

Gradient Coding with Iterative Block Leverage Score Sampling

Neophytos Charalambides^μ, Mert Pilanci^σ, and Alfred O. Hero III^μ

^μEECS Department University of Michigan ^σEE Department Stanford University

Email: neochara@umich.edu, pilanci@stanford.edu, hero@umich.edu

Abstract

Gradient coding is a method for mitigating straggling servers in a centralized computing network that uses erasure-coding techniques to distributively carry out first-order optimization methods. Randomized numerical linear algebra uses randomization to develop improved algorithms for large-scale linear algebra computations. In this paper, we propose a method for distributed optimization that combines gradient coding and randomized numerical linear algebra. The proposed method uses a randomized ℓ_2 -subspace embedding and a gradient coding technique to distribute blocks of data to the computational nodes of a centralized network, and at each iteration the central server only requires a small number of computations to obtain the steepest descent update. The novelty of our approach is that the data is replicated according to importance scores, called block leverage scores, in contrast to most gradient coding approaches that uniformly replicate the data blocks. Furthermore, we do not require a decoding step at each iteration, avoiding a bottleneck in previous gradient coding schemes. We show that our approach results in a valid ℓ_2 -subspace embedding, and that our resulting approximation converges to the optimal solution.

Index Terms

Low-Rank Approximations, Least-Squares Regression, Randomized Algorithms, Randomized Numerical Linear Algebra, Coded Computing, Stragglers, Erasure-Coding, Replication-Coding.

I. INTRODUCTION

In this work we bridge two disjoint areas, to accelerate first-order methods such as steepest descent distributively, while focusing on linear regression. Specifically, we propose a framework in which Randomized Numerical Linear Algebra (RandNLA) sampling algorithms can be used to devise Coded Computing (CC) schemes. We focus on the task of ℓ_2 -subspace embedding through an importance sampling procedure known as leverage score sampling; which scores measure the relative importance of each data point within a given dataset, and distributed gradient computation in the presence of stragglers; which is referred to as gradient coding (GC).

Traditional numerical linear algebra algorithms are deterministic. For example, inverting a full-rank matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ requires $O(N^3)$ arithmetic operations by performing Gaussian elimination, as does naive matrix multiplication. The fastest known algorithm which multiplies two $N \times N$ matrices, requires $O(N^\omega)$ operations, for $\omega < 2.372$ [1], [2]. Other important problems are computing the determinant, singular and eigenvalue decompositions, SVD, QR and Cholesky factorizations of large matrices.

Although these deterministic algorithms run in polynomial time and are numerically stable, their exponents make them prohibitive for many applications in scientific computing and machine learning, when N is in the order of millions or billions [3], [4]. To circumvent this issue, one can perform these algorithms on a significantly smaller approximation. Specifically, for a matrix $\mathbf{S} \in \mathbb{R}^{r \times N}$ with $r \ll N$, we apply the deterministic algorithm on the surrogate $\hat{\mathbf{M}} = \mathbf{S}\mathbf{M} \in \mathbb{R}^{r \times d}$ for $\mathbf{M} \in \mathbb{R}^{N \times d}$. The matrix \mathbf{S} is referred to as a dimension-reduction or a sketching matrix, and $\hat{\mathbf{M}}$ is a sketch of \mathbf{M} , which contains as much information about \mathbf{M} as possible. For instance, when computing the product of $\mathbf{A} \in \mathbb{R}^{N \times L}$ and $\mathbf{B} \in \mathbb{R}^{N \times M}$, we apply a carefully chosen $\mathbf{S} \in \mathbb{R}^{r \times N}$ on each matrix to get

$$\underbrace{\begin{pmatrix} \mathbf{A}^\top \end{pmatrix}}_{L \times N} \underbrace{\begin{pmatrix} \mathbf{B} \end{pmatrix}}_{N \times M} \approx \underbrace{\begin{pmatrix} \hat{\mathbf{A}}^\top \end{pmatrix}}_{L \times r} \underbrace{\begin{pmatrix} \hat{\mathbf{B}} \end{pmatrix}}_{r \times M}$$

where $\hat{\mathbf{A}} = \mathbf{S}\mathbf{A}$ and $\hat{\mathbf{B}} = \mathbf{S}\mathbf{B}$. Thus, naive matrix multiplication now requires $O(LMr)$ operations, instead of $O(LMN)$. Such approaches have been motivated by the Johnson-Lindenstrauss lemma [5], [6], and require low complexity.

A multitude of other problems, such as k -means clustering [7]–[9] and computing the SVD of a matrix [10]–[13], make use of this idea in order to accelerate computing accurate approximate solutions. We refer the reader to the following monographs and comprehensive surveys on the rich development of RandNLA [4], [14]–[21], an interdisciplinary field that exploits randomization as a computational resource, to develop improved algorithms for large-scale linear algebra problems.

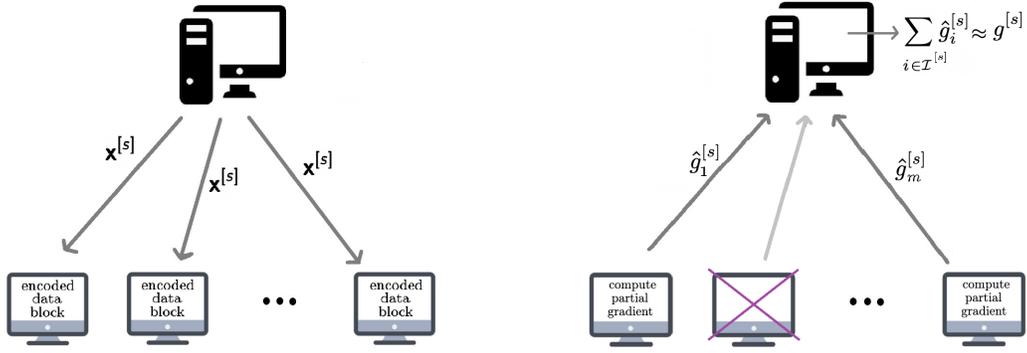


Fig. 1: Schematic of our approximate GC scheme, at iteration s . Each server has an encoded block of data, of which they compute the gradient once they receive the updated parameters vector $\mathbf{x}^{[s]}$. The central server then aggregates all the received partial gradients indexed by $\mathcal{I}^{[s]}$, i.e., $\{\hat{g}_i^{[s]}\}_{i \in \mathcal{I}^{[s]}}$, to approximate the gradient $g^{[s]}$. At each iteration we expect a different index set $\mathcal{I}^{[s]}$, which leads to iterative sketching.

The problem of ℓ_2 -subspace embedding, a form of spectral approximation of a matrix, has been extensively studied in RandNLA. The main techniques for constructing appropriate ℓ_2 -subspace embedding sketching matrices, are performing a random projection or row-sampling. Well-known choices of \mathbf{S} for reducing the effective dimension N to r include: i) *Gaussian projection*; for a matrix Θ where $\Theta_{ij} \sim \mathcal{N}(0, 1)$ define $\mathbf{S} = \frac{1}{\sqrt{r}} \Theta$, ii) *leverage score sampling*; sample with replacement r rows from the matrix according to its normalized leverage score distribution and rescale them appropriately, iii) *Subsampled Hadamard Transform (SRHT)*; apply a Hadamard transform and a random signature matrix to judiciously make the leverage scores approximately uniform and then follow similar steps to the leverage score sampling procedure.

In this paper, we first generalize ii) to appropriately sample submatrices instead of rows to attain an ℓ_2 -subspace embedding guarantee. We refer to such approaches as *block sampling*. Throughout this paper, sampling is done with replacement. Sampling blocks has been explored in block-iterative methods for solving systems of linear equations [22]–[25]. Our motivation in dealing with blocks rather than individual vectors, is the availability to invoke results that can be used to characterize the approximations of distributed computing networks, to speed up first-order methods, as sampling individual rows/columns is prohibitive in real-world environments. This in turn leads to an *iterative sketching* approach, which has been well studied in terms of second-order methods [26]–[28]. By iterative sketching, we refer to an iterative algorithm which uses a new sketching matrix $\mathbf{S}_{[s]}$ at each iteration, i.e., throughout such a procedure the sketches $(\tilde{\mathbf{S}}_{[1]}\mathbf{A}, \tilde{\mathbf{S}}_{[2]}\mathbf{A}, \dots, \tilde{\mathbf{S}}_{[r]}\mathbf{A})$ are obtained. The scenario where a single sketching matrix \mathbf{S} is applied before the iterative process, is referred to as the *sketch-and-solve paradigm* [29], which also induces bias towards the samples considered through \mathbf{S} .

Second, we propose a general framework which incorporates our sketching algorithm into a CC approach. This framework accommodates a central class of sketching algorithms, that of importance (block) sampling algorithms (e.g., *CUR* decomposition [30], *CR*-multiplication [31]). Coded computing is a novel computing paradigm that utilizes coding theory to effectively inject and leverage data and computation redundancy to mitigate errors and slow or non-responsive servers; known as *stragglers*, among other fundamental bottlenecks, in large-scale distributed computing. In our setting, the straggling effect is due to computations being communicated over erasure channels, whose erasure probability follows a probability distribution which is central to the CC probabilistic model. The seminal work of [32] which first introduced CC, focused on exact matrix-vector multiplication and data shuffling. More recent works deal with recovering good approximations, while some have utilized techniques from RandNLA, e.g., [31], [33]–[39]. Our results are presented in terms of the standard CC probabilistic model proposed in [32], though they extend to any computing network comprised of a central server and computational nodes, referred to as *servers*.

To mitigate stragglers, we appropriately encode and replicate the data blocks, which leads to accurate CC estimates. In contrast to previous works which simply replicate each computational task or data block the same number of times [40]–[43], we replicate blocks a number of times that corresponds to their *block leverage scores*. Consequently, this induces a non-uniform sampling distribution in the aforementioned CC model, which is an approximation to the normalized block leverage scores. Furthermore, this eliminates the need for a decoding step, which can be a high complexity bottleneck in CC. A drawback of using RandNLA techniques is that exact computations are not recoverable. For more details on the approximation error and computational complexity of CC, the reader is referred to the monographs [44], [45]. A CC schematic, focusing on approximate GC, is presented in Figure 1.

The central idea behind our approach is that non-uniform importance sampling can be emulated, by replicating tasks across the network's computational nodes, who communicate through an erasure channel with the central server. By emulate, we mean that after replicating elements according to their importance sampling distribution and then sampling uniformly at random (without replacement) from the set containing redundant elements, we mimic sampling with replacement from the initial set, according to the importance sampling distribution. The tasks' computation times follow a runtime distribution [32], which, along with a

prespecified gradient transmission *ending time* T , determine the number of replications per task across the network. Though similar ideas have been proposed [31], [34]–[36] where sketching has been incorporated into CC, this is the first time redundancy is introduced through RandNLA, as compared to compression, in order to reduce complexity and provide approximation guarantees. Although uniform replication is a powerful technique, it does not capture the non-equal relevance of different rows of the data matrix to the information in the dataset. We capture such information through replication and rescaling according to the block leverage scores. By performing uniform sampling within these blocks, we attain a spectral approximation of the data matrix. One can compare this method to doing an iterative ℓ_2 -subspace embedding sketching of \mathbf{A} in standard CC. The shortcoming of this approach, is that a large amount of servers may be required when the underlying sampling distribution is non-uniform. We are able to overcome this obstacle, by approximating the importance sampling distribution as close as possible, under the physical constraint imposed by the network that the number of servers is fixed.

In Appendix E we discuss how further compression can be attained by introducing weighting, while guaranteeing the same approximation error when first and second order methods are used for linear regression in the sketch-and-solve paradigm (Proposition 6 and Corollary 4). We also show that, in terms of the expected reduction in sketching dimension, that the benefit decreases as the block leverage score distribution approaches the uniform distribution (Theorem 8). This further justifies the fact that higher diversity in leverage scores leads to more accurate algorithms [46].

At each iteration of the proposed descent algorithm, all completed jobs that are received by the central server are aggregated to form the final gradient approximation. Thus, unlike many other CC schemes, the proposed method does not store completed jobs which will not be accounted for. Our method sacrifices accuracy for speed of computation, and the inaccuracy is quantified in terms of the resulting ℓ_2 -subspace embedding (Theorem 1). Specifically, the computations of the responsive servers will correspond to sampled block computation tasks enabled by leverage score sampling, as summarized in Algorithm 4. Approximate CC is of interest since it could lead to faster inexact but accurate solutions, at a reduced computational cost [44].

To summarize, our main contributions are:

- 1) propose *block leverage score sketching*, to accommodate block sampling for ℓ_2 -subspace embedding, which makes leverage score sampling techniques suitable for distributed computations,
- 2) provide theoretical guarantees for the sketching performance,
- 3) show the significance of weighting, when our *weighted* sketching algorithm is used in iterative first and second order methods,
- 4) propose the use of *expansion networks*, which use the sampling distribution to determine how to replicate and distribute blocks in the CC framework — this unifies the disciplines of RandNLA and CC — where replication and uniform sampling (without a random projection) result in a spectral approximation,
- 5) show how expansion networks can be used for approximate distributed *steepest descent* (SD), which approaches the optimal solution using unbiased gradient estimators in a similar manner to *batch stochastic steepest descent* (SSD),
- 6) experimental validation of the performance of our algorithm on artificial datasets with non-uniform induced sampling distributions.

The paper is organized as follows. In Subsection I-A we present related works, in terms of CC, and in Subsection I-B we give an overview of the approximate GC literature. In Section II we present the notation which will be used, and review necessary background. We also list the notational conventions used in the paper, in Appendix F. In Subsection III-A we present our sketching algorithm and its theoretical approximation guarantees. In Subsections III-B and III-C we give a framework for which our sketching algorithm and potentially other importance sampling algorithms, can be used to devise CC schemes. This is where we introduce redundancy through RandNLA, which has not been done before. In Subsection III-D we summarize our GC scheme, and in Subsection III-E we give a brief synopsis of our main results by tying everything together. A concise description of our distributed procedure can be found in Algorithm 4 of Subsection III-D. We conclude with experimental evaluations in Section IV on simulated data with highly non-uniform underlying sampling distributions, as empirical validation.

A. Related Work on CC

There are several related works [35], [36], [47], [48] to the GC approach we present in this paper. The authors of [47] propose replicating subtasks according to what they define as the priority of the job, while [36] and [48] incorporate sketching into CC. It is worth noting that even though the focus of this work is on first order gradient methods, our approach also applies to second-order methods, as well as approximate matrix products through the *CR*-multiplication algorithm [10], [31], [49]. We briefly discuss this in Section V.

The work of [47] deals with matrix-vector multiplication. Similar to our work, the authors of [47] also replicate the computational tasks a certain number of times; and stop the process at a prespecified instance, but they do not use block leverage scores. Here, the computation $\mathbf{A}\mathbf{x}$ for $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{x} \in \mathbb{R}^N$ is divided into C different tasks, prioritizing the smaller computations. The m servers are split into c groups, and are asked to compute one of the tasks $\mathbf{y}_j = (\sum_{i \in \mathcal{J}_j} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top) \mathbf{x}$, for $\mathbf{A} = \sum_{l=1}^N \sigma_l \mathbf{u}_l \mathbf{v}_l^\top$ the SVD representation of \mathbf{A} . Each task \mathbf{y}_j is computed by the servers of the respective group, and $\mathbb{N}_N = \bigsqcup_{j=1}^c \mathcal{J}_j$ is a disjoint partitioning of the rank-1 outer-products of the SVD representation. The size of the j^{th} task is $|\mathcal{J}_j| = p_j$, which in our work is determined by the normalized block scores. The scores in our proposed scheme is motivated and justified by RandNLA, in contrast to the selection of the sizes p_j which is not discussed in [47]. Furthermore, the scheme of [47] requires a separate

maximum distance separable code for each job y_j ; thus requiring multiple decodings, while we do not require a decoding step. Another drawback of [47] is that an integer program is formulated to determine the optimal ending time, which the authors do not solve explicitly. In contrast, we determine a solution for any desired ending time. Lastly, we note that the ℓ_2 -subspace embedding approximation guarantee of our method, depends on the ending time T .

In terms of sketching and RandNLA, the works of [35] and [48] utilize the *Count-Sketch* [50]; which relies on hashing. In “CodedSketch” [48], Count-Sketches are incorporated into the design of a variant of the improved “Entangled Polynomial Code” [51], to combine approximate matrix multiplication with straggler tolerance. The code approximates the submatrix blocks $\{\mathbf{C}_{i,j} : (i,j) \in \mathbb{N}_{k_1} \times \mathbb{N}_{k_2}\}$ of the final product matrix $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ (matrix \mathbf{C} is partitioned k_1 times across its rows, and k_2 times across its columns), with an accuracy that depends on $\|\mathbf{C}_{i',j'}\|_F$ for all $(i',j') \in \mathbb{N}_{k_1} \times \mathbb{N}_{k_2}$. This hinders its practical applications, as it requires accuracy guarantees without oracle knowledge of the outcome of each submatrix of the matrix product \mathbf{C} . This approach permits each block of \mathbf{C} to be approximately recovered, if only a subset of the servers complete their tasks.

In “OverSketch” [35], redundancy is introduced through additional Count-Sketches, to mitigate the effect of stragglers in distributed matrix multiplication. In particular, the Count-Sketches $\check{\mathbf{A}} = \mathbf{A}\mathbf{S}$ and $\check{\mathbf{B}} = \mathbf{S}^\top \mathbf{B}$ of the two inputs \mathbf{A} and \mathbf{B} are computed, and are partitioned into submatrices of size $b \times b$. The $b \times b$ submatrices of the final product $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ are then approximately calculated, by multiplying the corresponding row-block of $\check{\mathbf{A}}$ and column-block of $\check{\mathbf{B}}$, each of which is done by a single server. The “OverSketch” idea has also been extended to distributed Newton Sketching [36] for convex optimization problems.

B. Related Work on Approximate GC

Many approaches have been proposed for approximate GC, which introduces redundancy in order to accelerate the convergence rate of distributed optimization algorithms [34], [52]–[65]. These approaches can be grouped into two main categories:

- (1) *fixed coefficient decoding*: at each iteration the same decoding step takes place, which is a linear combination of the received computations with predetermined coefficients [52]–[58],
- (2) *optimal coefficient decoding*: at each iteration another least squares optimization problem is solved, in order to determine the optimal coefficients that will be used in the linear combination of the received computations, as part of the decoding [34], [58]–[62].

We note that optimal coefficient decoding does not necessarily lead to optimal convergence, though we use this term to be consistent with the literature.

The scheme proposed in this paper falls in the category of fixed coefficient decoding, instead of optimal coefficient decoding. Even though we do not consider a decoding step, our approach can be viewed as having a fixed decoding step, which we explain in Subsection III-F. The main drawback of optimal coefficient decoding schemes is that a decoding vector has to be determined at every iteration, which is usually done by solving a linear system of equations, putting an additional computational burden on the central server. In Subsection III-F we show how our gradient approximations can be quantified in terms of the optimal coefficient decoding error, and also discuss how approaches from both fixed and optimal coefficient decoding can be accelerated by reducing the total number of iterations.

Similarly to our data replication approach is the work of [53], which replicates each data point $\mathbf{A}_{(i)}$ proportionally to its norm $\|\mathbf{A}_{(i)}\|_2$, to attain what they call a *pairwise-balanced* distribution scheme. The “Stochastic Gradient Coding” algorithm of [53] also converges in expectation to the optimal solution with appropriately selected step-sizes. We use a similar analysis to that of [53] to show that our method achieves convergence rate of $O(1/S)$ of the parameters, where S is the total number of iterations which take place. Furthermore, we also show that our expected regret converges to zero at a rate of $O(1/\sqrt{S} + r/S)$, and, unlike [53], we quantify the error of our gradient approximation in terms of the optimal gradient decoding residual error. The allocation of the data points in the Stochastic Gradient Coding algorithm [53] is done randomly, and the servers need to compute a weighted sum of the gradients corresponding to each of their allocated points, while we compute a partial gradient corresponding to the block and rescale it according to its block leverage score. The encoding of our scheme corresponds to the weighted coefficients of [53], since both are rescalings of the individual and block partial gradients. A benefit of our proposed replication of the data according to the block leverage scores is that it is more appropriate for spectral approximations of the data matrix, as it preserves the subspace spanned by the basis of the data matrix \mathbf{A} . This in turn allows us to also draw comparisons to approximations from optimal coefficient decoding schemes (Theorem 4).

The authors of [54], [55] base their constructions on balanced incomplete block designs (BIBDs), which are combinatorial structures that allow resiliency to potential adversarial stragglers. A drawback of the methods proposed in [54], [55], is that they do not provide a convergence guarantee to the optimal solution \mathbf{x}^* of the linear system, for their proposed iterative approximate gradient based scheme. A similar line of work to [53] that also considers adversarial stragglers is that of [58], which considers “Fractional Repetition Codes” [41], [42] with approximate decoding. This work also proposes “Bernoulli Gradient Codes” and a regularized variant (rBGC), where each data point $\mathbf{A}_{(i)}$ is randomly assigned to a fixed number of distinct servers. This GC scheme is an approximation of the pairwise-balanced scheme of [53], and can be viewed as a case of SSD when all the data points have the same norm.

In [57], an optimal (in expectation) step-size ξ_s^* was determined for each iteration of GC methods with unbiased gradients, which can be applied by the central server. The idea proposed in [57] applies to all fixed coefficient decoding GC methods, where the step-size is modified in order to reduce the overall number of iterations needed for SD to converge. The optimal step-size ξ_s^* is determined by the central server at each iteration once it receives sufficiently many computations to do so, from the computational nodes. This computation replaces the decoding step performed by the central server, which is of high complexity.

Another line of work considers using low density matrices as encoding matrices for distributed optimization [64], [65]. In [64] the authors distribute the data to the servers using a Low Density Generator Matrix code. As with optimal coefficient decoding, a drawback of [64] is that at each iteration the central server has to run a decoding algorithm to decode the sum of the received partial gradients. The work of [65] focuses on linear loss functions, in which the data sent to the computational nodes is encoded through a Low Density Parity Check code. As is the case for all exact schemes, in [65] there is a threshold on the number of servers that need to respond in order to recover the exact gradient by performing the decoding step. In the CC literature, this is referred to as the *recovery threshold*, and has been extensively studied [66]. If this threshold is not met in the scheme of [65], the central server leverages the Low Density Parity Check code to compute an estimate of the gradient. To this extent, it is worth mentioning that both our proposed method and that of [53] are flexible in terms of the required number of responses, as in both approaches the gradient approximation degrades gracefully if more stragglers than expected occur. This is an important consideration, as in practice there is a great deal of variability in the number of stragglers over time.

It is important to also mention that there are many papers which consider distributed stochastic first-order methods which do not take account of stragglers [67]. There are also many works on CC that focus on the issue of numerical stability, and the problems of exact and approximate matrix multiplication, e.g., [66], [68]–[77]. We summarize the most relevant results from the works discussed above on approximate GC in Table I.

Comparison of Related Work			
Gradient Code	Decoding Coefficients	Residual Error	Convergence of $\mathbf{x}^{[s]}$
Expander Codes [52]	Fixed	—	Yes
Expander Codes [59]	Optimal	$Np^{\bar{r}-o(\bar{r})}$	Yes
Pairwise Balanced [53]	Fixed	—	Yes
BIBD [54]	Fixed & Optimal	—	No
rBGC [58]	Fixed	—	No
<i>Leverage Scores GC</i>	Fixed	$4\epsilon^2$	Yes

TABLE I: The second column indicates the category of the GC. The third column contains the average case normalized gradient residual error $\frac{\|g^{[s]} - \hat{g}^{[s]}\|_2^2}{\|\mathbf{A}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2}$, where $p = \frac{m-q}{m}$ is the fraction of stragglers, \bar{r} is the average replication factor, N is the number of data points in the data matrix \mathbf{x} , and ϵ is the ℓ_2 -subspace embedding error of our sketching algorithm. The fourth column, indicates whether a convergence proof of the parameters vector $\mathbf{x}^{[s]}$ is provided in the corresponding work. Our work is presented in the last row.

II. NOTATION AND BACKGROUND

We denote $\mathbb{N}_n := \{1, 2, \dots, n\}$, and $X_{\{n\}} = \{X_i\}_{i=1}^n$, where X is an arbitrary variable. We use \mathbf{A}, \mathbf{B} to denote real matrices, \mathbf{b}, \mathbf{x} real column vectors, \mathbf{I}_n the $n \times n$ identity matrix, $\mathbf{0}_{n \times m}$ and $\mathbf{1}_{n \times m}$ respectively the $n \times m$ all zeros and all ones matrices, and by \mathbf{e}_i the standard basis column vector whose dimension will be clear from the context. The largest eigenvalue of a matrix \mathbf{M} , is denoted by $\lambda_1(\mathbf{M})$. By $\mathbf{A}_{(i)}$ we denote the i^{th} row of \mathbf{A} , by $\mathbf{A}^{(j)}$ its j^{th} column, by \mathbf{A}_{ij} the value of \mathbf{A} 's entry in position (i, j) , and by x_i the i^{th} element of \mathbf{x} . The rounding function to the nearest integer is expressed by $\lfloor \cdot \rfloor$, i.e., $\lfloor a \rfloor = \lfloor a + 1/2 \rfloor$ for $a \in \mathbb{R}$. The cardinality of a set is expressed by $|\cdot|$, e.g., $|\{1, 2, 5, 9\}| = 4$. Disjoint unions are represented by \sqcup , e.g., $\mathbb{Z} = \{j : j \text{ is odd}\} \sqcup \{j : j \text{ is even}\}$, and we define \uplus as the union of multisets, e.g., $\{1, 2, 3\} \uplus \{3, 4\} = \{1, 2, 3, 3, 4\}$. The diagonal matrix with real entries $a_{\{n\}}$ is expressed as $\text{diag}(a_{\{n\}})$. The Moore–Penrose inverse of a rank m matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ with $n > m$, is defined as $\mathbf{M}^\dagger := (\mathbf{M}^\top \mathbf{M})^{-1} \mathbf{M}^\top$.

