Coded Elastic Computing

Yaoqing Yang, Pulkit Grover, Soummya Kar
Carnegie Mellon University
{yyaoqing, pgrover, soummyak}@andrew.cmu.edu

Matteo Interlandi, Saeed Amizadeh, Markus Weimer
Microsoft
{mainterl, saamizad, mweimer}@microsoft.com

Abstract

Cloud providers have recently introduced new offerings whereby spare computing resources are accessible at discounts compared to on-demand computing. Exploiting such opportunity is challenging inasmuch as such resources are accessed with low-priority and therefore can elastically leave (through preemption) and join the computation at any time. In this paper, we design a new technique called coded elastic computing enabling distributed computations over elastic resources. The proposed technique allows machines to leave the computation without sacrificing the algorithm-level performance, and, at the same time, adaptively reduce the workload at existing machines when new ones join the computation. Leveraging coded redundancy, our approach is able to achieve similar computational cost as the original (uncoded) method when all machines are present; the cost gracefully increases when machines are preempted and reduces when machines join. The performance of the proposed technique is evaluated on matrix-vector multiplication and linear regression tasks. In experimental validations, it can achieve exactly the same numerical result as the noiseless computation, while reducing the computation time by 46% when compared to non-adaptive coding schemes.

1 Introduction

New offerings from cloud-service providers allow exploiting under-utilized Virtual Machines (VMs) at a fraction of the original cost [1,3]. For example, Azure Batch provides low-priority machines at about one-fifth of the cost of ordinary virtual machines [4]. Similarly, Amazon Spot Instances provide machines at market price with a discount which can reach up to 90% with respect to regular on-demand prices [2]. Such offerings, however, have the drawback that machines can be preempted at any time if a high-priority job appears. This, in turn, will surface as a computation failure at the application level. While common distributed machine learning frameworks are already built with fault-tolerance [5,49], they often assume that failures are transient and rare. Due to this assumption, machine failures are often recovered by a “stop-the-world” scheme whereby the entire system is forced to wait until regular execution on the failure machines are restored from previous state (eventually on new machines). The above assumptions, however, do not necessarily hold for failures due to machines being preempted because (1) these failures are permanent and local data may not be accessible anymore; (2) several machines can be preempted altogether (up to 90% [46]); (3) these failures add up to transient failures, therefore leading to more frequent disruptions during computation; and (4) the computational framework may need to acquire additional machines to...
without moving data at the existing machines. We also compare with other baselines, such as ignore,
with the number of machines, which leads to linear scaling of theoretical computational cost. The
Apart from providing fault-tolerance when machines are preempted, coded elastic computing is
world scheme results in zero computation progress because, by the time a failure is recovered, a new
proposed technique in terms of the model generalization error.

In this paper, we first present a coded elastic matrix-vector multiplication algorithm to illustrate the
proposed technique in the multi-tenancy cluster at Microsoft
computation (where the number of machines remain fixed), while adaptively changing the workload
of experiments. The proposed technique achieves the exact computation result as the noiseless
main idea. Then, we present the generalizations of the proposed technique in broader applications,
among matrix-matrix multiplication, linear regression, master-free fully-distributed computing, and
training of deep neural networks (see Section 3). Finally, we validate the approach with a set
of available machines by selecting to use only a subset of the encoded data in a cyclic fashion. Apart from providing fault-tolerance when machines are preempted, coded elastic computing is
coded elastic computing framework can make the computational cost at each machine scale inversely
of machines needs to be dynamically adjusted during a learning task, such as when the number of
machines is tuned as a hyper-parameter, or when the machines have to be reallocated to achieve
fairness [33] between users or the specific need of some users at runtime.

The proposed coded elastic computing technique is tested in the multi-tenancy cluster at Microsoft
as an example of the Apache REEF EGC (Elastic Group Communication) framework. Apache
REEF is a library that helps develop distributed high-performance applications on top of cluster
resource managers such as YARN [66] and Mesos [33]. The Apache REEF project provides a set of
abstractions and reusable functional blocks to ease the process of building cloud-scale applications.
EGC is a distributed communication framework which extends Apache REEF by providing an API
allowing to implement elastic computations by chaining fault-tolerant MPI-like primitives. While
fault-tolerant distributed applications are usually conducted using the well-known MapReduce-style
computation model based on data shuffling [19,49], recent trends in machine learning show that
MPI-based computation [57] provides better performance with respect to the former. Nevertheless,
MPI-based applications are in general not fault tolerant, whereas EGC tries to unify the benefits of the
two worlds. Based on the EGC framework, we test the proposed technique for a coded implementation
of linear regression on a real dataset when machines can leave and join the computation. We show that
the coded elastic computing framework can make the computational cost at each machine scale inversely
with the number of machines, which leads to linear scaling of theoretical computational cost. The
proposed technique is also useful in other applications besides elastic computation when the number
of machines needs to be dynamically adjusted during a learning task, such as when the number of
machines is tuned as a hyper-parameter, or when the machines have to be reallocated to achieve
fairness [33] between users or the specific need of some users at runtime.

In order to deal with the aforementioned problems, in this paper we propose coded elastic computing:
a novel distributed learning framework allowing users to train their machine learning models over
preemptable machines. In our coded elastic computing framework machines are allowed to arbitrarily
join or leave during a distributed iterative learning task thanks to the introduction of redundancy
in the computation. Although coded elastic computing introduces redundancy using ideas from
error-correcting codes [22,24,25,37,40,45,63,73,76], and can let the computation continue when a
preemption failure happens like ordinary error-correcting codes, the way it utilizes the coded data
to deal with preemption is fundamentally different from existing works. More specifically, coded
elastic computing can flexibly change the workload of each machine at runtime based on the number

compensate, meaning that data has to be copied on the new machines which will likely become
stragglers for the running computation. In practice, we observed situations at scale where the stop-the-
world scheme results in zero computation progress because, by the time a failure is recovered, a new
failure occurs. This results in the necessity to build an elastic run-time framework [18,50] and related
failure-aware algorithms which can continue the computation and flexibly adapt in the presence of
failures. Another possible technique to deal with preemption is to view the preempted machines as erasure-type faults and ignore them. Although machine learning algorithms are robust to small transient faults, simply ignoring the computational results in these permanently-failed preempted machines may result in algorithmic-level performance loss [50]. The influence of ignoring the computation results for the preemption type of faults is also more severe than usual computation faults because the number of failures can be really large [46]. Similarly, even if the data are redundant in some applications, ignoring partial results may still lead to reduced confidence levels on the prediction accuracy. It may also not be desirable from a customer’s perspective who often requires the full dataset to be present during the entire training process in order to achieve the highest accuracy.

