





# Fast Server Learning Rate Tuning for Coded Federated Dropout

Giacomo Verardo , Daniel Barreira, Marco Chiesa , Dejan Kostic , and Gerald Q. Maguire Jr. 

KTH Royal Institute of Technology, Sweden  
{verardo, barreira, mchiesa, dmck, maguire}@kth.se

**Abstract.** In Federated Learning (FL), clients with low computational power train a common machine model by exchanging parameters via updates instead of transmitting potentially private data. Federated Dropout (FD) is a technique that improves the communication efficiency of a FL session by selecting a *subset* of model parameters to be updated in each training round. However, compared to standard FL, FD produces considerably lower accuracy and faces a longer convergence time. In this chapter, we leverage *coding theory* to enhance FD by allowing different sub-models to be used at each client. We also show that by carefully tuning the server learning rate hyper-parameter, we can achieve higher training speed while also reaching up to the same final accuracy as the no dropout case. Evaluations on the EMNIST dataset show that our mechanism achieves 99.6% of the final accuracy of the no dropout case while requiring  $2.43\times$  less bandwidth to achieve this level of accuracy.

**Keywords:** Federated Learning · Hyper-parameters tuning · Coding Theory

## 1 Introduction

In recent years, stricter regulations such as GDPR [20] have been introduced to preserve data privacy for end users; therefore, standard Machine Learning (ML) tasks, which require collecting information at centralized locations such as data-centers cannot be implemented without violating users' privacy. At the same time, the number of mobile phones has consistently grown and is estimated to reach 8.8 billion in 2026 [10]. For this reason, Federated Learning (FL) [18] has been proposed to train machine learning models *without* collecting private data from users' devices. In FL, a parameter server broadcasts a global ML model to low powered devices (clients), which in turn perform training over their own datasets. The model updates are sent from the clients to the server, which aggregates them and may start another FL round. Even in case of highly heterogeneous client datasets, it has been demonstrated that the model converges [17].

FL poses strict requirements both in terms of the amount of *bandwidth* required for exchanging models and the *computational* and *memory* resources required to perform training on large models on the clients' devices. The size of a model could be hundreds of MB [8]. Clients must download such large models using their available (often heterogeneous) bandwidth, use processing and memory resources to train on the model on their data, and re-transmit the updated models to the FL servers; therefore, optimizing

the bandwidth and computational overheads of running an FL session has become a topic of crucial importance.

Splitting a common, global model between clients and training it collaboratively has become imperative to reduce both memory and computational demands of an FL session. Federated Dropout (FD) is a technique that holds great promise to decrease the resource utilization of FL by pruning activations in the neural network, thus decreasing the number of variables to be exchanged and trained at the device side. Unlike standard centralized Dropout [24], where different dropped models are used at each training step, in FD a global model is divided into sub-models which are trained locally and then merged into an updated global model. Federated Dropout is orthogonal to message compression techniques, such as quantization [2] or sparsification [3], which also mitigate bandwidth overheads but do not reduce the computational power and memory needed at the client side.

Reaping the benefits of FD is not easy as it entails solving two main challenges:

- **Low accuracy.** FD learning may result in *lower accuracy* than traditional FL [7]. Intuitively, partially overlapping sub-models per client may improve performance. Recently it has been shown that selecting random sub-models for each client device may lead to better final accuracy [27]. However, it is not clear how sub-models should be selected, nor which is the orthogonality level which produce the optimal accuracy. Moreover, merging the sub-models into a new global model is a challenging task since it depends on their overlapping.
- **Slow convergence.** Although FD requires less bandwidth per round compared to traditional (no-dropout) FL, there is no guarantee that FD will converge rapidly. If FD requires significantly more rounds than FL to achieve high accuracy, then the promised bandwidth benefits would vanish.

