2018

C3LES: Codes for Coded Computation that Leverage Stragglers

Anindya B. Das
Iowa State University, abd149@iastate.edu

Aditya Ramamoorthy
Iowa State University, adityar@iastate.edu

C3LES: Codes for Coded Computation that Leverage Stragglers

Abstract
In distributed computing systems, it is well recognized that worker nodes that are slow (called stragglers) tend to dominate the overall job execution time. Coded computation utilizes concepts from erasure coding to mitigate the effect of stragglers by running "coded" copies of tasks comprising a job. Stragglers are typically treated as erasures in this process. While this is useful, there are issues with applying, e.g., MDS codes in a straightforward manner. Specifically, several applications such as matrix-vector products deal with sparse matrices. MDS codes typically require dense linear combinations of submatrices of the original matrix which destroy their inherent sparsity. This is problematic as it results in significantly higher processing times for computing the submatrix-vector products in coded computation. Furthermore, it also ignores partial computations at stragglers.
In this work, we propose a fine-grained model that quantifies the level of non-trivial coding needed to obtain the benefits of coding in matrix-vector computation. Simultaneously, it allows us to leverage partial computations performed by the straggler nodes. For this model, we propose and evaluate several code designs and discuss their properties.

Disciplines
Electrical and Computer Engineering | Systems and Communications | Theory and Algorithms

Comments
This is a pre-print of the article Das, Anindya B., Li Tang, and Aditya Ramamoorthy. "C3LES: Codes for Coded Computation that Leverage Stragglers." *arXiv preprint arXiv:1809.06242* (2018). Posted with permission.

C^3LES: Codes for Coded Computation that Leverage Stragglers

Anindya B. Das, Li Tang and Aditya Ramamoorthy
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50010, U.S.A.
{abd149,litang,adityar}@iastate.edu

***Abstract*—In distributed computing systems, it is well recognized that worker nodes that are slow (called stragglers) tend to dominate the overall job execution time. Coded computation utilizes concepts from erasure coding to mitigate the effect of stragglers by running "coded" copies of tasks comprising a job. Stragglers are typically treated as erasures in this process.**

While this is useful, there are issues with applying, e.g., MDS codes in a straightforward manner. Specifically, several applications such as matrix-vector products deal with sparse matrices. MDS codes typically require dense linear combinations of submatrices of the original matrix which destroy their inherent sparsity. This is problematic as it results in significantly higher processing times for computing the submatrix-vector products in coded computation. Furthermore, it also ignores partial computations at stragglers.

In this work, we propose a fine-grained model that quantifies the level of non-trivial coding needed to obtain the benefits of coding in matrix-vector computation. Simultaneously, it allows us to leverage partial computations performed by the straggler nodes. For this model, we propose and evaluate several code designs and discuss their properties.

***Index Terms*—Distributed computing, stragglers**

I. Introduction

Distributed computation plays a major role in several problems in machine learning. For example, large scale matrix-vector multiplication is repeatedly used in gradient descent which is typically used in high dimensional machine learning problems. The size of the underlying matrices makes it impractical to perform the computation on a single computer (both from a speed and a storage perspective). Thus, the computation is typically subdivided into smaller tasks that are run in parallel across multiple worker nodes.

In these systems the overall execution time is typically dominated by the speed of the slowest worker. Thus, the presence of stragglers (as these slow workers are called) can negatively impact the performance of distributed computation. In recent years, techniques from coding theory [1]–[4] have been used to mitigate the effect of stragglers for problems such as matrix-vector and matrix-matrix multiplication. For instance, the work of [1] proposes to partition the computation of $\mathbf{Ax}$ by first splitting $\mathbf{A}^T = [\mathbf{A}_1^T\ \mathbf{A}_2^T]^T$ into an equal number of rows and assigning three workers, the task of computing $\mathbf{A}_1\mathbf{x}$, $\mathbf{A}_2\mathbf{x}$ and $(\mathbf{A}_1 + \mathbf{A}_2)\,\mathbf{x}$, respectively. Evidently, the load on each node is half of the original job. Furthermore, it is easy to see that $\mathbf{Ax}$ can be recovered as soon as any two workers complete their tasks (with some minimal post-processing). Thus, this system is resilient to one straggler. The work of [3], poses the multiplication of two matrices in a form that is roughly equivalent to a Reed-Solomon code. In particular, each worker node's task (which is multiplying smaller submatrices) can be imagined as a coded symbol. As long as enough tasks are complete, the master node can recover the matrix product by polynomial interpolation.