In order to deal with the aforementioned problems, in this paper we propose coded elastic computing: a novel distributed learning framework allowing users to train their machine learning models over preemptable machines. In our coded elastic computing framework machines are allowed to arbitrarily join or leave during a distributed iterative learning task thanks to the introduction of redundancy in the computation. Although coded elastic computing introduces redundancy using ideas from error-correcting codes [22,24,25,37,40,45,63,73,76], and can let the computation continue when a preemption failure happens like ordinary error-correcting codes, the way it utilizes the coded data to deal with preemption is fundamentally different from existing works. More specifically, coded elastic computing can flexibly change the workload of each machine at runtime based on the number of available machines by selecting to use only a subset of the encoded data in a cyclic fashion. Apart from providing fault-tolerance when machines are preempted, coded elastic computing is also positively-elastic in that it can utilize the properties of the coded data to reduce the workload at existing machines flexibly when new machines join the computation. We will show that the coded elastic computing framework can make the computational cost at each machine scale inversely with the number of machines, which leads to linear scaling of theoretical computational cost. The proposed technique is also useful in other applications besides elastic computation when the number of machines needs to be dynamically adjusted during a learning task, such as when the number of machines is tuned as a hyper-parameter, or when the machines have to be reallocated to achieve fairness [33] between users or the specific need of some users at runtime.

The proposed coded elastic computing technique is tested in the multi-tenancy cluster at Microsoft
as an example of the Apache REEF EGC (Elastic Group Communication) framework. Apache
REEF is a library that helps develop distributed high-performance applications on top of cluster
resource managers such as YARN [66] and Mesos [33]. The Apache REEF project provides a set of
abstractions and reusable functional blocks to ease the process of building cloud-scale applications.
EGC is a distributed communication framework which extends Apache REEF by providing an API
allowing to implement elastic computations by chaining fault-tolerant MPI-like primitives. While
fault-tolerant distributed applications are usually conducted using the well-known MapReduce-style
computation model based on data shuffling [19,49], recent trends in machine learning show that
MPI-based computation [57] provides better performance with respect to the former. Nevertheless,
MPI-based applications are in general not fault tolerant, whereas EGC tries to unify the benefits of the
two worlds. Based on the EGC framework, we test the proposed technique for a coded implementation
of linear regression on a real dataset when machines can leave and join the computation. We show that
the coded elastic computing framework can make the computational cost at each machine scale inversely
with the number of machines, which leads to linear scaling of theoretical computational cost. The
proposed technique is also useful in other applications besides elastic computation when the number
of machines needs to be dynamically adjusted during a learning task, such as when the number of
machines is tuned as a hyper-parameter, or when the machines have to be reallocated to achieve
fairness [33] between users or the specific need of some users at runtime.

In this paper, we first present a coded elastic matrix-vector multiplication algorithm to illustrate the
main idea. Then, we present the generalizations of the proposed technique in broader applications,
among matrix-matrix multiplication, linear regression, master-free fully-distributed computing, and
the training of deep neural networks (see Section 3). Finally, we validate the approach with a set
of experiments. The proposed technique achieves the exact computation result as the noiseless
computation (where the number of machines remain fixed), while adaptively changing the workload

at each machine. The observed speedup compared to a non-adaptive coded computing baseline can reach up to 50%. The contributions of the paper are thus summarized in the following:

- We formalize the preemption problem in elastic computing and propose a coded elastic data-partitioning framework to deal with it.
- We design a computing technique that can adapt the workload at each machine in the presence of elastic events, without moving data at existing machines.
- We test the proposed technique using the Apache REEF EGC framework for linear regression and show advantage on real datasets over multiple baselines.

2 Resource-Elastic Coded Distributed Computing

In this section, we formally define computation elasticity (see Section 2.1), and present coded elastic techniques. We initially focus on the problem of matrix-vector multiplication for having a better theoretical understanding of coded elastic computing. Before presenting the algorithm in Section 2.4, we introduce the main idea of the paper in Section 2.2 and Section 2.3. Then, in Section 2.5.1 and Section 2.5.2, we analyze the proposed techniques and prove that they are indeed elastic according to our definition.

2.1 Definition of Computation Elasticity

We characterize elastic events whereby existing machines can be preempted, and new machines can be added to the computation. A preemption means that machines are taken away and the local data is lost. To formalize the notion of elastic events that we adopt in this paper, we state the following properties that are characteristic of these events:

Property 1. Which machine(s) to be preempted is decided by the resource allocator and is not known in advance.

Property 2. The preemption is permanent. However, new machines may join after some unknown time.

Property 3. If some machines leave or join, the other machines know immediately about which machines leave or join.

The second and third properties differentiate the elastic events from more commonly considered issues of faults and stragglers because (1) new machines can join the computation, and (2) one may adapt the computation scheme instantly after an elastic event and utilize the newly available resources.

Consider the case when a data matrix $X$ is stored distributedly. Denote by $P$ the initial number of machines. Denote by $P_{\text{max}}$ the maximum possible number of machines which equals to the overall size of the universe of machines handled by a cluster scheduler. A configuration point is a tuple $(n, m)$ in which $n$ is the number of machines and $m$ is the memory size. Note that $n$ changes over time, e.g., $n$ can exceed $P$. A computation policy of a task at a given configuration point $(n, m)$ is said to be achievable if the policy completes the task using $n$ machines, each with size $m$. A policy is said to be optimal if it obtains the optimal tuple $(e, u)$ simultaneously, where $e$ is the number of machine preemptions that can be tolerated (i.e., the exact result can be computed even if there are $e$ preemptions), and $u$ is the size of data that a machine actually selects to use ($u \leq m$). This means the memory of size $m$ can be used to store (coded) data, but during computation, we only access a part of it. Note that the optimal tuple $(e^*, u^*)$ depends on $(n, m)$. For a given configuration $(n, m)$, we want to minimize $u$ to reduce the memory access time and maximize $e$, the number of tolerable machine failures. Denote by $A_{n,m}$ the set of computing policies, and $A^*_{n,m} \subset A_{n,m}$ the set of optimal computing policies that obtain the optimal tuple $(e^*, u^*)$. One may think that the optimal tuple $(e^*, u^*)$ may not be unique. For example, there may exist a computing scheme that achieves $(e, u)$ such that $e > e^*$ and $u > u^*$. Note that our definition of $(e^*, u^*)$ is not in the sense of Pareto optimality. The $e^*$ and $u^*$ are information-theoretical optimal values that are obtained separately. If there is no scheme that achieves the optimal $e^*$ and $u^*$ at the same time, then in this case, $A^*_{n,m}$ is an empty set.

Definition 1. (transition compatibility) A pair of policies $(a, a')$ with $a \in A_{n,m}$ and $a' \in A_{n',m'}$ is said to be transition compatible if the policy $a$ can be transitioned to policy $a'$ without having to move or modify the data in the existing machines in the event of a configuration transition $(n,m) \rightarrow (n',m')$. 


The key to the above definition is that we discourage inter-machine data movement in order to make the elastic configuration transitions non-disruptive to ongoing computation tasks.

**Definition 2.** (fully transition compatibility) A family of policies $\mathcal{F} = \{a_{n,m}\}$, $n \in \mathcal{N}$, $m \in \mathcal{M}$, is said to be fully transition compatible if every pair $(a_{n,m}, a_{n',m'})$ in $\mathcal{F}$ are transition compatible.

**Definition 3.** (optimal fully transition compatibility) A family of fully transition compatible policies $\mathcal{F} = \{a_{n,m}\}$, $n \in \mathcal{N}$, $m \in \mathcal{M}$ is said to be optimal if all policies are optimal, i.e., each policy $a_{n,m} \in \mathcal{F}$ is in $\mathcal{A}_{n,m}^*$ and obtains the optimal tuple $(e^*, u^*)$ for the number of tolerable machine preemptions and the size of the selected data to use.