We propose novel techniques to improve the accuracy of FD *without* losing the inherent bandwidth savings offered by FD (*i.e.*, by improving convergence speed). To tackle the above challenges, we explore the following two ideas:

- **Applying coding theory for sub-model selection.** Building upon the idea of sending different models to different clients during one FD round, we deterministically compute sub-models and then examine whether they perform better than random sub-models. We draw inspirations from *coding theory* (specifically, from the Code Division Multiple Access (CDMA) problem where orthogonal codes enable simultaneous communication channels *without* interference). We employ Gold codes [11] and Constant Weight Codes (CWC) [5] as masks to drop units and create different sub-models. The intuition is that selecting sub-models using these mechanisms will produce higher accuracy than random selection.
- **Adaptive server learning rate.** We experimentally observe that the convergence speed of an FD session depends on a critical parameter called *server learning rate*, which determines how the weight updates from the clients are incorporated into the trained model. Thanks to the inherent bandwidth savings of FD, we propose to search for the best server learning rate at the beginning of an FD session. A key challenge is to avoid consuming all of the bandwidth saved due to FD when performing this search.

Based on the above ideas, we design a mechanism called Coded Federated Dropout (CFD), which we incorporate alongside existing state-of-the-art FL systems, such as FedAdam [21] and FedAvg [18]. By evaluating on the EMNIST62 dataset, we show that CFD increases the final accuracy of the trained models while preserving the bandwidth savings of FD. In summary, our contributions are:

1. We are the first to leverage coding theory to carefully select the sub-models used in each FL round.
2. We show that the optimal server learning rate in a traditional FL session differs from that of an FD session.
3. We design a technique to *quickly* search for the best server learning rate. Our evaluation shows that we can identify good learning rates in just hundreds of rounds.
4. We show that CFD with Gold Codes achieves *comparable accuracy* to no-dropout FL with  $2.43\times$  *less bandwidth* on the EMNIST dataset.
5. We show that minimizing the ‘‘cross-correlation’’ metric in Gold Codes produces better final accuracy than maximizing the ‘‘minimal distance’’ metric of CWC codes.

The main system parameters used throughout the rest of the chapter and the related notation are summarised in Table 5.

## 2 Background

**Federated Learning** An FL session is composed of one parameter server and multiple clients. At the beginning of an FL round, the server broadcasts a common global model  $\mathbf{w}_k^{(t)}$  to a fraction of the clients. At round  $t$ , each client  $k$  trains for a customizable number of epochs  $E$  and returns the update  $\Delta\mathbf{w}_k^{(t)}$  from the previously received weights:

$$\hat{\mathbf{w}}_k^{(t)} = \mathbf{w}_k^{(t)} - \eta_l \nabla_{\mathbf{w}} L(\mathbf{w}, D_k) \quad (1)$$

$$\Delta\mathbf{w}_k^{(t)} = \hat{\mathbf{w}}_k^{(t)} - \mathbf{w}_k^{(t)} \quad (2)$$

where  $L(\mathbf{w}, D_k)$  is the employed loss function, which depends on the model weights and the client dataset  $D_k$ . When employing Federated Averaging (FAVG, [18]), the originally proposed aggregation method, the parameter server computes new weights by averaging the updates and adding them to the previous global model:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta \sum_{j \in S(t)} p_j \Delta\mathbf{w}_j^{(t)} \quad (3)$$

$$p_j = \frac{|D_j|}{\sum_{j \in S(t)} |D_j|} \quad (4)$$

For FAVG,  $\eta$  is set to 1.0, whereas it can be different for other aggregation mechanisms. **Federated Dropout** One of the major issues of FL is the communication overhead. FD improves bandwidth efficiency by randomly dropping connections between adjacent neural network layers. Unlike standard dropout [24], FD is not employed as a regularisation tool. Instead, FD keeps a fixed percentage  $\alpha$  of activations, thus producing a

sub-model with a  $(1 - \alpha)^2$  parameters fraction of fully connected networks. The sub-model is trained at the client-side, while the aggregation procedure only involves those nodes that have been kept. Moreover, the required computational power and memory at the client are reduced. However, although the benefits in bandwidth efficiency are remarkable, the same set of weights is trained at each client per round.

In [4], the authors suggest adapting the selection of dropped nodes based on the loss for each client; however, they state that this approach is unsuitable for FL since it wastes too much memory at the server. Therefore, they propose an alternative technique which employs a single sub-model for all clients, which degrades convergence rate and final accuracy. In contrast to random dropout [27], we propose to use coding theory for mask selection to enhance the orthogonality of sub-models by producing different dropping masks for each client. Hence, we allow partially disjoint sub-models per clients per round to improve both convergence time and final accuracy for the same  $\alpha$ .