We partition vectors and matrices across their rows into K submatrices, so that any two submatrices either have the same number of rows or one of them has at most one additional row. Such a partitioning is always attainable through simple arithmetic [43]. For example, if we are to partition 100 rows into 7 blocks, 5 blocks are comprised of 14 rows and 2 blocks are comprised of 15 rows, and $100 = 5 * 14 + 2 * 15$. This allows us to assume that each node has the same computational load, and expected response times. For simplicity, we assume that K divides the number of rows N , denoted $K \mid N$. That is, for an ℓ_2 -subspace embedding of $\mathbf{A} \in \mathbb{R}^{N \times d}$ with target $\mathbf{b} \in \mathbb{R}^N$, we assume $K \mid N$ and the “size of each partition” is $\tau = N/K$.¹ We partition \mathbf{A} and \mathbf{b} across their rows:

$$\mathbf{A} = \left[\mathbf{A}_1^\top \cdots \mathbf{A}_K^\top \right]^\top \quad \text{and} \quad \mathbf{b} = \left[\mathbf{b}_1^\top \cdots \mathbf{b}_K^\top \right]^\top \quad (1)$$

¹If $K \nmid N$, we appropriately append zero vectors/entries until this is met. It is not required that all blocks have the same size, though we discuss this case to simplify the presentation. One can easily extend our results to blocks of varying sizes, and use the analysis from [42] to determine the optimal size of each partition.

where $\mathbf{A}_i \in \mathbb{R}^{\tau \times d}$ and $\mathbf{b}_i \in \mathbb{R}^\tau$ for all $i \in \mathbb{N}_K$. Partitions, are referred to as *blocks*. Throughout the paper we consider the case where $N \gg d$. For \mathbf{A} full-rank, its SVD is $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{N \times d}$ is its reduced left orthonormal basis.

Matrix \mathbf{A} represents a dataset \mathcal{D} of N samples with d features, and \mathbf{b} the corresponding labels of the data points. The partitioning (1) corresponds to K sub-datasets $\mathcal{D}_{\{K\}}$, i.e., $\mathcal{D} = \bigsqcup_{j=1}^K \mathcal{D}_j$. Our results are presented in terms of an arbitrary partition $\mathbb{N}_N = \bigsqcup_{\iota=1}^K \mathcal{K}_\iota$, for \mathbb{N}_N the index set of the rows of \mathbf{A} and \mathbf{b} . The index subsets $\mathcal{K}_{\{K\}}$ indicate which data samples are in each sub-dataset. By $\mathbf{A}_{(\mathcal{K}_\iota)}$, we denote the submatrix of \mathbf{A} comprised of the rows indexed by \mathcal{K}_ι . That is, for $\mathbf{I}_{(\mathcal{K}_\iota)}$ the restriction of \mathbf{I}_N to only include its rows corresponding to \mathcal{K}_ι , we have $\mathbf{A}_{(\mathcal{K}_\iota)} = \mathbf{I}_{(\mathcal{K}_\iota)} \cdot \mathbf{A}$. By \mathcal{K}_ι^i , we indicate that the ι^{th} block was sampled at trial i , i.e., the superscript i indicates the sampling trial and the subscript $\iota \in \mathbb{N}_K$ which block was sampled at that trial. An explicit example where this notation is used, can be found in Appendix B. Also, by $j(i)$ we denote the index of the submatrix which was sampled at the i^{th} sampling trial, i.e., $\mathcal{K}_{j(i)} = \mathcal{K}_{j(i)}^i$. The complement of \mathcal{K}_j is denoted by $\bar{\mathcal{K}}_j$, i.e., $\bar{\mathcal{K}}_j = \mathbb{N}_N \setminus \mathcal{K}_j$, for which $\mathbf{U}_{(\mathcal{K}_j)}^\top \mathbf{U}_{(\bar{\mathcal{K}}_j)} = \mathbf{I}_d - \mathbf{U}_{(\mathcal{K}_j)}^\top \mathbf{U}_{(\bar{\mathcal{K}}_j)}$. By $t \leftarrow T$, we mean T is a realization of the time variable t .

Sketching matrices are represented by \mathbf{S} and $\tilde{\mathbf{S}}_{[s]}$. The script $[s]$ indexes an iteration $s = 0, 1, 2, \dots$ which we drop when it is clear from the context. Throughout the paper, we will be reducing dimension N of \mathbf{A} to r , i.e., $\mathbf{S} \in \mathbb{R}^{r \times N}$. Sampling matrices are denoted by $\mathbf{\Omega} \in \{0, 1\}^{r \times N}$, and diagonal rescaling matrices by $\mathbf{D} \in \mathbb{R}^{N \times N}$.

A. Least Squares Approximation

Linear least squares approximation, abbreviated to least squares, is a technique to find an approximate solution to a system of linear equations that has no exact solution [78]. Consider the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, for which we want to find an approximation to the best-fitted

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \right\}, \quad (2)$$

where the objective function L_{ls} has gradient

$$g^{[s]} = \nabla_{\mathbf{x}} L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]}) = 2\mathbf{A}^\top (\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}). \quad (3)$$

We refer to the gradient of the block pair $(\mathbf{A}_i, \mathbf{b}_i)$ from (1) as the i^{th} *partial gradient*

$$g_i^{[s]} = \nabla_{\mathbf{x}} L_{ls}(\mathbf{A}_i, \mathbf{b}_i; \mathbf{x}^{[s]}) = 2\mathbf{A}_i^\top (\mathbf{A}_i \mathbf{x}^{[s]} - \mathbf{b}_i). \quad (4)$$

Existing exact methods find a solution vector \mathbf{x}^* in $O(Nd^2)$ time, where $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{b}$. In Subsection III-D we focus on approximating the optimal solution \mathbf{x}^* by using our methods, via distributive SD/SSD and iterative sketching. What we present also accommodates regularizers of the form $\gamma \|\mathbf{x}\|_2^2$, though to simplify the presentation of our expressions, we only consider (2).

B. Steepest Descent

When considering a minimization problem with a convex differentiable objective function $L: \mathbb{R}^d \rightarrow \mathbb{R}$, we select an initial $\mathbf{x}^{[0]} \in \mathbb{R}^d$ and repeat at iteration $s + 1$:

$$\mathbf{x}^{[s+1]} \leftarrow \mathbf{x}^{[s]} - \xi_s \cdot \nabla_{\mathbf{x}} L(\mathbf{x}^{[s]})$$

for $s = 0, 1, 2, \dots$, until a prespecified termination criterion is met. This iterative procedure is called steepest descent, also known as gradient descent. The parameter $\xi_s > 0$ is the corresponding step-size, which may be adaptive or fixed. To guarantee convergence of L_{ls} , one can select $\xi_s = 2/\sigma_{\max}(\mathbf{A})^2$ for all iterations, though this is too conservative.

C. Leverage Scores

In high dimensional least squares problems subsampling of the rows of \mathbf{A} is often used to reduce computational load. One commonly used method for subsampling uses what are known as *leverage scores* [79], [80] which characterize the importance of the data points (rows of \mathbf{A}). Common applications of leverage scores include: approximation by low rank [81], [82], approximate matrix multiplication [83], tensor products and approximations [84]–[86], matrix completion [87], kernel ridge regression [88], k -means clustering [89], graph sparsification [90], and Nyström-based low-rank approximations [91]. The leverage scores of \mathbf{A} measure the extent to which the vectors of its orthonormal basis \mathbf{U} are correlated with the standard basis, and define the key structural non-uniformity that underlies fast randomized matrix algorithms. Leverage scores are defined as $\ell_i := \|\mathbf{U}_{(i)}\|_2^2$, and are agnostic to any particular basis, as they are equal to the diagonal entries of the projection matrix $P_{\mathbf{A}} = \mathbf{A}\mathbf{A}^\dagger = \mathbf{U}\mathbf{U}^\top$. The normalized leverage scores $\pi_{\{N\}}$ of \mathbf{A} :

$$\pi_i := \|\mathbf{U}_{(i)}\|_2^2 / \|\mathbf{U}\|_F^2 = \|\mathbf{U}_{(i)}\|_2^2 / d \quad (5)$$

for each $i \in \mathbb{N}_N$, form a sampling probability distribution, as $\sum_{i=1}^N \pi_i = 1$ and $\pi_i \geq 0$ for all i . This induced distribution has proven to be useful in linear regression [15], [18], [80], [92], [93].

The *normalized block leverage scores*, introduced independently in [30], [34], are the sum of the normalized leverage scores of the subset of rows constituting each row-block of \mathbf{A} (1). Similar to leverage scores, these scores measure the correlation

between the data block's corresponding submatrices in \mathbf{U} and \mathbf{I}_N , as well as how much a particular row-block contributes to the approximation of \mathbf{A} . Analogous to (5), considering the partitioning of the dataset \mathcal{D} according to $\mathcal{K}_{\{K\}}$, the normalized block leverage scores of \mathbf{A} are defined as

$$\Pi_l := \|\mathbf{U}_{(\mathcal{K}_l)}\|_F^2 / \|\mathbf{U}\|_F^2 = \|\mathbf{U}_{(\mathcal{K}_l)}\|_F^2 / d = \sum_{j \in \mathcal{K}_l} \pi_j \quad (6)$$

for each $l \in \mathbb{N}_K$. This is a direct generalization of (5), as

$$\|\mathbf{U}_{(\mathcal{K}_l)}\|_F^2 = \left\| \left[(\mathbf{U}_{(\mathcal{K}_l)})_{(1)} \cdots (\mathbf{U}_{(\mathcal{K}_l)})_{(|\mathcal{K}_l|)} \right] \right\|_2^2.$$

Much like individual leverage scores, block leverage scores measure how much a particular block contributes to the approximation of the matrix \mathbf{A} [30].

A related notion is that of the *Frobenius block scores*, which, in the case of a partitioning as in (1), are $\|\mathbf{A}_\iota\|_F^2$ for each $\iota \in \mathbb{N}_K$. Such scores have been used for approximate matrix multiplication, e.g., *CR*-multiplication [12], [13], [49]. In our context, the block leverage scores of \mathbf{A} are the Frobenius block scores of \mathbf{U} .

A drawback of using leverage scores, is that calculating them requires $O(Nd^2)$ time. To alleviate this, one can instead settle for relative-error approximations which can be approximated much faster, e.g., [80] does so in $O(Nd \log N)$ time. In particular, one can consider approximate normalized scores $\tilde{\Pi}_{\{K\}}$ where $\tilde{\Pi}_i \geq \beta \Pi_i$ for all i , for some misestimation factor $\beta \in (0, 1]$. Since $\Pi_{\{K\}}$ and $\tilde{\Pi}_{\{K\}}$ are identical if and only if $\beta = 1$, a higher β implies the approximate distribution is more accurate. We specify that a misestimation factor β to $\Pi_{\{K\}}$ is for a specific approximate distribution $\tilde{\Pi}_{\{K\}}$, by denoting it as $\beta_{\tilde{\Pi}}$.

Approximate block sampling distributions to $\Pi_{\{K\}}$ are denoted by $\tilde{\Pi}_{\{K\}}$, and the distributions induced through *expansion networks*; which we introduce in Subsection III-B, are denoted by $\tilde{\Pi}_{\{K\}}$. We quantify the difference between distributions $\Pi_{\{K\}}$ and $\tilde{\Pi}_{\{K\}}$ by the distortion metric $d_{\Pi, \tilde{\Pi}} := \frac{1}{K} \sum_{i=1}^K |\Pi_i - \tilde{\Pi}_i|$, which is the ℓ_1 distortion between $\Pi_{\{K\}}$ and $\tilde{\Pi}_{\{K\}}$ [94].

D. Subspace Embedding

Our approach to approximating (2), is to apply an ℓ_2 -subspace embedding sketching matrix $\mathbf{S} \in \mathbb{R}^{r \times N}$ on \mathbf{A} . Recall that $\mathbf{S} \in \mathbb{R}^{r \times N}$ is a $(1 \pm \epsilon)$ ℓ_2 -subspace embedding of the column-space of \mathbf{A} , if

$$\Pr \left[(1 - \epsilon) \|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \epsilon) \|\mathbf{A}\mathbf{x}\|_2^2 \right] \geq 1 - \delta \quad (7)$$

for all $\mathbf{x} \in \mathbb{R}^d$, where $\delta \geq 0$ is the failure probability [18]. Notice that such an \mathbf{S} , is also a $(1 \pm \epsilon)$ ℓ_2 -subspace embedding of \mathbf{U} , as $\{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\} = \{\mathbf{U}\mathbf{y} : \mathbf{y} \in \mathbb{R}^d\}$. This implies that the event whose probability is bounded in (7) is equivalent to simultaneously satisfying

$$(1 - \epsilon) \|\mathbf{y}\|_2^2 = (1 - \epsilon) \|\mathbf{U}\mathbf{y}\|_2^2 \leq \|\mathbf{S}\mathbf{U}\mathbf{y}\|_2^2 \quad (8)$$

and

$$(1 + \epsilon) \|\mathbf{y}\|_2^2 = (1 + \epsilon) \|\mathbf{U}\mathbf{y}\|_2^2 \geq \|\mathbf{S}\mathbf{U}\mathbf{y}\|_2^2 \quad (9)$$

for all $\mathbf{y} \in \mathbb{R}^d$. The lower (8) and upper (9) bounds on $\|\mathbf{S}\mathbf{U}\mathbf{y}\|_2^2$ respectively imply

$$\mathbf{y}^\top (\mathbf{I}_d - (\mathbf{S}\mathbf{U})^\top \mathbf{S}\mathbf{U}) \mathbf{y} \leq \epsilon \|\mathbf{y}\|_2^2$$

and

$$\mathbf{y}^\top ((\mathbf{S}\mathbf{U})^\top \mathbf{S}\mathbf{U} - \mathbf{I}_d) \mathbf{y} \leq \epsilon \|\mathbf{y}\|_2^2$$

thus, a simplified condition for an ℓ_2 -subspace embedding of \mathbf{A} is

$$\Pr \left[\|\mathbf{I}_d - \mathbf{U}^\top \mathbf{S}^\top \mathbf{S}\mathbf{U}\|_2 \leq \epsilon \right] \geq 1 - \delta \quad (10)$$

for a small $\delta \geq 0$. Further details on the derivation of (10) can be found in Appendix F.

For the overdetermined system $\mathbf{A}\mathbf{x} = \mathbf{b}$, we require $r > d$, and in the sketch-and-solve paradigm the objective is to determine an $\hat{\mathbf{x}}$ that satisfies

$$(1 - \epsilon) \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2 \leq \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_2 \leq (1 + \epsilon) \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2 \quad (11)$$

where $\hat{\mathbf{x}}$ is an approximate solution to the modified least squares problem

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ L_{\mathbf{S}}(\mathbf{S}, \mathbf{A}, \mathbf{b}; \mathbf{x}) := \|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2 \right\}. \quad (12)$$

If (10) is met, we get with high probability the approximation characterizations:

- 1) $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_2 \leq \frac{1+\epsilon}{1-\epsilon} \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2 \leq (1 + O(\epsilon)) \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2$
- 2) $\|\mathbf{A}(\mathbf{x}^* - \hat{\mathbf{x}})\|_2 \leq \epsilon \|(\mathbf{I}_N - \mathbf{U}\mathbf{U}^\top)\mathbf{b}\|_2 = \epsilon \|\mathbf{b}^\perp\|_2$

where $\mathbf{b}^\perp = \mathbf{b} - \mathbf{A}\mathbf{x}^*$ is orthogonal to the column-span of \mathbf{A} , i.e., $\mathbf{A}^\top \mathbf{b}^\perp = \mathbf{0}_{d \times 1}$.

E. Coded Computing Probabilistic Model

In this subsection, we describe the GC problem and CC probabilistic model we will be considering, which were respectively introduced in [41] and [32]. In GC (Figure 1), there is a central server that shares the K disjoint subsets $\mathcal{D}_{\{K\}}$ of \mathcal{D} among m *homogeneous*² servers, to facilitate computing the solution of the following minimization problem, of the form (2):

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{j=1}^K L_{ls}(\mathcal{D}_j; \mathbf{x}) \right\}.$$

Since the objective function $L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x})$ is differentiable and additively separable, it follows that $g^{[s]} = \sum_{j=1}^K g_j^{[s]}$. The objective function's gradient is updated in a distributed manner while only requiring q servers to respond, i.e., it is robust to $m - q$ stragglers. This is achieved through an encoding of the computed partial gradients by the servers, and a decoding step once q servers have sent back their encoded computation. In our approach, we consider *approximate* GC by requesting updated partial gradients according to a sketched version of the objective problem given in (12).

The probabilistic computational model introduced in [32], which is the standard CC paradigm, is central to our framework. In this model one assumes the existence of a *mother runtime distribution* $F(t)$, with a corresponding probability density function $f(t)$. Let T_0 be the time it takes a single machine to complete its computation, and define the cumulative distribution function $F(t) := \Pr[T_0 \leq t]$. When the computation budget is τ , and the computations are partitioned into K subtasks and distributed, each of size $\frac{\tau}{N}$, the runtime distribution of each subtask is assumed to follow the distribution $\tilde{F}(t) := F(t\tau/N) = \Pr[T_i \leq t]$, where T_i is the random completion time of the i^{th} subtask. Moreover, since the m servers are homogeneous and the subtasks are of the same size, the distribution is the same for each $i \in \mathbb{N}_K$. In this work, we view the computations as being communicated to the central server over erasure channels, where the l^{th} server W_l has an erasure probability³

$$\phi(t) := 1 - \tilde{F}(t) = 1 - \Pr[W_l \text{ responds by time } t] \quad (13)$$

i.e., the probability that W_l is a straggler at time t is $\phi(t)$. All servers have the same erasure probability, as we are assuming they are homogeneous.

In our setting, there are two hyperparameters required for determining an expansion network. First, one needs to determine a time instance $t \leftarrow T$ after which the central server will stop receiving servers' computations. This may be decided by factors such as the system's limitations, number of servers, or an upper bound on the desired waiting time for the servers to respond. At time T , according to $\tilde{F}(t)$, the central server receives roughly $q(T) := \lfloor \tilde{F}(T) \cdot m \rfloor$ server computations, where m is the total number of computational nodes in the network. We refer to the prespecified time instance T after which the central server stops receiving computations, as the *ending time*. If the sketching procedure of the proposed sketching algorithm were to be carried out by a single server, there would be no benefit in setting T such that $q(T)\tau > N$, as the exact calculation could have taken place in the same amount of time. In distributed networks though there is no control over which servers respond, and it is not a major concern if $q(T)\tau$ is slightly above N , as this still results in acceleration of the computation. The trade-off between accuracy and waiting time t is captured in Theorem 1, for $q \leftarrow q(t)$ sampling trials. The second hyperparameter we need in order to design an expansion network is the block size τ , which is determined by the number of partitions K illustrated in (1). Together, $q(T)$ and τ determine the ideal number of servers needed for our framework to perfectly emulate sampling according to the normalized block leverage scores $\Pi_{\{K\}}$.

III. CODED COMPUTING FROM RANDNLA

In this section, we introduce our *block* leverage score sampling algorithm, which is more practical and can be carried out more efficiently than its vector-wise counterpart. Our ℓ_2 -subspace embedding result is presented in Theorem 1. By setting $\tau = 1$ and $\beta = 1$, we recover a known result for (exact) leverage score sampling.

In Subsection III-B we incorporate our block sampling algorithm into the CC probabilistic model described in Subsection II-E, in which we leverage task redundancy to mitigate stragglers. Specifically, we show how to replicate computational tasks among the servers, under the integer constraints imposed by the physical system and the desired waiting time. This is then used to approximate the gradient at each iteration in a way that emulates the sampling procedure of the sketching method presented in Algorithm 1. In Subsection III-C we further elaborate on when a perfect emulation is possible, and how emulated block leverage score sampling can be improved when it cannot be done perfectly, through the proposed networks. In Subsection III-D we present our GC approach and relate it to SD and SSD, which in turn implies convergence guarantees with appropriate step-sizes. In Subsection III-D, we also provide an algorithmic description of our GC procedure, to tie together the algorithms discussed in this section. Furthermore, at each iteration we have a different induced sketch, hence our procedure lies under the framework of iterative sketching. Specifically, we obtain gradients of multiple sketches of the data $(\tilde{\mathbf{S}}_{[1]}\mathbf{A}, \tilde{\mathbf{S}}_{[2]}\mathbf{A}, \dots, \tilde{\mathbf{S}}_{[n]}\mathbf{A})$ and iteratively

²This means that they have the same computational power, independent and identically distributed statistics for the computing time of similar tasks, and expected response time.

³This is also known as the *survival function*: $\phi(t) = \int_t^\infty (1 - \tilde{f}(u)) du = 1 - \tilde{F}(t) = \Pr[T_i > t]$, for $\tilde{f}(t)$ the PDF corresponding to $\tilde{F}(t)$. The function $\phi(t)$ is monotonically decreasing.

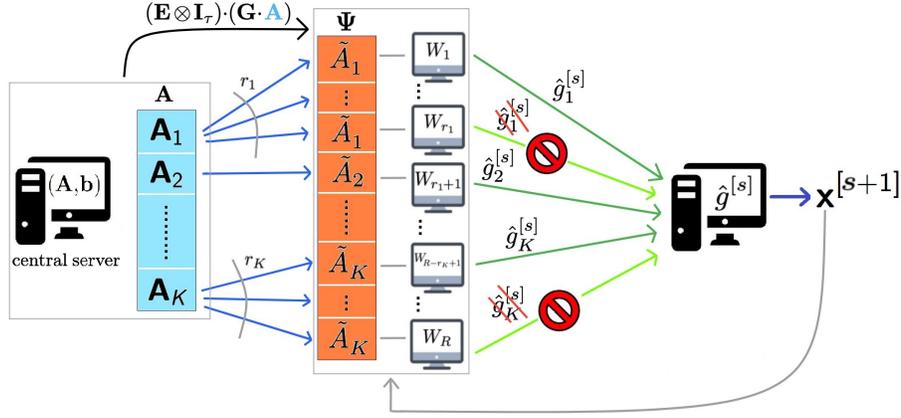


Fig. 2: Illustration of our GC approach, at iteration $s + 1$. The blocks of \mathbf{A} (and \mathbf{b}) are encoded through \mathbf{G} and then replicated through $\mathbf{E} \otimes \mathbf{I}_\tau$, where each block of the resulting Ψ is given to a single server, before the iterative SD procedure takes place. At the illustrated iteration, servers W_{r_1} and W_R are stragglers, and their computations are not received. The central server determines the estimate $\hat{g}^{[s]}$, and then shares $\mathbf{x}^{[s+1]}$ with all the computational nodes. The resulting estimate is the gradient of the induced sketch, i.e., $\hat{g}^{[s]} = \nabla_{\mathbf{x}} L_S(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]})$.

refine the solution, where n can be chosen logarithmic in N . A schematic of our approach is presented in Figure 2, and in Appendix B we provide a concrete example of the induced sketching matrices resulting from the iterative process.

A. Block Leverage Score Sampling

In the leverage score sketch [15], [18], [79], [80], [92] we sample with replacement r rows according $\pi_{\{N\}}$ (5), and then rescale each sampled row by $1/\sqrt{r\pi_{i_j}}$, where π_{i_j} is the sampling probability of the row that was selected at the j^{th} trial. Instead, we sample with replacement q blocks from (1) according to $\Pi_{\{K\}}$ (6), and rescale them by $1/\sqrt{q\Pi_{i_j}}$. The pseudocode of the block leverage score sketch is given in Algorithm 1, where we consider an approximate distribution $\tilde{\Pi}_{\{K\}}$ such that $\tilde{\Pi}_i \geq \beta\Pi_i$ for all i , for $\beta \in (0, 1]$ a dependent loss in accuracy [15], [80], [95]. The spectral guarantee of the sketching matrix $\tilde{\mathbf{S}}$ of Algorithm 1 is presented in Theorem 1. Iterative sketching in our distributed GC approach through Algorithm 1, corresponds to selecting a new sampling matrix $\tilde{\Omega}^{[s]}$ at each iteration through the servers' responses, i.e., $\tilde{\mathbf{S}}_{[s]} = \tilde{\mathbf{D}} \cdot \tilde{\Omega}^{[s]}$ for each s .