Our goal is thus to find conditions under which an optimal fully transition compatible family exists, and to provide explicit characterizations of the transition compatible families. We now present matrix-vector multiplication techniques that can provide a fully transition compatible family of optimal computation policies with fixed memory cost at each machine, i.e., when $m$ is fixed in different transition compatible policies $a_{n,m}$.

### 2.2 Coded Data Partitioning in the Presence of Preempted Machines

Assume that in the worst-case of preemption failures, there are at least $L$ machines that remain out of $P$ workers. We will show that the parameter $L$ is also equal to the recovery threshold, the meaning of which will be clear in this section (Section 2.2). In this paper, we consider repeatedly using the same data but with different input vectors. For matrix multiplications, it means that we compute $X_w$, for $t = 1, 2, \ldots$ for the same $X$. This computation primitive is applicable in a variety of scenarios, including training linear models [30, 72], PageRank [73], model-parallel deep neural networks [20, 21] and many machine learning algorithms at the inference stage.

We partition the data matrix $X$ into $L$ subsets (or equivalently, submatrices obtained by row-wise partition) $X_1, X_2, \ldots, X_L$ of equal size. If the total number of data points is not divisible by $L$, we can use zero-padding. The generator matrix $G_{P_{\text{max}} \times L} = (g_{s,k}), s = 1, \ldots, P_{\text{max}}$ is predetermined. We initially generate $P$ coded data matrices $X_s^\text{coded}, s = 1, 2, \ldots, P, (P > L)$, in which each matrix is a linear combination of the form:

$$X_s^\text{coded} = \sum_{k=1}^{L} g_{s,k} X_k$$

where each $g_{s,k}$ is a random but predetermined coefficient. The $P$ coded data matrices are distributed to $P$ workers. When the number of machines exceeds $P$, we generate new linear combinations based on $G_{P_{\text{max}} \times L}$. The generation can be done before the computation, and the data can be stored in the cloud which has a much larger size than the local fast memory.

**Lemma 2.1.** Suppose we want to compute the matrix-vector product $Xw$. Suppose the matrix $G_{P_{\text{max}} \times L} = (g_{s,k})$ satisfies the property that any $L \times L$ submatrix of $G_{P_{\text{max}} \times L}$ is full-rank. Then, any $L$ out of $n$ coded computation results obtained at the $n$ available machines $X^\text{coded}_s w, s = 1, 2, \ldots, n$ are sufficient to recover the original (uncoded) computation results $Xw$ regardless of the current number of machines $n$.

The rank condition in the lemma can be satisfied by a variety of choices of linear coefficients $g_{s,k}$, e.g., if $g_{s,k}$’s are i.i.d. Gaussian random variables. The recovery of the results is through solving $L$ linear systems of the form $X^\text{coded}_s w = \sum_{k=1}^{L} g_{s,k} X_k w$ for the $L$ different machines that successfully finish the computation. Lemma 2.1 is critical for the failure recovery. It essentially shows that no matter which machines are preempted, as long as the number of remaining machines is not smaller than $L$, the whole information of the original data is preserved in the remaining machines, and the computation results can be recovered. This is why we call the parameter $L$ the recovery threshold. The parameter $L$ is limited by the storage constraint at each machine. The more redundancy we can add to the data, the lower recovery threshold we need, and hence more failures we can tolerate. In our experiments, we consider the case when $P = P_{\text{max}}$, i.e., the maximum number of machines is the same as the overall number of machines, and use a redundancy factor of $P/L = 2$. Therefore, we can at maximum tolerate failures when half of the machines are preempted. Note that this is not necessary because one can generate more data and store the extra data on the cloud so that the

---

2 The parameter $L$, or a lower bound of $L$, is needed for exact computation. However, in many machine learning tasks, one can often optimize with a subset of data due to data redundancy. In that case, knowledge of $L$ is not necessary.
number of machines \( n \) can exceed the initial number of machines \( P \). An often-used coding technique is called systematic code, in which the linear coefficients satisfy

\[
g_{s,k} = 1_{\{s=k\}}, \quad \text{if} \quad s \leq L. \tag{2}
\]

In this case, the coded data at the first \( L \) machines \( X_s^{\text{coded}}, s = 1, 2, \ldots, L \) are the original data \( X_k, k = 1, 2, \ldots, L \). This can provide backward-compatibility to switch between coded computing and other uncoded ordinary computing techniques, and at the same time significantly reduce the cost of encoding the data at the preprocessing stage \([1]\).

### 2.3 Elastic Data Partitioning for Elastic Computation by Using Data In a Cyclic Way

According to Lemma 2.1 as long as the number of machines that are not preempted is greater or equal to \( L \), the remaining data using the coded data partitioning can preserve the whole information of the original data. However, when the number of machines is strictly larger than \( L \), it becomes redundant to use all the coded data because the data at \( L \) machines already preserve the whole information. One may think that this amount of waste is not significant. However, consider the case when the number of machines gradually increases from \( L \) to a large number. In this case, any \( L \) machines can provide the correct results, but we are not able to utilize the parallel gain if a fixed coding technique is used. The situation can be partially solved if we use some rateless coded techniques \([48]\). See Section 2.6 for a comparison between the proposed technique and rateless coded computing.

To positively utilize all the remaining machines and achieve the parallel computing capabilities of the extra machines, we select to use data in a cyclic fashion as shown in Figure 1. We use a systematic code (see (2)) by which the first \( L \) of the \( P \) coded blocks \( X_s^{\text{coded}}, s = 1, 2, \ldots, P \) are the original data \( X_k, k = 1, 2, \ldots, L \). In Figure 1(a), we use red to denote original data and blue to denote the remaining coded data. In this example, the initial number of machines is \( P = 6 \) and the recovery threshold is \( L = 3 \). From figure 1(b) to 1(e) we show how to continue the computation when machines are gradually preempted from 6 to 3 (machines correspond to the columns). The stored data remains fixed i.e., the same way as in Figure 1(a), but we further partition the data into smaller...
blocks and only select part of the data to use. Each machine is initially allocated a single subset of coded data $X^\text{coded} \in \mathbb{R}^{s \times L}$, $s = 1, 2, \ldots, P$, among which $L$ subsets are the original data. Each subset of data is represented as a column in any subfigure of Figure 1.

If no failures occur (see Figure 1(b)), to remove redundancy from the data, we partition each data block (column) into $P$ sub-blocks, and let each machine only use $L$ out of $P$ sub-blocks. By a sub-block of data, we mean one small rectangle in Figure 1(b). When there are $n$ machines and $n \neq P$ (see Figure 1(c)-(e)), we partition each data block into $n$ sub-blocks, and still let each node only use $L$ out of $n$ sub-blocks. If new machines join, they download the coded data previously used in some failed machines or some new linearly combined data based on $G_{P \times n \times L}$. All the machines, including the ones that just join, use elastic data partitioning based on the current number of available machines $n$.

There are two advantages of this type of data usage (1) the overall selected data to use is of the same size as the original data and the selected data across all remaining machines have the same size; and (2) the selected data preserve the whole information of the original data, according to Lemma 2.1. Thus, we can exactly recover the results while removing the redundancy in the way of using data. These two properties will be formally introduced and proved in Theorem 3.1.