**Code Division Multiple Access** Multiple access techniques address the problem of having multiple users communicate via a shared channel. Time and frequency division multiple access respectively splits the channel in time and frequency between users. In contrast, CDMA assigns a different code to each user and allows each of them to use the whole channel. If the codes are sufficiently orthogonal, the inter-user interference is low and transmission occurs with negligible error rate. CDMA has been extensively used in satellite [25] and mobile communication [15]. Gold [11] and Kasami [14] codes are examples of families of sequences designed for orthogonality.

**Adaptive Federated Optimization** [21] have proven that using a different learning rate for each parameter during aggregation can greatly improve FL model convergence. They propose 3 aggregation methods (FedAdam, FedAdagrad, and FedYogi), which replaces (3). Here we describe only the FedAdam algorithm, which performs well in all the datasets:

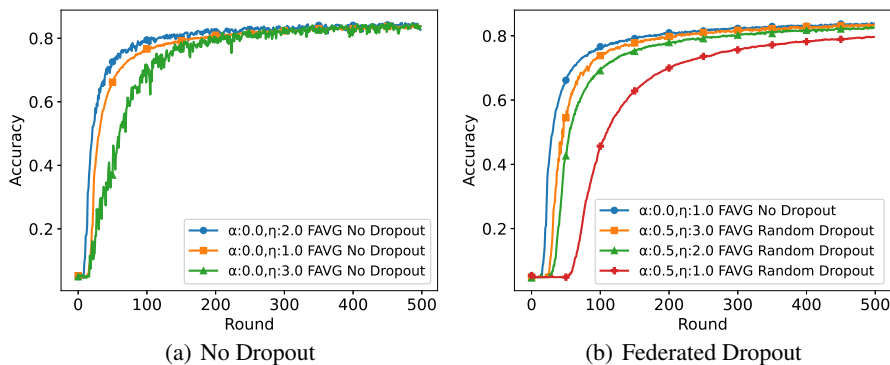
$$\Delta^t = \beta_1 \cdot \Delta^{t-1} + (1 - \beta_1) \cdot \sum_{j \in S(t)} p_j \Delta \mathbf{w}_j^{(t)} \quad (5)$$

$$\mathbf{v}^{(t)} = \beta_2 \cdot \mathbf{v}^{(t-1)} + (1 - \beta_2) \cdot \Delta^{(t)2} \quad (6)$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta \frac{\Delta^{(t)}}{\sqrt{\mathbf{v}^{(t)} + \tau}} \quad (7)$$

where  $\beta_1, \beta_2$  and  $\tau$  are hyper-parameters. The key insight is that training variables which have been trained less in the previous rounds will improve convergence. For this reason,  $v_t$  stores an indication of how much variables have been trained and is used to independently scale each component of the next update  $\Delta^{(t)}$ . However, the proposed optimization techniques require expensive hyper-parameters tuning and therefore a considerable amount of time, which might not be available in a FL session.

Other adaptive mechanisms have been proposed to correct the client drift due to the statistical heterogeneity in the clients datasets. [13] employ variance reduction, but this requires too much information to be stored server-side. [16] propose a trade-off between fairness and robustness of the global model.



**Fig. 1.** Increasing the server learning rate  $\eta$  to 3.0 is beneficial for random FD with different sub-models per client, while it is detrimental for the no dropout case.

### 3 Methodology

This section describes the proposed Coded Federated Dropout (CFD) method which performs both tuning of the server learning rate  $\eta$  (Sect. 3.1) and the selection of the sub-models sent to the clients (Sect. 3.2).

#### 3.1 Fast server learning rate adaptation

Similarly to centralized ML, increasing the server learning rate may lead to faster convergence, but further increasing the learning rate causes the objective function to diverge [28]. Fig. 1(a) empirically confirms that this is also the case for no-dropout FL where a high server learning rate of  $\eta = 3$  exhibits worse convergence than with  $\eta = 2$ . This result is based on the EMNIST62 dataset with more details in Sect. 4. Interestingly, Fig. 1(b) shows that in FD with random sub-models increasing the server learning rate to 3 leads to faster convergence. Moreover, a higher server learning rate in FD produces less oscillations in the accuracy across rounds compared to the no dropout case. This shows that the “best” server learning rate for FD may differ from the no-dropout case.