For such systems we can define a so-called recovery threshold, which is defined as the minimum value of τ, such that the master node can obtain the result as long as *any* τ workers complete their tasks. Thus, at the top level, in these systems stragglers are treated as the equivalent of erasures in coding theory, i.e., the assumption is that no useful information can be obtained from the stragglers.

While these are interesting ideas, there are certain issues that are ignored in the majority of prior work (see [5]–[7] for some exceptions). Firstly, several practical cases of matrix-vector or matrix-matrix multiplication involve sparse matrices. Using MDS coding strategies in a straightforward manner will often destroy the sparsity of the matrices being processed by the worker nodes. In fact, as noted in [7], this can cause the overall job execution time to actually go up rather than down. Secondly, in the distributed computation setting, we make the observation that it is possible to leverage partial computations that are performed by the stragglers. Thus, a slow worker may not necessarily be a useless worker.

A. Main Contributions

- In this work we present a more fine-grained model of the distributed matrix-vector multiplication that allows us to (i) leverage partial computations performed by stragglers and (ii) impose constraints on the extent to which coding is allowed in the solution. Our formulation leads to some new questions in the domain of code design that to our best knowledge have not been investigated systematically in the literature before.

This work was supported in part by the National Science Foundation (NSF) under grant CCF-1718470.

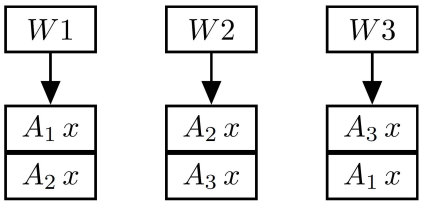

Fig. 1: Matrix A is divided into three submatrices. Each worker is assigned two of the submatrices.

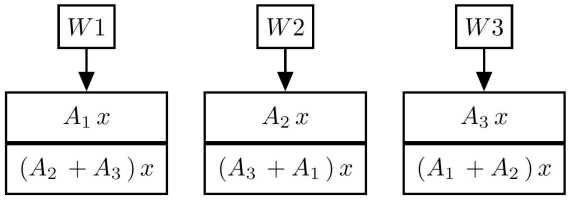

Fig. 2: Matrix A is divided into three submatrices. Each worker is assigned two submatrices one of which is coded.

- We present two models in our work. In the first model, the tasks assigned to the workers are uncoded, whereas in the second model we allow for a user specified fraction of coded tasks. In both cases, we present bounds on the amount of computation that the workers need to perform in the worst case and the straggler resilience of the system. We also present matching construction schemes in some cases. We emphasize that the uncoded model applies in general to *any* computation problem, and the bounds and constructions hold in significant generality for that case.

II. Problem Formulation

We consider a scenario where a master node has a matrix $\mathbf{A}$ and a vector $\mathbf{x}$ and needs to compute $\mathbf{Ax}$. The computation needs to be carried out in a distributed fashion over n nodes. Each node receives a certain fraction (denoted by γ) of the rows of $\mathbf{A}$ and the vector $\mathbf{x}$. The node is responsible for computing the product of its assigned submatrix and $\mathbf{x}$.

