

# Learning a Code: Machine Learning for Approximate Non-Linear Coded Computation

Jack Kosaian<sup>1</sup>, K.V. Rashmi<sup>1</sup>, and Shivaram Venkataraman<sup>2</sup>

<sup>1</sup>Computer Science Department, Carnegie Mellon University

<sup>2</sup>Microsoft Research

<sup>1</sup>{jkosaian, rvinayak}@cs.cmu.edu

<sup>2</sup>shivaram.venkataraman@microsoft.com

## Abstract

Machine learning algorithms are typically run on large scale, distributed compute infrastructure that routinely face a number of unavailabilities such as failures and temporary slowdowns. Adding redundant computations using coding-theoretic tools called “codes” is an emerging technique to alleviate the adverse effects of such unavailabilities. A code consists of an encoding function that proactively introduces redundant computation and a decoding function that reconstructs unavailable outputs using the available ones. Past work focuses on using codes to provide resilience for linear computations and specific iterative optimization algorithms. However, computations performed for a variety of applications including inference on state-of-the-art machine learning algorithms, such as neural networks, typically fall outside this realm. In this paper, we propose taking a *learning-based* approach to designing codes that can handle non-linear computations. We present carefully designed neural network architectures and a training methodology for learning encoding and decoding functions that produce *approximate* reconstructions of unavailable computation results. We present extensive experimental results demonstrating the effectiveness of the proposed approach: we show that the our learned codes can accurately reconstruct 64 – 98% of the unavailable predictions from neural-network based image classifiers (multi-layer perceptron and ResNet-18) on the MNIST, Fashion-MNIST, and CIFAR-10 datasets. To the best of our knowledge, this work proposes the first learning-based approach for designing codes, and also presents the first coding-theoretic solution that can provide resilience for any non-linear (differentiable) computation. Our results show that learning can be an effective technique for designing codes, and that learned codes are a highly promising approach for bringing the benefits of coding to non-linear computations.

## 1 Introduction

Machine learning serves as the backbone for a wide variety of cognitive tasks such as image classification, object recognition, and natural language processing. Today, applications can leverage state-of-the-art machine learning models by using cloud services that offer machine learning as a service [Azu, Goo, AWS]. To handle large traffic, such service providers typically use a distributed setup with a large number of interconnected servers (compute nodes). It is well-known that such a distributed compute infrastructure faces a number of unavailability events [Dea, RSG<sup>+</sup>14, SAP<sup>+</sup>13]. First, these clusters are typically built out of commodity components making failures the norm rather than the exception. Second, various factors including load imbalance and resource contention cause transient slowdowns. (Servers facing such temporary unavailability are called stragglers.) Both of these unavailabilities adversely affect service response time (latency).

A natural strategy for addressing unavailability in other domains such as communications and data storage has been through a proactive approach of adding *redundancy*: making use of extra resources upfront to aid in recovery from unavailability. The effectiveness of using redundancy to reduce latency in computer systems has been shown both theoretically [JLS14, GZD<sup>+</sup>15, SLR16, LK13, WJW14] as well as in practical systems [AGSS12, VMGS12, DB13, RCK<sup>+</sup>16]. A naive approach of adding redundancy is to replicate (that is, to have multiple copies), but this approach leads to significant resource overhead. A tool from the domain of coding theory, called *erasure codes* [RU08], provides a means for adding redundancy with significantly lesser overhead as compared to replication. Erasure codes

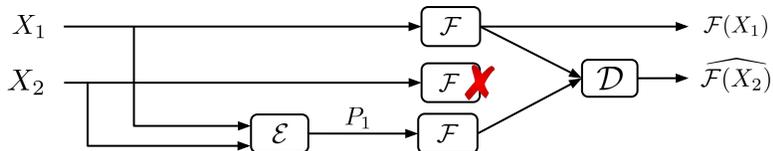


Figure 1: Coded computation with  $k = 2$  data inputs, and  $r = 1$  parity input generated by the encoding function  $\mathcal{E}$ . Here, the second output  $\mathcal{F}(X_2)$  is unavailable (denoted by a red cross). The decoding function  $\mathcal{D}$  acts on the available outputs  $\mathcal{F}(X_1)$  and  $\mathcal{F}(P_1)$  to produce an approximate reconstruction of the second output denoted by  $\widehat{\mathcal{F}(X_2)}$ .

have been successfully employed in communication [RU08], storage [PGK88, HDF, RSG<sup>+</sup>14, HSX<sup>+</sup>12, SAP<sup>+</sup>13], and distributed caching [RCK<sup>+</sup>16] systems to efficiently alleviate the impact of unavailabilities.

Coded computation is an emerging technique which extends the use of erasure codes to recover from *unavailability of computation*. Suppose there are  $k$  data inputs  $X_1, X_2, \dots, X_k$ , and suppose the goal is to apply a given function  $\mathcal{F}$  to these  $k$  data inputs, that is, to compute  $\mathcal{F}(X_1), \mathcal{F}(X_2), \dots, \mathcal{F}(X_k)$ . The computations  $\mathcal{F}(X_i)$  for different  $i \in \{1, \dots, k\}$  are performed on separate, unreliable devices, and hence each individual computation can straggle or fail arbitrarily. We let  $r$  represent a resilience parameter. The framework of coded computation involves two functions, an *encoding function*  $\mathcal{E}$  and a *decoding function*  $\mathcal{D}$ . First, the encoding function  $\mathcal{E}$  acts on the  $k$  data inputs  $X_1, X_2, \dots, X_k$  to generate  $r$  redundant inputs, called “parities,” which we denote as  $P_1, P_2, \dots, P_r$ . The given function  $\mathcal{F}$  is then applied on these  $(k + r)$  inputs (data and parity) on separate, unreliable devices that can fail or straggle arbitrarily. If any  $r$  or fewer outputs (out of the total  $(k + r)$  outputs) are unavailable, the decoding function  $\mathcal{D}$  is applied on all the available outputs to reconstruct the unavailable ones among  $\mathcal{F}(X_1), \mathcal{F}(X_2), \dots, \mathcal{F}(X_k)$ . Figure 1 illustrates the coded-computation framework. Given  $\mathcal{F}$ ,  $k$ , and  $r$ , the goal is to design the encoding function  $\mathcal{E}$  and the decoding function  $\mathcal{D}$  to enable reconstruction of unavailable outputs of  $\mathcal{F}$ .

Many recent works have employed erasure codes for coded computation of *linear* functions such as distributed matrix-vector multiplication [LLP<sup>+</sup>18, DCG16, LMAA16, YMAA17, WLS18, DCG17, RPPA17, MCJ18], and specific classes of iterative optimization algorithms [KSDY17, KSDY18, TLDK17]. However, to the best of our knowledge, none of the existing works are applicable for broader classes of *non-linear* computations, for example, when  $\mathcal{F}$  is a neural-network. While the fully-connected and convolutional layers common to neural-networks are linear, they are executed along with non-linearities such as activation functions and max-pooling, effectively making the overall function non-linear. These compelling applications serve as our motivation for designing erasure codes that can handle non-linear computations.

In the history of coding theory, advances in the design of codes have largely come about through human creativity, making use of handcrafted mathematical constructs. For coded computation of non-linear functions that arise in general tasks including machine learning applications and beyond, complex non-linear interactions make it challenging to handcraft erasure codes. In this paper, we propose to overcome this challenge via a novel approach for designing erasure codes, that of *learning codes*.

Learning an erasure code involves learning the encoding and the decoding functions ( $\mathcal{E}$  and  $\mathcal{D}$ ). Unlike the traditional approach in erasure coding, we allow the outputs of the decoding function to be an *approximation* of the unavailable outputs. Approximate outputs are sufficient for many applications, such as machine learning algorithms since many of these algorithms themselves are approximate. For any input  $X$  and a given function  $\mathcal{F}$ , we denote the (approximate) reconstruction of  $\mathcal{F}(X)$  as  $\widehat{\mathcal{F}(X)}$ . We make use of the ability of neural networks to perform universal function approximation by expressing the encoding and the decoding functions as neural networks. We train the neural networks for the encoding and the decoding functions in tandem via backpropagation directly through the given function  $\mathcal{F}$ .