Algorithm 1: Block Leverage Score Sketch

Input: $\mathbf{A} \in \mathbb{R}^{N \times d}$, $\tau = \frac{N}{K}$, $q = \frac{r}{\tau} > \frac{d}{\tau}$

Output: $\tilde{\mathbf{S}} \in \mathbb{R}^{r \times N}$, $\tilde{\mathbf{A}} \in \mathbb{R}^{r \times d}$

Initialize: $\Omega = \mathbf{0}_{q \times K}$, $\mathbf{D} = \mathbf{0}_{q \times q}$

Compute: (approximate) distribution $\tilde{\Pi}_{\{K\}}$ (6)

for $j = 1$ to q **do**

 sample with replacement i_j from \mathbb{N}_K , according to $\tilde{\Pi}_{\{K\}}$

$\Omega_{j,i_j} = 1$

$\mathbf{D}_{j,j} = \sqrt{\frac{\tau}{r\Pi_{i_j}}} = \sqrt{\frac{1}{q\Pi_{i_j}}}$

▷ equivalently $\Omega_{(j)} = \mathbf{e}_{i_j}^\top$

end

$\tilde{\Omega} \leftarrow \Omega \otimes \mathbf{I}_\tau$

$\tilde{\mathbf{D}} \leftarrow \mathbf{D} \otimes \mathbf{I}_\tau$

$\tilde{\mathbf{S}} \leftarrow \tilde{\mathbf{D}} \cdot \tilde{\Omega}$

▷ $\tilde{\mathbf{S}} = (\mathbf{D} \cdot \Omega) \otimes \mathbf{I}_\tau$

$\tilde{\mathbf{A}} \leftarrow \tilde{\mathbf{S}} \cdot \mathbf{A}$

Theorem 1. The sketching matrix $\tilde{\mathbf{S}}$ of Algorithm 1 is a $(1 \pm \epsilon)$ ℓ_2 -subspace embedding of \mathbf{A} , according to (10). Specifically, for $\delta > 0$ and $q = \Theta\left(\frac{d}{\tau} \log(2d/\delta)/(\beta\epsilon^2)\right)$:

$$\Pr \left[\|\mathbf{I}_d - \mathbf{U}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \mathbf{U}\|_2 \leq \epsilon \right] \geq 1 - \delta.$$

Proof. The main tool is a matrix Chernoff bound [18, Fact 1]. We define random matrices corresponding to the sampling process and bound their norm and variance, in order to apply the matrix Chernoff bound. The complete proof can be found in Appendix A. \square

Theorem 1 applies to more general problems than leverage score sampling. Specifically, one can apply a random projection to “flatten” the block leverage scores, i.e., to make them all approximately equal, and then sample with replacement uniformly at random. This is the main idea behind the analysis of the SRHT [96], [97]. The trade-off between such algorithms and Algorithm 1, is computing the leverage scores explicitly vs. the computational reduction due to applying a random projection. Sketching approaches similar to the SRHT which do not directly utilize the data, are referred to as “data oblivious sketches”. These approaches are better positioned for handling high velocity streams as well as highly unstructured and arbitrarily distributed data [98]. Multiplying the data by a random matrix spreads the information in the rows of the matrix, such that all rows are of equal importance, and the new matrix is “incoherent”. In Appendix C, we show when Algorithm 1 and the corresponding block sampling counterpart of the SRHT [56] achieve the same asymptotic guarantees for the same number of sampling trials q .

Next, we provide a sub-optimality result for non-iterative sketching of the block leverage score sketch for ordinary least squares

$$\tilde{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ L_S(\tilde{\mathbf{S}}, \mathbf{A}, \mathbf{b}; \mathbf{x}) := \|\tilde{\mathbf{S}}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2 \right\} \quad (14)$$

which follows from the results of [26]. The following Corollary is established by adapting the proof of [26, Theorem 1], which is based on a reduction from statistical minimax theory combined with information-theoretic bounds and an application of Fano’s inequality, yielding an upper bound to $\|\mathbb{E}[\tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top)^{-1} \tilde{\mathbf{S}}]\|_2$. The proof of Corollary 1 can be found in Appendix A.

Corollary 1. *For any full-rank data matrix $\mathbf{A} \in \mathbb{R}^{N \times d}$ with a noisy observation model $\mathbf{b} = \mathbf{A}\mathbf{x}^\bullet + \mathbf{w}$ where \mathbf{x}^\bullet is an arbitrary vector of length d and $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$, the optimal least squares solution \mathbf{x}^* of (2) has prediction error $\mathbb{E}[\|\mathbf{A}(\mathbf{x}^\bullet - \mathbf{x}^*)\|_2^2] \lesssim \frac{\sigma^2 d}{N}$. Furthermore, any estimate $\tilde{\mathbf{x}}$ based on the sketched system $(\tilde{\mathbf{S}}\mathbf{A}, \tilde{\mathbf{S}}\mathbf{b})$ produced by Algorithm 1 with sampling probabilities $\Pi_{\{K\}}$, has a prediction error lower bound of*

$$\mathbb{E}[\|\mathbf{A}(\mathbf{x}^\bullet - \tilde{\mathbf{x}})\|_2^2] \gtrsim \frac{\sigma^2 d}{\min\{r, N\}}. \quad (15)$$

Even though Corollary 1 considers sampling according to the exact block leverage scores, its proof can be modified to accommodate approximate sampling. Additionally, the above corollary holds for constrained least squares, though we do not explicitly state it, as this is not a focus of the work presented in this paper. From (15), it is clear that for a smaller r with $r < N$ the proposed approach produces a less efficient sketch and poorer approximation $\tilde{\mathbf{x}}$, though when considering a higher r approaching N , there is an improvement in the accuracy of $\tilde{\mathbf{x}}$ at the cost of a higher computational load.

B. Expansion Networks

The framework we propose emulates the sampling with replacement procedure of Algorithm 1 in distributed CC environments. Even though we focus on ℓ_2 -subspace embedding and descent methods in this paper, the proposed framework applies to any matrix algorithm which utilizes importance sampling with replacement. In contrast to other CC schemes in which RandNLA was used to compress the network, e.g., [31], [34], [39], here the networks are expanded according to $\Pi_{\{K\}}$. That is, the computations corresponding to the row-blocks of \mathbf{A} are replicated by the servers proportionally to $\Pi_{\{K\}}$. It is unlikely that we can exactly emulate this distribution, as the number of replications per task need to be integers. Instead, we mimic the exact probabilities with an induced distribution $\tilde{\Pi}_{\{K\}}$ through expansion networks, which are determined by $\tilde{F}(t)$ at a prespecified T , after which the central server stops receiving computations for that iteration.

The proposed method solves the minimization problem (19), whose approximate solution $\hat{r}_{\{K\}}$ (20) determines the number of replicas $r_{\{K\}}$ of each block in our expansion network. We note that (19) is a surrogate to the integer program (22), whose solution can achieve an accurate realizable distribution $\tilde{\Pi}_{\{K\}}$ to $\Pi_{\{K\}}$ through the distributed network, by appropriately replicating the blocks. Unfortunately, the integer program (22) is not always solvable. Nonetheless, when we have an approximation to (19) or (22), through uniform sampling we can minimize with high probability the ℓ_2 -subspace embedding condition for \mathbf{A} (10), up to a small error, given the integer constraints imposed by the physical system — $r_i \in \mathbb{Z}_+$ for all i and $R = \sum_{l=1}^K r_l$ such that $R \approx m$. In the CC context we want $m = R$, i.e., the total number of replicated blocks is equal to the number of servers. Next, we describe the desired induced distribution $\tilde{\Pi}_{\{K\}}$, in order to set up the optimization problem (19).

Assume without loss of generality that $\Pi_j \leq \Pi_{j+1}$ for all $j \in \mathbb{N}_{K-1}$, thus $r_j \leq r_{j+1}$. The sampling distribution through the expansion network translates to

$$\tilde{\Pi}_i := \Pr[\tilde{A}_i \text{ is sampled}] = r_i/R \approx \Pi_i \quad (16)$$

for all $i \in \mathbb{N}_K$ and $R = \sum_{l=1}^K r_l$, where \tilde{A}_i is the i^{th} encoded partition of our GC scheme. Our objective is to determine $r_{\{K\}}$ such that $\tilde{\Pi}_i \approx \Pi_i$ for all i . Furthermore, for an erasure probability determined by $\phi(t)$ at a specified time t (13), the probability that the computation corresponding to the i^{th} block is sampled with replacement through the erasure channels at time t is

$$\Pr[\text{sample } \tilde{A}_i \text{ through the channels}] = 1 - \phi(t)^{\rho_i(t)} \quad (17)$$

for some $\rho_i(t) \in \mathbb{R}_{>0}$,⁴ and the network emulates the sampling distribution $\Pi_{\{K\}}$ exactly when

$$\Pi_i = 1 - \phi(t)^{\rho_i(t)} \quad (18)$$

for all i .

The replications which take place can be interpreted as the task allocation through a directed bipartite graph $G = (\mathcal{L}, \mathcal{R}, \mathcal{E})$, where \mathcal{L} and \mathcal{R} correspond to the K encoded partitions $\tilde{A}_{\{K\}}$ and m servers respectively, where $\deg(x_i) = r_i$ for all $x_i \in \mathcal{L}$ and $\deg(y_j) = 1$ for all $y_j \in \mathcal{R}$, with $\{x_i, y_j\} \in \mathcal{E}$ only if the j^{th} server W_j is assigned \tilde{A}_i . This allocation is done only once, before the iterative procedure begins.

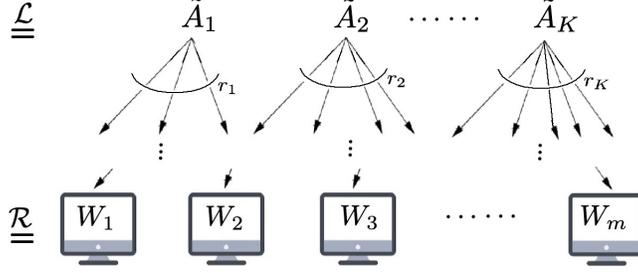


Fig. 3: Depiction of an expansion network as a bipartite graph, for $m = \sum_{l=1}^K r_l$.

Our goal is to determine $r_{\{K\}}$, which minimize the error in the emulated distribution $\bar{\Pi}_{\{K\}}$. Under the assumption that we have an integer number of replicas per block, from (17) and (18) we deduce that $\Pi_i \approx 1 - \phi(t)^{r_i}$ for $r_i \in \mathbb{Z}_+$, which lead to the minimization problem

$$\arg \min_{r_{\{K\}} \subseteq \mathbb{Z}_+} \left\{ \Delta_{\Pi, \bar{\Pi}} := \frac{1}{K} \sum_{i=1}^K |\Pi_i - (1 - \phi(t)^{r_i})| \right\} = \arg \left\{ \sum_{i=1}^K \min_{r_i \in \mathbb{Z}_+} \left\{ |\Pi_i - (1 - \phi(t)^{r_i})| \right\} \right\}. \quad (19)$$

Further note that $\Delta_{\Pi, \bar{\Pi}} \equiv d_{\Pi, \bar{\Pi}}$, where $\tilde{\Pi}_i = 1 - \phi(t)^{r_i}$ for each $i \in \mathbb{N}_K$. For our proposed distribution $\bar{\Pi}_{\{K\}}$, we may have $d_{\Pi, \bar{\Pi}} \neq \Delta_{\Pi, \bar{\Pi}}$. By combining (16) and (17), we then solve for the approximate replications $\hat{r}_{\{K\}}$ at time t :

$$\bar{\Pi}_i \approx \Pi_i = 1 - \phi(t)^{\rho_i(t)} \quad \implies \quad \hat{r}_i = \left\lfloor \frac{\log(1 - \Pi_i)}{\log(\phi(t))} \right\rfloor = \lfloor \rho_i(t) \rfloor \quad (20)$$

which result in the induced distribution $\bar{\Pi}_i = \hat{r}_i / \hat{R}$, for $\hat{R} := \sum_{l=1}^K \hat{r}_l$. In our context, we also require that $\hat{R} \approx m$.

Ideally, the above procedure would result in replication numbers $\hat{r}_{\{K\}}$ for which $\hat{R} = m$. This though is unlikely to occur, as $\Pi_{\{K\}}$ and R are determined by the data, and m is the fixed number of servers. To address this issue we redefine $\hat{r}_{\{K\}}$ to $\tilde{r}_{\{K\}}$ by $\tilde{r}_i = \hat{r}_i \pm \alpha_i$ for α_i small integers such that $\sum_{l=1}^K \tilde{r}_l = m$, and $\sum_{l=1}^K |\Pi_l - \tilde{r}_l/m|$ is minimal. If $m \gg \hat{R}$ for a large enough τ , we can set the number of replicas to be $\tilde{r}_i \approx \lfloor m/\hat{R} \rfloor \cdot \hat{r}_i$. Furthermore, the block size τ can be selected such that \hat{R} is approximately equal to the system's parameter m . We address the fact that \hat{R} is only approximately equal to m in Subsection III-C.

Lemma 1. *The approximation $\hat{r}_{\{K\}}$ according to (20) of the minimization problem (19) at time t , satisfies*

$$\Delta_{\Pi, \bar{\Pi}} \leq \left(1 - \sqrt{\phi(t)}\right) \cdot \left(\sum_{l=1}^K \phi(t)^{\min_{i \in \mathbb{N}_K} \{\hat{r}_i, \rho_i(t)\}} \right).$$

Proof. We break the proof into the cases where we round $\rho_i(t)$ to both the closest integers above and below. In either case, we know that $(\rho_i(t) - \hat{r}_i(t)) \in [-1/2, 1/2]$, for each $i \in \mathbb{N}_K$. Denote the respective individual summands of $\Delta_{\Pi, \bar{\Pi}}$ by Δ_i . In the case where $r_i = \lfloor \rho_i(t) \rfloor$, we have $\rho_i(t) = \hat{r}_i + \eta$ for $\eta \in [0, 1/2]$, hence

$$\begin{aligned} \Delta_i &= \left| (1 - \phi(t)^{\rho_i(t)}) - (1 - \phi(t)^{\hat{r}_i}) \right| \\ &= \left| \phi(t)^{\hat{r}_i} - \phi(t)^{\rho_i(t)} \right| \\ &= \left| \phi(t)^{\hat{r}_i} - \phi(t)^{\hat{r}_i + \eta} \right| \\ &= \left| \phi(t)^{\hat{r}_i} \cdot (1 - \phi(t)^\eta) \right| \\ &\leq \left| \phi(t)^{\hat{r}_i} \cdot (1 - \phi(t)^{1/2}) \right| \\ &= \phi(t)^{\hat{r}_i} \cdot \left(1 - \sqrt{\phi(t)}\right). \end{aligned}$$

⁴To be realizable, through replications, we need $\rho_i(t) \in \mathbb{Z}_+$.

Similarly, in the case where $r_i = \lceil \rho_i(t) \rceil$ we have $\hat{r}_i = \rho_i(t) + \eta$ for $\eta \in [0, 1/2]$, and

$$\Delta_i \leq \phi(t)^{\rho_i(t)} \cdot \left(1 - \sqrt{\phi(t)}\right).$$

Considering all summands, it follows that

$$\Delta_{\Pi, \bar{\Pi}} = \sum_{l=1}^K \Delta_l \leq \sum_{l=1}^K \left(\phi(t)^{\min_{i \in \mathbb{N}_K} \{\hat{r}_i, \rho_i(t)\}} \cdot \left(1 - \sqrt{\phi(t)}\right) \right).$$

□

We note that all terms involved in the upper bound of Lemma 1 are positive and strictly less than one. Furthermore, for a larger t we have a smaller $\phi(t)$, while for a smaller t we have a smaller \hat{r}_i for each i . This bound further corroborates the importance of the hyperparameter t and the distribution $F(t)$, in designing expansion networks.

The replication of blocks which takes place, can be described through a corresponding $R\tau \times K\tau$ binary *expansion matrix*:

$$\tilde{\mathbf{E}} = \mathbf{E} \otimes \mathbf{I}_\tau = \begin{pmatrix} \mathbf{1}_{r_1 \times 1} & & & \\ & \mathbf{1}_{r_2 \times 1} & & \\ & & \ddots & \\ & & & \mathbf{1}_{r_K \times 1} \end{pmatrix} \otimes \mathbf{I}_\tau \quad (21)$$

where $\mathbf{E} \in \{0, 1\}^{R \times K}$ is the adjacency matrix of the bipartite graph G portrayed in Figure 3 (up to a permutation of the rows/server indices). It follows that $(\tilde{\mathbf{E}} \cdot \mathbf{A}, \tilde{\mathbf{E}} \cdot \mathbf{b})$ are comprised of replicated blocks of the partitioning in (1), with replications according to $r_{\{K\}}$.

For the proposed networks, the multiplicative misestimation factor in Theorem 1 is $\beta_{\bar{\Pi}} = \min_{i \in \mathbb{N}_K} \{\Pi_i / \bar{\Pi}_i\} \leq 1$. In the case where $\tilde{R} = \sum_{l=1}^K \tilde{r}_l > m$ and $\tilde{\Pi}_i := \tilde{r}_i / \tilde{R}$, Algorithm 2 takes $\tilde{r}_{\{K\}}$ as an input and determines $r_{\{K\}}$ such that $R = \sum_{l=1}^K r_l = m$. The updated distribution $\bar{\Pi}_{\{K\}}$ where $\bar{\Pi}_i = r_i / R$ for each i , also has a more accurate misestimation factor, i.e., $\beta_{\bar{\Pi}} > \beta_{\tilde{\Pi}}$. To establish sampling guarantees in relation to $d_{\Pi, \bar{\Pi}}$, one would need to invoke an additive approximation error to the scores, i.e., $\tilde{\Pi}_i \leq \Pi_i + \epsilon$ for all i where $\epsilon \geq 0$ is a small constant [80], [99]. In our distributed networks, the additive error would be $\epsilon_{\bar{\Pi}} = \max_{i \in \mathbb{N}_K} \{|\Pi_i - \bar{\Pi}_i|\}$.

C. Optimally Induced Distributions

Recall that (19) is a surrogate to the integer program

$$r_{\{K\}}^* = \arg \min_{\substack{r_1, \dots, r_K \in \mathbb{Z}_+ \\ R = \sum_{l=1}^K r_l}} \left\{ d_{\Pi, \Pi^*} = \frac{1}{K} \sum_{i=1}^K |\Pi_i - \overbrace{r_i/R}^{\Pi_i^*}| \right\} \quad (22)$$

where R approximates the total number of servers. The solution $r_{\{K\}}^*$ approximates the block leverage score distribution $\Pi_{\{K\}}$ through expansion networks. Similar to $\Delta_{\Pi, \bar{\Pi}}$ from (19), the distortion metric d_{Π, Π^*} is a measure of closeness between the distributions $\Pi_{\{K\}}$ and $\bar{\Pi}_{\{K\}}$, under the network imposed constraints. In the case where $\Pi_{\{K\}} \not\subseteq (0, 1) \setminus \mathbb{Q}_+$, i.e., $\Pi_{\{K\}}$ are not necessarily all rational, the integer constraints of the physical network may make $\{r_i^*/R\}_{i=1}^K$ differ from $\Pi_{\{K\}}$. The integer program (22) cannot be solved exactly when $\Pi_{\{K\}} \not\subseteq (0, 1) \setminus \mathbb{Q}_+$, as we can always get finer approximations, e.g., through a continued fraction approximation. This is specific to the ending time T , which is not made explicit in (22), when considering erasures over the communication channels according to (13), which T we do not include in (22), to simplify notation. Furthermore, (22) can also be considered for centralized distributed settings which differ from the system model proposed in [32]. We note that solvers of (22) exist when R is fixed and we remove the constraint $R = \sum_{l=1}^K r_l$. The proof of Corollary 2 below is a constructive solution of (22) when $\Pi_{\{K\}} \subseteq [0, 1] \cap \mathbb{Q}_+$, in which case $\{r_i^*/R\}_{i=1}^K$ can be made exactly equal to $\Pi_{\{K\}}$, a condition we call *perfect emulation*.

Proposition 1. *A perfect emulation occurs when $d_{\Pi, \Pi^*} = 0$. This is possible if and only if $\Pi_i \in [0, 1] \cap \mathbb{Q}_+$ for all i and the denominators of $\Pi_{\{K\}}$ in reduced form are factors of R , i.e., $R \cdot \Pi_i \in \mathbb{Z}_+$.*

Proof. If $d_{\Pi, \Pi^*} = 0$, then $\Pi_i = \Pi_i^*$ for all $i \in \mathbb{N}_K$, thus $\Pi_{\{K\}}$ and $\Pi_{\{K\}}^*$ are the same sampling distributions. For the reverse direction, assume that for all i we have $\Pi_i = a_i/b_i$ for coprime integers $a_i, b_i \in \mathbb{Z}_+$, and that $R = \mu_i b_i$ for some $\mu_i \in \mathbb{Z}_+$; thus $R \cdot \Pi_i = \mu_i a_i \in \mathbb{Z}_+$. Let $r_i = \mu_i a_i$. It follows that $\Pi_i^* = \frac{r_i}{R} = \frac{\mu_i a_i}{\mu_i b_i} = \frac{a_i}{b_i} = \Pi_i$ for all i , hence $d_{\Pi, \Pi^*} = 0$. Now, assume for a contradiction that there is a $j \in \mathbb{N}_K$ for which $\Pi_j \in (0, 1) \setminus \mathbb{Q}_+$. Then, by definition, Π_j cannot be expressed as a fraction r_j/R for $r_j, R \in \mathbb{Z}_+$, thus $d_{\Pi, \Pi^*} \geq |\Pi_j - \Pi_j^*| > 0$. □

Corollary 2. *When $\Pi_{\{K\}} \subseteq [0, 1] \cap \mathbb{Q}_+$, we can solve (22), so that $d_{\Pi, \Pi^*} = 0$.*

Proof. From Proposition 1, the smallest R in order to attain $r_{\{K\}}$ for which $d_{\Pi, \Pi^*} = 0$ when considering $\Pi_i = a_i/b_i$ in reduced form, is the least common multiple $R = \text{lcm}(b_1, \dots, b_K)$. For each $i \in \mathbb{N}_K$, we then have $R = \mu_i b_i$ for $\mu_i \in \mathbb{Z}_+$, and $r_i = \mu_i a_i$. Hence $\Pi_i^* = \frac{r_i}{R} = \frac{\mu_i a_i}{\mu_i b_i} = \frac{a_i}{b_i} = \Pi_i$, for which $d_{\Pi, \Pi^*} = 0$. \square

Lemma 2. *If for a set of integers $\tilde{r}_{\{K\}}$ we have $\tilde{R} = \sum_{l=1}^K \tilde{r}_l$, $m = \tilde{R}$, and $\tilde{\Pi}_i = \tilde{r}_i/\tilde{R}$ for all $i \in \mathbb{N}_K$, then:*

$$d_{\Pi, \tilde{\Pi}} \geq \frac{1}{m} \cdot \min_{i \in \mathbb{N}_K} \{ \lfloor |m \cdot \Pi_i - \tilde{r}_i| \rfloor \}$$

and

$$d_{\Pi, \tilde{\Pi}} \leq \frac{1}{m} \cdot \max_{i \in \mathbb{N}_K} \{ \lceil |m \cdot \Pi_i - \tilde{r}_i| \rceil \}. \quad (23)$$

Proof. Let $\tilde{d}_i = |\Pi_i - \tilde{\Pi}_i| = |\Pi_i - \tilde{r}_i/m|$ for each i , hence

$$\tilde{r}_L := \frac{1}{m} \cdot \min_{i \in \mathbb{N}_K} \{ \lfloor |m \cdot \Pi_i - \tilde{r}_i| \rfloor \} \leq \tilde{d}_i$$

and

$$\tilde{r}_U := \frac{1}{m} \cdot \max_{i \in \mathbb{N}_K} \{ \lceil |m \cdot \Pi_i - \tilde{r}_i| \rceil \} \geq \tilde{d}_i$$

for all $i \in \mathbb{N}_K$. By rescaling the sum over all $\tilde{d}_{\{K\}}$ by $1/K$, we get

$$\tilde{r}_L = \frac{K \cdot \tilde{r}_L}{K} \leq \frac{1}{K} \cdot \sum_{i=1}^K \tilde{d}_i \leq \frac{K \cdot \tilde{r}_U}{K} = \tilde{r}_U$$

which completes the proof. \square

Next, we give a simple approximation to (22), for when we do not consider the erasure channel characterization through (18); nor an ending time T . Given $\Pi_{\{K\}}$, the replication numbers are $\tilde{r}_i = \lfloor \Pi_i/\Pi_1 \rfloor$ and $\tilde{R} = \sum_{l=1}^K \tilde{r}_l$. Further note that for more accurate approximations, we can select an integer $\nu > 1$ and take $\tilde{r}_i = \lfloor \nu \cdot \Pi_i/\Pi_1 \rfloor$. From the proof of Corollary 2, it follows that if $\Pi_{\{K\}} \subsetneq [0, 1] \cap \mathbb{Q}_+$ and $\nu = \mu_1 a_1$, we get $\tilde{r}_i = \Pi_i/\Pi_1 = \mu_i a_i \in \mathbb{Z}_+$, which solves (18). The drawback of designing an expansion network with this solution, is that as ν increases; \tilde{R} also increases.