We assume that the storage fraction γ can be expressed as ℓ/Δ where both ℓ and Δ are integers. In this work, we assume that $\mathbf{A}$ is large enough so that we can choose any large enough value of Δ. Following this, we partition the rows of $\mathbf{A}$ into Δ submatrices denoted $\mathbf{A}_1, \ldots, \mathbf{A}_\Delta$; we will also refer to these as the blocks of $\mathbf{A}$. Each node is assigned the equivalent of ℓ block rows. The assigned block rows can simply be subsets of $\{\mathbf{A}_1, \ldots, \mathbf{A}_\Delta\}$; in this case we call the solution "uncoded". Alternatively, the assigned block rows can be suitably chosen functions of $\{\mathbf{A}_1, \ldots, \mathbf{A}_\Delta\}$; in this case we call the solution "coded". Each worker node processes its assigned block rows sequentially from the top to the bottom. In particular, if a node is currently processing the i-the block row ($1 \leq i \leq l$), then it has already processed blocks 1 through $i-1$. As we shall show, the processing order matters in this problem.

We assume that each time a node computes the block product (with $\mathbf{x}$) it transmits the result to the master node. We enforce the requirement that the master node should be able to recover $\mathbf{Ax}$ as long it receives *any* Q block products from the worker nodes. This formulation subsumes treating stragglers as non-working nodes. Indeed, suppose that we want a system that is resilient to s stragglers. Then, a sufficient condition would be that $Q \leq (n-s)\ell$ in our system.

Example 1. Consider a system with $n = 3$ worker nodes with $\gamma = 2/3$. We partition $\mathbf{A}$ into $\Delta = 3$ row blocks and the assignment of blocks to each node is shown in Fig. 1 (this is an uncoded solution). We emphasize that the order of the computation also matters here, i.e., worker node 1 (for example) computes $\mathbf{A}_1\mathbf{x}$ first and then $\mathbf{A}_2\mathbf{x}$. For the specific assignment it is clear that the computation is successful as long as any four block products are returned. Thus, for this system $Q = 4$.

On the other hand, Fig. 2 demonstrates a coded solution, where the assignment in the second block rows of the workers are some suitably chosen functions of the elements of $\{\mathbf{A}_1\mathbf{x}, \mathbf{A}_2\mathbf{x}, \mathbf{A}_3\mathbf{x}\}$. For this assignment, it is obvious that the master can recover $\mathbf{Ax}$ as long as any three block products are returned by the workers, so in this system $Q = 3$.

For any time t, we let $w_i(t)$ represent the state of computation of the i-th worker node, i.e., $w_i(t)$ is a positive integer between 0 and ℓ which represents the number of block rows that have been processed by worker node i. Thus, our system requirement states as long as $\sum_{i=1}^{n} w_i(t) \geq Q$, the master node should be able to determine $\mathbf{Ax}$. As Δ is a parameter that can be chosen, our objective is to minimize the value of Q/Δ for such a system. This formulation minimizes the overall computation performed by the worker nodes.

Remark 1. It is important to note that the uncoded formulation applies to *any* computation job that can be subdivided into Δ tasks. In particular, the structure of matrix-vector multiplication is not important in this context. Thus, the discussion in the subsequent sections for the uncoded setup applies in significantly more generality.

III. Uncoded Scheme

The advantage of an uncoded scheme is that it does not require any computation from the master to recover $\mathbf{Ax}$ because the assignment to any worker is simply a subset of $\{\mathbf{A}_1\mathbf{x}, \mathbf{A}_2\mathbf{x}, \ldots, \mathbf{A}_\Delta\mathbf{x}\}$.

To avoid trivialities, we emphasize that each worker node only contains at most one copy of each block row. In what follows, we use r to represent the replication factor of each $\mathbf{A}_i$. Thus, each $\mathbf{A}_i$ appears r times across all worker nodes. We use the notation $\langle n, \ell, \Delta, r\rangle$-uncoded system to represent an uncoded system with the corresponding parameters.

Theorem 1. Consider an $\langle n, \ell, \Delta, r\rangle$-uncoded system. If the system needs to be resilient to s stragglers, then $r \geq s+1$ and $n\gamma = r$.