We propose a fast server learning rate adaptation method, which can also be extended to other parameters. At the beginning of training, we run Algorithm 1, which requires  $n_a$  adaptation steps. In each step (line 2), multiple FL sessions are launched in parallel from the same parameter server with different server learning rates  $\mathbb{H}$  and, in general, different clients subsets per round. We start our search using three  $\eta$  values during the first adaptation step (line 1) and reduce it to two server learning rates in the following adaptation steps (line 14). The goal of this search (lines 4 to 12) is to find the server learning rate that reaches a preconfigured accuracy target  $\gamma$  (lines 8-9) in the minimum number of rounds  $r^*$  (line 5). More specifically, in each round  $r$ , the server collects both the gradient update  $\Delta w_k^t$  and the accuracy of the model for each training

**Algorithm 1** Fast server learning rate adaptation

---

**Input:**  $w^0, \{D_k \forall k \in \{1, \dots, T\}\}$   
**Parameter:**  $\gamma^*, q, n_a, \eta_0, \Delta\eta$   
**Output:**  $\eta^*$

- 1:  $\mathbb{H} \leftarrow \{\eta_0, \eta_0 - \Delta\eta, \eta_0 + \Delta\eta\}$
- 2: **for**  $s = 0$  **to**  $n_a$  **do**
- 3:   **for**  $\eta \in \mathbb{H}$  **parallel do**
- 4:      $r \leftarrow 0$
- 5:     **while**  $r < r^*$  **do**
- 6:        $\bar{\gamma}^t, w^{(t+1)} \leftarrow \text{Round}(\{D_k\}, w^{(t)})$
- 7:        $\bar{\gamma} \leftarrow \frac{1}{q} \cdot \sum_{i=0}^{q-1} \bar{\gamma}^{t-i}$
- 8:       **if**  $\bar{\gamma} \geq \gamma^*$  **then**
- 9:          $r^* \leftarrow r, \eta^* \leftarrow \eta$
- 10:       wait for parallel search to end; go to line 15
- 11:       **end if**
- 12:        $r \leftarrow r + 1$
- 13:     **end while**
- 14:   **end for**
- 15:    $\Delta\eta \leftarrow \frac{\Delta\eta}{2}$
- 16:    $\mathbb{H} \leftarrow \{\eta^* - \Delta\eta, \eta^* + \Delta\eta\}$
- 17: **end for**

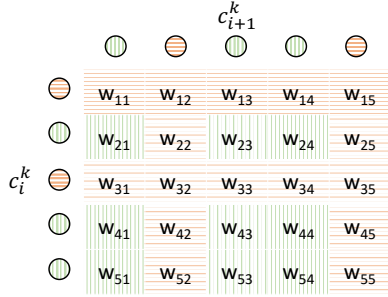
---

client in the round (line 6). Then, it computes the average of the median training accuracy  $\bar{\gamma}$  in the last  $q$  FL rounds (line 7). The median operation is performed in order to avoid the impact of outliers (*i.e.*, clients with too high or low training accuracies), while the average operation over the last rounds avoids sudden spikes. If one server learning rate  $\bar{\gamma}$  is higher than a predefined threshold  $\gamma^*$ , then we have found a new optimal server learning rate  $\eta^* = \eta$  that requires the new minimum number of rounds  $r^* = r$  to achieve the target accuracy (lines 8-9). For the next adaptation step, the new tentative server learning rates  $\mathbb{H}$  are chosen near  $\eta^*$  (lines 15-16) and the next adaptation step is performed. An adaptation step may also end when all the FL sessions produce  $r \geq r^*$  (line 5). Worth noting is that the search at lines 3 to 13 can be done in parallel to improve convergence speed.