We drop the constraint $R = \sum_{l=1}^K r_l$ of (22) and the assumption that $m = R$, by proposing a procedure in Algorithm 2 for determining $r_{\{K\}}$ from a given set $\tilde{r}_{\{K\}}$ (e.g., those proposed in (20)) to get the induced distribution $\{\tilde{\Pi}_i = r_i/m\}_{i=1}^K$, where $m = \sum_{l=1}^K r_l$. In Algorithm 2, $\chi = 1$ and $\chi = 0$ indicate whether $\tilde{R} \geq m$ or $\tilde{R} < m$ respectively.

Algorithm 2: Determine $r_{\{K\}}$ from $\tilde{r}_{\{K\}}$

Input: $m, \Pi_{\{K\}}, \tilde{r}_{\{K\}}, \tilde{R} = \sum_{i=1}^K \tilde{r}_i$

Output: replication numbers $r_{\{K\}}$

Initialize: $\tilde{d}_{\{K\}} = \left\{ \tilde{d}_i := \Pi_i - \tilde{r}_i/m \right\}_{i=1}^K, r_{\{K\}} = \tilde{r}_{\{K\}}, R = \tilde{R}, \chi = 1, \tilde{j} = 0$

if $R < m$ **then**

$\chi \leftarrow 0$

$\tilde{d}_{\{K\}} \leftarrow \left\{ -\tilde{d}_i \right\}_{i=1}^K$

$\triangleright \chi$ indicates: $\tilde{R} \geq m$ or $\tilde{R} < m$

end

while $R \neq m$ **do**

$j \leftarrow \arg \min_{i \in \mathbb{N}_K} \{ \tilde{d}_{\{K\}} \}$ (\blacklozenge)

if $(-1)^{x+1} \cdot (\Pi_j - \frac{1}{m} (r_j + (-1)^x)) > 0$ **and** $\tilde{j} \equiv j$ **then**

$\tilde{d}_j \leftarrow 1$

end

else

$r_j \leftarrow r_j + (-1)^x$

$R \leftarrow R + (-1)^x$

$\tilde{d}_j \leftarrow (-1)^{x+1} \cdot (\Pi_j - r_j/m)$

end

$\tilde{j} \leftarrow j$

end

Remark 1. *The objective of Algorithm 2 is to reduce the upper bound (23) when $m < \tilde{R}$, while guaranteeing that $\sum_{l=1}^K r_l = m$. In practice, the more concerning and limiting case is when $m < \tilde{R}$. The bottleneck of Algorithm 2 is retrieving the index j in the **while** loop, which takes $O(K)$ time. In order to reduce the number of instances we solve (\blacklozenge), we ensure that we only reduce*

the replica numbers $\hat{r}_{\{K\}}$ in the case where $R > m$; and increase them when $R < m$, by the inner **if** statement. Moreover, this is carried out once before sharing the replicated blocks. The more practical and realistic case is when $R > m$, as we can get a closer approximation with a greater R , and $\text{lcm}(b_1, \dots, b_K)$ will likely be large when $\Pi_{\{K\}} \subsetneq [0, 1] \cap \mathbb{Q}_+$. We note that the integers $\hat{r}_{\{K\}}$ of (20) and their sum \hat{R} are a byproduct of the prespecified ending time $t \leftarrow T$, the mother runtime distribution $F(t)$, and the block size τ , which can be selected so that $\hat{R} > m$.

Proposition 2. Assume we are given $\tilde{r}_{\{K\}}$ (not necessarily according to (20)), for which $m < \tilde{R} = \sum_{l=1}^K \tilde{r}_l$. Denote by $\tilde{\Pi}_{\{K\}}$ the corresponding sampling distribution $\left\{ \tilde{\Pi}_i = \tilde{r}_i / \tilde{R} \right\}_{i=1}^K$, for which $d_{\Pi, \tilde{\Pi}} \leq \tilde{\epsilon}$. Then, the output $r_{\{K\}}$ of Algorithm 2 produces an induced distribution $\{\bar{\Pi}_i = r_i / m\}_{i=1}^K$ which satisfies $d_{\Pi, \bar{\Pi}} < d_{\Pi, \tilde{\Pi}} \leq \tilde{\epsilon}$.

Proof. The case where $\tilde{R} > m$, corresponds to $\chi = 1$, in which case Algorithm 2 returns $r_{\{K\}}$ for which $r_i \leq \tilde{r}_i$ for all $i \in \mathbb{N}_K$. In this case, the optimization problem (\blacklozenge) assigns to j a partition index for which $\Pi_j < r_j / m$. Since $\tilde{d}_j = \Pi_j - r_j / m < 0$, we have $\Pi_j < r_j / m$. The **if** statement guarantees that we did not produce an r_j for which $\Pi_j > r_j / m$, when we previously had $\Pi_j < \tilde{r}_j / \tilde{R}$ (or $\Pi_j < r_j / R$ after a reassignment of \tilde{r}_j). Along with the fact that $\frac{r_j - 1}{R - 1} < \frac{r_j}{R}$, it follows that for the updated difference d'_j :

$$\begin{aligned} |d'_j| &= \left| \Pi_j - \frac{r_j - 1}{R - 1} \right| \\ &= \frac{r_j - 1}{R - 1} - \Pi_j \\ &< \frac{r_j}{R} - \Pi_j \\ &= \left| \Pi_j - \frac{r_j}{R} \right| \\ &= |\tilde{d}_j| \end{aligned}$$

i.e., at each iteration of the **else** statement, we decrease the distortion metric, thus $d_{\Pi, \bar{\Pi}} < d_{\Pi, \tilde{\Pi}} \leq \tilde{\epsilon}$.

The **else** statement is carried out $\tilde{R} - m$ times, producing $r_{\{K\}}$ for which $\sum_{l=1}^K r_l = \tilde{R} - (\tilde{R} - m) = m$, hence the normalizing integer for the distribution $\bar{\Pi}_{\{K\}}$ is $R = m$. \square

By incorporating Algorithm 2 into the proposed GC method, our procedure is more versatile in comparison to other methods when it comes to the code design parameters m and q . Other methods require a decoding recovery threshold of q responses in order to apply their decoding step, and the pre-selected parameters need to satisfy $(m - q + 1) \mid m$ [41], [43], [100]. In exact GC schemes, the average replication number of sub-datasets depends on the number of stragglers it can tolerate, as such codes are maximum distance separable. Specifically, to tolerate $(m - q)$ stragglers, each data block needs to be replicated at least $(m - q + 1)$ times across the servers. In contrast, our method can be deployed when this condition is not met, and does not require a recovery threshold.

D. GC with Leverage Score Sampling

Next, we derive the server computations of our GC scheme, so that the central server retrieves the gradient of the modified objective function $L_S(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]}) = \|\tilde{\mathbf{S}}_{[s]}(\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b})\|_2^2$:

$$\hat{g}^{[s]} := \nabla_{\mathbf{x}} L_S(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]}) \quad (24)$$

for $\tilde{\mathbf{S}}_{[s]}$ the sketching matrix according to Algorithm 1 and $\mathbf{x}^{[s]}$ our current parameters update; both at iteration s , to iteratively approximate the solution \mathbf{x}^* to (2). Furthermore, we show that with a diminishing step-size, our updates $\mathbf{x}^{[s]}$ converge in expectation to the optimal solution \mathbf{x}^* , and the expected regret of the least squares objective converges to zero. First though, in Algorithm 3 we provide the pseudocode of how SD is performed with iterative sketching. This corresponds to the ‘Iterative Procedure at Iteration s ’ in Algorithm 4.

The blocks of the proposed leverage score sampling procedure are those of the encoded data $\tilde{\mathbf{A}} := \mathbf{G} \cdot \mathbf{A}$ and $\tilde{\mathbf{b}} := \mathbf{G} \cdot \mathbf{b}$, for

$$\mathbf{G} = \text{diag}\left(\left\{1/\sqrt{q\bar{\Pi}_i}\right\}_{i=1}^K\right) \otimes \mathbf{I}_\tau \in \mathbb{R}_{\geq 0}^{N \times N}. \quad (25)$$

Specifically, the encoding carried out by the central server corresponds to the rescaling through \mathbf{G} . We partition both $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$ across their rows analogous to (1):

$$\tilde{\mathbf{A}} = \mathbf{G}\mathbf{A} = \left[\tilde{A}_1^\top \cdots \tilde{A}_K^\top \right]^\top \quad (26)$$

$$\tilde{\mathbf{b}} = \mathbf{G}\mathbf{b} = \left[\tilde{b}_1^\top \cdots \tilde{b}_K^\top \right]^\top \quad (27)$$

Algorithm 3: Steepest Descent with Iterative Sketching**Input:** $\mathbf{A} \in \mathbb{R}^{N \times d}$, $\mathbf{b} \in \mathbb{R}^N$, $\mathbf{x}^{[0]} \in \mathbb{R}^d$, $\tau, q = \frac{r}{\tau} > \frac{d}{\tau}$ **Output:** $\hat{\mathbf{x}} \in \mathbb{R}^d$ s.t. $\mathbf{A}\hat{\mathbf{x}} \approx \mathbf{b}$ **for** $s = 0, 1, 2, \dots$ **do**

- run Algorithm 1, to obtain $\tilde{\mathbf{S}}_{[s]} \in \mathbb{R}^{r \times N}$
 - ▷ approximate $\tilde{\Pi}_{\{K\}}$ only at $s = 0$
- $\hat{\mathbf{A}}_{[s]} \leftarrow \tilde{\mathbf{S}}_{[s]} \cdot \mathbf{A}$
- $\hat{\mathbf{b}}_{[s]} \leftarrow \tilde{\mathbf{S}}_{[s]} \cdot \mathbf{b}$
- $\hat{g}_{[s]} \leftarrow \nabla_{\mathbf{x}} L_{ls}(\hat{\mathbf{A}}_{[s]}, \hat{\mathbf{b}}_{[s]}; \mathbf{x}^{[s]}) \equiv \nabla_{\mathbf{x}} L_{\mathbf{S}}(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]})$
- **Select:** step-size $\xi_s > 0$
- **Update:** $\mathbf{x}^{[s+1]} \leftarrow \mathbf{x}^{[s]} - \xi_s \cdot \hat{g}_{[s]}$

end

where $\tilde{A}_i \in \mathbb{R}^{\tau \times d}$ and $\tilde{b}_i \in \mathbb{R}^\tau$ for all $i \in \mathbb{N}_K$. Furthermore, all the data across the expansion network after the scalar encoding, is contained in aggregated expanded and encoded matrix-vector pairs:

$$(\Psi, \vec{\psi}) := (\tilde{\mathbf{E}} \cdot \tilde{\mathbf{A}}, \tilde{\mathbf{E}} \cdot \tilde{\mathbf{b}}) \in \mathbb{R}^{R\tau \times d} \times \mathbb{R}^{R\tau}$$

as illustrated in Figure 2. For the encoded objective function $L_{\mathbf{G}}(\mathbf{x}) := \|\mathbf{G}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2$, we have:

- (i) $L_{\mathbf{G}}(\mathbf{x}) = \mathbf{x}^\top \left(\sum_{i=1}^K \tilde{A}_i^\top \tilde{A}_i \right) \mathbf{x} + \left(\sum_{i=1}^K (\tilde{b}_i^\top - 2\mathbf{x}^\top \tilde{A}_i^\top) \tilde{b}_i \right)$
- (ii) $\nabla_{\mathbf{x}} L_{\mathbf{G}}(\mathbf{x}) = 2 \sum_{i=1}^K \tilde{A}_i^\top (\tilde{A}_i \mathbf{x} - \tilde{b}_i)$
- (iii) $\nabla_{\mathbf{x}} L_{\mathbf{G}}(\mathbf{x}_{\mathbf{G}}^*) = 0 \implies \mathbf{x}_{\mathbf{G}}^* = \left(\sum_{i=1}^K \tilde{A}_i^\top \tilde{A}_i \right)^{-1} \cdot \left(\sum_{i=1}^K \tilde{A}_i^\top \tilde{b}_i \right)$.

We make use of (ii) to approximate the gradient distributively. Each server is provided with a partition $\tilde{\mathcal{D}}_j = (\tilde{A}_j, \tilde{b}_j)$, and computes the respective summand of the gradient from (ii), which is the encoded partial gradient on $\mathcal{D}_j = (\mathbf{A}_j, \mathbf{b}_j)$:

$$\hat{g}_i^{[s]} := \nabla_{\mathbf{x}} L_{ls}(\tilde{\mathcal{D}}_i; \mathbf{x}^{[s]}) = \nabla_{\mathbf{x}} L_{ls} \left(1/\sqrt{q\tilde{\Pi}_i} \cdot \mathbf{A}_i, 1/\sqrt{q\tilde{\Pi}_i} \cdot \mathbf{b}_i; \mathbf{x}^{[s]} \right) = \frac{1}{q\tilde{\Pi}_i} \cdot g_i^{[s]}. \quad (28)$$

Once a server computes its assigned partial gradient, it sends it back to the central server. When the central server receives q responses, it sums them in order to obtain the approximate gradient $\hat{g}^{[s]}$.

Denote the index multiset corresponding to the encoded pairs $(\tilde{A}_j, \tilde{b}_j)$ of the received computations at iteration s with $\mathcal{S}^{[s]}$, for which $|\mathcal{S}^{[s]}| = q$. The parameters vector update computed locally at the central server once sufficiently many servers respond is $\mathbf{x}^{[s+1]} \leftarrow \mathbf{x}^{[s]} - \xi_s \cdot \hat{g}^{[s]}$, where

$$\hat{g}^{[s]} = \sum_{i \in \mathcal{S}^{[s]}} \nabla_{\mathbf{x}} L_{ls}(\tilde{\mathcal{D}}_i; \mathbf{x}^{[s]}) = 2 \sum_{i \in \mathcal{S}^{[s]}} \tilde{A}_i^\top (\tilde{A}_i \mathbf{x}^{[s]} - \tilde{b}_i) \quad (29)$$

and ξ_s is an appropriate step-size. In the case where q is not determined a priori or varies at each iteration, the scaling corresponding to $1/\sqrt{q}$ in the encoding through \mathbf{G} could be done by the central server, after that iteration's computations are aggregated. We consider the case where q is the same for all iterations.

Next, we present the performance guarantees of the proposed GC method.

Theorem 2. *The proposed iterative block leverage score sketching method presented in Algorithm 3, results in a SSD procedure for minimizing $L_{ls}(\Psi, \vec{\psi}; \mathbf{x})$. Furthermore, at each iteration it produces an unbiased estimator of the least squares objective gradient (3), i.e., $\mathbb{E}[\hat{g}^{[s]}] = g^{[s]}$.*

Proof. The computations of the q fastest servers indexed by $\mathcal{I}^{[s]}$ (which corresponds to $\tilde{\Omega}^{[s]}$), are added to produce $\hat{g}^{[s]}$, and the sampling of Algorithm 1 is according to the block leverage score distribution $\tilde{\Pi}_{\{K\}}$. The application of $\tilde{\Omega}^{[s]}$ (in Algorithm 1) has a direct correspondence to the index set $\mathcal{I}^{[s]}$ of the $q(T)$ non-stragglers at iteration s , which can be viewed as drawing $\mathcal{I}^{[s]}$ uniformly at random from $\{\mathcal{I} \subseteq \mathbb{N}_m : |\mathcal{I}| = q(T)\}$. It follows that each $\mathcal{I}^{[s]}$ has equal chance of occurring, which is precisely the stochastic step of SSD, i.e., each group of q encoded block pairs has an equal chance of being selected.

Since the servers are homogeneous and respond independently of each other, it follows that at iteration s , each \hat{g}_i is received with probability $\bar{\Pi}_i$. Therefore

$$\begin{aligned}
\mathbb{E}[\hat{g}^{[s]}] &= \mathbb{E}\left[\sum_{i \in \mathcal{I}^{[s]}} \hat{g}_i^{[s]}\right] \\
&= \sum_{i \in \mathcal{I}^{[s]}} \mathbb{E}[\hat{g}_i^{[s]}] \\
&= \sum_{i \in \mathcal{I}^{[s]}} \sum_{j=1}^K \bar{\Pi}_j \cdot \hat{g}_j^{[s]} \\
&= q \cdot \sum_{j=1}^K \bar{\Pi}_j \cdot \hat{g}_j^{[s]} \\
&\stackrel{\text{b}}{=} q \cdot \sum_{j=1}^K \bar{\Pi}_j \cdot \frac{1}{q\bar{\Pi}_j} \cdot g_j^{[s]} \\
&= \sum_{j=1}^K g_j^{[s]} \\
&= g^{[s]}
\end{aligned}$$

where in b we invoked (28). \square

Lemma 3. *The optimal solution of the modified least squares problem $L_{ls}(\Psi, \vec{\psi}; \mathbf{x})$, is equal to the optimal solution \mathbf{x}^* of the least squares problem (2).*

Proof. Note that the modified objective function $L_{ls}(\Psi, \vec{\psi}; \mathbf{x})$ is $\|\tilde{\mathbf{E}}\mathbf{G} \cdot (\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2$. Denote its optimal solution by $x^* \in \mathbb{R}^d$. Further note that $\tilde{\mathbf{E}}$ is comprised of $\tau \times \tau$ identity matrices in such a way that it is full-rank, and \mathbf{G} corresponds to a rescaling of these \mathbf{I}_τ matrices, thus $\tilde{\mathbf{E}}\mathbf{G}$ is also full-rank. It then follows that

$$x^* = ((\mathbf{E}\mathbf{G}) \cdot \mathbf{A})^\dagger \cdot ((\mathbf{E}\mathbf{G}) \cdot \mathbf{b}) = \mathbf{A}^\dagger \cdot ((\mathbf{E}\mathbf{G})^\dagger \cdot (\mathbf{E}\mathbf{G})) \cdot \mathbf{b} = \mathbf{A}^\dagger \cdot \mathbf{I}_N \cdot \mathbf{b} = \mathbf{x}^*.$$

\square

Corollary 3. *The expected regret of the least squares objective resulting from Algorithm 3 with a diminishing step-size ξ_s , converges to zero at a rate of $O(1/\sqrt{s} + r/s)$, i.e.,*

$$\mathbb{E}[L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]}) - L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x}^*)] \leq O(1/\sqrt{s} + r/s).$$

Proof. This follows by directly applying Theorem 2 and Lemma 3 to [101, Theorem 6.3]. \square

Theorem 3. *Suppose that there exists a constant C so that $\|\nabla_{\mathbf{x}} L_{ls}(\mathbf{A}_{(i)}, \mathbf{b}_{(i)}; \mathbf{x})\|_2^2 \leq C$ for all $\mathbf{x} \in \mathbb{R}^d$ and $i \in \mathbb{N}_N$, and that we run our Algorithm 3 with step-size $\xi_s = 1/(\eta s)$ for a fixed $\eta > 0$. Then, the error after S iterations is bounded by*

$$\mathbb{E}[\|\mathbf{x}^{[S]} - \mathbf{x}^*\|] \leq \frac{4NrC^2}{S\eta^2q\Pi_1} = O(1/S).$$

Proof. Following a similar analysis to the proof of [53, Theorem 5], we have

$$\mathbb{E}\left[\|\nabla_{\mathbf{x}} L_{ls}(\mathbf{A}_{(i)}, \mathbf{b}_{(i)}; \mathbf{x})\|_2^2\right] \leq \mathbb{E}\left[\max_{\mathcal{S}^{[s]} \subseteq \mathbb{N}_K^q} \left\{ \sum_{i \in \mathcal{S}^{[s]}} \nabla_{\mathbf{x}} L_{ls}(\tilde{\mathbf{A}}_{(i)}, \tilde{\mathbf{b}}_{(i)}; \mathbf{x}) \right\}\right] \leq C^2 \cdot \frac{Nr}{q\Pi_1}.$$

By then applying [102, Lemma 1], it follows that

$$\mathbb{E}\left[\|\mathbf{x}^{[S]} - \mathbf{x}^*\|_2^2\right] \leq \frac{4NrC^2}{S\eta^2q\Pi_1} = O(1/S).$$

\square

The crucial aspect of our expansion network; incorporated in Theorem 2, which allowed us to use block leverage score sampling in Algorithm 1, is that uniform sampling of $L_{ls}(\Psi, \vec{\psi}; \mathbf{x}^{[s]})$ is $\beta_{\bar{\Pi}}$ -approximately equivalent to block sampling of $L_{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}; \mathbf{x}^{[s]})$ according to the block leverage scores of \mathbf{A} . Since the two objective functions are differentiable and additively separable, the resulting gradients are equal, under the assumption that we use the same $\mathbf{x}^{[s]}$ and sampled index set $\mathcal{S}^{[s]}$. Note that the index set of the sampled blocks from $L_{ls}(\Psi, \vec{\psi}; \mathbf{x}^{[s]})$, corresponds to an index multiset of the sampled blocks from $L_{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}; \mathbf{x}^{[s]})$, as in

the latter we are considering sampling with replacement. As previously mentioned, the main drawback is that in certain cases we need significantly more servers to accurately emulate $\Pi_{\{K\}}$.

The significance of Theorem 2, is that well-known established SD and SSD results directly apply to Algorithm 3 under the assumption that the approximate gradient is an unbiased estimator [103, Chapter 14]. Even though Algorithm 3 does not guarantee a descent step at every iteration, such stochastic descent methods are more common in practice when dealing with large datasets, as empirically they outperform regular SD. This is also validated in our experiments of Section IV.

Finally, we include a succinct description of the proposed approximate GC approach which utilizes block leverage score sampling, in the pseudocode of Algorithm 4, an illustration of which is also provided in Figure 2. Below, we also give an overview of how everything is tied together.

Recall that our GC objective is to distributively approximate the gradient $g^{[s]}$. We do so by computing the gradient $\hat{g}^{[s]}$ of the modified problem (12), where the sketching matrix $\mathbf{S} \leftarrow \tilde{\mathbf{S}}_{[s]}$ is the one proposed in Algorithm 1. In our distributed setting, the sketching matrix $\tilde{\mathbf{S}}_{[s]}$ is never explicitly computed, but we instead recover the gradient of the modified objective function (12), where $\mathbf{S} \leftarrow \tilde{\mathbf{S}}_{[s]}$ is induced by the proposed procedure. Specifically, to compute the sketch's gradient $\hat{g}^{[s]}$ we assume that the computational nodes are homogeneous, and that the central server expects a response from q servers after time T , and we replicate each encoded pair $(\tilde{A}_i, \tilde{b}_i)$ according to $r_{\{K\}}$ from Algorithm 2. The number of replications per task $r_{\{K\}}$ are determined such that uniformly sampling from the replicated multiset of the encoded block pairs

$$\bigoplus_{i=1}^K \left\{ \bigoplus_{l=1}^{r_i} \left\{ (\tilde{A}_l, \tilde{b}_l) \right\} \right\}$$

approximates leverage score sampling according to $\tilde{\Pi}_{\{K\}}$ from Algorithm 1. This requires that the cardinality in the above multiset union matches the total number of servers in the distributed network, i.e., $m = \sum_{j=1}^K r_j$. In order to recover an approximate solution to (2), the central server locally performs a SD update which is then communicated to every server, and this process is repeated until a prespecified termination criterion is met.

Algorithm 4: Approximate GC Approach Summary

Central Server:

- 1) partitions \mathbf{A}, \mathbf{b} according to (1), and decides q
- 2) estimates $\tilde{\Pi}_{\{K\}}$
- 3) determines $\tilde{r}_{\{K\}}$ from (20)
- 4) passes $m, \tilde{\Pi}_{\{K\}}, \tilde{r}_{\{K\}}, \tilde{R} = \sum_{l=1}^K \tilde{r}_l$ to Algorithm 2, to determine $r_{\{K\}}$, and sets $r_0 = 0$
- 5) applies the encodings according to (25), (26), (27)
- 6) delivers the encoded block $(\tilde{A}_i, \tilde{b}_i)$ to the servers indexed between $\left(1 + \sum_{j=0}^{i-1} r_j\right)$ and $\left(\sum_{j=0}^i r_j\right)$, for each $i \in \mathbb{N}_K$

The following iterative procedure corresponds to Algorithm 3

While the central server's termination criterion is not met, repeat:

Iterative Procedure at Iteration s :

- each server computes their corresponding partial gradient (28): $\hat{g}_i^{[s]} = \nabla_{\mathbf{x}} L_{l_s}(\tilde{A}_i, \tilde{b}_i; \mathbf{x}^{[s]})$
 - **Central Server:**
 - i) sums the q fastest responses indexed by $\mathcal{S}^{[s]}$, according to (29): $\hat{g}^{[s]} = \sum_{i \in \mathcal{S}^{[s]}} \hat{g}_i^{[s]}$
 $\triangleright \mathcal{S}^{[s]}$ determines $\tilde{\mathbf{S}}_{[s]}$ and $\hat{g}^{[s]} = \nabla_{\mathbf{x}} L_{\mathbf{S}}(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]})$
 - ii) performs the update: $\mathbf{x}^{[s+1]} \leftarrow \mathbf{x}^{[s]} - \xi_s \cdot \hat{g}^{[s]}$
 - iii) communicates $\mathbf{x}^{[s+1]}$ to all the servers
-

E. Synopsis

Here, we give a summary of the main theorems provided in the previous subsections. Firstly, as summarized in Subsection III-C, we mimic block leverage score sampling with replacement of (\mathbf{A}, \mathbf{b}) (from $L_{l_s}(\mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]})$) through uniform sampling, by approximately solving (19) through the implication of (20) (Lemma 1). This is done implicitly by communicating computations over erasure channels. Secondly, by Theorem 1 we know that the proposed block leverage score sketching matrices satisfy (10), where the approximate sampling distribution $\tilde{\Pi}_{\{K\}}$ is determined through the proposed expansion network associated with $\Pi_{\{K\}}$. Hence, at each iteration, we approach a solution $\hat{\mathbf{x}}^{[s]}$ of the induced sketched system

$$\tilde{\mathbf{S}}_{[s]} \left(\mathbf{A} \mathbf{x}^{[s]} \right) = \tilde{\mathbf{S}}_{[s]} \mathbf{b}$$

which $\mathbf{x}^{[s]}$ satisfies (11) with overwhelming probability. Thirdly, in Corollary 3 we showed that with a diminishing step-size ξ_s , the expected regret of the least squares objective converges to zero at a rate of $O(1/\sqrt{s} + r/s)$ [101], [104]. Lastly, by Theorem 3, our updates $\mathbf{x}^{[s]}$ converge to \mathbf{x}^* in expectation. A synopsis is given Figure 4.

$$\left\{ L_{\mathbf{S}}(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}) \text{ sol'ns} \right\} \xleftarrow{\text{Thm 1}} \left\{ \text{Solve } L_{l_s}(\Psi, \vec{\psi}; \mathbf{x}^{[s]}) \right\} \xrightarrow{\text{Thm 3}} \left\{ \text{With a diminishing } \xi_s: \right. \\ \left. \lim \mathbb{E}[\mathbf{x}^{[s]}] \rightarrow \mathbf{x}^* \right\}$$

Fig. 4: Synopsis of our main results.