### 2.4 Coded Elastic Computing for Matrix-vector Multiplications

We provide the detailed procedures of the coded elastic computing algorithm for the repeated matrix-vector multiplication problem $Xw_t, t = 1, 2, \ldots$ in Algorithm 1. We use $X^\text{coded}_{k,j}$ to represent the $j$-th sub-block of the data at the $k$-th machine. We will call $X^\text{coded}_{k,j}$’s with the same $j$ “the $j$-th group” of sub-blocks which correspond to the $j$-th row block in any subfigure of Figure 1. Note that the number of row blocks change with the number of preempted machines. We use $G_j$ to represent the collection of linear combination coefficients for the $j$-th group (row block) that are selected to use.

For example, for the first group (row block) in Figure 1(b), we have $G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix}$ because the three selected sub-blocks are $a_1, a_1 + 2a_2 + 3a_3$ and $a_1 + 4a_2 + 9a_3$. And for the last group (row block) in Figure 1(b), $G_6 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix}$, because the three selected sub-blocks are $f_1 + f_2 + f_3, f_1 + 2f_2 + 3f_3$, and $f_1 + 4f_2 + 9f_3$.

#### Algorithm 1 Coded Elastic Computing for Matrix-Vector Multiplication

**Input:** The data matrix $X$, the number of machines $P$, the recovery threshold $L$, the linear combination coefficients $g_{s,k}$’s in equation (1) and the sequence of input vectors $w_t, t = 1, 2, \ldots$.

**Preprocessing:** Partition the data $X$ into $L$ subsets and compute the coded subsets as in (1).

**Online computation:**

FOR each computation with input $w_t$:

- **Broadcast:** The master node sends $w_t$ to each worker.

  FOR each group index $j$:

  - **Gather:** The $k$-th worker computes $u_{t,k,j} = X^\text{coded}_{k,j}w_t$ and sends $u_{t,k,j}$ to the master. The master gathers vectors $u_{t,k,j}$ for all workers that use the $j$-th sub-block and obtains the matrix $u_{t,j}$ which contains the results for the $j$-th group (row block).

- **Decode:** The master node computes $u_{t,j}G^{-1}_j$ to obtain the results for the $j$-th group.

**Output:** The master node outputs $Xw_t$.

**IF Preemption/New Machines:** Change the selected data to use based on the current number of machines.

### 2.5 Analysis of Coded Elastic Computing: Achieving Optimal Fully Transition Compatibility

According to Definition 3, a fully compatible family of policies can support seamless transitions between any pairs of policies in the family. The coded elastic computing scheme provided in Algorithm 1 for matrix-vector multiplication gives a family of fully-compatible policies for fixed
memory cost at each machine. In Section 2.5.1, we analyze the memory cost and the number of
data points to be used at each existing machine. Then, in Section 2.5.2, we provide lower bounds on
these two quantities and show that Algorithm 1 achieves the optimal memory cost and the size of the
selected data. Thus, the coded computing policies in Algorithm 1 is a compatible family of optimal
policies, in that for the fixed storage size, each policy obtains the largest number of tolerable failures
and smallest size of the selected data to use (ε∗, u∗), and provides seamless transitions between each
other without moving data at existing machines.

2.5.1 Upper Bounds on the Storage Cost and the Size of the Selected Data to Use
Suppose the original data has N data points and all data points are in \( \mathbb{R}^d \), i.e., \( X \) has size \( N \times d \).
Recall that \( P_{\text{max}} \) is the maximum number of machines, \( L \) is the recovery threshold, and \( n \) is the
number of currently available machines. In the following theorem, by the size of the data we mean
the number of data points.

**Theorem 2.1.** Suppose the number of machines \( n \) satisfies \( L \leq n \leq P_{\text{max}} \) at any time, i.e., it is not
smaller than the recovery threshold \( L \) and not bigger than the maximum possible number of machines.
Then, the coded elastic computing algorithm (Algorithm 1) achieves the exact computation result of
\( Xw_t \) for all \( t \). The size of the data stored at each machine is \( N/L \). The size of the selected data
to use at each machine is \( N/n \) and is the same across different machines. The overall size of the
selected data is the same as the size of the original data.

**Proof.** Recall that we call the sub-blocks on the same row block (in Figure 1) of the \( P \) different
blocks of data a group, and we use \( X_{k,j}^{\text{ coded}} \) to represent the coded sub-block of data that is at the \( k \)-th
machine and belongs to the \( j \)-th group. Let \( X^j \) be the collections of original data that belongs to the
\( j \)-th group. For example, in Figure 1(b) \( X^0 \) represents the collection of original data \( \{f_1, f_2, f_3\}^\top \).

Then, from the cyclic way of using data, we can see that when each machine uses \( L \) sub-blocks, the
overall number of used sub-blocks in each group is \( L \). From Lemma 2.1, the results \( u_{t,k,j} = X_{k,j}^{\text{ coded}}w_t \)
for all the workers that use the coded data in the \( j \)-th group can be collected together to decode \( X^jw_t \).

The other claim can be seen from the figure, i.e., the area of the used data is always equal to the area
of the original data, and the area of the used data is the same across all remaining machines. We
prove this claim as follows. The size of the data in each of the \( L \) subsets is \( N/L \). Thus, the used data
at each machine is the number \( N/L \) · \( L/n \). There are \( n \) machines left, so the overall size of the
used data is \( N \times \frac{L}{n} \cdot \frac{L}{n} = N/n \). Which is the same as the original data.

Theorem 2.1 shows that our technique uses the same size of data as the original (uncoded) case. This
is desirable for memory-bound applications.

**Remark 1.** (Cost analysis) Recall that \( P \) is the number of workers, \( N \times d \) is the size of the matrix
\( X \). \( L \) is the recovery threshold, and \( n \) is the number of currently available machines. The encoding
(preprocessing step) is a one-time cost for online matrix-vector multiplications. The decoding by
solving a linear system at the master node has computational cost \( O(LN) \), because the linear system
for each group (row-block) of data involves \( L \) equations on \( L \) unknown sub-vectors of size \( N/L/n \)
(which is the height of each sub-block), and there are \( n \) such groups. Thus, the computational
size of the straight-forward matrix-vector multiplication is \( L^2 \cdot (N/L/n) \cdot n = NL \). The matrix-
multiplication step at each worker has cost \( O(dN/n) \). Thus, the decoding cost is smaller than the
computational cost at each worker as long as \( d = O(nL) \). Even if \( d < O(nL) \), we can partition
the machines into smaller groups and respectively code each group. The decoding complexity can
be further reduced to \( O(N \log^2 L) \) if Vandermonde systems are used [43], at the cost of numerical
stability. One thing to note is that the decoding complexity, even for the straight forward matrix-vector
multiplication method, is \( NL \) and is independent of the number of workers \( P \).

2.5.2 Lower Bounds on the Storage Cost and the Size of the Selected Data to Use
Here, we provide a fundamental limit which shows that the achievable scheme provided in 2.1 is
optimal in terms of the storage cost at each machine and the size of the actually used data at each
iteration, for a fixed number of machines and a fixed number of tolerable machine preemptions.
Before we present the theorem, we formalize the definition of the size of data using number of bits.
This is because in theory, we cannot store arbitrarily high-precision numbers.
Assumption 1. Suppose the entries of the matrix $X$ are i.i.d. random variables that take values in a finite set $S \subseteq \mathbb{R}$. Each of this random variable has entropy $H = \log |S|$. Thus, the overall entropy of all the data in $X$ is $NdH$.

