_t \leftarrow$  Armijo Step-Size Search( $f_{i_t}, \alpha_{\max}, x_t, t, \rho$ )
6:    $\eta_t = a\alpha_t$ 
7:    $g_t = \phi(m_t + \eta_t \nabla f_{i_t}(x_t))$ 
8:    $x_{t+1} \leftarrow x_t - g_t$ 
9:    $m_{t+1} = m_t + \eta_t \nabla f_{i_t}(x_t) - \phi(m_t + \eta_t \nabla f_{i_t}(x_t))$ 
10: end for
  
```

where η_t is the step-size, ϕ is a compressor in the class of biased compression operators, i_t is the index of the function sampled at time t , and m_t is the error value due to compression.

In Section V-A, we demonstrate with examples that SGD with Armijo step-size search as proposed in [9] does not generalize to the compressed setting, and can diverge at an exponential rate. A possible reason is that, in compressed SGD with feedback, the descent direction $\phi(m_t + \eta_t \nabla f_{i_t}(x_t))$ need not necessarily point towards a minimum, even if $\nabla f_{i_t}(x_t)$ points towards a minimum of f . Motivated by our discussion in Section III on scaling for GD with Armijo step-size search, we propose scaling as a fix for compressed SGD with Armijo step-size search and present the Compressed SGD with Armijo Step-Size Search and Scaling (CSGD-ASSS) algorithm.

At iteration t , in the CSGD-ASSS algorithm (Algorithm 2), we implement the Armijo step-size search (Algorithm 1) for the function f_{i_t} , which returns a value α_t that satisfies

$$f_{i_t}(x_t - \alpha_t \nabla f_{i_t}(x_t)) \leq f_{i_t}(x_t) - \sigma \alpha_t \|\nabla f_{i_t}(x_t)\|^2. \quad (9)$$

The step-size at iteration t is subsequently chosen as $\eta_t = a\alpha_t$, where a is a scaling factor. The algorithm computes the compressed gradient g_t with feedback error m_t , and scaled step-size η_t as

$$g_t = \phi(\eta_t \nabla f_{i_t}(x_t) + m_t). \quad (10)$$

The iterate and error for iteration $t + 1$ are updated respectively, as

$$x_{t+1} = x_t - g_t \quad (11)$$

$$m_{t+1} = \eta_t \nabla f_{i_t}(x_t) + m_t - \phi(\eta_t \nabla f_{i_t}(x_t) + m_t). \quad (12)$$

At iteration $t + 1$, the Armijo step-size search algorithm searches for the step-size starting from $\alpha_{\max} = \omega\alpha_t$, where ω is a parameter of CSGD-ASSS. Further explanation on the practical significance of ω is presented in Section V.

B. Convergence Analysis

We study the convergence of the CSGD-ASSS algorithm in the setting where the objective function is convex and strongly convex, under the interpolation condition (5). Additionally, for non-convex objectives, we examine convergence under the

TABLE II
SUMMARY OF PARAMETERS OF CSGD-ASSS

Parameter	Definition
α_{\max}	Initial value of step-size in Armijo step-size search
α_0	α_{\max} at iteration 1 of CSGD-ASSS
T	Number of iterations (time horizon)
ω	α_{\max} initialization parameter in Step 4 of CSGD-ASSS
a	Step-size scaling constant
m_t	Feedback error
ϕ	Compressor
γ	Compression factor of compressor
top_k	Compression operator top_k
ρ	Step-size decay parameter of Armijo search

strong growth condition (3). Notably, the convergence rates established under compression align with those derived in the uncompressed setting as detailed in [9].

Consider the minimization of the function f in the convex setting, with some f_i satisfying strong convexity with parameter $\mu_i > 0$. In this setting, we can establish the following *linear* convergence result.

Theorem 2 (CSGD-ASSS strongly convex): Let f_i be convex and L_i smooth for all $i \in [n]$, and μ_i strongly convex with $\mu_i > 0$ for some $i \in [n]$. Assume that the interpolation condition (5) is satisfied. Then there exists $\hat{a} > 0$ such that for $0 < a \leq \hat{a}$, the CSGD-ASSS algorithm with scaling a , parameter $\sigma \in (0, 1)$ and compression factor γ satisfies

$$E[\|x_t - x^*\|^2] \leq 2(\hat{\beta}(a))^t E[\|x_0 - x^*\|^2], \quad (13)$$

for some $\hat{\beta}(a) < 1$, where for $0 < \epsilon < \zeta$ and $\zeta \triangleq \frac{\sigma\gamma}{(2-\gamma)}$,

$$\hat{\beta}(a) \triangleq \max\{\beta_1(p(\epsilon), r(\epsilon), a), \beta_2(a)\},$$

$$\beta_1(p(\epsilon), r(\epsilon), a) \triangleq \left(\mu_{\max} a \alpha_{\max} + p(\epsilon) + (1 - \gamma)(1 + r(\epsilon)) \right),$$

$$\beta_2(a) = \left(1 - \frac{\bar{\mu}a(1 - \sigma)}{L_{\max}} \right), \quad \hat{a} = \zeta - \epsilon,$$

$$\mu_{\max} = \max_{i \in [n]} \mu_i, \quad \bar{\mu} = \frac{\sum_{i=1}^n \mu_i}{n}. \quad (14)$$

The values of $p(\epsilon), r(\epsilon)$ are obtained from the maximization of a convex problem, which is presented in the proof of the theorem in the Section VIII-B.

For the minimization of convex objective functions, we prove the following convergence result for CSGD-ASSS under the interpolation condition.

Theorem 3 (CSGD-ASSS convex): Let $f_i(x)$ be convex and L_i smooth for $i \in [n]$, and assume that the interpolation condition (5) is satisfied. Then there exists $\hat{a} > 0$ such that for $0 < a \leq \hat{a}$, the CSGD-ASSS algorithm with scaling a , parameter $\sigma \in (0, 1)$, and compression factor γ , satisfies

$$E \left[f \left(\frac{1}{T} \sum_{t=0}^{T-1} x_t \right) \right] - E[f(x^*)] \leq \frac{1}{\delta_1(a)T} (E[\|x_0 - x^*\|^2]), \quad (15)$$

where for any $0 < \epsilon < \zeta$, $\zeta \triangleq \frac{\sigma\gamma}{(2-\gamma)}$ and $L_{\max} \triangleq \max_i L_i$,

$$\begin{aligned} C_{\delta_1} &\triangleq \rho \frac{2(1-\sigma)}{L_{\max}}, \\ \delta_1(a) &\triangleq C_{\delta_1} \left(2a - \frac{a^2}{\sigma} - \frac{a^2}{\sigma p(\epsilon)} - (1-\gamma) \left(1 + \frac{1}{r(\epsilon)} \right) \frac{a^2}{\sigma} \right), \\ \hat{a} &= (\zeta - \epsilon). \end{aligned} \quad (16)$$

The exact expression for $p(\epsilon)$, $r(\epsilon)$ and the proof of the theorem are similar to the strongly convex case, and the details are presented in the Appendix. To provide convergence guarantees in the non-convex case, we use the strong growth condition (6).

Theorem 4 (CSGD-ASSS non-convex): Let f_i be non-convex, L_i smooth and let f_i $i \in [n]$ satisfy the strong growth condition (6). Let $\inf_x f_i(x) > -\infty$ for all $i \in [n]$, and let the Lipschitz constants L_i be upper bounded by L_{UP} . Then, there exists $\hat{a}, \hat{\alpha}$ such that for $0 < a \leq \hat{a}$ and $\alpha_{\max} \leq \hat{\alpha}$,

$$\frac{1}{T} \sum_{t=0}^{T-1} E[\|\nabla f(x_t)\|^2] \leq \frac{(E[f(x_0)] - E[f(\hat{x}_T)])}{\delta T}, \quad (17)$$

where \hat{x}_T is a perturbed iterate [23] obtained from $\{x_i\}_{i=1}^T$.

A discussion on the choice of \hat{a} , and $\hat{\alpha}$ is presented in the Appendix.

C. Analysis of Constants in Convergence Proofs

In this section, the properties of constants in Theorems 2, 3 are studied. In Theorem 2,

$$\hat{\beta}(a) = \max\{\beta_1(p(\epsilon), r(\epsilon), a), \beta_2(a)\}, \quad (18)$$

where $p(\epsilon), r(\epsilon)$ are variables. In order to establish a precise characterization of the parameter $\hat{\beta}(a)$, we provide the following lemma to bound $\beta_1(p, r, a)$.

Lemma 5: For any $0 \leq a < \zeta$, consider the optimization problem

$$\inf_{p,r} \beta_1(p, r, a) \quad (19)$$

$$a < 2 / \left(\frac{1}{\sigma} + \frac{1}{\sigma p} + \frac{(1-\gamma)(1+\frac{1}{r})}{\sigma} \right), \quad (20)$$

$$p \geq 0, r \geq 0. \quad (21)$$

The optimization problem has a solution p^*, r^* , and let the optimal value be denoted by $\beta_{\text{LB}}(a) = \beta_1(p^*, r^*, a)$. Further, if $a_1 < a_2$, then

$$\beta_{\text{LB}}(a_1) < \beta_{\text{LB}}(a_2). \quad (22)$$

Remark 1: The condition in Equation (21) is obtained from the proof of convergence of Theorem 2 presented in the Appendix.