Our approach is applicable for designing codes for imparting resilience to any differentiable non-linear function  $\mathcal{F}$ .<sup>1</sup> In this paper, we focus our attention on learning codes for machine learning models, specifically for the (often non-linear) computations during inference. We focus on inference as it is typically a user facing operation, and hence reducing the computation time during inference through failure and straggler mitigation has a significant impact on

<sup>1</sup>Although our approach is applicable for linear functions as well, we focus primarily on non-linear functions. There are several existing works (e.g., [LLP<sup>+</sup>18, DCG16, LMAA16, YMAA17, WLS18, DCG17, RPPA17, MCJ18]) that address only linear functions. These approaches may be more suitable for linear functions as they guarantee exact reconstruction of unavailable outputs.

service quality [Dan17]. In our evaluation, for the sake of concreteness, we particularly focus on functions  $\mathcal{F}$  that are neural network models, and use the term “base model” to refer to  $\mathcal{F}$ . However, we emphasize that our solution extends to any differentiable function, making it applicable to a large class of tasks in machine learning and beyond.

We evaluate our framework using two neural-network based image classifiers as base models (a multi-layer perceptron (MLP) and ResNet-18) using MNIST [LeC], Fashion-MNIST [XRV17], and CIFAR-10 [Ale] datasets. Our experimental results show that the proposed approach can *accurately reconstruct a significant fraction of the unavailable outputs*: for example, 98.87%, 92.06%, and 80.84% of ResNet-18 classifier outputs are accurately reconstructed on MNIST, Fashion-MNIST, and CIFAR-10 datasets respectively.

Our experimental results are highly promising for the following reasons. Consider the example application of handling failures and stragglers in distributed, machine-learning inference services. Inference services typically provide strong guarantees on response times (Service Level Agreements or SLAs) [Dan17]. Requests that face unavailability have prediction accuracy no better than random guessing in the absence of any corrective measures. Considering a distributed service employing ResNet-18 models, if say 10% of requests are unavailable, the overall prediction accuracy for CIFAR-10 will drop from 93.47% to 84.12%. Our learned codes can reconstruct the predictions for most of these unavailable cases and get close to the prediction accuracy of the underlying classifier at the cost of performing some redundant computation. Under the same scenario as considered above, learned codes can improve the overall prediction accuracy from 84.12% to 90.59% for CIFAR-10, and from 89.28% to 98.75% for MNIST by using only 20% redundant base model computations ( $k = 5, r = 1$ ).

A note on the scope of this paper: The goal of this work is to explore the feasibility of taking a learning-based approach for designing erasure codes to impart resilience to general non-linear computations; our focus is not on optimizing the encoding and decoding function architectures for computational efficiency.

The main contributions of this work are as follows:

1. To the best of our knowledge, we propose the first *learning-based* approach to designing erasure codes.
2. To the best of our knowledge, we propose the first coded-computation approach for providing resilience to *non-linear* functions, making it applicable to a large class of tasks in machine learning and beyond.
3. We carefully design neural network architectures and a training methodology for learning the encoding and decoding functions based on multilayer-perceptrons and dilated convolutional neural networks.
4. Through extensive evaluation on two neural-network based image classifiers (a multi-layer perceptron (MLP) and ResNet-18) using MNIST, Fashion-MNIST, and CIFAR-10 datasets, we show that our learned codes can accurately reconstruct 64 – 98% of the unavailable predictions.

## 2 Related Work

A host of recent works have explored using coding theoretic approaches to impart resilience to distributed linear computations such as matrix multiplication. Lee et al. [LLP<sup>+</sup>18] use a family of codes called “maximum-distance-separable” (MDS) codes to mitigate stragglers in distributed matrix-vector multiplication. In [DCG16], Dutta et al. propose Short-Dot codes to decompose long dot products that arise in certain matrix-vector multiplications into smaller products which facilitates parallel computation of such products. Li et al. [LMAA16] present a framework for navigating the tradeoff between computation time and communication time in coded computation schemes for matrix multiplication. Yu et al. [YMAA17] propose Polynomial Codes for distributed matrix multiplication, which reconstruct the full matrix multiplication result using the minimal number of results from workers. Sparse Codes are introduced by Wang et al. [WLS18] to exploit the sparsity of matrix operands in order to reduce decoding complexity in coded matrix-matrix multiplication. In [DCG17], Dutta et al. employ linear codes for resilient distributed convolution between two vectors. Reiszadeh et al. [RPPA17] propose a scheme to balance the load across compute nodes for coded, distributed matrix-multiplication by taking into account heterogeneity of compute resources. In [MCJ18], Mallick et al. propose using rateless codes for distributed matrix-vector multiplication in order to make use of partial work completed by straggling nodes. In comparison to the above works which are applicable to *only linear* computations, we present a learning-based approach that learns codes that can handle any differentiable *non-linear* computation.

In another direction in coded computation, several recent works present approaches to using codes for providing resilience to specific iterative optimization algorithms that are employed during training of machine learning algorithms.

Tandon et al. [TLDK17] propose a straggler mitigation scheme for data-parallel gradient descent which involves having multiple copies of the data across the worker nodes. Under this scheme, each worker node sends a carefully constructed linear combination of its computed gradients to a master node such that the master node can complete a gradient descent iteration without having to wait for results from all the worker nodes. In [KSDY17, KSDY18], Karakus et al. propose a coded-computation approach wherein both the data and labels of a training set are encoded, and the original optimization algorithm is directly run on the encoded training dataset. For specific optimization algorithms (e.g., gradient descent and L-BFGS) and machine learning tasks (e.g., ridge regression, matrix factorization, and logistic regression), the authors present code constructions that achieve stable convergence and reduced runtime as compared to replication-based approaches. In [MSM18], Maity et al. encode the second moment of the data matrix using LDPC codes in order to mitigate the effect of stragglers on gradient descent. The authors show that encoding the second moment reduces the number of aggregation steps necessary per training iteration compared to directly encoding the data matrix. In contrast to these lines of work that focus on specific iterative optimization algorithms that arise during the training phase of machine learning, the focus of our work is to add resilience through redundant computation to any differentiable non-linear computation that arise during the *inference* phase of machine learning.

Two recent works have explored taking a learning approach to designing decoding algorithms for *existing* error-correcting-codes employed in the domain of communication. Nachmani et al. [NML<sup>+</sup>18] propose using feed-forward and recurrent neural networks for decoding a family of codes called “block codes”. Kim et al. [KJR<sup>+</sup>18] show that recurrent neural networks can learn close-to-optimal decoding algorithms for several classes of well known codes employed in the domain of communication. In comparison with these works, we propose and establish the feasibility of taking a learning-based approach for the end-to-end design of codes, i.e., learning *both encoding and decoding* algorithms.

Another related line of work is on using neural networks for image compression and cryptography [TOH<sup>+</sup>16, TJZ<sup>+</sup>17, AA16]. While these lines of work are similar in spirit to learning an erasure code (transforming input data into alternate representation for later reconstruction), the overall goal, and thus the structure of the architecture and the training methodology differ significantly.

### 3 Learning a Code

In this section, we describe our proposed approach for learning erasure codes. Recall the coded computation setup (an example of which is illustrated in Figure 1): The encoding function  $\mathcal{E}$  acts on the  $k$  data inputs to create  $r$  parity inputs. The function  $\mathcal{F}$  is then applied on these  $(k + r)$  inputs (data and parity) on separate, unreliable devices that can fail or straggle arbitrarily. If any  $r$  or fewer of these outputs are unavailable, the decoding function  $\mathcal{D}$  is applied on all the available outputs to reconstruct the unavailable outputs corresponding to the  $k$  data inputs  $X_1, X_2, \dots, X_k$ . The goal is to learn the encoding function  $\mathcal{E}$  and the decoding function  $\mathcal{D}$  with the objective of minimizing a chosen loss function (which we discuss in more detail below in Section 3.1).

We use neural networks to learn the encoding and decoding functions. We find neural networks to be a natural choice for learning the encoding and decoding functions due to their ability to perform universal function approximation [Hor91].

In the remainder of this section, we first present our training methodology, and subsequently describe the neural network architectures for learning the encoding and decoding functions.