Assumption 2. For a certain computation policy, suppose each machine initially stores an array of finite-precision numbers in its memory. Each finite-precision number can be an arbitrary function of the original data. The overall number of finite-precision numbers that is stored is finite.

Note that although we use real-number computation all the time, the numbers that we deal with are always discrete, i.e., we conduct computation of finite-precision numbers in the real field. This validates Assumption 1. Also note that the number of possible combinations of floating point numbers that can be stored by a finite-length bit array is finite. Therefore, Assumption 2 is also valid. We need these two assumptions because if they do not hold, the system can concatenate all the real numbers in $X$ into one real number and only stores that particular real number. In that case, no bound on the storage is meaningful because one only needs to read this single number in memory to access all the information of $X$. We also need the assumption that the size of the stored numbers are finite because otherwise, we can enumerate all possible $Xw$ and store them.

Assumption 3. Suppose we do not alter the way that we store the data $X$ inside the memory after the computation begins. Even when a preemption-type failure happens, we do not move the data at existing machines.

Assumption 4. By selecting to use one number stored in the memory, we mean the algorithm reads the whole number (e.g., reading all digits if the number is stored as a floating-point number) from the memory for further processing. The array of finite-precision numbers can only be accessed one number at a time, meaning that one cannot access a function value of two numbers and claim that only one number is selected to use.

Theorem 2.2. Suppose the Assumptions 1-4 hold. Suppose we require the recovery of the exact computation result $Xw$. Then, the following fundamental limits hold.

(a) Denote the entropy of the encoded data at the $k$-th machine by $H_k$, $k = 1, 2, \ldots, n$. Then, to provide the tolerance to a maximum of $n - L$ failures (recall that $n$ is the current number of available machines), we have $\max_{k \in \{1, 2, \ldots, n\}} H_k \geq \frac{N}{L} \cdot dH$;

(b) The worst-case entropy of the actually used data (maximized with respect to the choice of $w$) has to be no less than $NdH$, or $\frac{N}{n} \cdot dH$ at each machine.

Remark 2. By comparing Theorem 2.1 and Theorem 2.2 we see that the coded elastic computing technique in Algorithm 1 achieves the fundamental limit because each data point has dimension $d$ and each entry has entropy $H$. We note a nuance here that the linear combinations in Algorithm 1 may make each encoded number have entropy larger than $H$. However, this can be solved if (i) we approximate the real-number computation using the computations in a finite field and assume each number in the data matrix $X$ has a uniform distribution over the finite field $[76]$, or (ii) we assume that each entry in the data matrix comes from a quantized version of the standard Gaussian-distribution, and assume that the encoding matrix has normalized rows, such that the linear combination result in (1) (before quantization) is also from the standard Gaussian distribution.

To make the point (ii) formal, suppose an arbitrary entry in the data matrix $X_k$ is sampled from i.i.d. standard Gaussian and quantized to $q(y_k)$ (here $y_k \sim \mathcal{N}(0, 1)$). A linear combination of the samples has the form $\sum_{k=1}^{L} g_{s,k} q(y_k)$ (see equation 1), where the linear combination coefficients satisfies $\sum_{k=1}^{L} g_{s,k}^2 = 1$. Then, the linear combination is stored as $q(\sum_{k=1}^{L} g_{s,k} y_k)$ in the $s$-th machine. Since the unquantized sum $\sum_{k=1}^{L} g_{s,k} y_k$ is a standard Gaussian random variable, its quantized version $q(\sum_{k=1}^{L} g_{s,k} y_k)$ has the same entropy as each $q(y_k)$. Therefore, we only need to show that $q(\sum_{k=1}^{L} g_{s,k} q(y_k))$ has an entropy that is close to $q(\sum_{k=1}^{L} g_{s,k} y_k)$. This can be proved by bounding the difference in the entropy, because the two random variables $q(\sum_{k=1}^{L} g_{s,k} q(y_k))$ and $q(\sum_{k=1}^{L} g_{s,k} y_k)$ have the same support, and their pmf’s can be made arbitrarily close when the quantization function $q$ has a quantization level $\Delta \rightarrow 0$.

Remark 3. Note that the claim (b) has to be stated in a worst-case way because for many choices of $w$, computing $Xw$ can be degenerated. For example, if we know in advance that $w$ only takes values in a very small finite set of vectors, we can compute $Xw$ for all possible $w$ and store these vectors. When $w$ is sparse, we also do not need to read the entire matrix $X$. Therefore, we indeed need to state the fundamental lower bound in terms of the worst-case $w$.  

8
Proof. Now, we prove statement (a). Suppose an arbitrary set of \( n - L \) machines fail. Since the algorithm is tolerant to any \( n - L \) failures, the master node can still recover exactly all the results \( \mathbf{Xw} \), no matter what \( \mathbf{w} \) is. Therefore, we can choose \( \mathbf{w} \) to be the elements of natural basis \( \mathbf{w} = \mathbf{e}_i, i = 1, 2, \ldots, d \), and collect all the results \( \mathbf{Xe}_1, \mathbf{Xe}_2, \ldots, \mathbf{Xe}_d = \mathbf{Xl}_d = \mathbf{X} \). Since data processing can only reduce entropy (from the data processing inequality), the entropy of the overall data stored at the remaining \( L \) machines is no less than the entropy of \( \mathbf{X} \) which is \( NdH \). This holds for any combination of \( L \) machines, i.e.,

\[
H(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_L}) \geq NdH, \forall 1 \leq i_1 < i_2 < \ldots < i_L \leq n. \tag{3}
\]

Adding up the above equation for all combinations of \( L \) out of \( n \) machines, and by plugging in 
\[
\sum_{j=1}^{L} H(\mathbf{X}_{i_j}) \geq H(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_L}), \text{ we have } \sum_{k=1}^{n} H_k \geq \frac{NdH}{L}.
\]