Remark 2: In Theorem 2, $\hat{\beta}(a) = \max\{\beta_1(p, r, a), \beta_2(a)\}$. Lemma 5 characterizes the infimum of $\beta_1(p, r, a)$ with respect to parameters p and r , thereby establishing a tight bound for $\hat{\beta}(a)$.

The proof of Lemma 5 is presented in the Appendix. From Lemma 5, it is evident that the smallest value $\beta_1(p, q, a)$ for a given scaling factor a , denoted as $\beta_{\text{LB}}(a)$, is a monotonically

increasing function of a . Simultaneously, from the definition of $\beta_2(a)$, its value monotonically decreases with a .

Consequently, as the parameter a increases from 0 to ζ , the value of $\beta_{\text{LB}}(a)$ increases, while the value of $\beta_2(a)$ decreases. This trade-off between $\beta_{\text{LB}}(a)$ and $\beta_2(a)$ leads to a decrease in the value of $\hat{\beta}(a)$, reaching a minimum before increasing. The precise value of a at which $\hat{\beta}(a)$ attains a minimum is contingent upon the parameters of the objective function to be minimized.

Similarly, the parameter $\delta_1(a)$ in Theorem 3 increases with respect to the scaling parameter a , attains a maximum, and decreases with a further increase in a . A characterization of the properties of $\delta_1(a)$ is presented in Lemma 14 in the Appendix.

Further, the upper limit ζ on the maximal value of the scaling parameter a in the proofs of convergence of CSGD-ASSS for convex and strong-convex objectives is directly proportional to the compression ratio γ .

V. EXPERIMENTAL RESULTS

The CSGD-ASSS algorithm is evaluated on the training of ResNet-18, ResNet-34, and DenseNet-121 neural nets on CIFAR-100 and CIFAR-10 datasets with top_k compression operator. Each layer, excluding those with fewer than 1000 parameters, is compressed. Using a batch size of 64, the networks are trained for 90 epochs.

Armijo step-size search uses $\sigma = 0.1$, and motivated by the analysis of GD with Armijo step-size search and scaling presented in Section III, we set the scaling factor a to be a multiple of σ . In our simulations, $a = 3\sigma$.

Experiments include compression rates of 1.5% and 10% for ResNet-34 and ResNet-18, and 4% and 10% for DenseNet-121. CSGD-ASSS is compared with non-adaptive compressed SGD in [4] using step-sizes 0.1, 0.05, 0.01. The reason for using the algorithm in [4] as a baseline is that prevalent *non-adaptive* compressed SGD algorithms with feedback are generally built on top_k gradient compression with memory feedback introduced in [4]. From Fig. 1, we observe that our proposed CSGD-ASSS algorithm (denoted by 3σ) outperforms fixed/non-adaptive step-size top_k compressed SGD with step-sizes 0.1, 0.05, 0.01, denoted by non-adap 0.1, non-adap 0.05, non-adap 0.01, respectively.

Note on computational complexity: In experiments, α_{\max} is 0.1 and is updated iteratively as $\alpha_{\max} = \omega \alpha_{t-1}$, with $\omega = 1.2$ and $\rho = 0.8$ ($\omega\rho < 1$). With this choice of parameters, CSGD-ASSS with Armijo step-size search computes, on average, fewer than two additional forward passes compared to non-adaptive compressed SGD on ResNets and DenseNets. Essentially, in practice, each iteration of Step 5 in CSGD-ASSS involves computing the gradient once, multiplying it with step-size α_t , and evaluating the stopping condition (Step 6) of the Armijo step-size search (Algorithm 1) twice. Evaluating the gradient is computationally more expensive than a forward pass in the SGD algorithm and in the federated learning [2] setup, forward pass steps can be parallelized in the worker nodes. Thus, in scenarios with large gradient sizes and multiple worker nodes, the primary bottleneck for implementing our algorithm

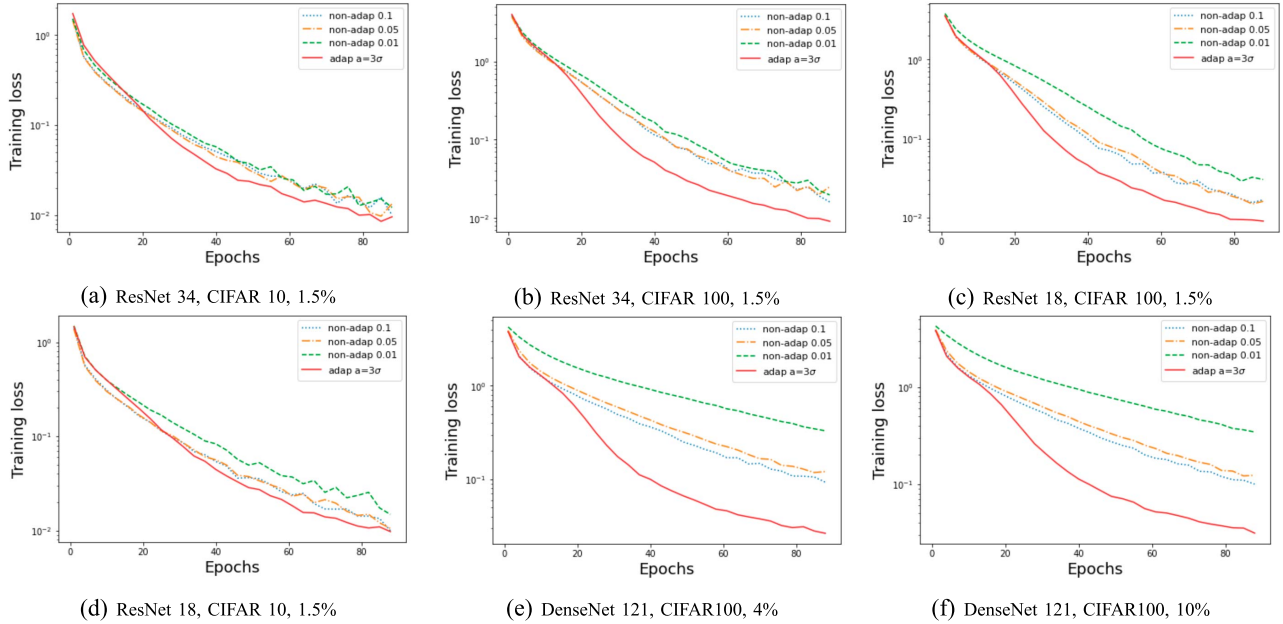


Fig. 1. Training loss of ResNets and DenseNets on CIFAR 10 and CIFAR 100 datasets at various levels of compression.

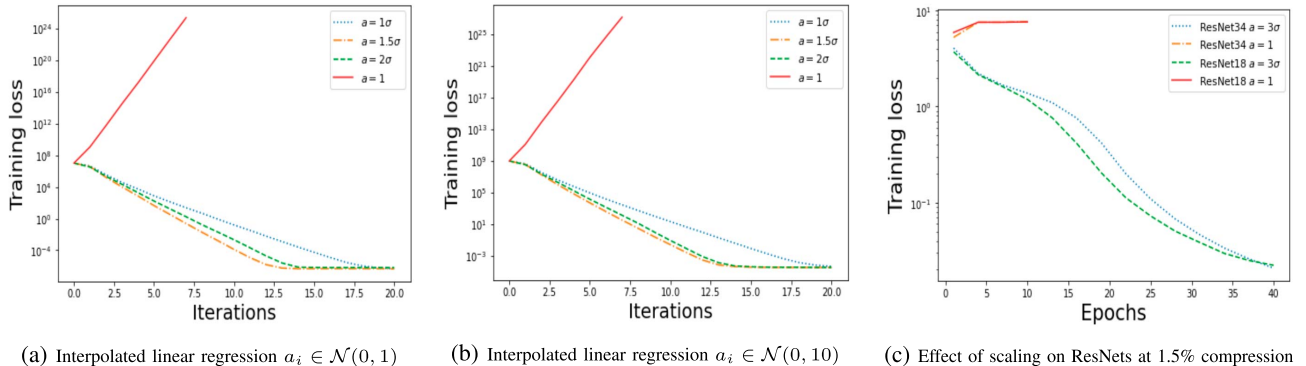


Fig. 2. Scaled vs non-scaled CSGD-ASSS at $\approx 1\%$ compression.

in federated/distributed learning, similar to other compressed SGD algorithms [3], [4], [5], [6], [7], continues to be the communication time.

A. The Necessity of Scaling for Convergence of CSGD-ASSS

In this section, we illustrate through examples that scaling is necessary for the convergence of CSGD-ASSS. We consider interpolated linear regression in the convex setting and neural networks in the non-convex setting as examples.