F. Approximate GC from ℓ_2 -subspace embedding

In this subsection, we quantify the approximation error of the recovered gradient using Algorithm 4. In conventional GC, the objective is to construct an encoding matrix $\mathbf{G} \in \mathbb{R}^{m \times K}$ and decoding vectors $\mathbf{a}_{\mathcal{I}} \in \mathbb{R}^{1 \times q}$, such that $\mathbf{a}_{\mathcal{I}} \mathbf{G}_{(\mathcal{I})} = \vec{\mathbf{1}}$ for any set of non-straggling servers \mathcal{I} . It follows that the optimal coefficient decoding vector for a set \mathcal{I} of size q in approximate GC [54], [55], [58] is

$$\mathbf{a}_{\mathcal{I}}^* = \arg \min_{\mathbf{a} \in \mathbb{R}^{1 \times q}} \{ \|\mathbf{a} \mathbf{G}_{(\mathcal{I})} - \vec{\mathbf{1}}\|_2^2 \} \implies \mathbf{a}_{\mathcal{I}}^* = \vec{\mathbf{1}} \mathbf{G}_{(\mathcal{I})}^\dagger \quad (30)$$

for $\mathbf{G}_{(\mathcal{I})}^\dagger = (\mathbf{G}_{(\mathcal{I})}^\top \mathbf{G}_{(\mathcal{I})})^{-1} \mathbf{G}_{(\mathcal{I})}^\top \in \mathbb{R}^{K \times q}$.

Proposition 3. *The error in the approximated gradient $\hat{g}^{[s]}$ of an approximate optimal coefficient decoding linear regression GC scheme $(\mathbf{G}, \mathbf{a}_{\mathcal{I}}^*)$, satisfies*

$$\|g^{[s]} - \hat{g}^{[s]}\|_2 \leq 2\sqrt{K} \cdot \text{err}(\mathbf{G}_{(\mathcal{I})}) \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2 \quad (31)$$

for $\text{err}(\mathbf{G}_{(\mathcal{I})}) := \|\mathbf{I}_K - \mathbf{G}_{(\mathcal{I})}^\dagger \mathbf{G}_{(\mathcal{I})}\|_2$.

Proof. Consider the optimal coefficient decoding vector of an approximate GC scheme $\mathbf{a}_{\mathcal{I}}^*$ (30). In the case where $q \geq K$, it follows that $\mathbf{a}_{\mathcal{I}}^* = \vec{\mathbf{1}} \mathbf{G}_{(\mathcal{I})}^\dagger$.

Let $\mathbf{g}^{[s]}$ be the matrix comprised of the transposed exact partial gradients at iteration s , i.e.,

$$\mathbf{g}^{[s]} := \begin{pmatrix} g_1^{[s]} & g_2^{[s]} & \dots & g_K^{[s]} \end{pmatrix}^\top \in \mathbb{R}^{K \times d}.$$

Then, for a GC encoding-decoding pair $(\mathbf{G}, \mathbf{a}_{\mathcal{I}})$ satisfying $\mathbf{a}_{\mathcal{I}} \mathbf{G}_{(\mathcal{I})} = \vec{\mathbf{1}}$ for any \mathcal{I} , it follows that

$$\mathbf{a}_{\mathcal{I}} \left(\mathbf{G}_{(\mathcal{I})} \mathbf{g}^{[s]} \right) = \vec{\mathbf{1}} \mathbf{g}^{[s]} = \sum_{j=1}^K (g_j^{[s]})^\top = (g^{[s]})^\top.$$

Hence, the gradient can be recovered exactly. Considering an optimal approximate scheme $(\mathbf{G}, \mathbf{a}_{\mathcal{I}}^*)$ which recovers the gradient estimate $\hat{g}^{[s]} = (\mathbf{a}_{\mathcal{I}}^* \mathbf{G}_{(\mathcal{I})}) \mathbf{g}^{[s]}$, the error in the gradient approximation is:

$$\begin{aligned} \|g^{[s]} - \hat{g}^{[s]}\|_2 &= \left\| (\vec{\mathbf{1}} - \mathbf{a}_{\mathcal{I}}^* \mathbf{G}_{(\mathcal{I})}) \mathbf{g}^{[s]} \right\|_2 \\ &= \left\| \vec{\mathbf{1}} (\mathbf{I}_K - \mathbf{G}_{(\mathcal{I})}^\dagger \mathbf{G}_{(\mathcal{I})}) \mathbf{g}^{[s]} \right\|_2 \\ &\leq \|\vec{\mathbf{1}}\|_2 \cdot \left\| \mathbf{I}_K - \mathbf{G}_{(\mathcal{I})}^\dagger \mathbf{G}_{(\mathcal{I})} \right\|_2 \cdot \|\mathbf{g}^{[s]}\|_2 \\ &\stackrel{\mathcal{L}}{\leq} \sqrt{K} \cdot \left\| \mathbf{I}_K - \mathbf{G}_{(\mathcal{I})}^\dagger \mathbf{G}_{(\mathcal{I})} \right\|_2 \cdot \|g^{[s]}\|_2 \\ &\stackrel{\mathcal{S}}{\leq} 2\sqrt{K} \cdot \underbrace{\left\| \mathbf{I}_K - \mathbf{G}_{(\mathcal{I})}^\dagger \mathbf{G}_{(\mathcal{I})} \right\|_2}_{\text{err}(\mathbf{G}_{(\mathcal{I})})} \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2 \end{aligned}$$

where \mathcal{L} follows from the facts that $\|\mathbf{g}^{[s]}\|_2 \leq \|g^{[s]}\|_2$ and $\|\vec{\mathbf{1}}\|_2 = \sqrt{K}$, and \mathcal{S} from (3) and sub-multiplicativity of matrix norms. \square

In Theorem 4, we show the accuracy of the approximate gradient of iterative GC approaches based on sketching techniques that satisfy (with high probability) the ℓ_2 -subspace embedding property $\|\mathbf{I}_d - \mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U}\|_2 \leq \epsilon$ from (10). This then holds true for the GC method of Algorithm 4, by Theorem 1 and the fact that we distributively emulate block leverage score sampling.

Theorem 4. *Assume that the induced sketching matrix \mathbf{S} from a GC scheme satisfies $\|\mathbf{I}_d - \mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U}\|_2 \leq \epsilon$ (with high probability). Then, the updated approximate gradient estimate $\hat{g}^{[s]}$ at any iteration, satisfies (with high probability):*

$$\|g^{[s]} - \hat{g}^{[s]}\|_2 \leq 2\epsilon \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2.$$

Specifically, it satisfies (31) with $\text{err}(\mathbf{G}_{(\mathcal{I})}) = \epsilon/\sqrt{K}$.

Proof. Consider $\hat{g}^{[s]}$ the approximated gradient in Algorithm 3 for linear regression with gradient $g^{[s]}$ (3). It follows that

$$\begin{aligned} \|g^{[s]} - \hat{g}^{[s]}\|_2 &= \|2\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}) - 2\mathbf{A}^\top(\mathbf{S}^\top\mathbf{S})(\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b})\|_2 \\ &= 2\|\mathbf{A}^\top(\mathbf{I}_N - \mathbf{S}^\top\mathbf{S})(\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b})\|_2 \\ &\leq 2\|\mathbf{A}\|_2 \cdot \|\mathbf{I}_N - \mathbf{S}^\top\mathbf{S}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2 \\ &= 2\|\mathbf{A}\|_2 \cdot \|\mathbf{U}^\top(\mathbf{I}_N - \mathbf{S}^\top\mathbf{S})\mathbf{U}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2 \\ &= 2\|\mathbf{A}\|_2 \cdot \|\mathbf{I}_d - \mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2 \\ &\stackrel{\text{b}}{\leq} 2\epsilon \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b}\|_2 \end{aligned}$$

where in b we make use of the assumption that \mathbf{S} satisfies (10). Our approximate GC approach therefore (with high probability) satisfies (31), with $\text{err}(\mathbf{G}_{(T)}) = \epsilon/\sqrt{K}$. \square

IV. EXPERIMENTS

In this section, we validate our theoretical results through numerical experiments. The minimum benefit of Algorithm 1 occurs when $\Pi_{\{K\}}$ is close to uniform. For this reason, we construct dataset matrices whose resulting sampling distributions and block leverage scores are non-uniform.

In our first experiment, we compare our iterative sketching approach, implemented as in Algorithm 3, to other iterative sketching techniques and to regular SD (without sketching). For this experiment, we generated $\mathbf{A} \in \mathbb{R}^{2000 \times 40}$ following a t -distribution, and standard Gaussian noise was added to an arbitrary vector from $\text{im}(\mathbf{A})$ to define \mathbf{b} . We considered $K = 100$ blocks, thus $\tau = 20$. The effective dimension $N = 2000$ was reduced to $r = 1000$, i.e., $q = 50$. We compared the iterative approach with exact block leverage scores, i.e., $\beta = 1$, against analogous approaches using the block-SRHT [57] and Gaussian sketches, and regular SD.

For Figure 5 we ran 600 iterations on six different instances for each approach, and varied ξ for each experiment by logarithmic factors of $\xi^\times = 2/\sigma_{\max}(\mathbf{A})^2$. The average log residual errors $\log_{10}(\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2/\sqrt{N})$ are depicted in Figure 5, and reported in Table II. In Figure 6 we observe the convergence of the different approaches, in the case where $\xi \approx 0.42$. In this case, our method (block-lvg) outperforms the Gaussian sketching approach and regular SD. The fact that the performance of the block-SRHT is similar to Algorithm 1, reflects the result of Proposition 5 in Appendix C.

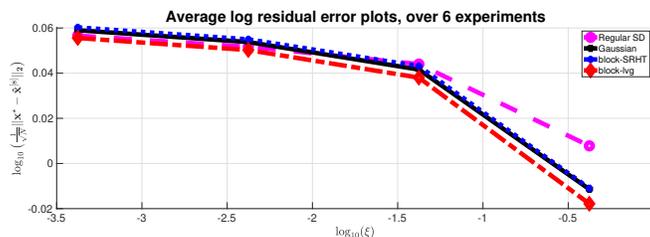


Fig. 5: Residual error for varying ξ_s .

Average log residual error: $\log_{10}(\ \mathbf{x}^* - \hat{\mathbf{x}}\ _2/\sqrt{N})$				
$\log_{10}(\xi)$	0.0004	0.0042	0.0421	0.4207
Regular SD	0.0566	0.0517	0.0440	0.0078
Gaussian	0.0590	0.0538	0.0416	-0.0114
block-SRHT	0.0603	0.0550	0.0431	-0.0110
block-lvg	0.0556	0.0502	0.0380	-0.0178

TABLE II: Average log residual errors, for six instances of SD with fixed steps, when performing Gaussian sketching with updated sketches, iterative block-SRHT and iterative block leverage score sketching, and uncoded SD.

We also considered the same experiment with \mathbf{A} drawn from a t -distribution, with optimal step-size

$$\xi_s^* = \langle \mathbf{A}g^{[s]}, \mathbf{A}\mathbf{x}^{[s]} - \mathbf{b} \rangle / \|\mathbf{A}g^{[s]}\|_2^2$$

at each iteration. From Figure 7, we observe that our iterative sketching approach outperforms Gaussian sketching with updated sketches, and iterative sketching is superior to non-iterative. Furthermore, we validate Theorem 2 and Lemma 3, as our iterative sketching approach and SSD have similar convergence. Additionally, it was observed that in some cases our iterative sketching method would outperform regular SD (and SSD). We also compared our method with updated step-sizes ξ_s^* , to iterative and non-iterative approaches according to the block leverage score sampling, block-SRHT, and Rademacher sketching methods [57], [78],

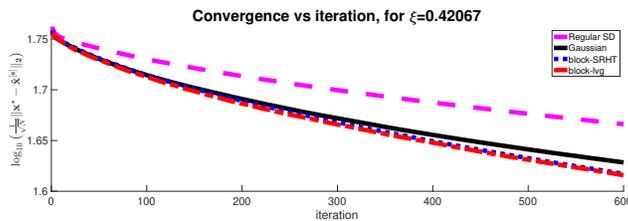
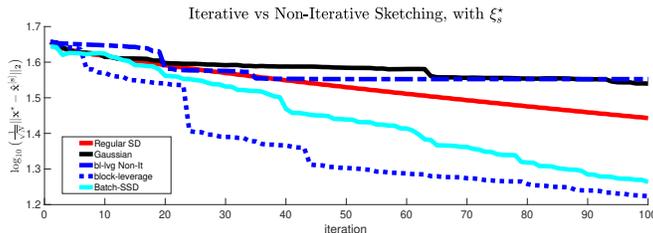


Fig. 6: log residual error convergence.

Fig. 7: log convergence with ξ_s^* .

[105], in which our corresponding approach again produced more accurate final approximations. Further experiments (not shown) were performed on various distributions for \mathbf{A} , in which similar results to those of Figures 5, 6 and 7 were observed.

The quality of our approximate block leverage sampling distribution $\bar{\Pi}_{\{K\}}$ was evaluated through an empirical mother runtime distribution taken from a distributed AWS experiment [106]. We considered the mother runtime distribution from the aforementioned experiment in which 500 computational servers were used to solve a large linear regression problem. The first step was to determine $\hat{r}_{\{K\}}$ according to (20), and then Algorithm 2 was performed to determine $r_{\{K\}}$ so that $\sum_{i=1}^K r_i = 500$, where $K = 100$. We achieved satisfactory approximations $\bar{\Pi}_{\{K\}}$ to $\Pi_{\{K\}}$, for varying ending times T (measured in seconds), i.e., $\beta_{\bar{\Pi}} \approx 1$ for each T . In Table III we present the resulting $q(T)$, misestimation factors $\beta_{\bar{\Pi}}$, and metric $\Delta_{\Pi, \bar{\Pi}}$, for each T .

Expansion Networks Misestimation Factors and Metrics							
$t \leftarrow T$	6	12	24	42	48	54	60
$q(T)$	142	226	230	230	262	267	499
$\Delta_{\Pi, \bar{\Pi}}$.8859	.9078	.9079	.9079	.9079	.9079	.9079
$\beta_{\bar{\Pi}}$.9407	.9720	.9727	.9727	.9727	.9727	.9727

TABLE III: Misestimations with AWS server completion times.

V. CONCLUSION AND FUTURE WORK

In this paper, we showed how one can exploit results from RandNLA to distributed CC, in the context of GC. By taking enough samples, or equivalently after waiting long enough, the approximation errors can be made arbitrarily small. In terms of CC, the advantages are that encodings correspond to a scalar multiplication, and no decoding step is required. Our proposed method has advantages relative to other CC approximation methods [47], [52]–[60], by incorporating information from our dataset into our scheme. These advantages include parallelization of block leverage score sketching. This is accomplished by leveraging the fact that non-straggling servers correspond to the sampled partitions of the proposed sketching algorithm.

Our proposed sketching also leads to faster algorithms. As observed experimentally in Section IV, our scheme outperforms Gaussian sketching and slightly outperforms block-SRHT sketching techniques. The improvements are also greater when adaptive step-sizes are considered. As was emphasized in the paper, iterative sketching has not been thoroughly studied in first order methods, which is what motivated our approach.

We quantified how much better the proposed solution is, both empirically and theoretically. The proposed method (Algorithm 3) compares favorably to other sketching methods in terms of empirical convergence rate and approximation error. As contrasted to other approximate GC methods, our algorithm uses data dependent spectral approximations, that enables the algorithm to quantify the residual error of the gradient approximation at each step. In terms of convergence guarantees, we obtain the same convergence rate to related works, e.g., [53], which matches the convergence rate of SSD.

Recently, papers have addressed cases where a certain subset of servers are persistent or adversarial stragglers [107], [108]. To adapt the theory developed here to the case where adversarial stragglers are present, would require modifying the replication across the non-persistent stragglers, in order to better approximate the sampling probability $\Pi_{\{K\}}$ through $\bar{\Pi}_{\{K\}}$. This is worthwhile future work.

Even though we focused on leverage score sampling for linear regression, other sampling algorithms and problems could benefit by designing analogous replication schemes. One such problem is the column subset selection problem, which can be used to compute partial SVD, QR decompositions, as well as low-rank approximations [15]. As for the sampling technique we studied, one can judiciously define a sampling distribution to approximate solutions to such problems [109], which are known to be NP-hard under the Unique Games Conjecture assumption [110].

Furthermore, existing block sampling algorithms can also benefit from the proposed expansion networks, e.g., CR -multiplication [31] and CUR decomposition [30]. For instance, a coded matrix multiplication algorithm of minimum variance can be designed, where the sampling distribution proposed in [31] is used to determine the replication numbers of the expansion network. In terms of matrix decompositions, the block leverage score algorithm of [30] can be used to distributively determine an additive ϵ -error decomposition of \mathbf{A} , in the CC setting. Another future direction is generalizing existing tensor product and factorization algorithms to block sampling, according to both approximate and exact sampling distributions, in order to make them practical for distributed environments.

ACKNOWLEDGMENT

The research presented in this paper was partially supported by the U.S. National Science Foundation (NSF) under grants NSF-2217058; NSF-2133484; NSF-2246157; ECCS-2037304 and DMS-2134248, the NSF CAREER Award under grant CCF-2236829, the U.S. Army Research Office under grant W911NF2310343; the U.S. Army Research Office Early Career Award under grant W911NF-21-1-0242, the U.S. Department of Energy under grant DE-NA0003921, and by the Office of Naval Research under grant N00014-24-1-2164. We also thank the anonymous reviewers and Aditya Ramamoorthy for their constructive comments and feedback.

APPENDIX A PROOFS OF SUBSECTION III-A

In this appendix, we provide the proof of Theorem 1 and Corollary 1. We first recall the following matrix Chernoff bound [18, Fact 1], and prove Proposition 4.

Theorem 5 (Matrix Chernoff Bound, [18, Fact 1]). *Let $\mathbf{X}_1, \dots, \mathbf{X}_q$ be independent copies of a symmetric random matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$, with $\mathbb{E}[\mathbf{X}] = \mathbf{0}_{d \times d}$, $\|\mathbf{X}\|_2 \leq \gamma$, $\|\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\|_2 \leq \sigma^2$. Let $\mathbf{Z} = \frac{1}{q} \sum_{i=1}^q \mathbf{X}_i$. Then $\forall \epsilon > 0$:*

$$\Pr \left[\|\mathbf{Z}\|_2 > \epsilon \right] \leq 2d \cdot \exp \left(-\frac{q\epsilon^2}{\sigma^2 + \gamma\epsilon/3} \right).$$

Proposition 4. *The sketching matrix $\tilde{\mathbf{S}}$ of Algorithm 1 with $\tilde{\Pi}_i \geq \beta \Pi_i$ for all i and $\beta \in (0, 1]$, guarantees*

$$\Pr \left[\|\mathbf{I}_d - \mathbf{U}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \mathbf{U}\|_2 > \epsilon \right] \leq 2d \cdot e^{-q\epsilon^2 \Theta(\beta/d)} \quad (32)$$

for any $\epsilon > 0$, and $q = r/\tau > d/\tau$.

Proof. [Proposition 4] In order to use Theorem 5, we first need to define an appropriate symmetric random matrix \mathbf{X} , whose realizations $\mathbf{X}_{\{q\}}$ correspond to the sampling procedure of Algorithm 1, and $\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} = \frac{1}{q} \sum_{i=1}^q \mathbf{X}_i$. The realization of the matrix random variable are

$$\mathbf{X}_i = \mathbf{I}_d - \left(\frac{\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)}}{\tilde{\Pi}_i} \right) = \mathbf{I}_d - \left(\frac{\sum_{l \in \mathcal{K}_i} \mathbf{U}_{(l)}^\top \mathbf{U}_{(l)}}{\tilde{\Pi}_i} \right)$$

where \mathcal{K}_i^i indicates the i^{th} block of $\mathbf{A} \in \mathbb{R}^{N \times d}$, which was sampled at trial i . This holds for all $l \in \mathbb{N}_K$. The expectation of the symmetric random matrix \mathbf{X} is

$$\begin{aligned} \mathbb{E}[\mathbf{X}] &\stackrel{\ddagger}{=} \mathbf{I}_d - \left(\sum_{j=1}^K \tilde{\Pi}_j \cdot \frac{\mathbf{U}_{(\mathcal{K}_j)}^\top \mathbf{U}_{(\mathcal{K}_j)}}{\tilde{\Pi}_j} \right) \\ &= \mathbf{I}_d - \sum_{j=1}^K \mathbf{U}_{(\mathcal{K}_j)}^\top \mathbf{U}_{(\mathcal{K}_j)} \\ &\stackrel{\#}{=} \mathbf{I}_d - \sum_{l=1}^N \mathbf{U}_{(l)}^\top \mathbf{U}_{(l)} \\ &= \mathbf{I}_d - \mathbf{I}_d \\ &= \mathbf{0}_{d \times d} \end{aligned}$$

where ‡ follows from the fact that each realization \mathbf{X}_i corresponding to each $\{\mathcal{K}_j^i\}_{j=1}^K$ of the random matrix is sampled with probability $\tilde{\Pi}_j$, and in † we simplify the expression in terms of rank-1 outer-product matrices. Furthermore, for $\{\tilde{\ell}_i\}_{i=1}^N$ the corresponding *approximate* leverage scores of \mathbf{A}

$$\begin{aligned} \|\mathbf{X}_i\|_2 &\stackrel{\ddagger}{\leq} \|\mathbf{I}_d\|_2 + \frac{\|\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)}\|_2}{\tilde{\Pi}_i} \\ &\stackrel{\diamond}{\leq} 1 + \frac{\sum_{l \in \mathcal{K}_i} \ell_l}{\left(\sum_{l \in \mathcal{K}_i} \tilde{\ell}_l\right)/d} \\ &= 1 + \frac{d \cdot \Pi_i}{\tilde{\Pi}_i} \\ &= 1 + \frac{d}{\beta} \end{aligned}$$

where in † we use the triangle inequality on the definition of \mathbf{X}_i , and in ‡ we use it on the sum of outer-products (the numerator of second summand).