#### 3.1 Training methodology

Recall that our overall architecture has three functions: the given function  $\mathcal{F}$  whose distributed execution is to be made resilient using the learned codes, the encoding function  $\mathcal{E}$ , and the decoding function  $\mathcal{D}$ . During training the goal is to train the parameters of the neural networks for the encoding and the decoding functions. Note that the given function  $\mathcal{F}$  is not modified during this training.

When the given function  $\mathcal{F}$  is a machine learning algorithm, we train the encoding and the decoding functions using the same training dataset (whenever available) that was used to train  $\mathcal{F}$ . When such a training dataset is not available, which will be the case for generic functions  $\mathcal{F}$  outside the realm of machine learning, one can instead generate a training dataset comprising pairs  $(X, \mathcal{F}(X))$  for various values of  $X$  in the domain of  $\mathcal{F}$ . Each sample for training the encoding and decoding functions uses a set of  $k$  (randomly chosen) inputs from the training dataset. For any sample, we perform a forward and a backward pass for each of  $\binom{k+r}{r}$  possible unavailability scenarios, except

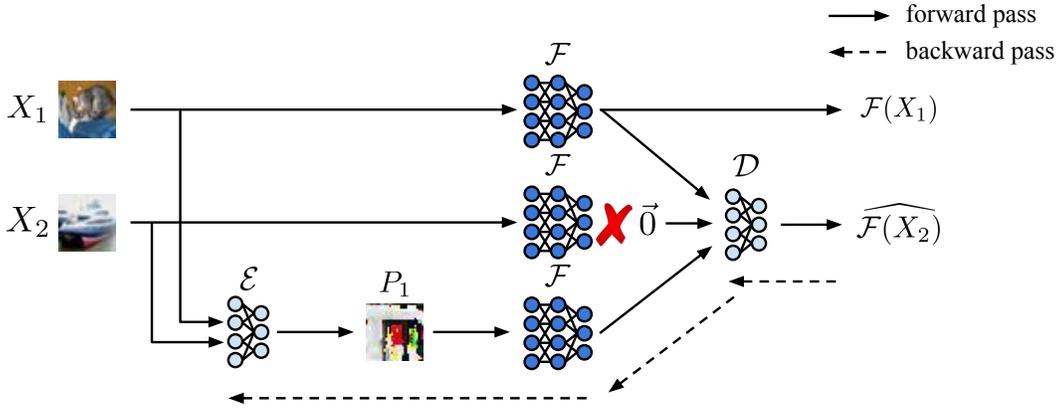


Figure 2: A forward and a backward pass in training the encoding and decoding functions for ( $k = 2, r = 1$ ) with the given function  $\mathcal{F}$  as a neural-network based image classifier. In the forward pass (solid left-to-right arrows), images  $X_1$  and  $X_2$  are fed as inputs to the neural network for the encoding-function  $\mathcal{E}$  to generate parity image  $P_1$ . Each of  $X_1, X_2$ , and  $P_1$  are fed through the given neural-network function  $\mathcal{F}$ . The available function outputs are fed as input to the neural network for the decoding function  $\mathcal{D}$ . The unavailable output  $\mathcal{F}(X_2)$  (denoted by a red cross) is replaced with a vector of zeros as input to  $\mathcal{D}$ . The decoding function  $\mathcal{D}$  produces an approximate reconstruction of  $\mathcal{F}(X_2)$ , denoted by  $\widehat{\mathcal{F}(X_2)}$ . The backward pass (dotted right-to-left arrows) propagates loss through  $\mathcal{D}$ ,  $\mathcal{F}$ , and  $\mathcal{E}$ . Only the parameters of  $\mathcal{E}$  and  $\mathcal{D}$  are updated (lightly shaded neural networks).

for the case where all unavailable outputs correspond to parity inputs (since the only role of parities is to aid in the reconstruction of unavailable outputs corresponding to the data inputs). Any iterative optimization algorithm, such as gradient descent and its variants, may be used for training.

A forward and a backward pass under our training method is illustrated in Figure 2. A forward pass involves the following steps. The  $k$  data inputs  $X_1, X_2, \dots, X_k$  are fed through the encoding function to generate  $r$  parity inputs  $P_1, P_2, \dots, P_r$ . Each of the  $(k + r)$  inputs (data and parity) are then fed through the given function  $\mathcal{F}$ . The resulting  $(k + r)$  outputs  $\mathcal{F}(X_1), \dots, \mathcal{F}(X_k), \mathcal{F}(P_1), \dots, \mathcal{F}(P_r)$  are fed through the decoding function  $\mathcal{D}$ , out of which no more than  $r$  are made unavailable (discussed in detail in Section 3.3.1). The decoding function outputs an (approximate) reconstruction for the unavailable function outputs among  $\mathcal{F}(X_1), \dots, \mathcal{F}(X_k)$ . The corresponding backward pass involves using any chosen loss function (discussed in detail below) for backpropagation through  $\mathcal{D}$ ,  $\mathcal{F}$ , and  $\mathcal{E}$ . We train the encoding and decoding functions in tandem via backpropagation of losses directly through  $\mathcal{F}$ . In other words, the parameters of the encoding and the decoding functions are updated by backpropagating through  $\mathcal{F}$ . Since training backpropagates directly through  $\mathcal{F}$ , this approach is applicable to any given differentiable function  $\mathcal{F}$ .

We consider two types of losses when training the encoding and the decoding functions:

1. *Loss with respect to function outputs:* Loss is computed between the function output  $\mathcal{F}(X)$  and its approximate reconstruction  $\widehat{\mathcal{F}(X)}$  produced by the decoding function. This approach can be employed for any given function  $\mathcal{F}$ .
2. *Loss with respect to true labels:* When  $\mathcal{F}$  is a machine learning algorithm, there is an additional option of calculating the loss using the true labels (when available in the training dataset). For example, consider  $\mathcal{F}$  to be a neural network for image classification, and let  $Y$  represent the true label for an input image  $X$ . Under this approach, the loss is computed between the true label  $Y$  and the label predicted using  $\widehat{\mathcal{F}(X)}$ .

The specific loss functions employed in our evaluation under both of the above approaches are discussed in Section 4.

We next move on to describing the neural network architectures for learning the encoding and decoding functions.

Layer	Layer Type
1	FC: $kn^2 \times kn^2$
2	FC: $kn^2 \times rn^2$

Layer	Layer Type
1	Conv: $3 \times 3$ , dilation: 1
2	Conv: $3 \times 3$ , dilation: 1
3	Conv: $3 \times 3$ , dilation: 2
4	Conv: $3 \times 3$ , dilation: 4
5	Conv: $3 \times 3$ , dilation: 8
6	Conv: $3 \times 3$ , dilation: 1
7	Conv: $1 \times 1$ , dilation: 1

Table 1: Neural network architectures for encoding functions employing fully-connected (FC) and convolutional (Conv) layers. All convolutional layers have stride of 1. In each network, ReLU activation functions are used after all but the final layer. The activation functions are omitted above for brevity.

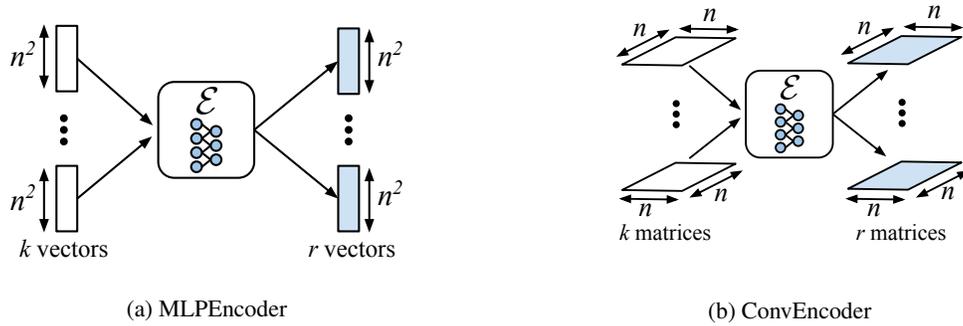


Figure 3: Encoding function architecture: (a) MLP Encoder flattens inputs into  $n^2$ -length vectors and produces  $n^2$ -length parity vectors. (b) Conv Encoder encodes over inputs in their original dimension, as  $n \times n$  matrices. The  $k$  input matrices are treated as  $k$  input channels.