In interpolated linear regression, we consider the loss function

$$f(x) = \frac{1}{n} \sum_{i=1}^n (\langle h_i, x \rangle - b_i)^2, \quad (23)$$

where $\exists x^*$ such that $\langle h_i, x^* \rangle = b_i, \forall i \in [n]$. In our experiments, we choose $n = 10000$, $\{h_i, x\} \in \mathbb{R}^{1024}$ and use top_k compression with compression ratio $\frac{k}{d} = 1\%$. The elements of

x^* are generated from $\mathcal{N}(0, 1)$ and the elements of a_i are distributed as $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 10)$ in Fig. 2(a) and 2(b), respectively. The loss increases exponentially without scaling in the convex setting. This clearly demonstrates that for compressed SGD with Armijo step-size search, scaling is a fundamental requirement and not just a proof technicality. In Fig. 2(c), the CSGD-ASSS algorithm is evaluated on ResNet-34 and ResNet-18 on CIFAR 100 dataset with scaling $a = 3\sigma$ and $a = 1$ (no scaling) at $\approx 1\%$ compression. The plot demonstrates that the conclusion regarding scaling also holds in (non-convex) neural network training tasks.

VI. CONCLUSION

We have motivated and presented a scaling technique for Armijo step-size search for compressed SGD with feedback, and used it to establish convergence for convex and non-convex objectives. In our experiments, we have shown that the proposed CSGD-ASSS algorithm outperforms non-adaptive compressed

SGD on ResNet-18 and ResNet-34 networks trained on CIFAR-100 and CIFAR-10 datasets at 1.5% compression. Through simulations, we have demonstrated that scaling is fundamental for the convergence of SGD with Armijo step-size search and compressed gradients with feedback. We have also shown additional simulation results and the proof of convergence of the CSGD-ASSS algorithm to local minima in distributed setting in [24]. The study of adaptive step-size methods for compressed SGD without feedback [18] is left for future work.

APPENDIX MATHEMATICAL PRELIMINARIES

We begin by stating some useful results, whose proofs are straightforward.

Lemma 6: For vectors $v_1, v_2 \in \mathbb{R}^d$, and $p > 0, r > 0$,

$$2\langle v_1, v_2 \rangle \leq p\|v_1\|^2 + \frac{1}{p}\|v_2\|^2. \quad (24)$$

$$\|v_1 + v_2\|^2 \leq (1+r)\|v_1\|^2 + \left(1 + \frac{1}{r}\right)\|v_2\|^2. \quad (25)$$

Lemma 7: [5] In CSGD-ASSS algorithm, $m_t = x_t - \hat{x}_t$.

Lemma 8: [6] For top_k compression operator with $\gamma = \frac{k}{d}$, and a vector $v_1 \in \mathbb{R}^n$,

$$\|v_1 - \text{top}_k(v_1)\|^2 \leq (1-\gamma)\|v_1\|^2. \quad (26)$$

Lemma 9: [9], [15] The Armijo step-size search stopping condition for a data-point i_t sampled at time t

$$f_{i_t}(x_t - \alpha_t \nabla f_{i_t}(x_t)) - f_{i_t}(x_t) \leq -\sigma \alpha_t \|\nabla f_{i_t}(x_t)\|^2 \quad (27)$$

is satisfied by all $\alpha_t \in [0, \frac{2(1-\sigma)}{L_{\max}}]$ where $L_{\max} = \max_{i \in [n]} \{L_i\}$.

Lemma 10: [15] The Armijo step-size search rule returns a step-size $\alpha_t \in [\tilde{\alpha}_{\min}\rho, \alpha_{\max}]$ where $\tilde{\alpha}_{\min} \triangleq \frac{2(1-\sigma)}{L_{\max}}$.

Remark 3: In our proofs, for the Armijo step-size search $\alpha_{\min} = \tilde{\alpha}_{\min}\rho$ and $\alpha_{\min} \leq \alpha_t \leq \alpha_{\max}$. The step-size η_t for CSGD-ASSS is obtained by $\eta_t = a\alpha_t$ and hence $\eta_t \in [\eta_{\min}, \eta_{\max}]$, where $\eta_{\min} = a\alpha_{\min}$ and $\eta_{\max} = a\alpha_{\max}$.

PROOF OF THEOREMS

A. Proof of Theorem 1

Proof: Let x_t be the iterate of GD with Armijo step-size search at iteration t . The iterate for the next iteration is

$$x_{t+1} = x_t - \alpha_t a \nabla f(x_t), \quad (28)$$

where α_t is the step-size returned by Armijo step-size search and a is the step-size scaling factor. Let x^* be a minimum of the objective f . Then,

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &= \|x_t - \alpha_t a \nabla f(x_t) - x^*\|^2 \\ &= \|x_t - x^*\|^2 + \alpha_t^2 a^2 \|\nabla f(x_t)\|^2 \\ &\quad - 2\alpha_t a \langle x_t - x^*, \nabla f(x_t) \rangle. \end{aligned} \quad (29)$$

From the convexity of f ,

$$\langle x^* - x_t, \nabla f(x_t) \rangle \leq f(x^*) - f(x_t). \quad (30)$$

Substituting Equation (30) in (29),

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 + \alpha_t^2 a^2 \|\nabla f(x_t)\|^2 \\ &\quad + 2\alpha_t a (f(x^*) - f(x_t)). \end{aligned} \quad (31)$$

Define the iterate \tilde{x}_{t+1} as

$$\tilde{x}_{t+1} = x_t - \alpha_t \nabla f(x_t). \quad (32)$$

Since α_t is the step-size returned by Armijo step-size search,

$$f(\tilde{x}_{t+1}) - f(x_t) \leq -\alpha_t \sigma \|\nabla f(x_t)\|^2. \quad (33)$$

Rearranging the terms in Equation (33), and multiplying by $\frac{a^2}{\sigma}$,

$$\alpha_t^2 a^2 \|\nabla f(x_t)\|^2 \leq \frac{a^2 \alpha_t}{\sigma} (f(x_t) - f(\tilde{x}_{t+1})). \quad (34)$$

Substituting (34) in (31) and using $f(\tilde{x}_{t+1}) \geq f(x^*)$,

$$\begin{aligned} \|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 + \frac{a^2 \alpha_t}{\sigma} (f(x_t) - f(x^*)) \\ &\quad + 2\alpha_t a (f(x^*) - f(x_t)), \\ &\quad \alpha_t \left(2a - \frac{a^2}{\sigma}\right) (f(x_t) - f(x^*)) \\ &\leq \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2. \end{aligned} \quad (35)$$

The factor $(2a - \frac{a^2}{\sigma}) > 0$ if and only if $a < 2\sigma$. Also, the step-size returned by Armijo step-size search is lower bounded by $\frac{2(1-\sigma)}{L}\rho$, as shown in Lemma 10. By choosing $a < 2\sigma$, lower bounding α_t , summing over the horizon T and using Jensen's inequality,

$$f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{\frac{2(1-\sigma)}{L}\rho(2a - \frac{a^2}{\sigma})T}. \quad (36)$$

□

B. Proof of Theorem 2

Proof: From the perturbed iterate analysis in [23], we have that $\hat{x}_{t+1} \triangleq \hat{x}_t - \eta_t \nabla f_{i_t}(x_t)$ where i_t is the data-point sampled at time t and $\hat{x}_0 = x_0$. Hence,

$$\begin{aligned} \|\hat{x}_{t+1} - x^*\|^2 &= \|\hat{x}_t - x^* - \eta_t \nabla f_{i_t}(x_t)\|^2 \\ &= \|\hat{x}_t - x^*\|^2 + \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2 \\ &\quad - 2\langle \hat{x}_t - x^*, \eta_t \nabla f_{i_t}(x_t) \rangle \\ &\quad + 2\langle x^* - x_t, \eta_t \nabla f_{i_t}(x_t) \rangle. \end{aligned} \quad (37)$$

If f_{i_t} is μ_{i_t} -strongly convex, then

$$\langle x^* - x_t, \nabla f_{i_t}(x_t) \rangle \leq f_{i_t}(x^*) - f_{i_t}(x_t) - \frac{\mu_{i_t}}{2} \|x^* - x_t\|^2. \quad (38)$$

If f_{i_t} is convex but not strongly convex, then $\mu_{i_t} = 0$. In the statement of the theorem, the only assumption is that $\exists i \in [n]$ such that $\mu_i > 0$.