Proof. It is evident that $r \geq s+1$ since we need at least one copy of each block row to be present even in the presence

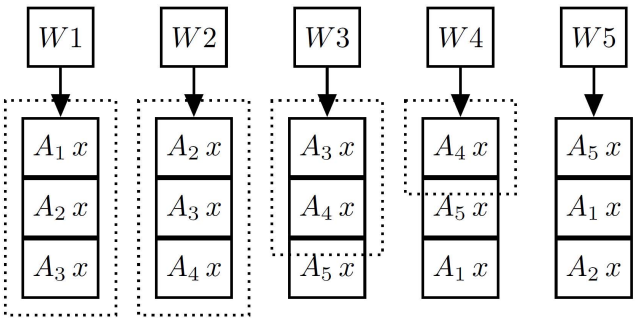

Fig. 3: A $\langle n, \ell, \Delta, r\rangle = \langle 5, 3, 5, 3\rangle$-uncoded system designed using Algorithm 1. The dotted blocks show the blocks that can be processed in the worst case without processing A_5.

of s stragglers. Next, the total number of symbols across all nodes equals Δr. A simple double counting argument, yields the equality $n\ell = \Delta r$ which further implies that $n\gamma = r$ by the definition of γ. ∎

Theorem 2. Consider an $\langle n, \ell, \Delta, r\rangle$-uncoded system. Then, $Q \geq \max(\Delta, \Delta r - \frac{r}{2}(l+1) + 1)$.

Proof. For the system under consideration the master node requires each block product, $\mathbf{A}_j\mathbf{x}$ where $j = 1, 2, ..., \Delta$ to be computed at least once by the n worker nodes. It is evident that the system needs to process at least Δ blocks, so that $Q \geq \Delta$. Let Q_j represent the maximum number of block rows that are processed in the worst case without obtaining $\mathbf{A}_j\mathbf{x}$ (see Fig. 3 for an example). It is evident in this case that $Q = \max_{j=1,...,\Delta} Q_j + 1$.

Our strategy is to calculate the average $\overline{Q} = \sum_{j=1}^{\Delta} Q_j/\Delta$ and use the simple bound $Q \geq \overline{Q} + 1$. Toward this end, note that for any uncoded solution, we can calculate $\sum_{j=1}^{\Delta} Q_j$ in a different way. For a worker i, there are ℓ assigned block rows and $\Delta - \ell$ do not appear in it. Thus, in the calculation of $\sum_{j=1}^{\Delta} Q_j$, worker node i contributes

$$\sum_{k=1}^{\ell}(k-1) + (\Delta - \ell)\ell,$$

which is clearly independent of i. Therefore,

$$\overline{Q} = n\left\lceil \frac{\sum_{k=1}^{\ell}(k-1) + (\Delta-\ell)\ell}{\Delta} \right\rceil$$
$$= n\ell - \frac{n\ell}{2\Delta}(\ell+1) \;=\; \Delta r - \frac{r}{2}(\ell+1),$$

where we used $n\ell = \Delta r$ in the last step above. ∎

Thus, Theorem 1 gives an upper bound on the number of stragglers that a system tolerates and Theorem 2 provides a lower bound on Q. Both these results can be treated as benchmarks for an uncoded scheme. We now propose a construction (see Algorithm 1) which meets both these bounds. The basic idea in Algorithm 1 is to set $\Delta = n$ and place the block rows in a cyclic fashion (see Fig. 3).

Algorithm 1: Cyclic Uncoded Scheme

Input : Matrix $\mathbf{A}$ and vector $\mathbf{x}$, n-number of worker nodes, replication factor r.

Set $\Delta = n$ and $\ell = r$. Partition $\mathbf{A}$ into Δ block rows $\mathbf{A}_1, \ldots, \mathbf{A}_\Delta$;

for $i \leftarrow 1$ **to** n **do**

  Assign $\mathbf{A}_i, \mathbf{A}_{i+1}, \ldots, \mathbf{A}_{i+\ell-1}$ from top to bottom (subscripts reduced modulo Δ) to worker node i

end

Output: $\langle n, \ell, \Delta, r\rangle$ uncoded system with stragger resilience $r-1$ and optimal Q/Δ.