## 3.2 Encoding function architectures

We consider two neural network architectures for learning the encoding function which are applicable to any differentiable function  $\mathcal{F}$ . For concreteness, we describe the proposed architectures below by setting the given function  $\mathcal{F}$  as a neural network for image classification over  $m$  classes, and use the term “base model” to refer to  $\mathcal{F}$ . For such an  $\mathcal{F}$ , each data input  $X$  is an  $n \times n$  pixel image. Each function output  $\mathcal{F}(X)$  is an  $m$ -length vector representing output from the last layer of the neural network classifier.

We now describe two neural network architectures for learning the encoding function. Recall that the encoding function acts on  $k$  data inputs to create  $r$  parity inputs. We first describe the architecture considering single-channel images as inputs, and consider multi-channel images in Section 3.2.3.

### 3.2.1 MLP Encoder

We first consider a simple 2-layer multilayer-perceptron (MLP) encoding function architecture, because it represents the basis for universal function approximation among neural networks [Hor91]. We call this encoding function architecture *MLP Encoder*. Under this architecture, the  $n \times n$  data inputs are flattened into  $n^2$ -length vectors, as illustrated in Figure 3a. The  $k$  flattened vectors from inputs  $X_1, X_2, \dots, X_k$ , are concatenated to form a single  $kn^2$ -length input vector to the MLP. The first fully-connected layer of the MLP produces a  $kn^2$ -length hidden vector. The second fully-connected layer produces an  $rn^2$ -length output vector, which represents the  $r$  parity inputs. Each layer used in MLP Encoder is outlined in Table 1a.

The fully-connected nature of the MLP allows for computation of arbitrary combinations from the  $kn^2$  total inputs with a small number of layers. While simple in design and effective for many scenarios (as will be shown in

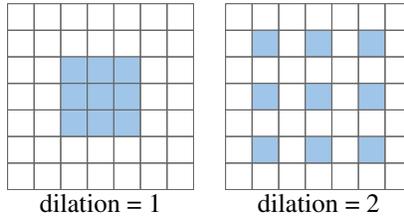


Figure 4:  $(3 \times 3)$  Dilated Convolution. Traditional convolution (left) operates on adjacent cells. Dilated convolution (right) considers cells spread apart from one another.

Section 4.2.2), the high parameter count of the fully-connected layers can lead to overfitting. We next describe an alternate encoding function architecture that avoids overfitting, which we call *ConvEncoder*.

### 3.2.2 ConvEncoder

The ConvEncoder architecture makes use of multiple convolutional layers as detailed in Table 1b. Unlike MLPEncoder, ConvEncoder computes over data inputs in their original  $n \times n$  representation. As depicted in Figure 3b, the  $k$  inputs to the encoding function are treated as  $k$  input channels to the first convolution layer. This is similar to feeding the RGB representations of an image to a convolutional neural network for image classification. We explain how the encoder handles multi-channel inputs in Section 3.2.3.

The traditional use of convolutional layers for image classification involves repeated downsampling of an input image to gradually expand the receptive field of convolutional filters. This approach works well when the output dimension of the network is significantly smaller than the input dimension, which is often the case for image classification. However, the encoding function of a code produces outputs that have the same dimension as the inputs (see Figure 3b). Hence, using convolutional layers with downsampling would necessitate subsequent upsampling to bring the outputs back to the input dimension. This has been shown to be inefficient in the context of image segmentation [Fis16]. To overcome this issue, we employ *dilated convolutions* [Fis16]. As shown in Figure 4, this approach increases the receptive field of a convolutional filter exponentially with linear increase in the number of layers.

Table 1b shows each layer of ConvEncoder. The first layer has  $k$  input channels and the final layer has  $r$  output channels, one for each parity to be produced. Each of the intermediate layers has  $20k$  input channels and  $20k$  output channels. We increase the receptive field of convolutions by increasing the dilation factor, borrowing this architecture from [Fis16], where it was used for image segmentation.

ConvEncoder uses less parameters than MLPEncoder but requires more layers to enable combinations of all input pixels. The lower parameter count compared to MLPEncoder helps avoid overfitting, as will be shown in Section 4.2.2.

### 3.2.3 Multi-channel input

It is common to represent colored images as having multiple channels. For example, a  $32 \times 32$  RGB image would consist of 3 channels, each  $32 \times 32$  in size, representing the pixel values of each of the red, green, and blue components. Our encoding function architectures handle multi-channel inputs by encoding across each channel independently. For example, an encoding function with  $k$  RGB images as inputs would encode across the  $k$  red channels to produce  $r$  “red” parity channels, and similarly for green and blue channels. The  $r$  “red”, “green”, and “blue” parity channels are combined together to create  $r$  parity “RGB” images.

## 3.3 Decoding function architecture

As in Section 3.2, for concreteness, we describe our decoding function architecture by setting the given function  $\mathcal{F}$  as a neural network for image classification over  $m$  classes. It is easy to re-purpose the proposed architecture for any differentiable function  $\mathcal{F}$ . Recall that we refer to  $\mathcal{F}$  as the base model. The base model output  $\mathcal{F}(X)$  for any input  $X$  is an  $m$ -length vector representing output from the last layer of the neural-network classifier. Further recall that the base model  $\mathcal{F}$  is applied on the  $(k + r)$  inputs  $X_1, X_2, \dots, X_k, P_1, \dots, P_r$  on separate, unreliable compute nodes that can fail or straggle arbitrarily. The decoding function  $\mathcal{D}$  operates on all the available base model outputs and reconstructs approximations of up to  $r$  unavailable base model outputs among  $\mathcal{F}(X_1), \mathcal{F}(X_2), \dots, \mathcal{F}(X_k)$ .

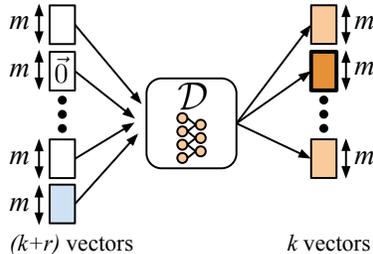


Figure 5: Decoding function architecture: The second input is unavailable and is set to a vector of zeros and the second (bolded) output represents its reconstruction. Parity inputs are shaded.

Layer	Layer Type
1	FC: $(k+r)m \times km$
2	FC: $km \times km$
3	FC: $km \times km$

Table 2: Neural network architecture for decoding function employing fully-connected (FC) layers. ReLU activation functions are used after all but the final layer. The activation functions are omitted above for brevity.

Figure 5 presents the overall architecture of our decoding process. The two key design choices for the decoding function architecture are: (a) representation of the unavailable base model outputs at the input layer of the neural network for the decoding function, and (b) the neural network architecture used for learning the decoding function.

### 3.3.1 Representing unavailability

A key design consideration for the decoding function is in the representation of the unavailable base model outputs at its input layer. We design the decoding function to take the  $(k+r)$  vectors of length  $m$ ,  $\mathcal{F}(X_1), \dots, \mathcal{F}(X_k), \mathcal{F}(P_1), \dots, \mathcal{F}(P_r)$ , as inputs. Some of these inputs to the decoding function could be unavailable. In place of any unavailable input, we insert a vector of all zeros. Note that an alternative approach is to provide the decoding function with only the (concatenated) available inputs. We chose the former as it allows us to learn a decoding function that depends on the relative position of the unavailable inputs; providing only the available inputs would hide this information. This approach is inspired by traditional (hand-crafted) erasure codes whose decoding functions leverage positional information. Correspondingly, the output of the decoding function maintains positional information and consists of  $k$  vectors  $\widehat{\mathcal{F}(X_1)}, \dots, \widehat{\mathcal{F}(X_k)}$ , each representing an approximate reconstruction of corresponding potentially unavailable function output.

### 3.3.2 Decoding function architecture

We design the neural network for learning the decoding function as a 3-layer MLP as described in Table 2. We use the raw outputs of the base model  $\mathcal{F}$  as input to the decoding function. Note that we do not convert such outputs to a probability distribution (via a softmax operation) as is typically done during training of classifiers.