From triangle inequality,

$$-\|x^* - x_t\|^2 \leq \|x_t - \hat{x}_t\|^2 - \frac{1}{2} \|\hat{x}_t - x^*\|^2. \quad (39)$$

Substituting (39) in (38), setting $m_t = x_t - \hat{x}_t$ (by Lemma 7), defining $e_t \triangleq f_{i_t}(x_t) - f_{i_t}(x^*)$ and using the resulting expression in (37), we have that

$$\begin{aligned} \|\hat{x}_{t+1} - x^*\|^2 &\leq \left(1 - \frac{\mu_{i_t}\eta_t}{2}\right) \|\hat{x}_t - x^*\|^2 + \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2 \\ &\quad - 2\eta_t e_t + \mu_{i_t}\eta_t \|m_t\|^2 \\ &\quad + 2\langle x_t - \hat{x}_t, \eta_t \nabla f_{i_t}(x_t) \rangle. \end{aligned} \quad (40)$$

From Lemma 6, we have that

$$\begin{aligned} 2\eta_t \langle x_t - \hat{x}_t, \nabla f_{i_t}(x_t) \rangle &\leq q_t \eta_t \|x_t - \hat{x}_t\|^2 \\ &\quad + \frac{\eta_t}{q_t} \|\nabla f_{i_t}(x_t)\|^2, \quad \forall q_t > 0. \end{aligned} \quad (41)$$

Substituting in (40),

$$\begin{aligned} \|\hat{x}_{t+1} - x^*\|^2 &\leq \left(1 - \frac{\mu_{i_t}\eta_t}{2}\right) \|\hat{x}_t - x^*\|^2 + \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2 \\ &\quad - 2\eta_t e_t + \mu_{i_t}\eta_t \|m_t\|^2 + q_t \eta_t \|x_t - \hat{x}_t\|^2 \\ &\quad + \frac{\eta_t}{q_t} \|\nabla f_{i_t}(x_t)\|^2. \end{aligned} \quad (42)$$

From the compression property in Definition 1 and the memory update step $m_{t+1} = (m_t + \eta_t \nabla f_{i_t}(x_t)) - \phi(m_t + \eta_t \nabla f_{i_t}(x_t))$ in the CSGD-ASSS algorithm, we have that

$$\begin{aligned} \|m_{t+1}\|^2 &= \|(m_t + \eta_t \nabla f_{i_t}(x_t)) - \phi(m_t + \eta_t \nabla f_{i_t}(x_t))\|^2 \\ &\leq (1 - \gamma) \|m_t + \eta_t \nabla f_{i_t}(x_t)\|^2. \end{aligned} \quad (43)$$

From Lemma 6,

$$\begin{aligned} \|m_t + \eta_t \nabla f_{i_t}(x_t)\|^2 &\leq (1 + r) \|m_t\|^2 \\ &\quad + \left(1 + \frac{1}{r}\right) \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2, \quad \forall r > 0. \end{aligned} \quad (44)$$

Substituting this in (43), we get that

$$\begin{aligned} \|m_{t+1}\|^2 &\leq (1 - \gamma)(1 + r) \|m_t\|^2 \\ &\quad + (1 - \gamma) \left(1 + \frac{1}{r}\right) \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2 \end{aligned} \quad (45)$$

for some $r > 0$. Adding (42) and (45), and setting $p \triangleq q_t a \alpha_t$ where $p \in \mathbb{R}^+$ (p can be set independent of t since q_t is arbitrary),

$$\begin{aligned} \|\hat{x}_{t+1} - x^*\|^2 &\leq -\|m_{t+1}\|^2 + \left(1 - \frac{\mu_{i_t}\eta_t}{2}\right) \|\hat{x}_t - x^*\|^2 \\ &\quad + \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2 - 2\eta_t e_t + \mu_{i_t}\eta_t \|m_t\|^2 \\ &\quad + \frac{p}{a\alpha_t} \eta_t \|x_t - \hat{x}_t\|^2 + \frac{a\alpha_t}{p} \eta_t \|\nabla f_{i_t}(x_t)\|^2 \\ &\quad + (1 - \gamma)(1 + r) \|m_t\|^2 \\ &\quad + (1 - \gamma) \left(1 + \frac{1}{r}\right) \eta_t^2 \|\nabla f_{i_t}(x_t)\|^2. \end{aligned} \quad (46)$$

Consider the Armijo step-size search stopping condition

$$f_{i_t}(\tilde{x}_{t+1}) - f_{i_t}(x_t) \leq -\sigma \alpha_t \|\nabla f_{i_t}(x_t)\|^2. \quad (47)$$

This implies,

$$\eta_t^2 \|\nabla f_{i_t}(x_t)\|^2 \stackrel{(\phi)}{\leq} \frac{a^2 \alpha_t}{\sigma} (f_{i_t}(x_t) - f_{i_t}(x^*)). \quad (48)$$

Note that (ϕ) holds due to the interpolation condition. Substituting (48) in (46),

$$\begin{aligned} \|\hat{x}_{t+1} - x^*\|^2 &\leq -\|m_{t+1}\|^2 + \left(1 - \frac{\mu_{i_t}\eta_t}{2}\right) \|\hat{x}_t - x^*\|^2 \\ &\quad + \left(\mu_{i_t}\eta_t + p + (1 - \gamma)(1 + r)\right) \|m_t\|^2 \\ &\quad - \alpha_t e_t \left(2a - \frac{a^2}{\sigma} - \frac{a^2}{\sigma p} - (1 - \gamma) \left(1 + \frac{1}{r}\right) \frac{a^2}{\sigma}\right). \end{aligned} \quad (49)$$

Let $\mu_{\max} \triangleq \max_{i \in [n]} \mu_i$. Using $\eta_t = a\alpha_t$ and $\alpha_t \leq \alpha_{\max}$, where α_{\max} is the starting value of the Armijo step-size search algorithm, we get

$$\begin{aligned} \mu_{i_t}\eta_t + p + (1 - \gamma)(1 + r) &\leq \mu_{\max} a \alpha_{\max} \\ &\quad + p + (1 - \gamma)(1 + r). \end{aligned} \quad (50)$$

Notice that in (49), p, r can be chosen arbitrarily. Define $\beta_1(p, r, a)$, $\tilde{a}_1(p, r)$ and $\tilde{a}_2(p, r, a)$ as

$$\begin{aligned} \beta_1(p, r, a) &\triangleq \left(\mu_{\max} a \alpha_{\max} + p + (1 - \gamma)(1 + r)\right) \\ \tilde{a}_1(p, r) &\triangleq 2 / \left(\frac{1}{\sigma} + \frac{1}{\sigma p} + \frac{(1 - \gamma)(1 + \frac{1}{r})}{\sigma}\right) \\ \tilde{a}_2(p, r, a) &\triangleq \left(2a - \frac{a^2}{\sigma} - \frac{a^2}{\sigma p} - (1 - \gamma) \left(1 + \frac{1}{r}\right) \frac{a^2}{\sigma}\right). \end{aligned} \quad (51)$$

If $a < \tilde{a}_1(p, r)$, then $\tilde{a}_2(p, r, a) > 0$. If p, r, a are chosen such that $\beta_1(p, r, a) < 1$ and $a < \tilde{a}_1(p, r)$, Equation (49) simplifies to

$$\begin{aligned} \|\hat{x}_{t+1} - x^*\|^2 &\leq -\|m_{t+1}\|^2 + \left(1 - \frac{\mu_{i_t} a \alpha_t}{2}\right) \|\hat{x}_t - x^*\|^2 \\ &\quad + \beta_1(p, r, a) \|m_t\|^2. \end{aligned} \quad (52)$$

The existence of such p, r, a is proven in Lemma 13. The step-size in the Armijo step-size search satisfies $\tilde{\alpha}_{\min} \rho \leq \alpha_t$. Using this property and taking expectation with respect to the data point i_t conditioned on the entire past therein,

$$\begin{aligned} E_{i_t} [\|\hat{x}_{t+1} - x^*\|^2 + \|m_{t+1}\|^2] &\leq \left(1 - \frac{E_{i_t}[\mu_{i_t}] a \alpha_{\min} \rho}{2}\right) \|\hat{x}_t - x^*\|^2 \\ &\quad + \beta_1(p, r, a) \|m_t\|^2. \end{aligned} \quad (53)$$

By the strong-convexity assumption in the theorem, $\exists i \in [n]$ such that $\mu_i > 0$. Hence,

$$\bar{\mu} = E_{i_t}[\mu_i] = \frac{\sum_{i=1}^n \mu_i}{n} > 0.$$

This implies that

$$0 < (\bar{\mu} a \tilde{\alpha}_{\min} \rho) / 2 = \bar{\mu} a \frac{1}{L_{\max}} (1 - \sigma) \rho < 1,$$

since $\bar{\mu} \leq L_{\max}$, the values $\sigma, \rho < 1$ and $a < 1$ by Lemma 13. Let

$$\beta_2(a) \triangleq \left(1 - \frac{\bar{\mu} a \tilde{\alpha}_{\min} \rho}{2}\right).$$

It follows that $0 < \beta_2 < 1$. Let $\hat{\beta} \triangleq \max\{\beta_1(p, r, a), \beta_2(a)\}$. Since $\beta_1(p, r, a) < 1$, $\beta_2(a) < 1$, this implies $\hat{\beta}(a) < 1$. Using this in (53) and taking expectation w.r.t the entire process, and recursion over t , we get

$$\begin{aligned} E[|\hat{x}_t - x^*|^2 + \|m_t\|^2] \\ \leq (\hat{\beta}(a))^t E[(|\hat{x}_0 - x^*|^2 + \|m_0\|^2)]. \end{aligned} \quad (54)$$

Since $\|m_t\| = \|x_t - \hat{x}_t\|$,

$$\begin{aligned} \|x_t - x^*\|^2 &\leq 2(\|x_t - \hat{x}_t\|^2) + 2(\|\hat{x}_t - x^*\|^2) \\ &= 2(\|m_t\|^2 + \|\hat{x}_t - x^*\|^2). \end{aligned} \quad (55)$$

At $t = 0$, $m_0 = 0$ and hence,

$$\begin{aligned} E[\|x_t - x^*\|^2] &\leq 2(E[\|\hat{x}_t - x^*\|^2 + \|m_t\|^2]) \\ &\leq 2(\hat{\beta}(a))^t E[(|\hat{x}_0 - x^*|^2)]. \end{aligned} \quad (56)$$

In the perturbed iterate analysis, $\hat{x}_0 = x_0$. Substituting this we get,

$$E[\|x_t - x^*\|^2] \leq 2(\hat{\beta})^t E[(\|x_0 - x^*\|^2)]. \quad (57)$$

Hence, we obtain the geometric convergence stated in the theorem. \square

To show the existence of p, r, a such that $\beta_1(p, r, a) < 1$ and $a < \tilde{a}_1(p, r)$ we prove Lemma 13, which is based on Lemmas 11 and 12 proved below.