Theorem 3. The Cyclic Uncoded Scheme (*cf.* Algorithm 1) is resilient to $(r-1)$ stragglers and meets the bound in Theorem 2.

Proof. It is evident from the cyclic nature of the construction that each block row appears r times in r different worker nodes. Thus, the scheme is resilient to $r-1$ stragglers. Next, we show that Q_j (defined in the proof of Theorem 2) in this construction is the same for all block rows $\mathbf{A}_j$, $j = 1, \ldots, \Delta$. To see this we note that the maximum number of block rows that can be processed without processing $\mathbf{A}_j$ can be calculated as follows.

- We can process the $(i-1)$ rows in worker node $[(j-i) \mod n] + 1$ for $i = 1, \ldots, r$.
- All the block rows in other $n-r$ workers can be processed as well.

Thus, we have

$$Q_j = [1 + 2 + 3 + ... + (\ell-1)] + (n-r)\ell$$
$$= \frac{(\ell-1)\ell}{2} + n\ell - \ell^2 \quad (\text{as } r = \ell)$$
$$= n\ell - \frac{\ell}{2}(\ell+1).$$

Again, using the fact that $n = \Delta$ and $\ell = r$, we have

$$Q_j = \Delta r - \frac{r}{2}(\ell+1).$$

Thus, Q_j is independent of j and therefore $Q = \Delta r - \frac{r}{2}(\ell + 1) + 1$. ∎

Example 2. As an example, consider Fig. 3, where we have $\Delta = n = 5$ and $\ell = r = 3$. The scheme is resilient to $(r-1) = 2$ stragglers and it can be verified that $\mathbf{Ax}$ can be computed once any $Q = 10$ block rows have been processed.

IV. Coded Scheme

We now explore coded schemes in our setting. As demonstrated in Example 1, the value of Q in the coded scenario can be strictly lesser than in the uncoded case. A simple solution to this problem is to use MDS (maximum distance separable) codes as was done in some of the inital papers [1] in this

area. Namely, one could use an $(n\ell, \Delta)$-MDS code for some value of Δ and $n\ell \geq \Delta$. The block rows $\mathbf{A}_1, \ldots, \mathbf{A}_\Delta$ will be combined using the corresponding generator matrix (with an appropriate mapping from the finite field to the real field). It is clear that in this case, the master node can compute $\mathbf{Ax}$ as long as any Δ rows are processed. However, such codes may require rather dense linear combination of the rows of $\mathbf{A}$ and this may incur a significant overhead in the computation performed by the worker nodes, especially in the practical case when $\mathbf{A}$ is sparse to begin with.

Accordingly, in this part of the work we are interested in examining schemes where we can specify the fraction of coded blocks. We now assume that each node receives a $\gamma = \gamma_u + \gamma_c$ fraction of the rows of $\mathbf{A}$, where γ_u and γ_c correspond to the uncoded and coded parts respectively. The replication factor for the uncoded portion is r_u and the number of uncoded block assignments in a worker is denoted as $\ell_u = \Delta\gamma_u$. This implies that $n\ell_u = \Delta r_u$. We let $\ell_c = \Delta\gamma_c$ represent the number of coded blocks in each worker.

For our bounds we assume that each processed coded block is useful to the master node. This can be ensured by our choice of coding coefficients that are obtained from Cauchy matrices of appropriate dimension. In what follows, we consider two different schemes. In one case the coded blocks appear at the bottom of each node while in the other case the coded blocks appear at the top. The first case leans towards easier decoding by the master node, whereas the second one aims to minimize the computation performed by the worker nodes. We use the notation $\langle n, \ell_u, \ell_c, \Delta, r_u\rangle$-bottom and $\langle n, \ell_u, \ell_c, \Delta, r_u\rangle$-top to refer to the corresponding systems. The value of Q for these systems will be denoted by Q_{cb} and Q_{ct} respectively (the subscripts are self-explanatory).