Then, we prove statement (b). Suppose \( (b) \) is not true. Then, it means that there exists a way to encode and store the encoded data in the memory of the machines, such that for any arbitrary vector \( \mathbf{w} \), computing \( \mathbf{Xw} \) only requires reading data of entropy strictly less than \( NdH \). Since the overall number of stored data is finite, we can assume that the overall number of stored numbers is \( p \). For an arbitrary subset \( S \) of the stored numbers such that \( H(S) < NdH \), we denote by \( \mathcal{P}_S \) the set of \( \mathbf{w} \) such that \( \mathbf{Xw} \) is able to be computed using only the numbers in \( S \). Then, we can see that \( \mathcal{P}_S \) for an arbitrary subset \( S \) is a linear subspace of \( \mathbb{R}^d \). This is because if \( \mathbf{Xw}_1 \) and \( \mathbf{Xw}_2 \) are both able to be computed using the numbers in \( S \), then \( a\mathbf{Xw}_1 + b\mathbf{Xw}_2 = \mathbf{X}(aw_1 + bw_2) \) is also able to be computed. Now, we prove that \( \mathcal{P}_S \) cannot be the entire \( \mathbb{R}^d \) that \( \mathbf{w} \) can take value from. This is because if \( \mathcal{P}_S \) is equal to \( \mathbb{R}^d \), then \( \mathbf{Xe}_1, \mathbf{Xe}_2, \ldots, \mathbf{Xe}_d \) are all able to be computed using the numbers in \( S \) (\( \mathbf{e}_1, \ldots, \mathbf{e}_d \) are the standard basis), which means \( \mathbf{X} \) itself is able to be computed using \( S \). This is a clear contradiction to the data-processing inequality because \( H(\mathbf{X}) = NdH \), while \( H(S) < NdH \). Thus, \( \mathcal{P}_S \) can at most be a linear space of dimension \( d - 1 \) in \( \mathbb{R}^d \). This means the union of \( \mathcal{P}_S \) for all \( S \) is a finite collection of linear spaces of dimension \( d - 1 \) in \( \mathbb{R}^d \), which cannot cover the entire \( \mathbb{R}^d \). Thus, there must exist \( \mathbf{w} \) in \( \mathbb{R}^d \) such that \( \mathbf{Xw} \) is not able to be computed by using data of entropy strictly less than \( NdH \).

\[\square\]

2.6 Related Works and the Comparison with Non-elastic Coded Computing Techniques

Coded computing is an emerging area to tackle stragglers [8, 11, 22, 27, 37, 38, 40, 42, 47, 48, 51, 52, 56, 58, 59, 61, 64, 70, 73, 77, 78], machine failures and soft errors [9, 21, 31, 36, 74], security issues and adversaries [13, 14, 16, 17, 71, 78], and communication bottlenecks [29, 32, 35, 45, 54, 58, 63, 65, 75]. It can handle both linear and nonlinear computations [21, 39, 63]. It is a significant advance on classical algorithm-based fault-tolerance (ABFT) techniques [34] and noisy computing [53, 65, 67], and often achieves scaling sense improvements. Our work is the first to address the system elasticity issue using coded computing.

Non-elastic coded computing cannot adapt the algorithmic procedures to exploit new machines. We consider the case when the number of machines increases from \( L \) (i.e., the recovery threshold) to infinity with fixed storage size at each machine. We can easily see that elastic coded computing can have lower and lower memory-access time by using less and less data at each machine, while ordinary coded computing cannot because the used data remains unchanged at each machine. However, each node can have more than one linear combinations [48]. E.g., suppose \( L=8 \) linear combinations suffice to recover the result, and each machine has 8 linear combinations. Then, when the number of machines increases from 1 to 8, rateless coded computing can move between the optimal points \( (n, m) = (1, 8), (2, 8), (4, 8), (8, 8) \), i.e., it achieves the optimal size of the selected data to use at some of the \( (n, m) \) configurations. But they cannot move beyond the point \( (8, 8) \) and use more machines, because each machine, in the rateless-coded case, cannot use smaller than 1 (a fraction) of linear combinations. From another perspective, for the rateless-coded scheme to achieve fully transition compatibility in the large scale, the number of linear combinations at each machine also has to grow with the number of (possible) machines, leading to scaling-sense higher decoding complexity than the elastic partitioning technique in Algorithm 1 (Note that the decoding cost in Algorithm 1 does not increase with the number of machines, and can be further reduced if parallel decoding is more efficient.)
allowed. See Section 3.3 for details). Another advantage of coded elastic computing compared to the rateless-coded scheme is that it allows non-uniform partitioning of each row-strip (see Figure 1) and can have much more fine-grained task scheduling. Regarding the communication efficiency of coded elastic computing, see Section 3.3 for a fully-distributed version of coded elastic computing.

For task-level failures or stragglers, dynamic task allocation is useful [68][69]. The task scheduler can choose which tasks to replicate, relaunch based on task profiles, and delay the relaunching to save time [7][8][10][28][62]. However, unlike task relaunching and allocation, recovering from machine preemptions requires both restoring the machine states and downloading the data, which is time-consuming. The situation gets worse if elastic events are frequent.

3 Applications of the Coded Elastic Computing

The cyclic way of elastic data partitioning applies to general coded computing techniques proposed thus far and is not limited to matrix-vector multiplications.

3.1 Matrix-Matrix Multiplications

First, we consider the application of matrix-matrix multiplications. We consider an online version: we store an encoded version of the matrix $A$ at $P$ machines and compute the matrix multiplication $AB$ for different $B$’s. It is shown [24][25][77] that the storage-optimal technique for coded robust matrix multiplications is the MatDot scheme proposed in [24][25]. This technique partitions the matrix $A$ column-wise and the matrix $B$ row-wise, and stores linearly combined submatrices of $A$ and $B$ at each machine. In an online setting with elastic machine preemptions, we encode and store $A$ at the initial stage, and do not move $A$ anymore. When we receive $B$, we still partition the matrix $B$ row-wise but linearly combine them using the knowledge of the availability of the machines. The advantage is that we do not need to use the polynomial-based codes and can thus avoid possible numerical issues. At the same time, since the availability of the machines are known before we encode $B$, we can also remove the factor of 2 in the recovery threshold of MatDot codes. We can also use the same computational time cost as the uncoded case, which is similar to what we can achieve in the matrix-vector case.

More precisely, suppose we partition the matrix $A$ into $L$ column blocks $A_1, \ldots, A_L$, and encode these blocks initially into $P$ blocks $A^\text{coded}_s$, $s = 1, 2, \ldots, P$, where $P$ is the initial number of machines:

$$A^\text{coded}_s = \sum_{k=1}^{L} g_{s,k} A_k. \quad (4)$$

Denote by $G_{P_{\text{max}} \times L}$ the predetermined encoding matrix $[g_{s,k}]$, which is of size $P_{\text{max}} \times L$. Again, the extra linear combinations $A^\text{coded}_s$, $s = P + 1, P + 2, \ldots, P_{\text{max}}$ can be generated offline before the computation and stored on the cloud. Assume there are currently $n$ machines that remain. Similar to Algorithm 1, we partition each coded sub-matrix $A^\text{coded}_s$ into $n$ submatrices. However, here, we partition each $A^\text{coded}_s$ column-wise to $A^\text{coded}_{s,i}$, $i = 1, 2, \ldots, n$. This is mathematically equivalent to partitioning each uncoded submatrix $A_s$ into $A_{s,i}$, $i = 1, 2, \ldots, n$, and compute

$$A^\text{coded}_{s,i} = \sum_{k=1}^{L} g_{s,k} A_{k,i}, \quad i = 1, 2, \ldots, n. \quad (5)$$

The subscript $s$ in $A_{s,i}$ belongs to a subset with size $n$ of the set $\{1, 2, \ldots, P_{\text{max}}\}$, which corresponds to the $n$ available machines after the preemption failures. We again select to use these submatrices $A_{s,i}$ in a cyclic fashion, just as shown in Figure 1. For example, consider the case when $P = 6$, $L = 3$, $n = 4$ and the 2nd and the 4th machines are preempted, which is exactly the same as shown in Figure 1(d). Then, we use $A_{1,1}, A_{1,2}, A_{1,3}$ at the 1st machine, $A_{3,2}, A_{3,3}, A_{3,4}$ at the 3rd machine, and so on. Denote by $S_i$ the set of the indices of all the machines that use $A_{s,i}$. For example, for $i = 1$, in the above example, $S_1 = \{1, 5, 6\}$. Again, similar to Section 2.4, denote by $G_i$ the submatrix of $G$ with row indices in $S_i$. Each $G_i$ is a $L \times L$ matrix. Denote by

$$H_i = (G_i^\top)^{-1}, \quad (6)$$

and assume $H_i$ has element $h_{l,k}, l = 1, 2, \ldots, L, k = 1, 2, \ldots, L$. 