We now upper bound the variance of the copies of \mathbf{X} :

$$\begin{aligned} \|\mathbb{E}[\mathbf{X}_i^\top \mathbf{X}_i]\|_2 &= \left\| \mathbb{E} \left[\left(\mathbf{I}_d - \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} / \tilde{\Pi}_i \right) \right)^\top \left(\mathbf{I}_d - \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} / \tilde{\Pi}_i \right) \right) \right] \right\|_2 \\ &= \left\| \mathbf{I}_d - 2 \cdot \mathbb{E} \left[\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} / \tilde{\Pi}_i \right) \right] + \mathbb{E} \left[\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2 / \tilde{\Pi}_i^2 \right] \right\|_2 \\ &= \left\| 2 \left(\mathbf{I}_d - \overbrace{\mathbb{E} \left[\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} / \tilde{\Pi}_i \right) \right]}^{\mathbf{I}_d} \right) - \mathbf{I}_d + \mathbb{E} \left[\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2 / \tilde{\Pi}_i^2 \right] \right\|_2 \\ &= \left\| \mathbb{E} \left[\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2 / \tilde{\Pi}_i^2 \right] - \mathbf{I}_d \right\|_2 \\ &= \left\| \left(\sum_{i=1}^K \Pi_i \cdot \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2 / \tilde{\Pi}_i^2 \right) - \mathbf{I}_d \right\|_2 \\ &\stackrel{\ddagger}{\leq} \left\| \left(\sum_{i=1}^K \frac{1}{\beta} \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2 / \tilde{\Pi}_i \right) - \mathbf{I}_d \right\|_2 \\ &= \left\| \left(\sum_{i=1}^K \frac{d}{\beta} \cdot \frac{\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2}{\|\mathbf{U}_{(\mathcal{K}_i)}\|_F^2} \right) - \mathbf{I}_d \right\|_2 \\ &\leq \left\| \left(\sum_{i=1}^K \frac{d}{\beta} \cdot \frac{\left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2}{\|\mathbf{U}_{(\mathcal{K}_i)}\|_2^2} \right) - \mathbf{I}_d \right\|_2 \\ &\stackrel{\ddagger}{=} \left\| \frac{d}{\beta} \cdot \left(\sum_{i=1}^K \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right)^2 \right) - \mathbf{I}_d \right\|_2 \\ &\stackrel{\ddagger}{\leq} \left\| \frac{d}{\beta} \left(\sum_{i=1}^K \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right) \left(\mathbf{I}_d - \mathbf{U}_{(\bar{\mathcal{K}}_i)}^\top \mathbf{U}_{(\bar{\mathcal{K}}_i)} \right) \right) - \mathbf{I}_d \right\|_2 \\ &= \left\| \frac{d}{\beta} \left(\overbrace{\sum_{i=1}^K \mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)}}^{\mathbf{I}_d} \right) - \frac{d}{\beta} \left(\sum_{i=1}^K \left(\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} \right) \left(\mathbf{U}_{(\bar{\mathcal{K}}_i)}^\top \mathbf{U}_{(\bar{\mathcal{K}}_i)} \right) \right) - \mathbf{I}_d \right\|_2 \\ &= \left\| (d/\beta - 1) \cdot \mathbf{I}_d - \frac{d}{\beta} \left(\sum_{i=1}^K \mathbf{U}_{(\mathcal{K}_i)}^\top \overbrace{\left(\mathbf{U}_{(\mathcal{K}_i)} \mathbf{U}_{(\bar{\mathcal{K}}_i)}^\top \right)}^{\mathbf{0}_{d \times d}} \mathbf{U}_{(\bar{\mathcal{K}}_i)} \right) \right\|_2 \\ &= \|(d/\beta - 1) \cdot \mathbf{I}_d\|_2 \\ &= d/\beta - 1 \end{aligned}$$

where in † we used the fact $\Pi_i/\tilde{\Pi}_i \leq 1/\beta$, in ‡ that $\|\mathbf{U}_{(\mathcal{K}_i)}\|_2^2 = 1$, and in ‡ that $\mathbf{U}_{(\mathcal{K}_i)}^\top \mathbf{U}_{(\mathcal{K}_i)} = \mathbf{I}_d - \mathbf{U}_{(\bar{\mathcal{K}}_i)}^\top \mathbf{U}_{(\bar{\mathcal{K}}_i)}$ for each i .

According to Theorem 5, we substitute $\gamma = 1 + d$ and $\sigma^2 = d/\beta - 1$ to get

$$\begin{aligned} \frac{1}{q} \sum_{i=1}^q \mathbf{X}_i &= \frac{1}{q} \sum_{i=1}^q \left(\mathbf{I}_d - \frac{\mathbf{U}^\top(\mathcal{K}_{j(i)}) \mathbf{U}(\mathcal{K}_{j(i)})}{\tilde{\Pi}_{j(i)}} \right) \\ &= \mathbf{I}_d - \frac{1}{q} \left(\sum_{i=1}^q \frac{\mathbf{U}^\top(\mathcal{K}_{j(i)}) \mathbf{U}(\mathcal{K}_{j(i)})}{\tilde{\Pi}_{j(i)}} \right) \\ &= \mathbf{I}_d - \mathbf{U}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \mathbf{U} \end{aligned}$$

where the last equality follows from the definition of $\tilde{\mathbf{S}}$. Putting everything together into Theorem 5, we get that

$$\Pr [\|\mathbf{I}_d - \mathbf{U}^\top \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \mathbf{U}\|_2 > \epsilon] \leq 2d \cdot e^{-qe^2\Theta(\beta/d)}.$$

□

Proof. [Theorem 1] By substituting

$$q = \Theta \left(\frac{d}{\tau} \log(2d/\delta) / (\beta\epsilon^2) \right)$$

in (32) and taking the complementary event, we attain the desired statement. □

Before we prove Corollary 1, we introduce the notion of *block α -balanced*, which is a generalization of *α -balanced* from [26]. The sampling distribution $\Pi_{\{K\}}$ is said to be *block α -balanced*, if

$$\max_{i \in \mathbb{N}_K} \{\Pi_i\} \leq \frac{\alpha}{N/\tau} = \frac{\alpha}{K}$$

for some constant α independent of K and q . Furthermore, in our context, if the individual leverage scores $\pi_{\{N\}}$ are α -balanced for α independent of N and r , then the block leverage scores $\Pi_{\{K\}}$ are block α -balanced, as

$$\max_{i \in \mathbb{N}_K} \{\Pi_i\} \leq \tau \cdot \max_{j \in \mathbb{N}_N} \{\pi_j\} \leq \tau \cdot \frac{\alpha}{N} = \frac{\alpha}{N/\tau} = \frac{\alpha}{K}. \quad (33)$$

Proof. [Corollary 1] From the proof of [26, Theorem 1], we simply need to show that

$$\|\mathbb{E}[\tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top)^{-1} \tilde{\mathbf{S}}]\|_2 \leq \eta \cdot \frac{r}{N}$$

for $\tilde{\mathbf{S}}$ a single sketch produced in Algorithm 1, and an appropriate constant η independent of N and r . We assume that the individual leverage scores $\pi_{\{N\}}$ are α -balanced, where α is a constant independent of N and r . By (33), it follows that the block leverage scores $\Pi_{\{K\}}$ are block α -balanced, i.e., $\Pi_i \leq \Pi_K \leq \frac{\alpha}{K}$ for all $i \in \mathbb{N}_{K-1}$.

A direct computation shows that

$$\begin{aligned} (\tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top)^{-1} &= \left((\mathbf{D} \cdot \boldsymbol{\Omega} \otimes \mathbf{I}_\tau) \cdot (\boldsymbol{\Omega}^\top \mathbf{D}^\top \otimes \mathbf{I}_\tau) \right)^{-1} \\ &= \left((\mathbf{D} \cdot \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}^\top \mathbf{D}^\top) \otimes \mathbf{I}_\tau \right)^{-1} \\ &= (\mathbf{D} \cdot \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}^\top \mathbf{D}^\top)^{-1} \otimes \mathbf{I}_\tau \end{aligned}$$

and consequently

$$\begin{aligned} \tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top)^{-1} \tilde{\mathbf{S}} &= (\boldsymbol{\Omega}^\top \cdot \mathbf{D}^\top \otimes \mathbf{I}_\tau) \cdot \left((\mathbf{D} \cdot \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}^\top \cdot \mathbf{D}^\top)^{-1} \otimes \mathbf{I}_\tau \right) \cdot (\mathbf{D} \cdot \boldsymbol{\Omega} \otimes \mathbf{I}_\tau) \\ &= \left(\boldsymbol{\Omega}^\top \cdot \mathbf{D}^\top \cdot (\mathbf{D} \cdot \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}^\top \cdot \mathbf{D}^\top)^{-1} \otimes \mathbf{I}_\tau \right) \cdot (\mathbf{D} \cdot \boldsymbol{\Omega} \otimes \mathbf{I}_\tau) \\ &= \underbrace{\left(\boldsymbol{\Omega}^\top \cdot \mathbf{D}^\top \cdot (\mathbf{D} \cdot \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}^\top \cdot \mathbf{D}^\top)^{-1} \cdot \mathbf{D} \cdot \boldsymbol{\Omega} \right)}_{:= Z \in \mathbb{R}_{\geq 0}^{K \times K}} \otimes \mathbf{I}_\tau \end{aligned}$$

where $Z = \sum_{\iota=1}^q Z_\iota$, for $\{Z_\iota\}_{\iota=1}^q$ rank-1 outer-product matrices of size $K \times K$ corresponding to each sampling trial, through Ω . For each sampling trial, we know that the i^{th} block is sampled with probability Π_i . Furthermore, if the i^{th} block is sampled at iteration ι , it follows that

$$\begin{aligned} Z_\iota &= \mathbf{e}_i \cdot \sqrt{\frac{1}{q\Pi_i}} \cdot \left(\mathbf{e}_i^\top \cdot \left(\sqrt{\frac{1}{q\Pi_i}} \right)^2 \cdot \mathbf{e}_i \right)^{-1} \cdot \sqrt{\frac{1}{q\Pi_i}} \cdot \mathbf{e}_i^\top \\ &= \mathbf{e}_i \cdot \sqrt{\frac{1}{q\Pi_i}} \cdot \left(\sqrt{q\Pi_i} \right)^2 \cdot \sqrt{\frac{1}{q\Pi_i}} \cdot \mathbf{e}_i^\top \\ &= \mathbf{e}_i \cdot \mathbf{e}_i^\top \end{aligned}$$

hence

$$\mathbb{E} \left[\tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top)^{-1} \tilde{\mathbf{S}} \right] = \mathbb{E}_\Omega \left[\sum_{j=1}^K \mathbf{e}_j \cdot \mathbf{e}_j^\top \right] \otimes \mathbf{I}_\tau = \left(\sum_{j=1}^K \mathbb{E}_\Omega [\mathbf{e}_j \cdot \mathbf{e}_j^\top] \right) \otimes \mathbf{I}_\tau = \mathbf{Q} \otimes \mathbf{I}_\tau$$

where $\mathbf{Q} = \text{diag}(\{h_j\}_{j=1}^q)$, for $h_i = 1 - (1 - \Pi_i)^q$ the probability that the i^{th} block was sampled at least once. Since we assume that the blocks $\mathbf{A}_{\{K\}}$ are indexed in ascending order according their block leverage scores, i.e., $\Pi_i \leq \Pi_{i+1}$ for all $i \in \mathbb{N}_{K-1}$, it follows that $h_i \leq h_K$ for all i , thus

$$\mathbb{E} \left[\tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top)^{-1} \tilde{\mathbf{S}} \right] = \text{diag}(\{h_j\}_{j=1}^q) \otimes \mathbf{I}_\tau \preceq h_K \cdot \mathbf{I}_N = (1 - (1 - \Pi_K)^q) \cdot \mathbf{I}_N \preceq q\Pi_K \cdot \mathbf{I}_N.$$

Consequently, since $\Pi_{\{K\}}$ are block α -balanced, we have

$$\left\| \mathbb{E} \left[\tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top)^{-1} \tilde{\mathbf{S}} \right] \right\|_2 \leq q \cdot \Pi_K \leq \alpha \cdot \frac{q}{K} = \alpha \cdot \frac{r}{N}.$$

This completes the proof, as we can take $\eta = \alpha$. \square

APPENDIX B CONCRETE EXAMPLE OF INDUCED SKETCHING

In this appendix, we give a simple demonstration of the induced sketching matrices, through our proposed GC approach of Algorithm 4. For simplicity, we will consider exact sampling, with $\Pi_{\{K\}} = \{3/20, 3/20, 4/20, 5/20, 5/20\}$ for $K = 5$, an arbitrary large block size of τ , and $m = 20$. The optimal replication numbers resulting from this distribution are $r_{\{K\}}^* = \{3, 3, 4, 5, 5\}$, hence $m = R$. In order to obtain a reduced dimension r which is 60% of the original N , we carry out $q = 3$ sampling trials at each iteration.

Let the resulting index multisets corresponding to the encoded pairs $(\tilde{A}_j, \tilde{b}_j)$ of the first four iterations, along with the resulting estimated gradients, be:

$$\begin{aligned} 1) \mathcal{S}^{[1]} = \{1, 4, 5\} &\implies \hat{g}^{[1]} = \nabla_{\mathbf{x}} L_{\mathcal{S}} \left(\tilde{\mathbf{S}}_{[1]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[1]} \right) = \hat{g}_1^{[1]} + \hat{g}_4^{[1]} + \hat{g}_5^{[1]} \\ 2) \mathcal{S}^{[2]} = \{3, 5, 5\} &\implies \hat{g}^{[2]} = \nabla_{\mathbf{x}} L_{\mathcal{S}} \left(\tilde{\mathbf{S}}_{[2]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[2]} \right) = \hat{g}_3^{[2]} + \hat{g}_5^{[2]} + \hat{g}_5^{[2]} \\ 3) \mathcal{S}^{[3]} = \{2, 4, 5\} &\implies \hat{g}^{[3]} = \nabla_{\mathbf{x}} L_{\mathcal{S}} \left(\tilde{\mathbf{S}}_{[3]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[3]} \right) = \hat{g}_2^{[3]} + \hat{g}_4^{[3]} + \hat{g}_5^{[3]} \\ 4) \mathcal{S}^{[4]} = \{4, 1, 4\} &\implies \hat{g}^{[4]} = \nabla_{\mathbf{x}} L_{\mathcal{S}} \left(\tilde{\mathbf{S}}_{[4]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[4]} \right) = \hat{g}_4^{[4]} + \hat{g}_1^{[4]} + \hat{g}_4^{[4]}. \end{aligned}$$

Considering $\mathcal{S}^{[1]}$ and $\mathcal{S}^{[2]}$ the first two iterations of this example, the index subsets are $\{\mathcal{K}_1^1, \mathcal{K}_4^2, \mathcal{K}_5^3\}$ and $\{\mathcal{K}_3^1, \mathcal{K}_5^2, \mathcal{K}_5^3\}$ respectively. Then, the corresponding induced block leverage score sketching matrices of Algorithm 1, are:

$$\begin{aligned} 1) \tilde{\mathbf{S}}_{[1]} &= \begin{pmatrix} 1/\sqrt{3\Pi_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\sqrt{3\Pi_4} & 0 \\ 0 & 0 & 0 & 0 & 1/\sqrt{3\Pi_5} \end{pmatrix} \otimes \mathbf{I}_\tau \\ 2) \tilde{\mathbf{S}}_{[2]} &= \begin{pmatrix} 0 & 0 & 1/\sqrt{3\Pi_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/\sqrt{3\Pi_5} \\ 0 & 0 & 0 & 0 & 1/\sqrt{3\Pi_5} \end{pmatrix} \otimes \mathbf{I}_\tau \\ 3) \tilde{\mathbf{S}}_{[3]} &= \begin{pmatrix} 0 & 1/\sqrt{3\Pi_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\sqrt{3\Pi_4} & 0 \\ 0 & 0 & 0 & 0 & 1/\sqrt{3\Pi_5} \end{pmatrix} \otimes \mathbf{I}_\tau \\ 4) \tilde{\mathbf{S}}_{[4]} &= \begin{pmatrix} 0 & 0 & 0 & 1/\sqrt{3\Pi_4} & 0 \\ 1/\sqrt{3\Pi_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\sqrt{3\Pi_4} & 0 \end{pmatrix} \otimes \mathbf{I}_\tau \end{aligned}$$

each of which are of size $(3\tau) \times (5\tau)$.

APPENDIX C
COMPARISON TO THE BLOCK-SRHT

An alternative view point is that the matrix

$$\tilde{\mathbf{U}}_{\text{exp}} := \Psi \cdot (\Sigma \mathbf{V}^\top)^{-1}$$

has uniform Frobenius block scores, which further justifies the proposed GC approach for homogeneous servers. This resembles the main idea behind the SRHT [78], [96] and different variants known as *block-SRHT* [56], [111], where a random projection is applied to the data matrix to *flatten* its leverage scores, i.e., make them close to uniform. These two techniques for flattening the corresponding block scores are vastly different, and the proximity of the corresponding block scores are measured and quantified differently. In contrast to the block-SRHT, the quality of the approximation in our case depends on the misestimation factor $\beta_{\tilde{\Pi}}$, and is not quantified probabilistically.

We note that since Ψ has repeated blocks from the expansion, the scores we consider in Lemma 4 are not the block leverage scores of Ψ . The Frobenius block scores of $\tilde{\mathbf{U}}_{\text{exp}}$, are in fact the corresponding block leverage scores of $\tilde{\mathbf{A}}$, which are replicated in the expansion. Moreover, note that the closer $\beta_{\tilde{\Pi}}$ is to 1, the closer the sampling distribution according to the Frobenius block scores of $\tilde{\mathbf{U}}_{\text{exp}}$; which we denote by $\tilde{Q}_{\{R\}}$, is to being exactly uniform. We denote the uniform sampling distribution by $\mathcal{U}_{\{R\}}$.

Lemma 4. *When $\tilde{\Pi}_{\{K\}} = \Pi_{\{K\}}$, the sampling distribution $\tilde{Q}_{\{R\}}$ is uniform. When $\tilde{\Pi}_\iota \geq \beta_{\tilde{\Pi}} \Pi_\iota$ for $\beta_{\tilde{\Pi}} = \min_{i \in \mathbb{N}_K} \{\Pi_i / \tilde{\Pi}_i\} \in (0, 1)$ and all $\iota \in \mathbb{N}_K$, the resulting distribution $\tilde{Q}_{\{R\}}$ is approximately uniform, and satisfies $d_{\mathcal{U}, \tilde{Q}} \leq 1 / (R\beta_{\tilde{\Pi}})$.*

Proof. Let $\mathbf{U} = [\mathbf{U}_1^\top \ \cdots \ \mathbf{U}_K^\top]^\top$ denote the corresponding blocks of \mathbf{U} according the partitioning of \mathcal{D} . Without loss of generality, assume that the data points within each partition are consecutive rows of \mathbf{A} , and $\mathbf{U}_\iota \in \mathbb{R}^{\tau \times d}$ for all $\iota \in \mathbb{N}_K$.

From (21) and (25), it follows that

$$\begin{aligned} \Psi &= \tilde{\mathbf{E}} \cdot \tilde{\mathbf{A}} \\ &= (\mathbf{E} \otimes \mathbf{I}_\tau) \cdot (\mathbf{G} \cdot \mathbf{U} \Sigma \mathbf{V}^\top) \\ &= (\mathbf{E} \otimes \mathbf{I}_\tau) \cdot \left[\mathbf{U}_1^\top / \sqrt{q\tilde{\Pi}_1} \ \cdots \ \mathbf{U}_K^\top / \sqrt{q\tilde{\Pi}_K} \right]^\top \cdot \Sigma \mathbf{V}^\top \\ &=: (\mathbf{E} \otimes \mathbf{I}_\tau) \cdot \left[\tilde{\mathbf{U}}_1^\top \ \cdots \ \tilde{\mathbf{U}}_K^\top \right]^\top \cdot \Sigma \mathbf{V}^\top \\ &=: \underbrace{\left[\tilde{\mathbf{U}}_1^\top \ \cdots \ \tilde{\mathbf{U}}_1^\top \ \tilde{\mathbf{U}}_2^\top \ \cdots \ \tilde{\mathbf{U}}_2^\top \ \cdots \ \tilde{\mathbf{U}}_K^\top \ \cdots \ \tilde{\mathbf{U}}_K^\top \right]^\top}_{\tilde{\mathbf{U}}_{\text{exp}} \in \mathbb{R}^{R\tau \times d}} \cdot \Sigma \mathbf{V}^\top. \end{aligned}$$

Note that $\tilde{\mathbf{U}}_{\text{exp}} \Sigma \mathbf{V}^\top$ is not the SVD of Ψ . For the normalizing factor of $\frac{q}{Rd}$:

$$\tilde{Q}_\iota = \frac{q}{Rd} \cdot \|\tilde{\mathbf{U}}_\iota\|_F^2 = \frac{q}{Rd} \cdot \frac{\|\mathbf{U}_\iota\|_F^2}{q\tilde{\Pi}_\iota} = \frac{\Pi_\iota}{R\tilde{\Pi}_\iota} \leq \frac{1}{R\beta_{\tilde{\Pi}}}$$

therefore

$$\sum_{i=1}^R |\tilde{Q}_i - 1/R| \triangleq \frac{R}{R\beta_{\tilde{\Pi}}} = \frac{1}{\beta_{\tilde{\Pi}}}$$

where \triangleq follows from the fact that $|\tilde{Q}_i - 1/R| \leq 1 / (R\beta_{\tilde{\Pi}})$ for each $i \in \mathbb{N}_R$. After normalizing by $1/R$ according to the distortion metric, we deduce that $d_{\mathcal{U}, \tilde{Q}} \leq 1 / (R\beta_{\tilde{\Pi}})$.

In the case where $\tilde{\Pi}_{\{K\}} = \Pi_{\{K\}}$, we have

$$\tilde{Q}_\iota = \frac{q}{Rd} \cdot \|\tilde{\mathbf{U}}_\iota\|_F^2 = \frac{q}{Rd} \cdot \frac{\|\mathbf{U}_\iota\|_F^2}{q\Pi_\iota} = \frac{\Pi_\iota}{R\Pi_\iota} = \frac{1}{R}$$

for all $\iota \in \mathbb{N}_K$, thus $\tilde{Q}_{\{K\}} = \mathcal{U}_{\{K\}}$. □

Finally, in Proposition 5 we show when the block leverage score sampling sketch of Algorithm 1 and the block-SRHT of [56] have the same ℓ_2 -subspace embedding guarantee. We first recall the corresponding result to Theorem 1, of the block-SRHT.

Theorem 6 ([56, Theorem 7]). *The block-SRHT $\mathbf{S}_{\tilde{\mathbf{H}}}$ is an ℓ_2 -subspace embedding of \mathbf{A} . For $\delta > 0$ and $q = \Theta\left(\frac{d}{\tau} \log(Nd/\delta)\right) \cdot \log(2d/\delta)/\epsilon^2$:*

$$\Pr \left[\|\mathbf{I}_d - \mathbf{U}^\top \mathbf{S}_{\tilde{\mathbf{H}}}^\top \mathbf{S}_{\tilde{\mathbf{H}}} \mathbf{U}\|_2 \leq \epsilon \right] \geq 1 - \delta.$$

Proposition 5. Let $\beta = 1$. For $\delta = e^{\Theta(1)}/(Nd)$, the sketches of Algorithm 1 and the block-SRHT of [56] achieve the same asymptotic ℓ_2 -subspace embedding guarantee, for the same number of sampling trials q .

Proof. For $\delta = e^{\Theta(1)}/(Nd)$, the two sketching methods have the same q , and both satisfy the ℓ_2 -subspace embedding property with error probability $1 - \delta$. \square

APPENDIX D CONTRACTION RATE OF BLOCK LEVERAGE SCORE SAMPLING

In this appendix we quantify the contraction rate of our method on the error term $\mathbf{x}^{[s]} - \mathbf{x}^*$, which further characterizes the convergence of SD after applying our method. The contraction rate is compared to that of regular SD.

Recall that the contraction rate of an iterative process given by a function $h(x^{[s]})$ is the constant $\gamma \in (0, 1)$ for which at each iteration we are guaranteed that $h(x^{[s+1]}) \leq \gamma \cdot h(x^{[s]})$, therefore $h(x^{[s]}) \leq \gamma_s \cdot h(x^{[0]})$. Let ξ be a fixed step-size, $\tilde{\mathbf{S}}_{[s]}$ the induced sketching matrix of Algorithm 1 at iteration s , and define $\mathbf{B}_{SD} = (\mathbf{I}_d - 2\xi \cdot \mathbf{A}^\top \mathbf{A})$ and $\mathbf{B}_s = (\mathbf{I}_d - 2\xi \cdot \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{A})$. We note that the contraction rates could be further improved if one also incorporates an optimal step-size. It is also worth noting that when weighting from Appendix E is introduced, we have the same contraction rate and straggler ratio.

Lemma 5. For $\tilde{\mathbf{S}}$ the sketching matrix of Algorithm 1, we have $\mathbb{E} [\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}] = \mathbf{I}_N$.

Proof. Similar to the proof of Proposition 4, we define a symmetric random matrix \mathbf{Y} , whose realizations correspond to the sampled and rescaled submatrices of Algorithm 1. The realizations are

$$\mathbf{Y}_i = \frac{\mathbf{I}_{(\mathcal{K}_i)}^\top \mathbf{I}_{(\mathcal{K}_i)}}{q\tilde{\Pi}_i} = \frac{\sum_{l \in \mathcal{K}_i} \mathbf{e}_l \mathbf{e}_l^\top}{q\tilde{\Pi}_i}.$$

After q sampling trials, we have $\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} = \sum_{i=1}^q \mathbf{Y}_i$. It follows that

$$\begin{aligned} \mathbb{E} [\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}] &= \mathbb{E} \left[\sum_{i=1}^q \mathbf{Y}_i \right] \\ &= \sum_{i=1}^q \mathbb{E} [\mathbf{Y}_i] \\ &= q \cdot \left(\sum_{j=1}^K \tilde{\Pi}_j \cdot \frac{\mathbf{I}_{(\mathcal{K}_j)}^\top \mathbf{I}_{(\mathcal{K}_j)}}{q\tilde{\Pi}_j} \right) \\ &= \sum_{j=1}^K \mathbf{I}_{(\mathcal{K}_j)}^\top \mathbf{I}_{(\mathcal{K}_j)} \\ &= \sum_{l=1}^N \mathbf{e}_{(l)} \mathbf{e}_{(l)}^\top \\ &= \mathbf{I}_N. \end{aligned}$$

\square

Theorem 7. The contraction rate of our GC approach through the expected sketch $\tilde{\mathbf{S}}_{[s]}$ at each iteration, is equal to the contraction rate of regular SD. Specifically, for $e_s := \mathbf{x}^{[s]} - \mathbf{x}^*$ the error at iteration s and $\gamma_{SD} = \lambda_1(\mathbf{B}_{SD})$ the contraction rate of regular SD, we have $\|\mathbb{E}[e_{s+1}]\|_2 \leq \gamma_{SD} \cdot \|e_s\|_2$.