Lemma 11: Consider the following convex optimization problem with variables $s, z, \psi \in \mathbb{R}$ and $0 < \psi < 1$.

$$\begin{aligned} \min_{s, z} g(s, z) &= \frac{1}{s} + \psi \frac{1}{z} \\ \text{s.t. } (s + \psi(1 + z)) &\leq 1, \\ s \geq 0, z \geq 0 \end{aligned} \quad (58)$$

The function $g(s, z)$ attains the minimum value $\frac{(1+\psi)^2}{(1-\psi)}$ at $(s^*, z^*) = (\frac{1-\psi}{1+\psi}, \frac{1-\psi}{1+\psi})$.

Proof: To solve the convex optimization problem, we consider the dual of the problem and apply the approach of Proposition 3.3.1 in [22]. The constraints $s \geq 0$ and $z \geq 0$ are inactive since $s = 0$ or $z = 0$ implies the objective $g(s, z) = \infty$. Since only $(s + \psi(1 + z)) \leq 1$ can be active, all local minima are regular. Consider the Lagrangian

$$\begin{aligned} L(s, z, \lambda_1, \lambda_2, \lambda_3) &= g(s, z) + \lambda_1(s + \psi(1 + z) - 1) \\ &\quad + \lambda_2(-s) + \lambda_3(-z) \end{aligned} \quad (59)$$

The values s^*, z^* are a local minimum of the convex problem (58) if $\exists \lambda_1^*, \lambda_2^*, \lambda_3^*$ such that $\lambda_2^* = 0, \lambda_3^* = 0$ and $\nabla_{s, z} L(s^*, z^*, \lambda_1^*, \lambda_2^*, \lambda_3^*) = 0$.

Solving for $L(s^*, z^*, \lambda_1^*, \lambda_2^*, \lambda_3^*) = 0$ we get $s^* = z^*$ and $\lambda_1^* = \frac{1}{s^*}$ at any local minimum. Hence $\lambda_1^* > 0$ which implies that the constraint $(s^* + \psi(1 + z^*)) \leq 1$ is active or tight. Substituting $s^* = z^*$ in the active constraint, we get

$$s^* + \psi(1 + s^*) = 1. \quad (60)$$

This implies that

$$s^* = z^* = \frac{1 - \psi}{1 + \psi}. \quad (61)$$

Hence there is a unique local minimum. The minimum value of the objective $g(s, z)$ subject to the constraints is

$$\begin{aligned} g(s^*, z^*) &= \frac{1}{s^*} + \psi \frac{1}{z^*} \\ &= \frac{1}{s^*} (1 + \psi) \\ &= \frac{(1 + \psi)^2}{1 - \psi}. \end{aligned} \quad (62)$$

\square

Lemma 12: Let $f_i(x)$, $i \in [n]$ be a set of μ_i strong-convex functions s.t $\mu_i > 0$ for some $i > 0$ and $\mu_{\max} \triangleq \max_{i \in [n]} \mu_i$. Then $\mu_{\max} \leq \xi$, $\forall \xi > 0$ is justified.

Proof: The case where $\xi \geq \mu_{\max}$ is trivial. Hence consider the case where $\xi < \mu_{\max}$. Without loss of generality, let $\xi < \mu_1$ where μ_1 is the strong convexity constant of f_1 . This implies that $\nabla^2 f_1(x) \succeq \mu_1 I$. Since $\xi < \mu_1$, it also implies that $\nabla^2 f_1(x) \succeq \xi I$. Hence f_1 is ξ strongly convex. It is thus justified to assume $\mu_1 = \xi$ and set $\mu_{\max} \leq \xi$. \square

We show the following lemma based on Lemmas 11 and 12.

Lemma 13: Let $\zeta \triangleq \frac{\sigma\gamma}{(2-\gamma)}$. For all $0 < \epsilon < \zeta$ and $a \leq \zeta - \epsilon$, $\exists (p(\epsilon), r(\epsilon))$ such that $\beta_1(p(\epsilon), r(\epsilon), a) < 1$ and $a < \tilde{a}_1(p(\epsilon), r(\epsilon))$, where $p(\epsilon), r(\epsilon)$ are functions of ϵ .

Proof: The functions $\beta_1(p, r)$ and $\tilde{a}_1(p, r)$ have been defined as

$$\begin{aligned} \beta_1(p, r) &= \left(\mu_{\max} a \alpha_{\max} + p + (1 - \gamma)(1 + r) \right), \\ \tilde{a}_1(p, r) &= 2 / \left(\frac{1}{\sigma} + \frac{1}{\sigma p} + \frac{(1 - \gamma)(1 + \frac{1}{r})}{\sigma} \right). \end{aligned} \quad (63)$$

Consider the convex problem

$$\begin{aligned} \sup_{p, r} \tilde{a}_1(p, r) &= 2\sigma / \left(1 + \frac{1}{p} + (1 - \gamma) \left(1 + \frac{1}{r} \right) \right), \\ \text{s.t. } (p + (1 - \gamma)(1 + r)) &< 1, \\ p \geq 0, r \geq 0. \end{aligned} \quad (64)$$

Equivalently, the problem can be stated as

$$\begin{aligned} \inf_{p, r} \left(\frac{1}{p} + (1 - \gamma) \left(1 + \frac{1}{r} \right) \right), \\ \text{s.t. } (p + (1 - \gamma)(1 + r)) &< 1, \\ p \geq 0, r \geq 0. \end{aligned} \quad (65)$$

Using Lemma 11 with $\psi = (1 - \gamma)$, the infimum is $\frac{(2-\gamma)^2}{\gamma} + (1 - \gamma)$ and the infimum is attained at a point (p, r) such that $(p + (1 - \gamma)(1 + r)) = 1$. Hence, by substituting the infimum of (65) in (64),

$$\zeta \triangleq \sup_{p, r} \tilde{a}_1(p, r) = \frac{\sigma\gamma}{(2-\gamma)}, \quad (66)$$

subject to the constraints. By the continuity of the function $2\sigma / \left(1 + \frac{1}{p} + (1 - \gamma)(1 + \frac{1}{r}) \right)$ at $p > 0, r > 0$, and since $(p + (1 - \gamma)(1 + r)) < 1, p \geq 0, r \geq 0$ is a region in the first quadrant and ζ is attained at (p^*, r^*) such that $(p^* + (1 - \gamma)(1 + r^*)) = 1$, for all $\epsilon > 0$, $\exists (p(\epsilon), r(\epsilon))$ and $0 < \delta < 1$ such that

$$\begin{aligned} a_1(p(\epsilon), r(\epsilon)) &> \zeta - \epsilon, \\ (p(\epsilon) + (1 - \gamma)(1 + r(\epsilon))) &= 1 - \delta. \end{aligned} \quad (67)$$

Let $\mu_{\max} \leq \frac{(\delta-\tau)}{\alpha_{\max}\zeta}$ for some $0 < \tau < \delta$. This assumption is justified by Lemma 12. Hence,

$$\begin{aligned} \beta_1(p(\epsilon), r(\epsilon), a) &\leq \left(\mu_{\max}\zeta\alpha_{\max} + p(\epsilon) + (1-\gamma)(1+r(\epsilon)) \right) \\ &\leq (1-\tau) \\ &< 1. \end{aligned} \quad (68)$$

To conclude, we have shown that for any $a \leq \zeta - \epsilon$, $\exists (p(\epsilon), r(\epsilon))$ such that $\tilde{a}_1(p(\epsilon), r(\epsilon)) > \zeta - \epsilon$ and $\beta_1(p(\epsilon), r(\epsilon), a) < 1$. This implies that $\exists (p(\epsilon), r(\epsilon))$ s.t.

$$\begin{aligned} a &< \tilde{a}_1(p(\epsilon), r(\epsilon)), \\ \beta_1(p(\epsilon), r(\epsilon), a) &< 1. \end{aligned} \quad (69)$$

□

Remark 4: In the proof of Theorem 2, for all $a < \zeta$, $\exists p, r$ such that $a < \tilde{a}_1(p, r)$ and $\beta_1(p, r, a) < 1$ by Lemma 13 and hence (52) of Theorem 2 holds.