10
When we get the matrix $AB$ without moving data at the existing machines. At the same time, the computation time cost is the same as distributing the matrix-matrix multiplication task to $nL$ machines.

Note that

Now, we show that when we do a reduction on all the partial results $A_{s_{i,i},i}B_{s_{i,i},i}$, we can indeed get $AB$.

**Lemma 3.1.** If we use the cyclic partitioning technique to determine each $S_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,L}\}$ and encode $B$ according to (7), we have

$$
\sum_{i=1}^{n} \sum_{l=1}^{L} A_{s_{i,i},i}B_{s_{i,i},i} = AB.
$$

**Proof.** Note that

$$
A_{s_{i,i},i}B_{s_{i,i},i} = (\sum_{k=1}^{L} g_{s_{i,i},k}A_{k,i}) (\sum_{k=1}^{L} h_{l,k}B_{k,i}),
$$

where (a) holds because $[g_{s_{i,i},1}, g_{s_{i,i},2}, \ldots, g_{s_{i,i},L}]$ is the $s_{i,i}$-th row in $G$, which is the $l$-th row in $G_i$ (recall that $G_i$ is the submatrix of $G$ with row indices in $S_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,L}\}$). Thus, by adding up the above equation for all $l$, we have that for each $i$,

$$
\sum_{i=1}^{n} \sum_{l=1}^{L} A_{s_{i,i},i}B_{s_{i,i},i} = \sum_{i=1}^{L} [A_{1,i}, A_{2,i}, \ldots, A_{L,i}] (G_i)_{l-th \ row} (H_i)_{l-th \ row} [B_{1,i}^{T}, B_{2,i}^{T}, \ldots, B_{L,i}^{T}]^{T},
$$

where (a) holds because $H_i = (G_i^{T})^{-1}$ (see equation (6)). Adding up the above equation for all $i$, we have

$$
\sum_{i=1}^{n} \sum_{i=1}^{L} A_{s_{i,i},i}B_{s_{i,i},i} = \sum_{i=1}^{n} \sum_{k=1}^{L} A_{k,i}B_{k,i} = AB.
$$

**Remark 4.** (Cost analysis) We can see that the computational cost for this algorithm comes from two parts: (1) the encoding of $B$ using (7), and (2) the computational cost of $A_{s_{i,i},i}B_{s_{i,i},i}$. Assume the matrix $A$ has size $d_A \times N$ and $B$ has size $N \times d_B$, where $d_A, d_B = \Theta(N)$. Then, encoding $B$ has complexity $nL \cdot (d_B \times N/L \cdot n) = LD_B N = \Theta(LN^2)$. Computation of a single $A_{s_{i,i},i}B_{s_{i,i},i}$ has complexity $d_A \times N/L \times d_B$, and each machine computes $L$ different $A_{s_{i,i},i}B_{s_{i,i},i}$, which means the overall complexity is $d_A \times N/L \times n \times d_B \times L = d_A d_B N/n$. Note that this is in the order of $\Theta(N^3)$ and the complexity is the same as distributing the matrix-matrix multiplication task to $n$ machines. The encoding time is much less than the computation time per worker if $LD_B N \ll d_A d_B N/n$, or $Ln \ll d_A$. If the encoding time is much smaller than the matrix-multiplication time, we again achieve linear scaling of the theoretical computational complexity in the number of available machines $n$ without moving data at the existing machines. At the same time, the computation time cost is the same at each worker machine as the uncoded case, which means that the factor 2 in the MatDot codes [24] can be removed.\footnote{It was shown that the optimal recovery threshold in MatDot is $2L - 1$, which means that the computational cost at each machine is twice that of the uncoded case for the same number of machines. The scheme here}
3.2 Coded Elastic Computing for Linear Models

Then, we focus on the application of coded computing for linear regression [40, 44]. For the ease of presentation, we consider vanilla gradient descent (we also use line search in the experiment validation for all competing techniques), in which the full matrix $X$ is used at each iteration. The technique developed here naturally generalizes to stochastic gradient descent and other generalized linear models such as logistic regression. Consider the linear objective function:

$$f(w; X, y) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + h(w).$$

(12)

The vanilla distributed gradient descent has the form $w_{t+1} = w_t - \eta g_t$ and $g_t = X^T (Xw_t - y) + \partial_h h(w_t)$. When the data matrix $X$ is large, the most time-consuming part is the computation of $X^T (Xw_t - y)$. We thus extend Algorithm 1 in the following way to compute $X^T (Xw_t - y)$. It is nothing but a combination of Algorithm 1 and the matrix-matrix multiplication algorithm in Section 3.1.

- Compute $X^T w_t$ (where $X^T$ is the data in the $j$-th group, or the $j$-th row block in Figure 1 across different machines) for all group-index $j$ in an elastic way using Algorithm 1.
- The master computes $z_j = X^T w_t - y_j$ for all group-index $j$, where $y_j$ are the labels corresponding to the data points in the $j$-th group;
- The master re-encodes $z_j$’s using the (pre-computed) inverse generator matrix $H_j = (G_j^{-1})^T$ to obtain $(G_j^{-1})^T z_j$, and scatters the results to the workers that use the $j$-th group of data;
- Since data at workers are encoded using $G_j$, the reduced results from all the workers are

$$\sum_j (X^j)^T G_j^T (G_j^{-1})^T z_j = \sum_j (X^j)^T (X^j w_t - y_j) = X^T (Xw_t - y).$$

(13)

Note that in the above extension of Algorithm 1, the workers also utilize the data as in Figure 1. The experiment results of coded elastic computing in linear models are provided in Section 4.

3.3 Fully Distributed Coded Elastic Computing

Coded elastic computing is not restricted to a master-worker setting. In this section, we consider a fully distributed coded elastic matrix-vector multiplication technique that is a trivial generalization of Algorithm 1. The advantage of a fully distributed framework is that communication among workers can be overlapped and the communication to the master does not become the single bottleneck in the limit of a large number of machines. Consider an application of iterative matrix-vector multiplication

$$w_{t+1} = f(Xw_t),$$

(14)

where $X$ is a square data matrix, and $f(\cdot)$ is an entry-wise low-complexity operation on the vector $Xw_t$. We again consider the example shown in Figure 1 but apart from the 6 worker machines, there is no master-node. In each iteration, each machine computes its own matrix-vector multiplication based on the selected data to use. For example, the 2nd machine computes $b_2 w_t, c_2 w_t, d_2 w_t$, and stacks the results into one vector. Then, the results at each machine can be broadcast to all the other machines using an all-gather communication (using a bucketing algorithm [12]). The communication is also done in a cyclic way so that all the communications can be maximally overlapped (i.e., all machines can communicate at the same time). In this way, the communication bandwidth can be reduced when compared to the master-worker framework. Moreover, the overall communication time in this scheme is the same as the uncoded scheme using the same number of machines, because the selected size of data at each machine is the same as the uncoded case. The decoding of the intermediate results at each iteration is done independently at each machine. The decoding can also be conducted distributedly for each group of data (each row-block), i.e., each machine only takes care of the decoding of each row-block. This can significantly reduce the decoding time, but requires

---

1 Note that the example in Figure 1 partitions the data row-wise. The column-wise partition can use the elastic coding scheme in Section 3.1 which is essentially an elastic dot-product scheme.
all machines to exchange different messages twice and carry out an all-to-all communication. Also, the smaller communication is possible at the cost of more storage or computation in the individual machines.