Proof. For a fixed step-size 2ξ , our SD parameter update at iteration $s + 1$ is

$$\mathbf{x}^{[s+1]} \leftarrow \mathbf{x}^{[s]} - 2\xi \cdot \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} (\mathbf{A} \mathbf{x}^{[s]} - \mathbf{b}),$$

where for regular SD we have $\tilde{\mathbf{S}}_{[s]} \leftarrow \mathbf{I}_N$. At iteration $s+1$, the error e_s of the previous iteration is not random, hence $\mathbb{E}[e_s] = e_s$. By substituting the expression of \mathbf{B}_s , it follows that

$$\begin{aligned}
e_{s+1} &= \mathbf{x}^{[s+1]} - \mathbf{x}^* \\
&= \left(\mathbf{x}^{[s]} - 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \left(\mathbf{A} \mathbf{x}^{[s]} - \mathbf{b} \right) \right) - \mathbf{x}^* \\
&= \mathbf{x}^{[s]} - 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{A} \mathbf{x}^{[s]} + 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{b} - \mathbf{x}^* \\
&= \mathbf{B}_s \mathbf{x}^{[s]} - \left(\mathbf{x}^* - 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{b} \right) \\
&= \mathbf{B}_s \mathbf{x}^{[s]} - \left(\mathbf{x}^* - 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \left(\mathbf{A} \mathbf{x}^* + \mathbf{b}^\perp \right) \right) \\
&= \mathbf{B}_s \left(\mathbf{x}^{[s]} - \mathbf{x}^* \right) - 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{b}^\perp \\
&= \mathbf{B}_s e_s - 2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{b}^\perp
\end{aligned} \tag{34}$$

and by Lemma 5

$$\begin{aligned}
\mathbb{E} \left[2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{b}^\perp \right] &= 2\xi \mathbf{A}^\top \mathbb{E} \left[\tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \right] \mathbf{b}^\perp \\
&= 2\xi \mathbf{A}^\top \mathbf{b}^\perp \\
&= \mathbf{0}_{d \times 1}
\end{aligned} \tag{35}$$

as \mathbf{b}^\perp lies in the kernel of \mathbf{A}^\top , and

$$\begin{aligned}
\mathbb{E} [\mathbf{B}_s] &= \mathbf{I}_d - 2\xi \mathbf{A}^\top \mathbb{E} \left[\tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \right] \mathbf{A} \\
&= \mathbf{I}_d - 2\xi \mathbf{A}^\top \mathbf{A} \\
&= \mathbf{B}_{SD}.
\end{aligned} \tag{36}$$

From (34), (35) and (36), it follows that

$$\begin{aligned}
\mathbb{E} [e_{s+1}] &= \mathbb{E} [\mathbf{B}_s e_s] - \mathbb{E} \left[2\xi \mathbf{A}^\top \tilde{\mathbf{S}}_{[s]}^\top \tilde{\mathbf{S}}_{[s]} \mathbf{b}^\perp \right] \\
&= \mathbb{E} [\mathbf{B}_s] \cdot e_s \\
&= \mathbf{B}_{SD} \cdot e_s.
\end{aligned}$$

This results in the inequality

$$\|\mathbb{E}[e_{s+1}]\|_2 \leq \lambda_1(\mathbb{E}[\mathbf{B}_s]) \cdot \|e_s\|_2 = \lambda_1(\mathbf{B}_{SD}) \cdot \|e_s\|_2$$

which implies the contraction rate of the expected sketch through Algorithm 1:

$$\gamma_{s+1} = \lambda_1(\mathbf{B}_{SD}).$$

By replacing \mathbf{B}_s with \mathbf{B}_{SD} in (34) and $\tilde{\mathbf{S}}_{[s]} \leftarrow \mathbf{I}_N$, we conclude that the contraction rate of SD is $\gamma_{SD} = \lambda_1(\mathbf{B}_{SD})$. \square

We conclude this appendix by stating the expected ratio of dimensions r to N , and stragglers to servers.

Remark 2. *The expected ratio of the reduced dimension r to the original dimension N is $q(T)\tau/N$, and the expected straggler to servers ratio is $(m-q)/m$. Since we stop receiving computations at a time instance T ; we expect that $q \leftarrow q(T)$ computations are received, hence there are $m-q$ stragglers. Therefore, the straggler to servers ratio is $(m-q)/m$. The expected ratio between the two dimensions is immediate from the fact that $r = q\tau$ is the new reduced dimension, in the case where $q \leq K$.*

APPENDIX E WEIGHTED BLOCK LEVERAGE SCORE SKETCHING

So far, we have considered sampling with replacement according to the normalized block leverage scores, to reduce the effective dimension N of \mathbf{A} and \mathbf{b} to $r = q\tau$. In this appendix, we show that by *weighting* the sampled blocks according to the sampling which has taken place through Ω for the construction of the sketching matrix $\tilde{\mathbf{S}}$, we can further compress the data matrix \mathbf{A} , and get the same results when first and second order optimization methods are used to solve (12). The weighting we propose is more beneficial with non-uniform distributions, as we expect the sampling with replacement to capture the importance of the more influence blocks. The *weighted sketching matrix* $\tilde{\mathbf{S}}_w$ we propose, is a simple extension to $\tilde{\mathbf{S}}$ of Algorithm 1. For simplicity, we do not consider iterative sketching, though similar arguments and guarantees can be derived.

The main idea is to not keep repetitions of blocks which were sampled multiple times, but rather weigh each block by the number of times it was sampled. By doing so, we retain the *weighted sketch* $\tilde{\mathbf{S}}_w \mathbf{A}$ of size $\bar{q}\tau \times d$, for \bar{q} the number of distinct blocks that were sampled.⁵ Additionally, the gradient and Hessian of $L_{\tilde{\mathbf{S}}_w}(\tilde{\mathbf{S}}_w, \mathbf{A}, \mathbf{b}; \mathbf{x})$ are respectively equal to those of $L_{\tilde{\mathbf{S}}}(\tilde{\mathbf{S}}, \mathbf{A}, \mathbf{b}; \mathbf{x})$, and

⁵In practice, for highly non-uniform $\Pi_{\{K\}}$ we expect $\bar{q}\tau \ll r = q\tau$. The sketch $\tilde{\mathbf{S}}_w \mathbf{A}$ could therefore be stored in much less space than $\tilde{\mathbf{S}} \mathbf{A}$, and the system of equations $\tilde{\mathbf{S}}_w(\mathbf{A} - \mathbf{b}) = \mathbf{0}_{\bar{q}\tau \times 1}$ could have significantly fewer equations than $\tilde{\mathbf{S}}(\mathbf{A} - \mathbf{b}) = \mathbf{0}_{q\tau \times 1}$.

are unbiased estimators of the gradient and Hessian of $L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x})$.

To achieve the weighting algorithmically, we count how many times each of the distinct \bar{q} blocks were sampled, and at the end of the sampling procedure we multiply the blocks by their corresponding *weight*. We initialize a weight vector $\mathbf{w} = \mathbf{0}_{1 \times K}$, and in the sampling procedure whenever the i^{th} partition is drawn, we update its corresponding weight: $\mathbf{w}_i \leftarrow \mathbf{w}_i + 1$. It is clear that once q trials are been carried out, we have $\|\mathbf{w}\|_1 = q$ for $\mathbf{w} \in \mathbb{N}_0^{1 \times K}$.

Let \mathcal{S} denote the index multiset observed after the sampling procedure of Algorithm 1, and $\bar{\mathcal{S}}$ the set of indices comprising \mathcal{S} . That is, \mathcal{S} has cardinality q and may have repetitions, while $\bar{\mathcal{S}} = \mathbb{N}_K \cap \mathcal{S}$ has cardinality $\bar{q} = |\bar{\mathcal{S}}| \leq q$ with no repetitions. We denote the ratio of the two sets by $\zeta := q/\bar{q} = \|\mathbf{w}\|_1/\|\mathbf{w}\|_0 \geq 1$, which indicates how much further compression we get by utilizing the fact that blocks may be sampled multiple times. The corresponding weighted sketching matrix $\tilde{\mathbf{S}}_{\mathbf{w}}$ of $\tilde{\mathbf{S}}$ is then

$$\tilde{\mathbf{S}}_{\mathbf{w}} = \overbrace{\text{diag} \left(\left\{ \sqrt{\mathbf{w}_j / (q\Pi_j)} \right\}_{j \in \bar{\mathcal{S}}} \right)}^{\bar{\mathbf{W}}_{1/2} \in \mathbb{R}_{\geq 0}^{\bar{q} \times \bar{q}}} \cdot \mathbf{I}_{(\bar{\mathcal{S}})} \otimes \mathbf{I}_{\tau} \in \mathbb{R}^{\bar{q}\tau \times N} \quad (37)$$

where $\mathbf{I}_{(\bar{\mathcal{S}})} \in \{0, 1\}^{\bar{q} \times K}$ is the restriction of \mathbf{I}_K to the rows indexed by $\bar{\mathcal{S}}$.

For simplicity, assume that the sampling matrices which are devised for $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{S}}_{\mathbf{w}}$ follow the ordering of the sampled blocks in order of the samples, i.e., if $(\boldsymbol{\Omega}_{(i)})_j = 1$ then $(\boldsymbol{\Omega}_{(i+1)})_l = 1$ for $l \geq j$, and equivalently $\mathcal{S}_l \leq \mathcal{S}_{l+1}$ and $\bar{\mathcal{S}}_l < \bar{\mathcal{S}}_{l+1}$ for all valid l . We restrict the sampling matrix $\boldsymbol{\Omega} \in \{0, 1\}^{q \times K}$ to its unique rows, by applying $\bar{\boldsymbol{\Omega}} \in \{0, 1\}^{\bar{q} \times q}$:

$$\bar{\boldsymbol{\Omega}}_{ij} = \begin{cases} 1 & \text{for } i = 1 \text{ and } j = \mathcal{S}_1 \\ 1 & \text{if } \mathcal{S}_j > \mathcal{S}_{j-1} \text{ for } j \in \mathbb{N}_q \setminus \{1\} \\ 0 & \text{otherwise} \end{cases}$$

to the left of the sampling matrix $\boldsymbol{\Omega}$ of Algorithm 1, i.e., $\boldsymbol{\Omega}_{\mathbf{w}} := \bar{\boldsymbol{\Omega}} \cdot \boldsymbol{\Omega} \in \{0, 1\}^{\bar{q} \times K}$. This sampling matrix then satisfies

$$(\boldsymbol{\Omega}_{\mathbf{w}})_{ij} = \begin{cases} 1 & \text{if } j = \bar{\mathcal{S}}_j \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j \in \mathbb{N}_{\bar{q}}.$$

Let $\tilde{\mathbf{w}} = \mathbf{w}_{|\bar{\mathcal{S}}} \in \mathbb{Z}_+^{1 \times \bar{q}}$ be the restriction of \mathbf{w} to its nonzero elements; hence $\|\tilde{\mathbf{w}}\|_1 = \|\mathbf{w}\|_1 = |\mathcal{S}| = q$, and define the rescaling diagonal matrix $\tilde{\mathbf{W}}_{1/2} = \text{diag} \left(\left\{ \sqrt{\tilde{\mathbf{w}}_i} \right\}_{i=1}^{\bar{q}} \right)$. We then have the following relationship

$$\begin{aligned} \tilde{\mathbf{S}}_{\mathbf{w}} &= \overbrace{\left(\tilde{\mathbf{W}}_{1/2} \cdot \bar{\boldsymbol{\Omega}} \cdot \mathbf{D} \cdot \bar{\boldsymbol{\Omega}}^\top \right)}^{=\tilde{\mathbf{W}}_{1/2}} \cdot \boldsymbol{\Omega}_{\mathbf{w}} \otimes \mathbf{I}_{\tau} \\ &= (\tilde{\mathbf{W}}_{1/2} \otimes \mathbf{I}_{\tau}) \cdot (\boldsymbol{\Omega}_{\mathbf{w}} \otimes \mathbf{I}_{\tau}) \end{aligned} \quad (38)$$

when $\bar{\mathcal{S}} = \mathbb{N}_K \cap \mathcal{S}$.

As previously noted, $\tilde{\mathbf{S}}_{\mathbf{w}}$ has ζ times less rows than $\tilde{\mathbf{S}}$. Hence, the required storage space for the sketch $\tilde{\mathbf{S}}_{\mathbf{w}}\mathbf{A}$ drops by a multiplicatively factor of ζ , and the required operations are reduced analogously, according to the computation. The weighted sketching matrix $\tilde{\mathbf{S}}_{\mathbf{w}}$ has the following guarantees, which imply that the proposed weighting will not affect first or second order iterative methods which are used to approximate (12).

Proposition 6. *The resulting gradient and Hessian of the modified least squares problem (12) when sketching with $\tilde{\mathbf{S}}$ of Algorithm 1, are respectively identical to the resulting gradient and Hessian when sketching with $\tilde{\mathbf{S}}_{\mathbf{w}}$ presented in (37) and (38).*

Proof. From Algorithm 1, the assumption on the ordering of the elements in \mathcal{S} and $\bar{\mathcal{S}}$, and the construction of $\tilde{\mathbf{S}}$, we have

$$\begin{aligned} \tilde{\mathbf{S}}_{\mathbf{w}}^\top \tilde{\mathbf{S}}_{\mathbf{w}} &= \left((\boldsymbol{\Omega}_{\mathbf{w}}^\top \tilde{\mathbf{W}}_{1/2}^\top) \otimes \mathbf{I}_{\tau} \right) \cdot \left((\tilde{\mathbf{W}}_{1/2} \cdot \boldsymbol{\Omega}_{\mathbf{w}}) \otimes \mathbf{I}_{\tau} \right) \\ &= \left((\boldsymbol{\Omega}^\top \mathbf{D}^\top) \otimes \mathbf{I}_{\tau} \right) \cdot \left((\mathbf{D} \cdot \boldsymbol{\Omega}) \otimes \mathbf{I}_{\tau} \right) \\ &= \tilde{\mathbf{S}}^\top \tilde{\mathbf{S}}. \end{aligned}$$

Let $\mathcal{T} = \bigsqcup_{j \in \mathcal{S}} \mathcal{K}_j$ and $\tilde{\mathcal{T}} = \bigsqcup_{j \in \tilde{\mathcal{S}}} \mathcal{K}_j$, thus $\tilde{\mathcal{T}}$ is contained in \mathcal{T} when both are viewed as multisets. Considering the objective function $L_{\mathbf{S}}(\tilde{\mathbf{S}}, \mathbf{A}, \mathbf{b}; \mathbf{x})$ of (12), the equivalence of gradients is observed through the following computation

$$\begin{aligned}
\nabla_{\mathbf{x}} L_{\mathbf{S}}(\tilde{\mathbf{S}}, \mathbf{A}, \mathbf{b}; \mathbf{x}) &= 2\mathbf{A}^\top \left(\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \right) (\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= 2 \sum_{l \in \mathcal{T}} \mathbf{A}_{(l)}^\top \cdot \mathbf{D}_{ll}^2 \cdot (\mathbf{A}_{(l)}\mathbf{x} - \mathbf{b}_l) \\
&= 2 \sum_{j \in \tilde{\mathcal{T}}} \tilde{\mathbf{w}}_j \cdot \mathbf{A}_{(j)}^\top \cdot \mathbf{D}_{jj}^2 \cdot (\mathbf{A}_{(j)}\mathbf{x} - \mathbf{b}_j) \\
&= 2 \sum_{j \in \tilde{\mathcal{T}}} \mathbf{A}_{(j)}^\top \cdot (\tilde{\mathbf{W}}_{1/2})_{jj}^2 \cdot (\mathbf{A}_{(j)}\mathbf{x} - \mathbf{b}_j) \\
&= 2\mathbf{A}^\top \left(\tilde{\mathbf{S}}_{\mathbf{w}}^\top \tilde{\mathbf{S}}_{\mathbf{w}} \right) (\mathbf{A}\mathbf{x} - \mathbf{b}) \\
&= \nabla_{\mathbf{x}} L_{\mathbf{S}}(\tilde{\mathbf{S}}_{\mathbf{w}}, \mathbf{A}, \mathbf{b}; \mathbf{x}).
\end{aligned}$$

Recall that the Hessian of the least squares objective function (2) is $\nabla_{\mathbf{x}}^2 L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x}) = 2\mathbf{A}^\top \mathbf{A}$. Considering the modified objective function (12) and our sketching matrices, it follows that

$$\begin{aligned}
\nabla_{\mathbf{x}}^2 L_{\mathbf{S}}(\tilde{\mathbf{S}}, \mathbf{A}, \mathbf{b}; \mathbf{x}) &= 2\mathbf{A}^\top \left(\tilde{\mathbf{S}}^\top \tilde{\mathbf{S}} \right) \mathbf{A} \\
&= 2 \sum_{l \in \mathcal{T}} \mathbf{A}_{(l)}^\top \cdot \mathbf{D}_{ll}^2 \cdot \mathbf{A}_{(l)} \\
&= 2 \sum_{j \in \tilde{\mathcal{T}}} \tilde{\mathbf{w}}_j \cdot \mathbf{A}_{(j)}^\top \cdot \mathbf{D}_{jj}^2 \cdot \mathbf{A}_{(j)} \\
&= 2 \sum_{j \in \tilde{\mathcal{T}}} \mathbf{A}_{(j)}^\top \cdot (\tilde{\mathbf{W}}_{1/2})_{jj}^2 \cdot \mathbf{A}_{(j)} \\
&= 2\mathbf{A}^\top \left(\tilde{\mathbf{S}}_{\mathbf{w}}^\top \tilde{\mathbf{S}}_{\mathbf{w}} \right) \mathbf{A} \\
&= \nabla_{\mathbf{x}}^2 L_{\mathbf{S}}(\tilde{\mathbf{S}}_{\mathbf{w}}, \mathbf{A}, \mathbf{b}; \mathbf{x})
\end{aligned}$$

which completes the proof. \square

Corollary 4. *At each iteration, the gradient and Hessian of the weighted sketch system of equations $\tilde{\mathbf{S}}_{\mathbf{w}}(\mathbf{A} - \mathbf{b}) = \mathbf{0}_{q\tau \times 1}$, are unbiased estimators of the gradient and Hessian of the original system $(\mathbf{A} - \mathbf{b}) = \mathbf{0}_{N \times 1}$.*

Proof. Denote the gradient and Hessian of the weighted sketch at iteration s by $\hat{g}_{\mathbf{w}}^{[s]}$ and $\hat{H}_{\mathbf{w}}^{[s]}$ respectively. By Proposition 6 we know that $\hat{g}_{\mathbf{w}}^{[s]} = \hat{g}^{[s]}$, and by Theorem 2 that $\mathbb{E}[\hat{g}^{[s]}] = g^{[s]}$. Hence $\mathbb{E}[\hat{g}_{\mathbf{w}}^{[s]}] = g^{[s]}$.

Following the same notation as in the proof of Theorem 2, the Hessian $\hat{H}^{[s]} = \nabla_{\mathbf{x}}^2 L_{\mathbf{S}}(\tilde{\mathbf{S}}_{[s]}, \mathbf{A}, \mathbf{b}; \mathbf{x}^{[s]})$ is

$$\hat{H}^{[s]} = 2 \sum_{i \in \mathcal{I}^{[s]}} \frac{1}{q\tilde{\Pi}_i} \mathbf{A}_i^\top \mathbf{A}_i$$

thus

$$\begin{aligned}
\mathbb{E}[\hat{H}^{[s]}] &= 2\mathbb{E} \left[\sum_{i \in \mathcal{I}^{[s]}} \frac{1}{q\tilde{\Pi}_i} \mathbf{A}_i^\top \mathbf{A}_i \right] \\
&= 2 \sum_{i \in \mathcal{I}^{[s]}} \sum_{j=1}^K \tilde{\Pi}_j \frac{1}{q\tilde{\Pi}_j} \mathbf{A}_j^\top \mathbf{A}_j \\
&= 2q \cdot \sum_{j=1}^K \frac{1}{q} \mathbf{A}_j^\top \mathbf{A}_j \\
&= 2\mathbf{A}^\top \mathbf{A}
\end{aligned}$$

which is precisely the Hessian of (2). By Proposition 6, it follows that $\mathbb{E}[\hat{H}_{\mathbf{w}}^{[s]}] = 2\mathbf{A}^\top \mathbf{A}$, which completes the proof. \square

Geometrically, from the point of view of adding vectors, the partial gradients of the partitions sampled will be scaled accordingly to their weights. Therefore, the partial gradients \hat{g}_i with higher weights have a greater influence in the direction of the resulting

gradient \hat{g} . This was also the fundamental idea behind our sketching and GC techniques, as the partitions sampled multiple times are of greater importance.

Next, we quantify the expected dimension of the weighted sketch $\tilde{\mathbf{S}}_{\mathbf{w}}\mathbf{A}$. This shows the dependence on $\tilde{\Pi}_{\{K\}}$, and further justifies that we attain a higher compression factor ζ when the block leverage scores are non-uniform. We note that when the data points are i.i.d. and N is significantly larger than d , then the leverage scores will be approximately equal; and use of these scores to determine dominant rows will offer little or no benefit.

Theorem 8. *The expected reduced dimension of $\tilde{\mathbf{S}}_{\mathbf{w}}\mathbf{A}$ is*

$$\left(K - \sum_{i=1}^K (1 - \tilde{\Pi}_i)^q\right) \cdot \tau$$

which is maximal when $\tilde{\Pi}_{\{K\}}$ is uniform.

Proof. It suffices to determine the expected number of distinct blocks \mathbf{A}_i which are sampled after q trials when carrying out Algorithm 1. The probability of not sampling \mathbf{A}_i at a given trial is $(1 - \tilde{\Pi}_i)$, hence not sampling \mathbf{A}_i at any trial occurs with probability $(1 - \tilde{\Pi}_i)^q$, since the trials are identical and independent. Thus, the expected number of distinct blocks being sampled is

$$\begin{aligned} \mathbb{E}[\bar{q}] &= \sum_{i=1}^K 1 \cdot \Pr[\mathbf{A}_i \text{ was sampled at least once}] \\ &= \sum_{i=1}^K (1 - \Pr[\mathbf{A}_i \text{ was not sampled at any trial}]) \\ &= \sum_{i=1}^K \left(1 - (1 - \tilde{\Pi}_i)^q\right) \\ &= K - \sum_{i=1}^K (1 - \tilde{\Pi}_i)^q. \end{aligned}$$

Thus, the expected reduced dimension is $\tau \cdot \mathbb{E}[\bar{q}]$.

Let $Q(\tilde{\Pi}_{\{K\}}) := \sum_{i=1}^K (1 - \tilde{\Pi}_i)^q$, and introduce the Lagrange multiplier $\lambda > 0$ to the constraint $R(\tilde{\Pi}_{\{K\}}) = \left(\sum_{i=1}^K \tilde{\Pi}_i - 1\right)$, to get the Lagrange function

$$\begin{aligned} \mathcal{L}(\tilde{\Pi}_{\{K\}}, \lambda) &:= Q(\tilde{\Pi}_{\{K\}}) + \lambda \cdot R(\tilde{\Pi}_{\{K\}}) \\ &= \sum_{i=1}^K \left(\lambda \cdot \tilde{\Pi}_i + (1 - \tilde{\Pi}_i)^q\right) - \lambda \end{aligned}$$

for which

$$\frac{\partial \mathcal{L}(\tilde{\Pi}_{\{K\}}, \lambda)}{\partial \tilde{\Pi}_i} = \lambda - q(1 - \tilde{\Pi}_i)^{q-1} = 0 \quad (39)$$

$$\implies \tilde{\Pi}_i = 1 - (\lambda/q)^{1/(q-1)} \quad \text{and} \quad \lambda = q(1 - \tilde{\Pi}_i)^{q-1} \quad (40)$$

for all $i \in \mathbb{N}_K$, and

$$\frac{\partial \mathcal{L}(\tilde{\Pi}_{\{K\}}, \lambda)}{\partial \lambda} = \sum_{i=1}^K \tilde{\Pi}_i - 1 = 0. \quad (41)$$

Note that the uniform distribution $\mathcal{U}_{\{K\}} = \{\tilde{\Pi}_i = 1/K\}_{i=1}^K$ is a solution to (41). We will now verify that $\mathcal{U}_{\{K\}}$ satisfies (39). From (40), for $\tilde{\Pi}_{\{K\}} \leftarrow \mathcal{U}_{\{K\}}$ we have $\lambda = q(1 - 1/K)^{q-1} > 0$, which we substitute into (39):

$$\lambda - q(1 - \tilde{\Pi}_i)^{q-1} = q(1 - 1/K)^{q-1} - q(1 - 1/K)^{q-1} = 0.$$

Hence, $\mathcal{U}_{\{K\}}$ is the solution to both (39) and (41).