C. Proof of Theorem 3

Proof: For CSGD-ASSS, when the objective is convex, Equation (49) holds with $\mu_{i_t} = 0$. Define $\tilde{a}_1(p, r)$ and $\tilde{a}_2(p, r, a)$ as in Equation (51). From Lemma 13, $\exists (p(\epsilon), r(\epsilon))$ such that

$$\begin{aligned} (p(\epsilon) + (1-\gamma)(1+r(\epsilon))) &= 1 - \delta < 1, \\ \tilde{a}_1(p(\epsilon), r(\epsilon)) &> \zeta - \epsilon. \end{aligned} \quad (70)$$

Choosing p, r as $p(\epsilon), r(\epsilon)$ and a scale factor $a < \zeta - \epsilon$ implies that $\tilde{a}_2(p(\epsilon), r(\epsilon), a) > 0$. Substituting $p(\epsilon), r(\epsilon), \tilde{a}_2(p(\epsilon), r(\epsilon), a)$ in Equation (49) with $\mu_{i_t} = 0$ and summing over horizon T and averaging proves the theorem. □

D. Proof of Theorem 4

Let f_i be L_i smooth and let f_i $i \in [n]$ satisfy the strong growth condition (6). Let $L_i \leq L_{\text{UP}} \forall i \in [n]$ and $\nu \leq \nu_{\text{UP}}$. Then, there exists $\hat{a}, \hat{\alpha}$ such that for $0 < a \leq \hat{a}$ and $\alpha_{\max} \leq \hat{\alpha}$,

$$\frac{1}{T} \sum_{t=0}^{T-1} E[\|\nabla f(x_t)\|^2] \leq \frac{(E[f(x_0)] - E[f(\hat{x}_T)])}{\delta T}, \quad (71)$$

where \hat{x}_T is a perturbed iterate [23] obtained from $\{x_i\}_{i=0}^{T-1}$ and

$$\begin{aligned} \delta &= \left[\left(\eta_{\max} + \frac{\eta_{\min} p}{1+p} \right) - \left(\nu(\eta_{\max} - \eta_{\min}) + \nu L \eta_{\max}^2 \right. \right. \\ &\quad \left. \left. + \left(\nu \eta_{\max}^2 \theta (1-\gamma) \left(1 + \frac{1}{r} \right) \right) \right) \right], \end{aligned} \quad (72)$$

$$\hat{a} = \frac{1}{2} \min \left\{ \frac{[(\frac{p}{p+1}) + 1] L_{\text{UP}}}{2(1-\sigma)\nu_{\text{UP}}(L_{\text{UP}} + \theta G)}, \frac{\theta \epsilon}{\beta L_{\text{UP}}^2 + 2(1-\sigma)p L_{\text{UP}}} \right\}, \quad (73)$$

UB(a)

$$\begin{aligned} &\frac{-(\nu_{\text{UP}} - 1)}{+ \sqrt{(\nu_{\text{UP}} - 1)^2 + (4\nu_{\text{UP}}(L_{\text{UP}} + \theta G)a(\nu_{\text{UP}} + \frac{p}{1+p})\frac{2(1-\sigma)}{L_{\text{UP}}})\rho}}, \\ &\triangleq \frac{-(\nu_{\text{UP}} - 1)}{2a\nu_{\text{UP}}(L_{\text{UP}} + \theta G)}, \end{aligned} \quad (74)$$

$$\hat{\alpha} = \min\{UB(\hat{a}), \beta\}, G \triangleq (1-\gamma) \left(1 + \frac{1}{r} \right), L = \frac{\sum_{i=1}^n L_i}{n} \quad (75)$$

for any $p, r, \theta, \beta > 0$, $\tilde{\alpha}_{\min} = \frac{2(1-\sigma)}{L_{\max}}$, $L_{\max} = \max_i \{L_i\}$, $\epsilon < \gamma$ and ν is the strong growth constant and ν_{UP} is an upper bound on the strong growth constant.

Proof: By the perturbed iterate analysis framework, let $\{\hat{x}_t\}_{t \geq 0}$ be the virtual sequence generated as

$$\hat{x}_{t+1} = \hat{x}_t - \eta_t \nabla f_{i_t}(x_t). \quad (76)$$

Using the L -Lipschitz smoothness of gradient of f and (76),

$$\begin{aligned} f(\hat{x}_{t+1}) &\leq f(\hat{x}_t) + \langle \nabla f(\hat{x}_t), -\eta_t \nabla f_{i_t}(x_t) \rangle \\ &\quad + \frac{L}{2} \|\eta_t \nabla f_{i_t}(x_t)\|^2. \end{aligned} \quad (77)$$

Using the identity $-2\langle a, b \rangle = \|a - b\|^2 - \|a\|^2 - \|b\|^2$, the bounds on step-size $\eta_t \in [\eta_{\min}, \eta_{\max}]$ and taking expectation E_{i_t} with respect to the data point i_t conditioned on x_t, i_{t-1} and the entire past therein,

$$\begin{aligned} 2E_{i_t}[f(\hat{x}_{t+1}) - f(\hat{x}_t)] &\leq \eta_{\max} E_{i_t}[\|\nabla f(\hat{x}_t) - \nabla f_{i_t}(x_t)\|^2] \\ &\quad - \eta_{\min} \|\nabla f(\hat{x}_t)\|^2 \\ &\quad - \eta_{\min} E_{i_t}[\|\nabla f_{i_t}(x_t)\|^2] \\ &\quad + L\eta_{\max}^2 E_{i_t}[\|\nabla f_{i_t}(x_t)\|^2]. \end{aligned} \quad (78)$$

As $E_{i_t}[\nabla f_{i_t}(x)] = \nabla f(x)$,

$$\begin{aligned} E_{i_t}[\|\nabla f(\hat{x}_t) - \nabla f_{i_t}(x_t)\|^2] &= \|\nabla f(\hat{x}_t)\|^2 + E_{i_t}[\|\nabla f_{i_t}(x_t)\|^2] \\ &\quad - 2\langle \nabla f(\hat{x}_t), \nabla f(x_t) \rangle. \end{aligned} \quad (79)$$

The following inner product can be rewritten as

$$\begin{aligned} -2\eta_{\max} \langle \nabla f(\hat{x}_t), \nabla f(x_t) \rangle &= \eta_{\max} \|\nabla f(\hat{x}_t) - \nabla f(x_t)\|^2 \\ &\quad - \eta_{\max} \|\nabla f(\hat{x}_t)\|^2 \\ &\quad - \eta_{\max} \|\nabla f(x_t)\|^2. \end{aligned} \quad (80)$$

Substituting (79) in (78), and using the expression (80) and Lipschitz smoothness of ∇f ,

$$\begin{aligned} 2E_{i_t}[f(\hat{x}_{t+1}) - f(\hat{x}_t)] &\leq (\eta_{\max} - \eta_{\min} + L\eta_{\max}^2) E_{i_t}[\|\nabla f_{i_t}(x_t)\|^2] \\ &\quad + \eta_{\max} L^2 [\|\hat{x}_t - x_t\|^2] - \eta_{\max} \|\nabla f(x_t)\|^2 \\ &\quad - \eta_{\min} \|\nabla f(\hat{x}_t)\|^2. \end{aligned} \quad (81)$$

Using Lemma 6, for any $p > 0$,

$$\begin{aligned} \|\nabla f(x_t)\|^2 &\leq (1+p) \|\nabla f(x_t) - \nabla f(\hat{x}_t)\|^2 \\ &\quad + \left(1 + \frac{1}{p} \right) \|\nabla f(\hat{x}_t)\|^2. \end{aligned} \quad (82)$$

Using the Lipschitz smoothness of ∇f and rearranging terms in (82) and substituting in (81), and using the strong growth condition $E_{i_t}[\|\nabla f_{i_t}(x)\|^2] \leq \nu\|\nabla f(x)\|^2$,

$$\begin{aligned} & \left(\eta_{\max} + \frac{\eta_{\min}p}{1+p} - (\eta_{\max} - \eta_{\min} + L\eta_{\max}^2)\nu \right) (\|\nabla f(x_t)\|^2) \\ & \leq 2E_{i_t}[f(\hat{x}_t) - f(\hat{x}_{t+1})] \\ & \quad + (\eta_{\max}L^2 + \eta_{\min}pL^2)(\|\hat{x}_t - x_t\|^2). \end{aligned} \quad (83)$$

Bounding η_t by η_{\max} in (45) and taking expectation w.r.t data-point i_t and mutiplying by $\theta > 0$ and adding to (83),

$$\begin{aligned} & \left(\eta_{\max} + \frac{\eta_{\min}p}{1+p} - (\eta_{\max} - \eta_{\min} + L\eta_{\max}^2)\nu \right) (\|\nabla f(x_t)\|^2) \\ & \leq -\theta E_{i_t}[\|m_{t+1}\|^2] + 2E_{i_t}[f(\hat{x}_t) - f(\hat{x}_{t+1})] \\ & \quad + (\eta_{\max}L^2 + \eta_{\min}pL^2)(\|\hat{x}_t - x_t\|^2) \\ & \quad + \theta(1-\gamma)(1+r)\|m_t\|^2 \\ & \quad + \theta(1-\gamma) \left(1 + \frac{1}{r} \right) \eta_{\max}^2 (E_{i_t}[\|\nabla f(x_t)\|^2]). \end{aligned} \quad (84)$$

Using the strong growth condition and taking expectation w.r.t the entire process,

$$\begin{aligned} & \left[\left(\eta_{\max} + \frac{\eta_{\min}p}{1+p} \right) - \left(\nu(\eta_{\max} - \eta_{\min}) + \nu L\eta_{\max}^2 \right. \right. \\ & \quad \left. \left. + \left(\nu\eta_{\max}^2\theta(1-\gamma) \left(1 + \frac{1}{r} \right) \right) \right) \right] E[\|\nabla f(x_t)\|^2] \\ & \leq 2E[f(\hat{x}_t) - f(\hat{x}_{t+1})] \\ & \quad + (\eta_{\max}L^2 + \eta_{\min}pL^2 + \theta(1-\gamma)(1+r))E[\|m_t\|^2] \\ & \quad - \theta E[\|m_{t+1}\|^2]. \end{aligned} \quad (85)$$