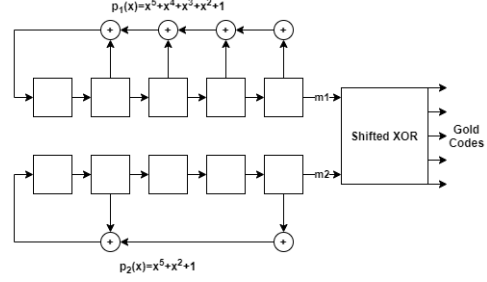
This algorithm reduces the number of rounds compared to testing all possible server learning rates using full FL sessions. In particular, since sessions are aborted when  $r \geq r^*$ , the overhead introduced by each adaptation step is limited. Assuming the parallel search is synchronized round-by-round, the additional overhead in number of rounds of our algorithm is:

$$3 \cdot r_0^* + 2 \sum_{i=1}^{n_a} r_i^* - r^* \quad (8)$$

where  $r_i^*$  is the minimum number of rounds at the end of the adaptive step  $i$ . The first term accounts for the first adaptive step, the second terms for the following adaptive steps, and the third term for the spared training rounds when running the full simulation with  $\eta = \eta^*$ .



**Fig. 2.** Federated Dropout for masks  $c_i^k=[0,1,0,1,1]$  and  $c_{i+1}^k=[1,0,1,1,0]$  reduces the weights to be trained from 25 to 9.



**Fig. 3.** Gold code generation for two preferred feedback polynomials. The output is a set of  $2^5 + 1 = 32$  sequences of length  $2^5 - 1 = 31$ .

Algorithm 1 selects the optimal  $\eta^*$  in terms of number of rounds to reach the target accuracy. The Round function at line 6 represents the underlying FL mechanism being used to compute the trained model, for instance, FAVG or FedAdam.

We argue that running full simulations to achieve the same level of granularity takes  $T \times$  the number of rounds per simulation, where  $T$  is the number of tried server learning rates. This value is strictly greater than the bound provided by equation (8).

### 3.2 Coded Federated Dropout

We reduce the size of the model by dropping weights from each layer by associating to each client  $k$  and model layer  $i$  in the FL round a binary mask vector  $c_i^k \in \mathbb{R}^{N_i}$ . A unit is dropped or kept when the component  $c_{ij}^k$  is equal to 0 or 1 respectively. For adjacent fully connected layers (Fig. 2) the dropped weights can be straightforwardly obtained by eliminating rows and columns corresponding to the dropped units from the previous and following layer respectively. As in standard FD, we only drop a fraction  $\alpha$  of nodes per layer  $i$ , which produces the same model size for all clients. For instance, in Fig. 2 we have  $N_i = N_{i+1} = 5$  and  $\alpha = 2/5$  and therefore only  $N_i \cdot N_{i+1} \cdot (1 - \alpha)^2 = 25 \cdot \frac{9}{25} = 9$  weights should be transmitted instead of  $N_i \cdot N_{i+1} = 25$ .

The problem to be solved is to obtain a matrix  $C_i$  per layer  $i$  with the following properties:

- each row of  $C_i$  is a codeword  $c_i^k \in \mathbb{R}^{N_i}$ ; and
- the Hamming weight (*i.e.*, the number of ones in the codeword) of each row in  $C_i$  is equal to  $N_i \cdot (1 - \alpha)$ ; and
- the number of rows in  $C_i$  is greater or equal than the number of clients per round  $M$ .

We consider 4 methods to compute  $C_i$ : (i) same random codeword for each client (baseline Federated Dropout), (ii) different random codeword for each client (proposed contemporaneously by [27]), (iii) Gold sequences, and (iv) CWCs. While the first two are straightforward, the other two provide different levels of orthogonality between the

**Table 1.** Preferred polynomial pairs for Gold codes generation.

Degree	Sequence length	Polynomial 1	Polynomial 2
5	31	$1 + x^2 + x^5$	$1 + x^2 + x^3 + x^4 + x^5$
6	63	$1 + x^6$	$1 + x^1 + x^2 + x^5 + x^6$
7	127	$1 + x^3 + x^7$	$1 + x^1 + x^2 + x^3 + x^7$
9	511	$1 + x^4 + x^9$	$1 + x^3 + x^4 + x^6 + x^9$
10	1023	$1 + x^3 + x^{10}$	$1 + x^2 + x^3 + x^8 + x^{10}$
11	2047	$1 + x^2 + x^5 + x^8 + x^{11}$	$1 + x^2 + x^{11}$

dropped models. Since sub-models are trained independently and then aggregated, having partially non overlapping sub-models reduces the impact of heterogeneous updates.