## 4 Evaluation

As discussed in Section 3, we evaluate our approach of learning codes for imparting resilience to non-linear computations by setting the base model  $\mathcal{F}$  as inference on neural-network based image classifiers. For any input  $X$ ,  $\mathcal{F}(X)$  represents the output from the last layer of the neural network used as the base model. We start by describing our experimental setup and then present results using two neural-network based image classifiers as base models (a multi-layer perceptron (MLP) and ResNet-18) on the MNIST [LeC], Fashion-MNIST [XRV17], and CIFAR-10 [Ale] datasets. Finally, we present a more detailed analysis of the accuracy attained by the learned codes and the quality of the predictions obtained from the reconstructed outputs.

Base Model	MNIST	Fashion-MNIST	CIFAR-10
ResNet-18	0.9920	0.9285	0.9347
Base-MLP	0.9793	0.8947	-

Table 3: Test accuracies for the base models used in our experiments on the MNIST, Fashion-MNIST, and CIFAR-10 datasets.

## 4.1 Experimental setup

We implement all the encoding and decoding function architectures as well as the training methodology using PyTorch [Pyt]. Since (to the best of our knowledge) this work presents the first training methodology for learning codes for coded computation, we experiment with several loss functions and architectures, and consider multiple accuracy metrics. We describe our experimental setup below.

### 4.1.1 Loss functions used in training

As discussed in Section 3.1, we use two approaches for calculating the loss when training the neural networks for the encoding and the decoding functions: (a) calculating the loss with respect to the base model output and (b) calculating the loss with respect to the true label (when available in the training dataset). When calculating the loss with respect to the base model output, we experiment with two different loss functions: (a) mean-squared error (denoted by *MSE-Base*) and (b) KL-divergence (denoted by *KL-Base*) between  $\widehat{\mathcal{F}(X)}$  and  $\mathcal{F}(X)$ . When calculating loss with respect to the true labels of the underlying task, we use the cross-entropy between  $\widehat{\mathcal{F}(X)}$  and the true label of  $X$  (denoted by *XENT-Label*).

### 4.1.2 Base models

We experiment with two neural network architectures as base models: Base-MLP and ResNet-18. Base-MLP is a 3-layer multilayer-perceptron used for the MNIST and Fashion-MNIST datasets containing three fully-connected layers with dimensions  $784 \times 200$ ,  $200 \times 100$ , and  $100 \times 10$  with ReLU activation functions following all but the final layer. We choose an MLP model due to its simplicity and its reported success on MNIST [Yan98]. ResNet-18 [HZRS16] is an 18-layer state-of-the-art neural network for image classification consisting of convolutional, pooling, and fully-connected layers.<sup>2</sup> We choose to use ResNet-18 for two reasons: (a) it has been shown to provide high classification accuracy on both CIFAR-10 and Fashion-MNIST, and (b) it is a significantly more complex model than Base-MLP and thus provides a good alternative evaluation point for our proposed approach. Table 3 shows the classification accuracies of the base models. We do not use Base-MLP as a base model for CIFAR-10 as similar architectures have been shown to achieve low accuracy [LMK15].

### 4.1.3 Encoding and decoding function architectures

Recall that, in Section 3.2 we presented two architectures for learning the encoding function: MLP Encoder and Conv Encoder. Then, in Section 3.3, we presented an MLP-based architecture for learning the decoding function. We present experimental results for all the proposed architectures.

### 4.1.4 Parameters and training details

We perform experiments for all combinations of the configuration settings discussed above for  $k = 2$  and  $k = 5$  with  $r = 1$ . We focus on  $r = 1$  because this corresponds to the case of typical unavailability faced in today’s data centers as shown from measurements on Facebook’s data analytics cluster [RSG<sup>+</sup>13, RSG<sup>+</sup>14]. With  $r = 1$ , the parameter settings with  $k = 2$  and  $k = 5$  correspond to 50% and 20% redundant computation, respectively.

Training uses minibatches of 64 samples for  $k = 2$  and 32 samples for  $k = 5$ . Each sample in the minibatch consists of  $k$  images from the dataset drawn randomly without replacement (i.e., no image is used more than once per epoch). Thus each minibatch for  $k = 2$  consists of 128 images and for  $k = 5$  consists of 160 images from the dataset. The encoding and decoding functions are trained in tandem using the Adam optimizer [Die15] with learning rate of

<sup>2</sup>We use the ResNet-18 model described at <https://github.com/zalando-research/fashion-mnist>

0.001 and L2-regularization of  $1 \times 10^{-5}$ . The weights for the convolutional layers are initialized via uniform Xavier initialization [GB10] and weights for the fully-connected layer are initialized according to  $\mathcal{N}(0, 0.01)$ . All bias values are initialized to zero.

#### 4.1.5 Accuracy metrics

We measure the accuracy of the reconstructed output with respect to the machine learning task at hand using the following two metrics:

1. *Recovery-accuracy*: This metric measures the accuracy of the reconstructed output based on its ability to recover the label predicted by the base model output. For example, when  $\mathcal{F}$  is a classifier, for any input  $X$ , a reconstructed output  $\widehat{\mathcal{F}(X)}$  is considered accurate if the classes predicted using  $\widehat{\mathcal{F}(X)}$  and  $\mathcal{F}(X)$  are identical. More formally, let  $\mathcal{C}(\cdot)$  denote the argmax operator (which is typically used to predict the class label from the output layer of a neural network classifier). For an input  $X$ , a reconstructed output  $\widehat{\mathcal{F}(X)}$  is considered accurate if  $\mathcal{C}(\widehat{\mathcal{F}(X)}) = \mathcal{C}(\mathcal{F}(X))$ . This metric helps decouple the accuracy of the learned code in its ability to reconstruct unavailable base model outputs and the classification accuracy of the base model itself.
2. *Overall-accuracy*: This metric measures the accuracy of the reconstructed output based on the true label. For example, when  $\mathcal{F}$  is a classifier, for any input  $X$  with true label  $Y$ , a reconstructed output  $\widehat{\mathcal{F}(X)}$  is considered accurate if the class predicted using  $\widehat{\mathcal{F}(X)}$  and  $Y$  are identical. More formally, using the terminology defined above, a reconstructed output  $\widehat{\mathcal{F}(X)}$  is considered accurate if  $\mathcal{C}(\widehat{\mathcal{F}(X)}) = Y$ .

In the results presented, for both the metrics, we calculate the aggregate accuracy by averaging the accuracy over all unavailability scenarios. If unavailability statistics are known, one can instead weigh different unavailability scenarios based on the statistics.

## 4.2 Experimental results

As discussed above, we have performed experiments for a wide range of configuration settings. To avoid clutter, we focus our discussion below on a subset of the configuration settings. The remaining experiments also have similar results as the ones discussed below and hence are relegated to appendices.

### 4.2.1 Main results

We begin by discussing experiments in which training is performed using the loss with respect to the base model outputs; we focus on recovery-accuracy as the accuracy metric. Table 4 presents the results on test datasets for all combinations of datasets, base models, and parameter  $k$ . For clarity, we show the results only for the encoding function architecture and training loss function which achieved the highest recovery-accuracy on the training dataset. The results for all encoding function architectures and training loss functions are available in Table 5 in Appendix A.

The results in Table 4 show that our proposed approach can *accurately reconstruct a significant fraction of the unavailable predictions*. For example, for a ResNet-18 classifier on MNIST and Fashion-MNIST, the learned code can accurately reconstruct 95.71% and 82.77% of the unavailable outputs, respectively, with only 20% redundant base model computations (corresponding to  $k = 5$ ). Moreover, even on a more complex dataset such as CIFAR-10, our learned code can accurately reconstruct 80.74% of the unavailable outputs (corresponding to  $k = 2$ ).

We relegate the overall-accuracy attained in our experiments to the appendix. We find that the overall-accuracy attained by our learned codes differs only marginally from the recovery-accuracy. The overall-accuracy metrics for all the experiments are available in Table 5 in Appendix A. In the rest of the paper, we use the term “accuracy” to refer to changes in both recovery-accuracy and overall-accuracy.

As mentioned earlier, our focus is on designing codes that can impart resilience for non-linear computations. There are several existing approaches (e.g., [LLP<sup>+</sup>18, DCG16, LMAA16, YMAA17, WLS18, DCG17, RPPA17, MCJ18]) that address linear computations. For the sake of completeness, we include results for learning the encoding and decoding functions for a linear base model in Appendix B.