By the second derivative test, since $\partial^2 \mathcal{L}(\tilde{\Pi}_{\{K\}}) / \partial \tilde{\Pi}_i^2 = q(q-1)(1 - \tilde{\Pi}_i)^{q-2}$ is positive for $\tilde{\Pi}_i = 1/K$, we conclude that $Q(\mathcal{U}_{\{K\}}) \leq Q(\tilde{\Pi}_{\{K\}})$ for any $\tilde{\Pi}_{\{K\}} \neq \mathcal{U}_{\{K\}}$. This implies that $\mathbb{E}[\bar{q}]$ is maximal when $\tilde{\Pi}_{\{K\}} = \mathcal{U}_{\{K\}}$, and so is the expected reduced dimension of $\tilde{\mathbf{S}}_{\mathbf{w}}\mathbf{A}$. \square

We further note that $\mathbb{E}[\bar{q}]$ from the proof of Theorem 8, is trivially minimal in the degenerate case where $\tilde{\Pi}_\ell = 1$ for a single $\ell \in \mathbb{N}_K$, and $\tilde{\Pi}_j = 0$ for every $j \in \mathbb{N}_K \setminus \{\ell\}$. This occurs in the case where $\mathbf{A}_j = \mathbf{0}_{\tau \times d}$ for each j , and \bar{q} is therefore exactly

$$K - \sum_{j \neq \ell} (1 - \tilde{\Pi}_j)^q = K - \sum_{j \neq \ell} 1^q = K - (K - 1) = 1.$$

APPENDIX F

TABLE OF NOTATIONAL CONVENTIONS AND ℓ_2 -SUBSPACE EMBEDDING CONDITION

In this appendix, we collect the main notational conventions used in this paper, and show how to derive the ℓ_2 -subspace embedding condition (10) from (7). This definition of an ℓ_2 -subspace embedding and the corresponding condition, can also be found in [18].

Let $\mathbf{Z} = \mathbf{I}_d - (\mathbf{S}\mathbf{U})^\top(\mathbf{S}\mathbf{U})$, and $\mathbf{y} \in \mathbb{R}^d$ be an arbitrary vector. In the bounds (8) and (9), note that $\epsilon\|\mathbf{y}\|_2^2 > 0$ and $\mathbf{y}^\top\mathbf{Z}\mathbf{y} = -(\mathbf{y}^\top(-\mathbf{Z})\mathbf{y})$, so we can work with the absolute value $|\mathbf{y}^\top\mathbf{Z}\mathbf{y}|$ in order to take care of both inequalities. That is, together both imply that $0 \leq |\tilde{\mathbf{y}}^\top\mathbf{Z}\tilde{\mathbf{y}}| \leq \epsilon\|\tilde{\mathbf{y}}\|^2$, where for now we let $\tilde{\mathbf{y}}$ be a unit vector. Then, this requirement is satisfied for all vectors \mathbf{y} by scaling, which implies $\|\mathbf{Z}\| \leq \epsilon$.

REFERENCES

- [1] R. Peng and S. Vempala, "Solving Sparse Linear Systems Faster than Matrix Multiplication," in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2021, pp. 504–521.
- [2] V. V. Williams, Y. Xu, Z. Xu, and R. Zhou, "New Bounds for Matrix Multiplication: from Alpha to Omega," in *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2024, pp. 3792–3835.
- [3] M. W. Mahoney, L. Lim, and G. E. Carlsson, "Algorithmic and Statistical Challenges in Modern Large-Scale Data Analysis are the Focus of MMDS 2008," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 2, pp. 57–60, 2008.
- [4] M. W. Mahoney, "Algorithmic and Statistical Perspectives on Large-Scale Data Analysis," *Combinatorial Scientific Computing*, pp. 427–469, 2012.
- [5] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Contemp. Math.*, vol. 26, 1984, pp. 189–206.
- [6] S. Dasgupta and A. Gupta, "An Elementary Proof of a Theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.
- [7] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized Dimensionality Reduction for k -means Clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, 2014.
- [8] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu, "Dimensionality Reduction for k -Means Clustering and Low Rank Approximation," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 2015, pp. 163–172.
- [9] N. Charalambides, "Dimensionality Reduction for k -means Clustering," *arXiv preprint arXiv:2007.13185*, 2020.
- [10] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering in Large Graphs and Matrices," in *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, 1999, pp. 291–299.
- [11] —, "Clustering Large Graphs via the Singular Value Decomposition," *Machine learning*, vol. 56, no. 1-3, pp. 9–33, 2004.
- [12] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication," *SIAM Journal on Computing*, vol. 36, no. 1, pp. 132–157, 2006.
- [13] —, "Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix," *SIAM Journal on computing*, vol. 36, no. 1, pp. 158–183, 2006.
- [14] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.
- [15] M. W. Mahoney, "Lecture Notes on Randomized Linear Algebra," *arXiv preprint arXiv:1608.04481*, 2016.
- [16] C. Musco and C. Musco, "Projection-Cost-Preserving Sketches: Proof Strategies and Constructions," *arXiv preprint arXiv:2004.08434*, 2020.
- [17] S. Wang, "A practical guide to randomized matrix computations with matlab implementations," *arXiv preprint arXiv:1505.07570*, 2015.
- [18] D. P. Woodruff, "Sketching as a Tool for Numerical Linear Algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.
- [19] P. Drineas and M. W. Mahoney, "RandNLA: Randomized Numerical Linear Algebra," *Communications of the ACM*, vol. 59, no. 6, pp. 80–90, 2016.
- [20] R. Kannan and S. Vempala, "Randomized algorithms in numerical linear algebra," *Acta Numerica*, vol. 26, pp. 95–135, 2017.
- [21] R. Murray, J. Demmel, M. W. Mahoney, N. B. Erichson, M. Melnichenko, O. A. Malik, L. Grigori, P. Luszczek, M. Dereziński, M. E. Lopes *et al.*, "Randomized Numerical Linear Algebra: A Perspective on the Field With an Eye to Software," *arXiv preprint arXiv:2302.11474*, 2023.
- [22] T. Elfving, "Block-iterative methods for consistent and inconsistent linear equations," *Numerische Mathematik*, vol. 35, no. 1, pp. 1–12, 1980.
- [23] M. H. Gutknecht, "Block Krylov Space Methods for Linear Systems With Multiple Right-hand Sides: An Introduction," *Modern Mathematical Models, Methods and Algorithms for Real World Systems*, 2006.
- [24] D. Needell and J. A. Tropp, "Paved with good intentions: analysis of a randomized block Kaczmarz method," *Linear Algebra and its Applications*, vol. 441, pp. 199–221, 2014.
- [25] E. Rebrova and D. Needell, "On block Gaussian sketching for the Kaczmarz method," *Numerical Algorithms*, pp. 1–31, 2020.
- [26] M. Pilanci and M. J. Wainwright, "Iterative Hessian Sketch: Fast and Accurate Solution Approximation for Constrained Least-Squares," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1842–1879, 2016.
- [27] —, "Newton Sketch: A Linear-time Optimization Algorithm with Linear-Quadratic Convergence," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 205–245, 2017.
- [28] J. Lacotte, S. Liu, E. Dobriban, and M. Pilanci, "Optimal Iterative Sketching with the Subsampled Randomized Hadamard Transform," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [29] T. Sarlós, "Improved Approximation Algorithms for Large Matrices via Random Projections," in *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*. IEEE, 2006, pp. 143–152.
- [30] U. Oswal, S. Jain, K. S. Xu, and B. Eriksson, "Block CUR: Decomposing Matrices Using Groups of Columns," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 360–376.
- [31] N. Charalambides, M. Pilanci, and A. O. Hero, "Approximate Weighted CR Coded Matrix Multiplication," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5095–5099.
- [32] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding Up Distributed Machine Learning Using Codes," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1514–1529, 2017.
- [33] B. Bartan and M. Pilanci, "Distributed sketching for randomized optimization: Exact characterization, concentration, and lower bounds," *IEEE Transactions on Information Theory*, vol. 69, no. 6, pp. 3850–3879, 2023.
- [34] N. Charalambides, M. Pilanci, and A. O. Hero, "Weighted Gradient Coding with Leverage Score Sampling," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5215–5219.
- [35] V. Gupta, S. Wang, T. Courtade, and K. Ramchandran, "OverSketch: Approximate Matrix Multiplication for the Cloud," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 298–304.
- [36] V. Gupta, S. Kadhe, T. Courtade, M. W. Mahoney, and K. Ramchandran, "OverSketched Newton: Fast Convex Optimization for Serverless Systems," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 288–297.

LIST OF NOTATION	
m	number of computational nodes
q	number of sampled blocks/responsive nodes
\mathbb{N}_n	set of integers $\{1, 2, \dots, n\}$
$[s]$	used as a superscript or subscript, to denote the iteration number 's' of the iterative process
$X_{\{n\}}$	set $\{X_1, \dots, X_n\}$
\mathbf{A}	dataset matrix of size $N \times d$
$\mathbf{A}_{(i)}/\mathbf{A}^{(i)}$	i^{th} row/column of \mathbf{A}
\mathbf{b}	corresponding $d \times 1$ label vector to \mathbf{A}
\mathbf{S}	arbitrary sketching matrix of size $r \times N$
$\tilde{\mathbf{S}}$	sketching matrix of Algorithm 1
Ω	row sampling matrix of size $r \times N$
$\tilde{\Omega}$	block sampling matrix for \mathbf{S}
\mathcal{D}	dataset corresponding to \mathbf{A}, \mathbf{b}
\mathcal{D}_i	sub-dataset corresponding to $\mathbf{A}_i, \mathbf{b}_i$
N	number of data points in \mathcal{D}
K	number of partitions of the data matrix \mathbf{A}
τ	size of the data partitions, $\tau = N/K$
$\tilde{\mathbf{A}} = [\tilde{A}_1^\top \dots \tilde{A}_K^\top]^\top$	encoding of blocks $\mathbf{A}_{\{K\}}$
$\tilde{\mathbf{b}} = [\tilde{b}_1^\top \dots \tilde{b}_K^\top]^\top$	encoding of blocks $\mathbf{b}_{\{K\}}$
$L_{ls}(\mathbf{A}, \mathbf{b}; \mathbf{x})$	least squares objective function $\ \mathbf{A}\mathbf{x} - \mathbf{b}\ _2^2$
$L_S(\mathbf{S}, \mathbf{A}, \mathbf{b}; \mathbf{x})$	modified least squares $\ S(\mathbf{A}\mathbf{x} - \mathbf{b})\ _2^2$, for S a specified encoding or sketching matrix
\mathbf{x}^*	optimal solution to (2)
$\hat{\mathbf{x}}$	approximated solution to (2) through (12)
$\tilde{\mathbf{x}}$	optimal solution to (14)
$\mathcal{T}^{[s]}$	index set of the q fastest servers at iteration s
$\mathcal{S}^{[s]}$	index multiset of sampled blocks at iteration s
$g^{[s]}$	gradient $g^{[s]} = 2\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{[s]} - \mathbf{b})$ (3)
$g_i^{[s]}$	partial gradient corresponding to \mathcal{D}_i (4)
$\hat{g}^{[s]}$	approximated gradient at iteration s (24)
$\hat{g}_i^{[s]}$	appr. partial gradient of \mathcal{D}_i at iteration s (28)
ξ_s	SD step-size at iteration s
ϵ	ℓ_2 -subspace embedding error
δ	ℓ_2 -subspace embedding failure probability
$\Pi_{\{K\}}$	block leverage score sampling distribution
$\tilde{\Pi}_{\{K\}}$	approximate distribution to $\Pi_{\{K\}}$
$\bar{\Pi}_{\{K\}}$	induced sampling distribution through expansion networks
$\beta_{\bar{\Pi}}$	misestimation factor of $\bar{\Pi}$ to Π
$d_{\Pi, Q}$	distortion metric between $\Pi_{\{K\}}$ and $Q_{\{K\}}$
$\Delta_{\Pi, \bar{\Pi}}$	normalized distortion metric (19)
$F(t)$	mother runtime distribution of the computational network
$\tilde{F}(t)$	runtime distribution of computational tasks
$\phi(t)$	survival function $\phi(t) := 1 - F(t)$
T	ending time, at which the central server stops receiving computations
$q(T)$	number of received computations at time T , i.e., $q(T) := \lfloor \tilde{F}(T) \cdot m \rfloor$
$\hat{r}_{\{K\}}$	approximate replication numbers through (20)
$r_{\{K\}}^*$	optimal replication numbers to (22)
$\tilde{r}_{\{K\}}$	approximate replication numbers
$r_{\{K\}}$	replication numbers produced by Algorithm 2
R	$R = \sum_{i=1}^K r_i$
\mathbf{E}	expansion matrix (21)
\mathbf{G}	encoding generator matrix (25)

TABLE IV: Common notation list used throughout the paper.

- [37] T. Jahani-Nezhad and M. A. Maddah-Ali, "CodedSketch: A Coding Scheme for Distributed Computation of Approximated Matrix Multiplication," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4185–4196, 2021.
- [38] B. Tegin, E. Hernandez, S. Rini, T. M. Duman *et al.*, "Straggler mitigation through unequal error protection for distributed approximate matrix multiplication," *arXiv preprint arXiv:2103.02928*, 2021.
- [39] M. Rudow, N. Charalambides, A. O. Hero III, and K. Rashmi, "Compression-Informed Coded Computing," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 2177–2182.
- [40] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica, "Improving MapReduce Performance in Heterogeneous Environments," in *USENIX Association*, ser. OSDI'08, 2008, p. 29–42.
- [41] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient Coding: Avoiding Stragglers in Distributed Learning," in *International Conference on Machine Learning*, 2017, pp. 3368–3376.
- [42] N. Charalambides, H. Mahdaviifar, and A. O. Hero, "Numerically Stable Binary Gradient Coding," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2622–2627.
- [43] N. Charalambides, H. Mahdaviifar, and A. O. Hero III, "Generalized Fractional Repetition Codes for Binary Coded Computations," *arXiv preprint arXiv:2109.10484*, 2021.
- [44] S. Li and S. Avestimehr, "Coded Computing," *Foundations and Trends® in Communications and Information Theory*, vol. 17, no. 1, pp. 1–148, 2020. [Online]. Available: <http://dx.doi.org/10.1561/0100000103>
- [45] J. S. Ng, W. Y. B. Lim, N. C. Luong, Z. Xiong, A. Asheralieva, D. Niyato, C. Leung, and C. Miao, "A Comprehensive Survey on Coded Distributed Computing: Fundamentals, Challenges, and Networking Applications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1800–1837, 2021.
- [46] D. Papailiopoulos, A. Kyriillidis, and C. Boutsidis, "Provable Deterministic Leverage Score Sampling," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 997–1006.
- [47] N. S. Ferdinand and S. C. Draper, "Anytime Coding for Distributed Computation," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 954–960.
- [48] T. Jahani-Nezhad and M. A. Maddah-Ali, "CodedSketch: Coded Distributed Computation of Approximated Matrix Multiplication," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 2489–2493.
- [49] P. Drine and R. Kannan, "Fast Monte-Carlo Algorithms for Approximate Matrix Multiplication," in *Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, 2001, pp. 452–459.
- [50] M. Charikar, K. Chen, and M. Farach-Colton, "Finding Frequent Items in Data Streams," *Theoretical Computer Science*, vol. 312, no. 1, pp. 3–15, 2004.
- [51] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Straggler Mitigation in Distributed Matrix Multiplication: Fundamental Limits and Optimal Coding," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1920–1933, 2020.
- [52] N. Raviv, I. Tamo, R. Tandon, and A. G. Dimakis, "Gradient Coding from Cyclic MDS Codes and Expander Graphs," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7475–7489, 2020.
- [53] R. Bitar, M. Wootters, and S. El Rouayheb, "Stochastic Gradient Coding for Straggler Mitigation in Distributed Learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, pp. 277–291, 2020.
- [54] S. Kadhe, O. O. Koyluoglu, and K. Ramchandran, "Gradient Coding Based on Block Designs for Mitigating Adversarial Stragglers," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 2813–2817.
- [55] A. Sakorikar and L. Wang, "Soft BIBD and Product Gradient Codes," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 2, pp. 229–240, 2022.
- [56] N. Charalambides, H. Mahdaviifar, M. Pilanci, and A. O. Hero, "Orthonormal Sketches for Secure Coded Regression," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 826–831.
- [57] N. Charalambides, H. Mahdaviifar, M. Pilanci, and A. O. Hero III, "Iterative Sketching for Secure Coded Regression," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 148–161, 2024.
- [58] Z. Charles, D. Papailiopoulos, and J. Ellenberg, "Approximate gradient coding via sparse random graphs," *arXiv preprint arXiv:1711.06771*, 2017.
- [59] M. Glasgow and M. Wootters, "Approximate Gradient Coding with Optimal Decoding," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 855–866, 2021.
- [60] Z. Charles and D. Papailiopoulos, "Gradient Coding Using the Stochastic Block Model," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1998–2002.
- [61] H. Wang, Z. Charles, and D. Papailiopoulos, "ErasureHead: Distributed Gradient Descent without Delays Using Approximate Gradient Coding," *arXiv preprint arXiv:1901.09671*, 2019.
- [62] S. Wang, J. Liu, and N. Shroff, "Fundamental Limits of Approximate Gradient Coding," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 3, pp. 1–22, 2019.
- [63] L. Chen, H. Wang, Z. Charles, and D. Papailiopoulos, "DRACO: Byzantine-resilient Distributed Training via Redundant Gradients," *arXiv preprint arXiv:1803.09877*, 2018.
- [64] S. Horii, T. Yoshida, M. Kobayashi, and T. Matsushima, "Distributed Stochastic Gradient Descent Using LDGM Codes," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1417–1421.
- [65] R. K. Maity, A. Singh Rawa, and A. Mazumdar, "Robust Gradient Descent via Moment Encoding and LDPC Codes," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2734–2738.
- [66] S. Dutta, M. Fahim, F. Haddadpour, H. Jeong, V. Cadambe, and P. Grover, "On the Optimal Recovery Threshold of Coded Matrix Multiplication," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 278–301, 2020.
- [67] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951. [Online]. Available: <https://doi.org/10.1214/aoms/1177729586>
- [68] S. Dutta, V. Cadambe, and P. Grover, "Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products," in *Advances In Neural Information Processing Systems*, 2016, pp. 2100–2108.
- [69] F. Haddadpour and V. R. Cadambe, "Codes for Distributed Finite Alphabet Matrix-Vector Multiplication," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1625–1629.
- [70] B. Bartan and M. Pilanci, "Polar Coded Distributed Matrix Multiplication," *arXiv preprint arXiv:1901.06811*, 2019.
- [71] M. V. Jamali, M. Soleymani, and H. Mahdaviifar, "Coded Distributed Computing: Performance Limits and Code Designs," in *2019 IEEE Information Theory Workshop (ITW)*, 2019, pp. 1–5.
- [72] H. Jeong, A. Devulapalli, V. R. Cadambe, and F. P. Calmon, " ϵ -Approximate Coded Matrix Multiplication Is Nearly Twice as Efficient as Exact Multiplication," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 845–854, 2021.
- [73] S. Kiani and S. C. Draper, "Successive Approximation Coding for Distributed Matrix Multiplication," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 2, pp. 286–305, 2022.
- [74] A. M. Subramaniam, A. Heidarzadeh, and K. R. Narayanan, "Random Khatri-Rao-Product Codes for Numerically-Stable Distributed Matrix Multiplication," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 253–259.
- [75] A. B. Das, A. Ramamoorthy, and N. Vaswani, "Efficient and Robust Distributed Matrix Computations via Convolutional Coding," *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 6266–6282, 2021.
- [76] A. Ramamoorthy and L. Tang, "Numerically stable coded matrix computations via circulant and rotation matrix embeddings," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1712–1717.

- [77] A. B. Das and A. Ramamoorthy, "Coded sparse matrix computation schemes that leverage partial stragglers," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1570–1575.
- [78] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, "Faster Least Squares Approximation," *Numerische mathematik*, vol. 117, no. 2, pp. 219–249, 2011.
- [79] P. Ma, M. W. Mahoney, and B. Yu, "A Statistical Perspective on Algorithmic Leveraging," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 861–911, 2015.
- [80] P. Drineas, M. Magdon-Ismaïl, M. W. Mahoney, and D. P. Woodruff, "Fast Approximation of Matrix Coherence and Statistical Leverage," *Journal of Machine Learning Research*, vol. 13, no. Dec, pp. 3475–3506, 2012.
- [81] P. Drineas and M. W. Mahoney, "Relative-error cur matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844–881, 2008.
- [82] M. W. Mahoney and P. Drineas, " CUR matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.
- [83] N. Charalambides, "Coding Theory and Randomized Sketching for Distributed Optimization," Ph.D. dissertation, University of Michigan, 2023.
- [84] M. Fahrback, M. Ghadiri, and T. Fu, "Fast Low-Rank Tensor Decomposition by Ridge Leverage Score Sampling," *arXiv preprint arXiv:2107.10654*, 2020.
- [85] L. Ma and E. Solomonik, "Fast and Accurate Randomized Algorithms for Low-rank Tensor Decompositions," *Advances in neural information processing systems*, vol. 34, pp. 24 299–24 312, 2021.
- [86] D. Woodruff and A. Zandieh, "Leverage Score Sampling for Tensor Product Matrices in Input Sparsity Time," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 933–23 964.
- [87] X. Huang, W. Liu, and B. Du, "Matrix Completion with Noise via Leveraged Sampling," *arXiv preprint arXiv:2011.05885*, 2020.
- [88] A. Alaoui and M. W. Mahoney, "Fast Randomized Kernel Ridge Regression with Statistical Guarantees," *Advances in neural information processing systems*, vol. 28, 2015.
- [89] C. Boutsidis, P. Drineas, and M. W. Mahoney, "Unsupervised Feature Selection for the k -means Clustering Problem," in *Advances in Neural Information Processing Systems*, 2009, pp. 153–161.
- [90] D. A. Spielman and N. Srivastava, "Graph Sparsification by Effective Resistances," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1913–1926, 2011.
- [91] C. Musco and C. Musco, "Recursive Sampling for the Nystrom Method," *Advances in neural information processing systems*, vol. 30, 2017.
- [92] A. Eshragh, F. Roosta, A. Nazari, and M. W. Mahoney, "LSAR: Efficient Leverage Score Sampling Algorithm for the Analysis of Big Time Series Data," *Journal of Machine Learning Research*, vol. 23, no. 22, pp. 1–36, 2022. [Online]. Available: <http://jmlr.org/papers/v23/20-247.html>
- [93] J. Bourgain, S. Dirksen, and J. Nelson, "Toward a unified theory of sparse dimensionality reduction in euclidean space," in *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, 2015, pp. 499–508.
- [94] B. Isik, T. Weissman, and A. No, "An information-theoretic justification for model pruning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3821–3846.
- [95] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Sampling algorithms for ℓ_2 regression and applications," in *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 2006, pp. 1127–1136.
- [96] N. Ailon and B. Chazelle, "Approximate Nearest Neighbors and the Fast Johnson–Lindenstrauss Transform," in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 2006, pp. 557–563.
- [97] —, "The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors," *SIAM Journal on computing*, vol. 39, no. 1, pp. 302–322, 2009.
- [98] A. Munteanu, S. Omlor, and D. P. Woodruff, "Oblivious Sketching for Logistic Regression," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7861–7871.
- [99] M. B. Cohen, Y. T. Lee, C. Musco, C. Musco, R. Peng, and A. Sidford, "Uniform Sampling for Matrix Approximation," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, 2015, pp. 181–190.
- [100] W. Halbawi, N. Azizan, F. Salehi, and B. Hassibi, "Improving distributed gradient descent using Reed-Solomon codes," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2027–2031.
- [101] S. Bubeck, "Convex Optimization: Algorithms and Complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015. [Online]. Available: <http://dx.doi.org/10.1561/22000000050>
- [102] A. Rakhlin, O. Shamir, and K. Sridharan, "Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization," *arXiv preprint arXiv:1109.5647*, 2011.
- [103] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [104] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal Distributed Online Prediction Using Mini-Batches," *Journal of Machine Learning Research*, vol. 13, no. 1, 2012.
- [105] D. Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.
- [106] B. Bartan and M. Pilanci, "Straggler Resilient Serverless Computing Based on Polar Codes," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 276–283.
- [107] H. Cao, Q. Yan, X. Tang, and G. Han, "Adaptive Gradient Coding," *IEEE/ACM Transactions on Networking*, vol. 30, no. 2, pp. 717–734, 2022.
- [108] B. Buyukates, E. Ozfatura, S. Ulukus, and D. Gündüz, "Gradient Coding With Dynamic Clustering for Straggler-Tolerant Distributed Learning," *IEEE Transactions on Communications*, 2022.
- [109] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An Improved Approximation Algorithm for the Column Subset Selection Problem," in *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2009, pp. 968–977.
- [110] A. Civril, "Column Subset Selection Problem is UG-hard," *Journal of Computer and System Sciences*, vol. 80, no. 4, pp. 849–859, 2014.
- [111] O. Balabanov, M. Beaupère, L. Grigori, and V. Lederer, "Block subsampled randomized Hadamard transform for low-rank approximation on distributed architectures," *arXiv preprint arXiv:2210.11295*, 2022.