3.4 Application to Deep Neural Networks

The proposed coded elastic data partitioning directly applies to the coded training of model-parallel deep neural networks. The coded computation utilizes a novel technique called PolyDot, which adopts a 2D partitioning on the weight matrices of a neural network and encodes the weight matrices using polynomial-based codes. The way to apply coded elastic data partitioning to the PolyDot technique is exactly the same as Section 2.3:

- In PolyDot, each machine has one encoded submatrix of the weight matrix. Partition each encoded submatrix in Polydot into \( n \) small submatrices. Recall that \( n \) represents the number of currently available machines.
- Let each machine select to compute the matrix-vector multiplication using only \( L \) out of \( n \) small submatrices. Recall that \( L \) is the recovery threshold.
- The selection is done in a cyclic way as well to ensure that there are enough encoded computation results to decode the original computation results.

Note that the matrix-vector multiplication in requires both forward pass and backward pass. This means that for the weight matrix \( W \), we need to compute matrix-vector multiplications in two ways, i.e., \( Wx \) and \( W^\top x \). Therefore, in the forward pass, the partitioning of the encoded weight submatrix at each machine should be done in a row-wise way, which is similar to Section 2.3. In the backward pass, the partitioning should be done in a column-wise way because one needs to compute \( W^\top x \).

4 Implementation and Experimental Evaluation

The proposed coded elastic computing technique has been implemented on top of Apache REEF.Elastic Group Communication (EGC) framework. REEF EGC provides an API allowing to implement elastic computations by chaining fault-tolerant MPI-like primitives. In this short paper we assess the performance of our elastic code computing approach through 2 mini-benchmarks.

Matrix-vector mini-benchmark. In this mini-benchmark, we test that indeed the time cost decreases linearly with the increase in the number of machines available. We mimic an elastic computing environment on Amazon EC2 by using different numbers of M4.large instances to compute the same matrix-vector product \( Xw \). The master node is a M4.4xlarge machine. The matrix is randomly generated and with size \( 100000 \times 10000 \), and it is partitioned initially into 10 submatrices of size \( 10000 \times 10000 \). Then, they are encoded into 20 submatrices of the same size, and each submatrix is stored at one machine (for a total of 20 machines). To mimic the elastic events, we change the number of available machines by injecting artificial failures. The maximum number of failures is 10. The per-iteration overall time (including both communication and computation) is shown in Figure 2(a). The result is averaged using 20 independent trials. As we can see, the coded elastic computing technique can utilize the extra machines when the number of machines increases (see the blue bars). We also compare it with a non-adaptive coded computation baseline. It can be seen that when the number of machines increases, the adaptive scheme can provide increasing speedup when compared to the non-adaptive coded computing scheme. The maximum observed speedup in our experiment is 46%.

Linear model mini-benchmark. In this experiment, we test a coded implementation of linear regression using line-search-based batch gradient descent (the same setting as the baseline). We run the test over 20 machines on a Microsoft internal multi-tenancy cluster. Each data point in the dataset has 3352 features, and we sample 10000 data for training and 10000 data for testing. We

\[ \text{https://github.com/interesaaat/reef/tree/elastic-sync} \]

Note that in the non-adaptive scheme, we do not utilize the coding to deal with stragglers because the amount of stragglers in m4 instances is small. The per-iteration time increases slightly with the number of machines because of the extra communication overhead.
Figure 2: Mini-benchmarks experiments (results normalized due to confidentiality). The per-iteration time in (a) and (b) includes decoding and communication.

generate random failures and allow REEF EGC to reschedule new machines when failures occur. We start with Figure 2(b) where we plot the time for each iteration. In theory, when all the workers are present, the computational cost per iteration should be the same as the uncoded case. However, the coded method (all) has slight overhead due to coding and a second round of communication for coded linear regression using matrix-vector multiplications. The coded method (half) shows the cost when only half of the workers are running, which is, as expected, twice the cost of the uncoded method. In Figure 2(c) and 2(d) we report the generalization error and we compare our coded elastic computing technique with three baselines, namely noiseless (no failure), ignore the failure and continue, and an existing algorithm called Elastic Distr-BGD [50]. The coded method can achieve the same convergence behavior as the noiseless case, while the ignore method achieves worst generalization error even for different regularization parameters (the other techniques all use the same regularization coefficient 0.001). In Figure 2(c) we show 5 different experiments on Distr-BGD using the same failure probability but different realizations. The convergence of Distr-BGD depends on when a failure occurs and can lead to different algorithm performance. This is because the Distr-BGD keeps using previous gradient vectors at the failed machines, and this can (1) lead to overfitting, and (2) make the optimization miss the minimum point. In the plot of Distr-BGD, the valley part is due to overfitting, and the sudden change to near flat loss growth is because when the gradient descent has missed the optimal point of empirical training loss, the fixed gradient at the failure nodes makes the line search choose the smallest step size. In some cases, the Distr-BGD

\[ \text{Note that although a fine-grained grid search is possible to find the best regularization coefficient, it is very time consuming in practice.} \]
works extremely well because the fixed gradients act like momentum and can improve the speed of convergence.

From the experiment results, we can see that the coded elastic computing technique can obtain the same convergence behavior as ordinary gradient-descent-based algorithms but can elastically allocate the workload based on the number of available machines without moving data around.

5 Conclusions

The coded elastic computing framework presented in this paper can deal with new cloud offerings where machines can leave and join during the computation. Our framework handles the elastic events in a positive way, meaning that when machines leave, it shifts the computation to the remaining workers, and when new machines join the computation, it actively reduces the workload of existing machines without the reallocation of data. We prove that the coded elastic computing technique can achieve the same memory-access cost as the noiseless case, and hence is optimal for memory-bound applications. Using experiments in both Amazon EC2 and on a Microsoft multi-tenancy cluster, we show that the coded elastic computing technique can achieve the same convergence behavior as if no failure occurs, and can dynamically adjust working loads respect to the number of remaining workers. The proposed technique can be applied to coded matrix-vector, matrix-matrix multiplications and linear regression, and potentially other applications where the large-scale matrix operations are the bottleneck.

6 Acknowledgment

We sincerely thank Professor Michael Mahoney, Professor Kannan Ramchandran, Professor Viveck Cadambe, Professor Joseph E. Gonzalez, Vipul Gupta, Swanand Kadhe, Zhewei Yao, Sanghamitra Dutta, and Haewon Jeong for helpful discussions and suggestions.

References