Consider the term $\eta_{\max}L^2 + \eta_{\min}pL^2 + \theta(1-\gamma)(1+r)$ and an $\epsilon < \gamma$. Then,

$$(1-\gamma)(1+r) \leq 1-\gamma+r \leq (1-\epsilon) \quad (86)$$

iff $r \leq \gamma - \epsilon$. Set $r \leq \gamma - \epsilon$ and $\eta_{\max}L^2 + \eta_{\min}pL^2 \leq \theta\epsilon$. The inequality $\eta_{\max}L^2 + \eta_{\min}pL^2 \leq \theta\epsilon$ holds iff

$$a \leq \frac{\theta\epsilon}{(\alpha_{\max}L^2 + \alpha_{\min}pL^2)}. \quad (87)$$

Hence, setting $r = \gamma - \epsilon$ and a satisfying (87), implies

$$(\eta_{\max}L^2 + \eta_{\min}pL^2 + \theta(1-\gamma)(1+r)) \leq \theta. \quad (88)$$

If $\delta > 0$, and a satisfies (87), (85) simplifies as

$$\begin{aligned} \delta E[\|\nabla f(x_t)\|^2] & \leq 2E[f(\hat{x}_t) - f(\hat{x}_{t+1})] \\ & \quad + \theta(E[\|m_t\|^2] - E[\|m_{t+1}\|^2]). \end{aligned} \quad (89)$$

Since $m_0 = 0$ and $\hat{x}_0 = x_0$, summing (89) over the horizon T and averaging,

$$\frac{1}{T} \sum_{t=0}^{T-1} E[\|\nabla f(x_t)\|^2] \leq 2 \frac{(E[f(\hat{x}_0)] - E[f(\hat{x}_T)])}{\delta T}. \quad (90)$$

By definition, $L_{\max} \leq L_{\text{UP}}$. Hence L_{\max} in Lemma 9 can be replaced by L_{UP} and $\tilde{\alpha}_{\min}$ in Lemma 10 can be replaced by $\frac{2(1-\sigma)}{L_{\text{UP}}}$.

The conditions under which $\delta > 0$, can be stated in 2 cases depending on the value of α_{\max} .

1) Case 1: $\alpha_{\max} \leq \frac{2(1-\sigma)}{L_{\text{UP}}}\rho$. In this case $\alpha_{\max} = \alpha_{\min}$ by Lemma 9 and this implies $\eta_{\max} = \eta_{\min}$. Hence $\delta > 0$ iff

$$a \leq \frac{[(\frac{p}{p+1}) + 1]L_{\text{UP}}}{2(1-\sigma)\nu(L + \theta G)}. \quad (91)$$

2) Case 2: $\alpha_{\max} > \frac{2(1-\sigma)}{L_{\text{UP}}}\rho$. In this case $\delta > 0$ iff

$$\alpha_{\max} \leq \text{UB}(a). \quad (92)$$

An interval $\alpha_{\max} \in (\frac{2(1-\sigma)\rho}{L_{\text{UP}}}, \text{UB}(a)]$ exists if $\frac{2(1-\sigma)\rho}{L_{\text{UP}}} \leq \text{UB}(a)$. The bound $\text{UB}(a)$ is a monotonically decreasing function of a and the interval exists if

$$a \leq \frac{[(\frac{p}{p+1}) + 1]L_{\text{UP}}}{2(1-\sigma)\nu(L + \theta G)}. \quad (93)$$

Combining the bounds on a for $\delta > 0$ and $\frac{1}{T}$ convergence from (87),(93) and using $L \leq L_{\text{UP}}$, and always using $\alpha_t \leq \beta$ in CSGD-ASSS for some pre-specified fixed $\beta > 0$,

$$a \leq \min \left\{ \frac{[(\frac{p}{p+1}) + 1]L_{\text{UP}}}{2(1-\sigma)\nu_{\text{UP}}(L_{\text{UP}} + \theta G)}, \frac{\theta\epsilon}{\beta L_{\text{UP}}^2 + 2(1-\sigma)pL_{\text{UP}}} \right\}. \quad (94)$$

Hence, setting $a = \hat{a}$ in $\text{UB}(a)$,

$$\hat{\alpha} = \min\{\text{UB}(\hat{a}), \beta\}.$$

Hence proved. \square

Remark 5: Similar to proofs for the uncompressed SGD setting [9], CSGD-ASSS only requires an estimate of an upper bound L_{UP} on the Lipschitz constant L_i and an upper bound ν_{UP} on the strong growth constant ν for non-convex objectives. The bounds need not be tight to prove $O(\frac{1}{T})$ convergence rate. However, for convex and strong-convex objectives, the algorithm does not require any bounds on problem parameters.

Remark 6: The dependency of δ on function parameters in the proof of convergence of CSGD-ASSS for non-convex objectives is highly non-linear and left for future work.

Remark 7: The proof of Theorem 4 does not explicitly utilize the stopping condition, Step 6, of the Armijo step-size search. Nevertheless, the dependency is subtle. The proof necessitates that the step-sizes of the algorithm at each iteration are lower-bounded by η_{\min} . This lower bound is assured for CSGD-ASSS, as indicated in Remark 3. However, the proof is applicable to a broader class of algorithms with compressed gradients.

Remark 8: The function $\text{UB}(a)$ monotonically decreases with respect to a . This can be proved by establishing the negativity of its derivative with respect to a . As a approaches infinity, the limit of $\text{UB}(a)$ converges to 0. Conversely, when a tends to 0, the limit is defined as $\text{UB}(0^+, \nu)$ is $\frac{2(\nu + \frac{p}{1+p})}{\nu-1}$ for some $p > 0$. Notably, if $\nu \rightarrow 1$, then $\text{UB}(0^+, \nu) \rightarrow \infty$.

LEMMAS FOR ANALYSIS OF CONSTANTS IN
CONVERGENCE PROOFS

A. Proof of Lemma 5

Proof: From Lemmas 15, 16, if there exists $p \geq 0$, $r \geq 0$ such that $\beta_1(p, r, a) = D$ if and only if $a \in [0, \hat{a}_{\beta_1}(D)]$ (The definition of $\hat{a}_{\beta_1}(D)$ is presented in Lemma 16). Hence there exists $\beta_{LB}(a)$ such that if $a_1 < a_2$, then $\beta_{LB}(a_1) < \beta_{LB}(a_2)$. \square

B. Analysis of Parameters in Theorem 3

Lemma 14: Let $p, r \in [0, \infty)$ and consider the function $\delta_{2,(p,r)}(a)$ defined as

$$\delta_{2,(p,r)}(a) \triangleq \left(2a - \frac{a^2}{\sigma} - \frac{a^2}{\sigma p} - (1-\gamma) \left(1 + \frac{1}{r} \right) \frac{a^2}{\sigma} \right). \quad (95)$$

Let $\mathcal{R}(\delta_{2,(p,r)})$ be the set of roots of the function $\delta_{2,(p,r)}(a)$. Then,

$$\mathcal{R}(\delta_{2,(p,r)}) = \left\{ 0, \frac{2\sigma}{\left(1 + \frac{1}{p} + (1-\gamma)\left(1 + \frac{1}{r}\right)\right)} \right\}. \quad (96)$$

The function $\delta_{2,(p,r)}(a)$ attains its maximum value at $\hat{a}_{\delta_{2,(p,r)}}$, where

$$\hat{a}_{\delta_{2,(p,r)}} = \frac{\sigma}{\left(1 + \frac{1}{p} + (1-\gamma)\left(1 + \frac{1}{r}\right)\right)}, \quad (97)$$

$$\delta_{2,(p,r)}(\hat{a}_{\delta_{2,(p,r)}}) = \frac{\sigma}{\left(1 + \frac{1}{p} + (1-\gamma)\left(1 + \frac{1}{r}\right)\right)}. \quad (98)$$

The proof of the lemma is straightforward and can be shown by setting the derivative of $\delta_{2,(p,r)}(a)$ to 0. For the proof of convergence of CSGD-ASSS for convex objectives in Theorem 3, parameters $p(\epsilon), r(\epsilon)$ satisfy $p(\epsilon) + (1-\gamma)(1+r(\epsilon)) < 1$. For a given choice of $p(\epsilon), r(\epsilon)$, the value of $\delta_{2,(p(\epsilon),r(\epsilon))}(a)$ increases with a when $a \leq \hat{a}_{\delta_{2,(p(\epsilon),r(\epsilon))}}$ and decreases with a when $a > \hat{a}_{\delta_{2,(p(\epsilon),r(\epsilon))}}$. Since $\delta_1(a) = \rho \frac{2(1-\sigma)}{L_{max}} \delta_{2,(p(\epsilon),r(\epsilon))}(a)$, it is concave and attains its maximum value at $\hat{a}_{\delta_{2,(p(\epsilon),r(\epsilon))}}$.