A. Coded blocks at the bottom

In this case, the results of Section III (See Theorem 2) immediately imply that $Q_{cb} \geq \max(\Delta, \Delta r_u - \frac{r_u}{2}(\ell+1)+1)$ (the subscript denotes that the coded blocks appear at the bottom). This follows by applying the previous arguments to the uncoded part of the solution. A construction that meets these bounds is outlined in Algorithm 2. The algorithm uses a Cauchy matrix of dimension $n\ell_c \times \Delta$.

Theorem 4. The scheme in Algorithm 2 satisfies $Q_{cb} = \max(\Delta, \Delta r_u - \frac{r_u}{2}(\ell_u+1)+1)$. Furthermore, it is resilient to $\left\lfloor \frac{n^2\gamma_c + n\gamma_u - 1}{n\gamma_c + 1} \right\rfloor$ stragglers.

Proof. We need to show that for any pattern of Q_{cb} blocks the master node can decode $\mathbf{Ax}$. Towards this end, from the discussion in Section III, we know that any pattern of Q_{cb} uncoded blocks allows the recovery of Δ distinct blocks. In other words for any computation state vector $\mathbf{w}(t) = [w_1(t)\ w_2(t)\ \ldots\ w_n(t)]$ such that $w_i(t) \leq \ell_u$ and $\sum_{i=1}^n w_i(t) \geq Q_{cb}$ the master node can decode. Now, consider a vector $\mathbf{w}'(t)$ such that (w.l.o.g.) $w'_1(t) \ldots w'_\alpha(t) \geq \ell_u + 1$ and $w'_{\alpha+1}(t), \ldots, w'_n(t) \leq \ell_u$ and $\sum_{i=1}^n w'_i(t) \geq Q_{cb}$, i.e., the first α worker nodes process coded blocks whereas the others do not. It is not too hard to determine a different vector $\tilde{\mathbf{w}}(t)$ with the following properties.

$$\tilde{w}_i(t) = \begin{cases} \ell_u & 1 \leq i \leq \alpha, \\ w'_i(t) + \beta_i & \alpha+1 \leq i \leq n, \end{cases}$$

where β_i's are positive integers such that $w'_i(t) + \beta_i \leq \ell_u$ and $\sum_{i=1}^n \tilde{w}_i(t) = Q_{cb}$. Thus, $\tilde{\mathbf{w}}(t)$ corresponds to a pattern of Q_{cb} uncoded blocks that recovers Δ distinct blocks.

Now, we compare the vectors $\mathbf{w}'(t)$ and $\tilde{\mathbf{w}}(t)$. Let the uncoded symbols in $\mathbf{w}'(t)$ be denoted by the set $\mathcal{A}$. Then the set of uncoded symbols in $\tilde{\mathbf{w}}(t)$ can be expressed as $\mathcal{A} \cup \mathcal{B}$ where the set $\mathcal{B}$ results from the transformation above. It is evident that for computation state vector $\mathbf{w}'(t)$ the master node has $\sum_{i=1}^{\alpha} (w'_i(t) - \ell_u)$ equations with $\Delta - |\mathcal{A}|$ variables. Now,

$$\sum_{i=1}^{\alpha} (w'_i(t) - \ell_u) \geq |\mathcal{B}| \geq |\mathcal{B} \setminus \mathcal{A}| = \Delta - |\mathcal{A}|.$$

In particular, this establishes that we have at least as many equations as variables. As, any square submatrix of a Cauchy matrix is invertible, we have the required result.

To establish the straggler resilience of our construction we identify the set of worker nodes that contain the least number of uncoded blocks. With some work (the details appear in the full version of the paper) it can be established that any k workers have at least $\min(\ell_u + k - 1, \Delta)$ uncoded blocks. For our construction, these correspond to picking k consecutive worker nodes. Thus, we are trying to determine the minimum value of k such that

$$\ell_u + (k-1) + k(\ell - \ell_u) \geq \Delta$$

which further implies

$$k \geq \frac{n - \ell_u + 1}{\ell - \ell_u + 1} = \frac{n - n\gamma_u + 1}{n\gamma - n\gamma_u + 1}$$

Algorithm 2: Cyclic Coded at Bottom Scheme

Input : Matrix $\mathbf{A}$ and vector $\mathbf{x}$, n-number of worker nodes, total storage capacity fraction γ, replication factor for uncoded portion r_u.