Dataset	Base Model	$k$	Recovery-accuracy	Encoding Function Architecture	Training Loss Func.
MNIST	Base-MLP	2	0.9885	MLPEncoder	MSE-Base
		5	0.9485	ConvEncoder	KL-Base
	ResNet-18	2	0.9904	ConvEncoder	XENT-Label
		5	0.9571	ConvEncoder	KL-Base
Fashion-MNIST	Base-MLP	2	0.9215	MLPEncoder	KL-Base
		5	0.8364	ConvEncoder	XENT-Label
	ResNet-18	2	0.9242	ConvEncoder	XENT-Label
		5	0.8277	MLPEncoder	XENT-Label
CIFAR-10	ResNet-18	2	0.8074	ConvEncoder	MSE-Base
		5	0.6431	ConvEncoder	MSE-Base

Table 4: Recovery-accuracy (for  $r = 1$ ) on test data with the training loss function and the encoding function architecture chosen based on highest training recovery-accuracy.

#### 4.2.2 Effect of configuration settings and parameters

We next discuss how the accuracy attained by the learned code differs under certain parameter settings and configurations.

**Value of parameter  $k$ .** Across all datasets, base models, and encoding function architectures, we find that test accuracy is significantly higher when  $k = 2$  than when  $k = 5$ . We believe this is because for  $k = 5$ , a single parity needs to pack information about 5 input images, whereas for  $k = 2$ , a single parity contains information about only 2 inputs images. Note that with  $r$  fixed, the value of  $k$  controls the amount of redundant base model computation. For  $r = 1$ , having  $k = 2$  corresponds to 50% redundant base model computation and having  $k = 5$  corresponds to 20% redundant base model computation. The above observation hints towards a fundamental tradeoff between recovery-accuracy and the amount of redundant computation. The difference between  $k = 2$  and  $k = 5$  is more pronounced for the Fashion-MNIST and CIFAR-10 datasets, which we attribute to the increased complexity of the dataset.

**Effect of base model complexity.** In our experiments, we find that the complexity of the base model does not have an adverse effect on the accuracy of the learned code. As discussed in Section 4.1.2, ResNet-18 is a significantly more complex model than Base-MLP, including many more layers of non-linearities. Despite this higher complexity, we see that the learned codes achieve similar accuracies for both Base-MLP and ResNet-18 (in Table 4 see accuracy achieved for the two base models for the MNIST and Fashion-MNIST datasets). This is very promising, since it suggests that the proposed approach is effective even for complex base models.

**Encoding function architectures.** For the MNIST and Fashion-MNIST datasets, there is little difference in the accuracies attained by the two proposed neural network encoding function architectures, MLPEncoder and ConvEncoder. The difference between the two architectures comes to fore in the more complex dataset CIFAR-10, where ConvEncoder greatly outperforms MLPEncoder. MLPEncoder’s high parameter count causes it to overfit and plateau at low accuracy on CIFAR-10, while ConvEncoder is able to reach significantly higher accuracy. Table 5 in Appendix A contains a direct comparison between the accuracies attained by both of the encoding function architectures.

**Loss functions used in training.** We found only marginal difference between using the different means of calculating loss discussed in Section 4.1.1. Table 5 in Appendix A contains results for all configurations with both loss functions.

#### 4.2.3 Detailed analysis of accuracy and quality of predictions

We next take a deeper look at the recovery-accuracy attained on the configurations discussed above and analyze cases where the predicted class from reconstructed outputs does not match that from the base model outputs.

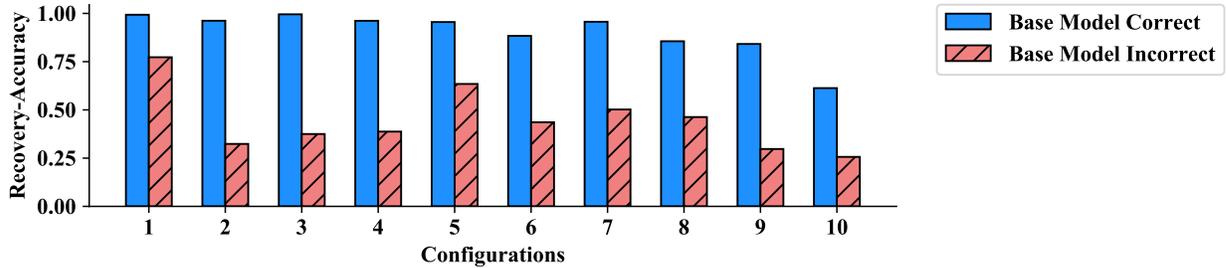


Figure 6: Recovery-accuracy attained on samples which the base model correctly classifies (“Base Model Correct”) and those which the base model incorrectly classifies (“Base Model Incorrect”). The configuration for each pair of bars is available in Table 4 – configuration number corresponds to the row number in the table.

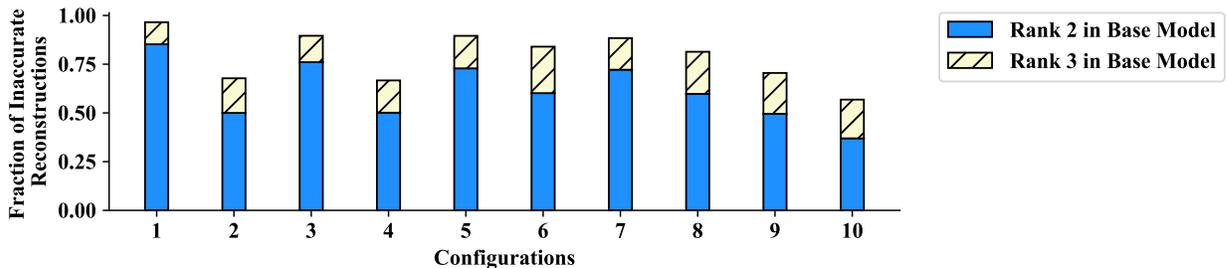


Figure 7: Fraction of inaccurate reconstructions that are rank 2 and 3 in the base model output. The configuration for each bar is available in Table 4 – configuration number corresponds to the row number in the table.

**Recovery-accuracy stratified based on accuracy of the base model.** In our experiments, interestingly, the learned codes achieve a significantly higher recovery-accuracy on the set of samples that the base model classifies correctly as compared to the set of samples that the base model classifies incorrectly. Figure 6 shows the recovery-accuracy on these two sets of samples in all the configurations of experiments listed in Table 4. We see that the learned codes achieve, on average, 2.19 times higher recovery-accuracy on the set of samples that the base model classifies correctly (“Base Model Correct” in Figure 6) as compared to the set of samples that the base model classifies incorrectly (“Base Model Incorrect” in Figure 6). Thus, the recovery-accuracy of the learned codes is higher on samples where it indeed matters more to reconstruct accurately.

**Analysis of errors in the learned code.** Here we analyze how poor is the class predicted from inaccurate reconstructions. Specifically, we look at the samples for which the reconstructed output is inaccurate with respect to the base model output (as defined under recovery-accuracy in Section 4.1.5), and analyze how far the resulting predicted class label is from the label predicted by the base model output. We quantify the quality of the label predicted from an inaccurate reconstruction by its rank in the base model output. A rank of 2 means that the class predicted using the reconstruction was ranked second in the base model output.<sup>3</sup> Figure 7 shows the fraction of inaccurate reconstructions which lead to predicted labels that have rank 2 and rank 3 in the base model output for the configurations considered in Table 4. We see that, on average, 61.29% of the inaccurate reconstructions result in a class prediction that is the second best in the base model output. Furthermore, on average, 79.12% of the inaccurate reconstructions result in a class prediction among the top 3 predictions of the base model output. Thus, even when the class prediction resulting from a reconstructed output does not match that of the base model output, the predicted class is not far off.

<sup>3</sup>Note that rank 1 is unattainable since we are analyzing only those instances for which the predicted class from the reconstruction does not match that of the base model output.