ADDITIONAL LEMMAS

Lemma 15: If there exists p_1, r_1, a_1 and $(1-\gamma) < D < 1$ such that $\beta_1(p_1, r_1, a_1) = D$, then for any $0 \leq a_2 < a_1$, there exists p_2, r_2 such that $\beta_2(p_2, r_2, a_2) = D$

Proof: From Theorem 2,

$$\beta_1(p, r, a) = \left(\mu_{max} a \alpha_{max} + p + (1-\gamma)(1+r) \right). \quad (99)$$

Since $\beta_1(p_1, r_1, a_1) = D$,

$$D = \left(\mu_{max} a_1 \alpha_{max} + p_1 + (1-\gamma)(1+r_1) \right) \quad (100)$$

$$= \left(\mu_{max} a_2 \alpha_{max} + \left(p_1 + \frac{\mu_{max}(a_1 - a_2)\alpha_{max}}{2-\gamma} \right) \right) \quad (101)$$

$$+ (1-\gamma) \left(1 + r_1 + \frac{\mu_{max}(a_1 - a_2)\alpha_{max}}{2-\gamma} \right) \quad (102)$$

Setting p_2, r_2 as

$$p_2 = \left(p_1 + \frac{\mu_{max}(a_1 - a_2)\alpha_{max}}{2-\gamma} \right), \quad (104)$$

$$r_2 = \left(r_1 + \frac{\mu_{max}(a_1 - a_2)\alpha_{max}}{2-\gamma} \right), \quad (105)$$

satisfies the conditions of the lemma. \square

Lemma 16: Let $\gamma < 1$ and $(1-\gamma) < D < 1$. Consider the optimization problem

$$\begin{aligned} & \sup_{0 \leq a < \zeta} a & (107) \\ & \text{s.t. } \beta_1(p, r, a) = D \\ & a < 2 / \left(\frac{1}{\sigma} + \frac{1}{\sigma p} + \frac{(1-\gamma)\left(1 + \frac{1}{r}\right)}{\sigma} \right) \\ & p \geq 0, r \geq 0. \end{aligned}$$

The optimization problem has a solution p^*, r^*, a^* and let the optimal value be $\hat{a}_{\beta_1}(D) = a^*$. Then,

$$D_1 < D_2 \implies \hat{a}_{\beta_1}(D_1) < \hat{a}_{\beta_2}(D_2). \quad (108)$$

Proof: Let $d < D$ and $p + (1-\gamma)(1+r) = d$. Since $\beta_1(p, r, a) = D$,

$$\mu_{max} a \alpha_{max} = D - d. \quad (109)$$

$$\implies a = \frac{D - d}{\mu_{max} \alpha_{max}} \quad (110)$$

Consider the optimization problem

$$\sup 2 / \left(\frac{1}{\sigma} + \frac{1}{\sigma p} + \frac{(1-\gamma)\left(1 + \frac{1}{r}\right)}{\sigma} \right) \quad (111)$$

$$\text{s.t. } p + (1-\gamma)(1+r) = d, \quad (112)$$

$$p \geq 0, r \geq 0. \quad (113)$$

Similar to the proof of Lemma 13, the solution to the above optimization problem defined as $\zeta(d)$ is

$$\zeta(d) \triangleq \frac{2\sigma}{2-\gamma} \left(1 - \frac{2-\gamma}{d+1} \right). \quad (114)$$

It is straightforward to see that the solution to the optimization problem in the statement of the lemma $\hat{a}_{\beta_1}(D)$ can be solved for by equating Equations (110), (114) and solving for d and substituting the resulting d in Equation (110) or (114). Equating Equations (110), (114), we get

$$\frac{D}{\mu_{max} \alpha_{max}} = \frac{d}{\mu_{max} \alpha_{max}} + \frac{2\sigma}{2-\gamma} \left(1 - \frac{2-\gamma}{d+1} \right) \quad (115)$$

The R.H.S of Equation (115) is monotonically increasing function of d . Hence, if for D_1, D_2 , the solution to the optimization problem in the lemma is d_1, d_2 respectively,

$$D_1 < D_2 \implies d_1 < d_2. \quad (116)$$

The equation $\zeta(d)$ is monotonically increasing in d and hence

$$\hat{a}_{\beta_1}(D_1) < \hat{a}_{\beta_1}(D_2). \quad (117)$$

\square

REFERENCES

- [1] A. M. Subramaniam, A. Magesh, and V. V. Veeravalli, "Adaptive step-size methods for compressed SGD," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [2] P. Kairouz et al., "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.
- [3] N. Dryden, T. Moon, S. A. Jacobs, and B. Van Essen, "Communication quantization for data-parallel training of deep neural networks," in *Proc. 2nd Workshop Mach. Learn. HPC Environ. (MLHPC)*, 2016, pp. 1–8.
- [4] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," 2017, *arXiv:1704.05021*.
- [5] S. U. Stich, J. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, 2018, pp. 4448–4459.
- [6] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. dé-Buc, E. Fox, and R. Garnett, Eds., 2019.
- [7] V. Gandikota, D. Kane, R. Kumar Maity, and A. Mazumdar, "VQSGD: Vector quantized stochastic gradient descent," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, vol. 130, A. Banerjee and K. Fukumizu, Eds., PMLR, 13–15 Apr. 2021, pp. 2197–2205. [Online]. Available: <https://proceedings.mlr.press/v130/gandikota21a.html>
- [8] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018. Accessed: May 7, 2024. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/file/314450613369e0ee72d0da7f6fee773c-Paper.pdf>
- [9] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien, "Painless stochastic gradient: Interpolation, line-search, and convergence rates," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. dé-Buc, E. Fox, and R. Garnett, Eds., New York, USA: Curran Associates, Inc., 2019. Accessed: May 7, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/2557911c1bf75c2b643afb4ecbfc8ec2-Paper.pdf
- [10] B. Dubois-Taine, S. Vaswani, R. Babanezhad, M. Schmidt, and S. Lacoste-Julien, "SVRG meets AdaGrad: Painless variance reduction," 2021, *arXiv:2102.09645*.
- [11] S. Vaswani, F. Kunstner, I. H. Laradji, S. Y. Meng, M. Schmidt, and S. Lacoste-Julien, "Adaptive gradient methods converge faster with over-parameterization (and you can do a line-search)," 2020, *arXiv:2006.06835*.
- [12] R. Bassily, M. Belkin, and S. Ma, "On exponential convergence of SGD in non-convex over-parametrized learning," 2018, *arXiv:1811.02564*.
- [13] C. Liu and M. Belkin, "MaSS: An accelerated stochastic method for over-parametrized learning," 2018, *arXiv:1810.13395*.
- [14] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, J. Dy and A. Krause, Eds., PMLR, 10–15 Jul. 2018, pp. 3325–3334. [Online]. Available: <https://proceedings.mlr.press/v80/ma18a.html>
- [15] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.
- [16] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., PMLR, 9–15 Jun. 2019, pp. 322–332. [Online]. Available: <https://proceedings.mlr.press/v97/arora19a.html>
- [17] A. Montanari and Y. Zhong, "The interpolation phase transition in neural networks: Memorization and generalization under lazy training," 2020, *arXiv:2007.12826*.
- [18] P. Richtarik, I. Sokolov, and I. Fatkhullin, "EF21: A new, simpler, theoretically better, and practically faster error feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., New York, USA: Curran Associates, Inc., 2021, pp. 4384–4396.
- [19] M. Schmidt and N. L. Roux, "Fast convergence of stochastic gradient descent under a strong growth condition," 2013, *arXiv:1308.6370*.
- [20] M. V. Solodov, "Incremental gradient algorithms with stepsizes bounded away from zero," *Comput. Optim. Appl.*, vol. 11, pp. 23–35, Oct. 1998.
- [21] P. Tseng, "An incremental gradient (-projection) method with momentum term and adaptive stepsize rule," *SIAM J. Optim.*, vol. 8, no. 2, pp. 506–531, 1998.
- [22] D. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1995.
- [23] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2202–2229, 2017.
- [24] A. M. Subramaniam, A. Magesh, and V. V. Veeravalli, "Adaptive step-size methods for compressed SGD," 2022, *arXiv:2207.10046*.