Set $\Delta = n$, $\ell_u = r_u$, $\ell = \gamma\Delta$, $\ell_c = \ell - \ell_u$. Determine a Cauchy matrix $\mathcal{C}$ of dimension $n\ell_c \times \Delta$;
Partition $\mathbf{A}$ into Δ block rows $\mathbf{A}_1, \ldots, \mathbf{A}_\Delta$;
for $i \leftarrow 1$ **to** n **do**
  Assign $\mathbf{A}_i, \mathbf{A}_{i+1}, \ldots, \mathbf{A}_{i+\ell_u-1}$ from top to bottom (subscripts reduced modulo Δ) to worker node i;
  for $j \leftarrow 1$ **to** ℓ_c **do**
    Define $T = \{i, i+1, \ldots, i+\ell_u-1\} \mod \Delta$;
    Pick a row $\mathbf{c}$ of $\mathcal{C}$ and assign coded block $\sum_{k=1}^{\Delta} \mathbf{c}_k \mathbb{1}_{k \notin T} \mathbf{A}_k$;
    Remove $\mathbf{c}$ from $\mathcal{C}$, so $\mathcal{C} \leftarrow \mathcal{C} \backslash \{\mathbf{c}\}$;
  end
end

Output: $\langle n, \ell_u, \ell_c, \Delta, r_u\rangle$-bottom system.

as $n = \Delta$. So, if the system is resilient to s stragglers then

$$s \leq \left\lfloor n - \frac{n - n\gamma_u + 1}{n\gamma - n\gamma_u + 1} \right\rfloor = \left\lfloor \frac{n^2\gamma_c + n\gamma_u - 1}{n\gamma_c + 1} \right\rfloor.$$

■

Remark 2. We emphasize that the construction in Algorithm 2 is not isomorphic to a MDS code, even if $r_u = 1$.

Example 3. Consider again the setting of Example 2 where $\Delta = n = 5$, $\gamma = \frac{3}{5}$. Suppose that we set $\gamma_u = \frac{2}{5}$ and $\gamma_c = \frac{1}{5}$. This scheme is resilient to $\left\lfloor \frac{n^2\gamma_c + n\gamma_u - 1}{n\gamma_c + 1} \right\rfloor = 3$ stragglers and it can be verified that $\mathbf{Ax}$ can be computed once any $Q = \Delta r_u - \frac{r_u}{2}(\ell_u + 1) + 1 = 8$ block rows have been processed. Thus, we can conclude that introducing a single coded block in each worker (at the bottom), helps to improve both Q and the straggler resilience of the system as compared to an uncoded system.

B. Coded blocks at the top

The situation is quite different when we consider the placement of the coded blocks at the top of the worker nodes. In this case a given worker node only processes uncoded blocks after having processed ℓ_c coded blocks. Thus, if x coded blocks have been processed by the worker nodes before starting work on the uncoded blocks, it suffices if *any* $\Delta - x$ distinct uncoded blocks are processed. This weaker requirement allows us to potentially improve the Q_{ct}/Δ ratio as compared to the previous constructions.

We now develop a lower bound on Q_{ct}. Suppose that we consider an arbitrary set of β worker nodes that process all their blocks and another set of worker nodes that only contribute x coded blocks. Evidently, in this case the total number of coded blocks is $x + \ell_c\beta$. Let $\mathcal{A}$ denote the set of distinct uncoded blocks obtained from the chosen set of β worker nodes. We note that

$$Q_{ct} \geq x + \ell\beta + 1, \text{ when}$$
$$x + \ell_c\beta + |\mathcal{A}| < \Delta.$$

This is because, we do not have enough equations to decode the $\Delta - |\mathcal{A}|$ unknowns.

Next, we use another averaging argument. We calculate the average size of $\mathcal{A}$ when considering all possible $\binom{n}{\beta}$ worker nodes via a double counting argument.