## 5 Conclusion

Coded computation is an emerging technique which makes use of coding-theoretic tools to impart resilience against failures and stragglers in distributed computation. However, the applicability of current techniques to general computation, including machine learning algorithms, is limited due to the lack of codes that can handle non-linear functions. In this paper, we propose a novel learning-based approach for designing erasure codes that approximate unavailable outputs for any differentiable non-linear function. We present carefully designed neural network architectures and a training methodology for learning the encoding and decoding functions. We show that our learned codes can accurately reconstruct up to 98.85%, 92.15%, and 80.74% of the unavailable class predictions from image classifiers for MNIST, Fashion-MNIST, and CIFAR-10 datasets, respectively. These results are highly promising as they show the potential of using learning-based approaches for designing erasure codes and herald a new direction for coded computation by handling general non-linear computations.

## References

- [AA16] Martín Abadi and David G. Andersen. Learning to Protect Communications with Adversarial Neural Cryptography. *ArXiv e-prints*, October 2016.
- [AGSS12] Ganesh. Ananthanarayanan, Ali. Ghodsi, Scott. Shenker, and Ion. Stoica. Why Let Resources Idle? Aggressive Cloning of Jobs with Dolly. In *USENIX HotCloud*, June 2012.
- [Ale] Alex Krizhevsky and Vinod Nair and Geoffrey Hinton. The CIFAR-10 and CIFAR-100 Datasets. <https://www.cs.toronto.edu/~kriz/cifar.html>.
- [AWS] AWS Machine Learning. <https://aws.amazon.com/machine-learning/>. Last accessed 24 May 2018.
- [Azu] Azure Machine Learning. <https://azure.microsoft.com/en-us/overview/machine-learning/>. Last accessed 24 May 2018.
- [Dan17] Daniel Crankshaw and Xin Wang and Guilio Zhou and Michael J. Franklin and Joseph E. Gonzalez and Ion Stoica. Clipper: A Low-Latency Online Prediction Serving System. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2017.
- [DB13] Jeffrey Dean and Luiz André Barroso. The Tail at Scale. *Communications of the ACM*, 56(2):74–80, 2013.
- [DCG16] Sanghamitra Dutta, Viveck Cadambe, and Pulkrit Grover. Short-dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products. In *Advances In Neural Information Processing Systems (NIPS)*, 2016.
- [DCG17] Sanghamitra Dutta, Viveck Cadambe, and Pulkrit Grover. Coded Convolution for Parallel and Distributed Computing Within a Deadline. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.
- [Dea] Jeff Dean. Software Engineering Advice from Building Large-Scale Distributed Systems. <https://static.googleusercontent.com/media/research.google.com/en/people/jeff/stanford-295-talk.pdf>. Last accessed 24 May 2018.
- [Die15] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [Fis16] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations (ICLR)*, 2016.
- [GB10] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Proceedings of Machine Learning Research. PMLR, 2010.

- [Goo] Google Cloud AI. <https://cloud.google.com/products/machine-learning/>. Last accessed 24 May 2018.
- [GZD<sup>+</sup>15] Kristen Gardner, Samuel Zbarsky, Sherwin Doroudi, Mor Harchol-Balter, and Esa Hyttia. Reducing Latency via Redundant Requests: Exact Analysis. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):347–360, 2015.
- [HDF] HDFS RAID. <http://www.slideshare.net/ydn/hdfs-raid-facebook>. Last accessed 24 May 2018.
- [Hor91] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, pages 251–257, 1991.
- [HSX<sup>+</sup>12] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin. Erasure Coding in Windows Azure Storage. In *Proc. USENIX Annual Technical Conference (ATC)*, 2012.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [JLS14] Gauri Joshi, Yanpei Liu, and Emina Soljanin. On the Delay-Storage Trade-Off in Content Download From Coded Distributed Storage Systems. *IEEE JSAC*, (5):989–997, 2014.
- [KJR<sup>+</sup>18] Hyeji Kim, Yihan Jiang, Ranvir B. Rana, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Communication Algorithms via Deep Learning. In *International Conference on Learning Representations (ICLR)*, 2018.
- [KSDY17] Can Karakus, Yifan Sun, Suhas Diggavi, and Wotao Yin. Straggler Mitigation in Distributed Optimization Through Data Encoding. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [KSDY18] Can Karakus, Yifan Sun, Suhas Diggavi, and Wotao Yin. Redundancy Techniques for Straggler Mitigation in Distributed Optimization and Learning. *arXiv preprint arXiv:1803.05397*, March 2018.
- [LeC] Yann LeCun. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.
- [LK13] Guanfeng Liang and Ulas C. Kozat. FAST CLOUD: Pushing the Envelope on Delay Performance of Cloud Storage with Coding. *arXiv:1301.1294*, January 2013.
- [LLP<sup>+</sup>18] Kangwook Lee, Maximilian Lam, Ramtin Pedarsani, Dimitris Papailiopoulos, and Kannan Ramchandran. Speeding Up Distributed Machine Learning Using Codes. *IEEE Transactions on Information Theory*, July 2018.
- [LMAA16] Songze Li, Mohammad Ali Maddah-Ali, and A Salman Avestimehr. A Unified Coding Framework for Distributed Computing With Straggling Servers. In *2016 IEEE Globecom Workshops (GC Wkshps)*, 2016.
- [LMK15] Zhouhan Lin, Roland Memisevic, and Kishore Reddy Konda. How Far Can We Go Without Convolution: Improving Fully-Connected Networks. November 2015.
- [MCJ18] Ankur Mallick, Malhar Chaudhari, and Gauri Joshi. Rateless Codes for Near-Perfect Load Balancing in Distributed Matrix-Vector Multiplication. *arXiv preprint arXiv:1804.10331*, 2018.
- [MSM18] Raj Kumar Maity, Ankit Singh Rawat, and Arya Mazumdar. Robust Gradient Descent via Moment Encoding with LDPC Codes. *ArXiv e-prints*, May 2018.
- [NML<sup>+</sup>18] Eliya Nachmani, Elad Marciano, Loren Lugosch, Warren J. Gross, David Burshtein, and Yair Be’ery. Deep Learning Methods for Improved Decoding of Linear Codes. *IEEE Journal of Selected Topics in Signal Processing*, pages 119–131, February 2018.
- [PGK88] David A. Patterson, Garth Gibson, and Randy H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *Proc. ACM SIGMOD International Conference on Management of Data*, June 1988.

- [Pyt] Pytorch. <https://pytorch.org/>. Last accessed 1 June 2018.
- [RCK<sup>+</sup>16] KV Rashmi, Mosharaf Chowdhury, Jack Kosaian, Ion Stoica, and Kannan Ramchandran. EC-Cache: Load-Balanced, Low-Latency Cluster Caching with Online Erasure Coding. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.
- [RPPA17] Amirhossein Reisizadeh, Saurav Prakash, Ramtin Pedarsani, and Salman Avestimehr. Coded Computation Over Heterogeneous Clusters. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.
- [RSG<sup>+</sup>13] K. V. Rashmi, Nihar B. Shah, Dikang Gu, Hairong Kuang, Dhruba Borthakur, and Kannan Ramchandran. A solution to the network challenges of data recovery in erasure-coded distributed storage systems: A study on the Facebook warehouse cluster. In *Proc. USENIX HotStorage*, June 2013.
- [RSG<sup>+</sup>14] KV Rashmi, Nihar B Shah, Dikang Gu, Hairong Kuang, Dhruba Borthakur, and Kannan Ramchandran. A Hitchhiker’s Guide to Fast and Efficient Data Reconstruction in Erasure-Coded Data Centers. In *ACM SIGCOMM*, 2014.
- [RU08] Tom Richardson and Ruediger Urbanke. *Modern coding theory*. Cambridge university press, 2008.
- [SAP<sup>+</sup>13] Mahesh Sathiamoorthy, Megasthenis Asteris, Dimitris Papailiopoulos, Alexandros G Dimakis, Ramkumar Vadali, Scott Chen, and Dhruba Borthakur. XORing Elephants: Novel Erasure Codes for Big Data. In *VLDB Endowment*, 2013.
- [SLR16] Nihar B Shah, Kangwook Lee, and Kannan Ramchandran. When do Redundant Requests Reduce Latency? *IEEE Transactions on Communications*, 64(2):715–722, 2016.
- [TJZ<sup>+</sup>17] Wen Tao, Feng Jiang, Shengping Zhang, Jie Ren, Wuzhen Shi, Wangmeng Zuo, Xun Guo, and Debin Zhao. An End-to-End Compression Framework Based on Convolutional Neural Networks. In *2017 Data Compression Conference (DCC)*, 2017.
- [TLDK17] Rashish Tandon, Qi Lei, Alexandros G Dimakis, and Nikos Karampatziakis. Gradient coding: Avoiding stragglers in distributed learning. In *International Conference on Machine Learning (ICML)*, 2017.
- [TOH<sup>+</sup>16] George Toderici, Sean M. O’Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable Rate Image Compression with Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2016.
- [VMGS12] Ashish Vulimiri, Oliver. Michel, P. Brighten. Godfrey, and Scott Shenker. More is Less: Reducing Latency via Redundancy. In *ACM HotNets*, 2012.
- [WJW14] Da Wang, Gauri Joshi, and Gregory Wornell. Efficient Task Replication for Fast Response Times in Parallel Computation. In *SIGMETRICS*, 2014.
- [WLS18] Sinog Wang, Jiashang Liu, and Ness Shroff. Coded Sparse Matrix Multiplication. In *International Conference on Machine Learning (ICML)*, 2018. To Appear.
- [XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-Mnist: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- [Yan98] Yann LeCun and Léon Bottou and Yoshua Bengio and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 1998.
- [YMAA17] Qian Yu, Mohammad Maddah-Ali, and Salman Avestimehr. Polynomial Codes: An Optimal Design for High-Dimensional Coded Matrix Multiplication. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