Consider a bipartite graph G, whose vertex set is $\mathcal{U} \cup \mathcal{V}$. Each element of $\mathcal{U}$ is the set of uncoded blocks contained in a particular set of β workers. Thus, the cardinality of $\mathcal{U}$ is $|\mathcal{U}| = \binom{n}{\beta}$. The set $\mathcal{V}$ is the set of all possible uncoded blocks so that $|\mathcal{V}| = \Delta$. There is an edge between $u \in \mathcal{U}$ and $v \in \mathcal{V}$ if $v \in u$. The degree of $v \in \mathcal{V}$ in G can be computed by observing that there are $\binom{n-r_u}{\beta}$ sets that do not contain v. Therefore, the degree of v is $\binom{n}{\beta} - \binom{n-r_u}{\beta}$. Thus, the average degree of the nodes in $\mathcal{U}$ is given by

$$\bar{d}_{\mathcal{U}} = \Delta \times \left[1 - \frac{\binom{n-r_u}{\beta}}{\binom{n}{\beta}}\right].$$

It follows that if

$$x + \ell_c\beta + \bar{d}_{\mathcal{U}} < \Delta,$$

there is at least one choice of β worker nodes that will not allow the decoding of $\mathbf{Ax}$. Our lower bound on Q_{ct} can be derived by solving the following optimization problem.

$$\begin{aligned} &\text{maximize} \quad x + \ell\beta + 1 \\ &\text{subject to} \quad (x + \ell_c\beta) < \Delta \left[\frac{\binom{n-r_u}{\beta}}{\binom{n}{\beta}}\right]. \end{aligned} \tag{1}$$

Example 4. We can consider a $\langle n, \ell_u, \ell_c, \Delta, r_u\rangle = \langle 15, 3, 1, 15, 3\rangle$-top system and derive the bound $Q_{ct} \geq 18 > \Delta = 15$ by solving the optimization problem in (1). The optimal setting turns out to be $x = 1$ and $\beta = 4$.

Example 5. Continuing our discussion of the setting in Examples 2 and 3, if we assign the coded blocks at the top rows of the workers and apply cyclic scheme for the uncoded portion, we can show that $Q_{ct} = 6$ while being resilient to $s = 3$ stragglers. Thus, moving the coded rows to the top provides the best construction in terms of Q and straggler resilience.

V. Conclusion

In this paper we have formulated a new model for distributed coded matrix-vector multiplication. Our model allows us to leverage partial work performed by stragglers while controlling the level to which coding is utilized in the solution. We propose lower bounds on the required computation from the worker nodes in the worst case and present matching constructions in certain cases. Our proposed model demonstrates that the ordering of the computations within different worker nodes plays an important role in the overall job execution time. This in turn leads to new (to our best knowledge) code design problems that should be interesting to investigate.

References

[1] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. on Info. Th.*, vol. 64, no. 3, pp. 1514–1529, 2018.

[2] S. Dutta, V. Cadambe, and P. Grover, "Short-dot: Computing large linear transforms distributedly using coded short dot products," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2100–2108.

[3] Q. Yu, M. Maddah-Ali, and S. Avestimehr, "Polynomial codes: an optimal design for high-dimensional coded matrix multiplication," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 4403–4413.

[4] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *Proc. of Intl. Conf. on Machine Learning (ICML)*, 2017.

[5] S. Kiani, N. Ferdinand, and S. C. Draper, "Exploitation of stragglers in coded computation," in *IEEE Intl. Symposium on Info. Th.*, 2018, pp. 1988–1992.

[6] A. Mallick, M. Chaudhari, and G. Joshi, "Rateless codes for near-perfect load balancing in distributed matrix-vector multiplication," preprint, 2018, [Online] Available: https://arxiv.org/abs/1804.10331.

[7] S. Wang, J. Liu, and N. Shroff, "Coded sparse matrix multiplication," in *Proc. of Intl. Conf. on Machine Learning (ICML)*, 2018.