## Appendix A Full Experimental Results

Table 5 contains experimental results for all configurations considered in Section 4.

Dataset	Base Model	$k$	Training Loss Function	MLPEncoder		ConvEncoder	
				Recovery Accuracy	Overall Accuracy	Recovery Accuracy	Overall Accuracy
MNIST	Base-MLP	2	KL-Base	0.9769	0.9758	0.9831	0.9854
			MSE-Base	0.9885	0.9776	0.9767	0.9770
			XENT-Label	0.9737	0.9768	0.9769	0.9893
		5	KL-Base	0.9371	0.9340	0.9485	0.9518
			MSE-Base	0.9480	0.9424	0.9339	0.9357
			XENT-Label	0.9251	0.9232	0.9474	0.9533
	ResNet-18	2	KL-Base	0.9742	0.9760	0.9836	0.9854
			MSE-Base	0.9788	0.9806	0.9887	0.9888
			XENT-Label	0.9774	0.9796	0.9904	0.9925
		5	KL-Base	0.9460	0.9466	0.9571	0.9585
			MSE-Base	0.9349	0.9359	0.9415	0.9433
			XENT-Label	0.9401	0.9407	0.9171	0.9178
Fashion-MNIST	Base-MLP	2	KL-Base	0.9215	0.8800	0.9128	0.9080
			MSE-Base	0.8484	0.8196	0.8471	0.8253
			XENT-Label	0.9107	0.8808	0.9036	0.9185
		5	KL-Base	0.8275	0.7997	0.8300	0.8153
			MSE-Base	0.7133	0.6987	0.7302	0.7193
			XENT-Label	0.8259	0.8037	0.8364	0.8282
	ResNet-18	2	KL-Base	0.9002	0.8845	0.9206	0.9031
			MSE-Base	0.8960	0.8815	0.8982	0.8892
			XENT-Label	0.8947	0.8880	0.9242	0.9164
		5	KL-Base	0.8219	0.8133	0.8033	0.7960
			MSE-Base	0.7726	0.7672	0.7939	0.7885
			XENT-Label	0.8277	0.8203	0.8303	0.8248
CIFAR-10	ResNet-18	2	KL-Base	0.4293	0.4283	0.7889	0.8002
			MSE-Base	0.4107	0.4116	0.8074	0.8204
			XENT-Label	0.4284	0.4238	0.7980	0.8106
		5	KL-Base	0.1889	0.1895	0.5368	0.5382
			MSE-Base	0.1913	0.1936	0.6431	0.6466
			XENT-Label	0.1874	0.1890	0.5224	0.5287

Table 5: Recovery-accuracy and overall-accuracy for  $r = 1$  for all configuration settings.

Recall that results presented in Section 4.2.1 did not consider all parameter settings and configurations. We briefly highlight some relevant configuration comparisons made available through the full results presented in Table 5.

**Overall-accuracy metric:** Looking at the “Recovery Accuracy” and “Overall Accuracy” columns of Table 5, there is little difference between the two metrics, when holding architecture, parameters, and other configuration settings constant. We believe that the similarity of these two metrics can in part be explained by the observation in Section 4.2.3 that the recovery-accuracy attained on samples which are correctly classified by the base model is often significantly higher than that attained on samples which are incorrectly classified by the base model.

Encoding Function Architecture	k	Recovery-accuracy	Overall-accuracy
MLPEncoder	2	0.9831	0.9279
	5	0.9817	0.9270
ConvEncoder	2	0.9899	0.9295
	5	0.9869	0.9260

Table 6: Test recovery-accuracy and overall-accuracy for logistic regression base model on the MNIST dataset with  $r = 1$  and using KL-divergence as the loss function.

**Difference between training loss functions.** Results with “XENT-Label” as the training loss function in Table 5 correspond to those configurations for which training calculated loss via cross-entropy between the reconstructed output  $\widehat{\mathcal{F}(X)}$  and the true label of  $X$ . The recovery-accuracy and overall-accuracy for the XENT-Label configurations are very similar to those of the corresponding configurations with KL-Base and MSE-Base (which calculate the KL-divergence and MSE, respectively, between  $\widehat{\mathcal{F}(X)}$  and  $\mathcal{F}(X)$ ).

There are two configurations for which we observe significant difference in the accuracies attained using each loss function. First, for Fashion-MNIST with Base-MLP as the base model and MLPEncoder as the encoding function architecture, we find that using MSE-Base leads to a decrease in test recovery-accuracy and overall-accuracy compared to both KL-Base and XENT-Label. Second, when training ResNet-18 base models on CIFAR-10 with ConvEncoder for  $k = 5$ , we find that using MSE-Base leads to roughly 0.10 increase in test recovery-accuracy and overall-accuracy compared to using KL-Base and XENT-Label.

## Appendix B Multinomial Logistic Regression

In this section, we evaluate our learned codes on a multinomial logistic regression problem on the MNIST dataset. The overall calculation in multinomial logistic regression is of the form  $\sigma(AX + b)$  for parameters  $A$  and  $b$ , data  $X$ , and with  $\sigma$  being a softmax operator. Recall from Section 3.3.2 that the inputs to our neural network decoding function are the raw outputs of the base model (prior to any softmax operation), which are not converted to a probability distribution. As such, the available inputs to the decoding function are  $\mathcal{F}(X) = (AX + b)$ . The softmax operation is applied to reconstructed outputs of the decoder. For the MNIST dataset, the base model thus consists of parameter matrix  $A \in \mathbb{R}^{10 \times 784}$  and a vector  $b \in \mathbb{R}^{10}$ . The value 10 corresponds to the number of classes in the MNIST dataset. Each  $28 \times 28$  input image from the MNIST dataset is flattened to form a 784 length vector  $X$ .

We train the base model described above on the MNIST dataset. The trained base model achieves an accuracy of 0.9283 on the MNIST test set. Table 6 shows the recovery-accuracies achieved on the test set over this base model by each of the encoding function architectures described in Section 3.2 with  $r = 1$ ,  $k$  being 2 and 5, and using KL-divergence as the loss function. In all cases, the proposed codes are able to achieve a high recovery-accuracy. We note that  $\mathcal{F}(X) = (AX + b)$  can be (trivially) transformed into a linear function by juxtaposing the matrix  $A$  and the vector  $b$  and appending 1 to vector  $X$ . Hence even though our approach provides high recovery-accuracy, the existing approaches that address only linear functions [LLP<sup>+</sup>18, DCG16, LMAA16, YMAA17, WLS18, DCG17, RPPA17, MCJ18] are perhaps more suitable for this particular  $\mathcal{F}$  as these approaches guarantee exact reconstruction of unavailable outputs. However, note that these existing approaches are applicable only for linear functions while our goal is to handle non-linear functions.