In CFD we exploit the different masks per client. At the beginning of each training round, we compute one matrix  $C_i$  for each layer  $i$ . Client  $k$  is assigned the  $k$ -th row from each  $C_i$  and the correspondent weights are extracted from the global model. If such a matrix is burdensome to be computed or if it would be the same after the generation process, then the rows and columns of the matrix are shuffled instead. In that way, the excess codewords can be employed when the number of rows of  $C_i$  is greater than  $M$ .

Applying coded masks to neural networks is not straightforward, since it requires different approaches according to the kind of employed layers. For *fully connected* layers, it is sufficient to apply dropout as in Fig. 2. For *convolutional* layers, we experimentally observed that dropping entire filters rather than individual weights achieves better performance. *Long Short-Term Memory* layers [12], which are composed of multiple gates, require that the same mask is employed for each set of weights.

After each sub-model has been trained, each weight in the global model is updated by averaging the contribution of each sub-model which contained that weight.

**Gold codes** Codewords orthogonality may be defined by means of cross-correlation. The correlation between two real binary sequences  $u^1$  and  $u^2$  of length  $L_u$  is a function of the shift  $l$ :

$$R(u_1, u_2, l) = \frac{1}{L_u} \sum_{j=1}^{L_u} u_j^1 \cdot u_{(j+l) \bmod L_u}^2 \quad (9)$$

$$u_j^1, u_j^2 \in \{-1, 1\} \quad \forall j = 1, \dots, L_u$$

Gold codes are generated from two Linear Feedback Shift Registers (LFSR) with suitable feedback polynomials and initial conditions. The LFSRs produce two  $m$ -sequences, which are then circularly shifted and element-wise xored to produce all the sequences in the family. The size  $n_{\text{LFSR}}$  of the LFSR determines the code length  $2^{n_{\text{LFSR}}} - 1$ , the number of codewords in the set  $2^{n_{\text{LFSR}}} + 1$ , and the upper bounds for the maximum cross-correlation in the set:

$$\max_l \left| R(c_i^{k_1}, c_i^{k_2}, l) \right| = 2^{\lfloor (n_{\text{LFSR}} + 2)/2 \rfloor} + 1 \quad (10)$$

$$\forall k_1, k_2 \in \{1, \dots, (2^{n_{\text{LFSR}}} + 1)\}, \quad k_1 \neq k_2$$



**Algorithm 2** Constant Weight Code Generation**Input:**  $N_i, M, \alpha, t_{max}$ **Output:**  $C_i$ 


---

```

1: Let  $t = 0$ .
2:  $d_{min} = N_i, C_i = \{\}$ 
3:  $F_i = \{f \in \{0, 1\}^{N_i} : w_h(f) = (1 - \alpha) \cdot N_i\}$ , f added in lexicographic order
4: Add random codeword from  $F_i$  to  $C_i$ 
5: while  $t < t_{max}$  do
6:    $F_i = \{f \in F_i : d_h(f, c) \geq d_{min} \forall c \in C_i\}$ 
7:   if  $|F_i| \neq 0$  then
8:     Add first sequence from  $F_i$  to  $C_i$ 
9:   else
10:     $d_{min} = d_{min} - 2, C_i = \{\}$ 
11:     $F_i = \{f \in \{0, 1\}^{N_i} : w_h(f) = (1 - \alpha) \cdot N_i\}$ , f added in lexicographic order
12:    Add random codeword from  $F_i$  to  $C_i$ 
13:   end if
14:   if  $|C_i| = M$  then
15:     return  $C_i$ 
16:   end if
17: end while
18: return  $C_i$ 

```

---

After computing the sequences, we concatenate them as row vectors in matrix  $C_i$ , which will then have  $2^{n_{LFSR}} + 1$  rows and  $2^{n_{LFSR}}$  columns. For our experiments, we employ the preferred polynomials pairs listed in Table 1. The table also contains the correspondent length of the generated gold sequences. In order to have a suitable length for the model layers, we usually pad each resulting sequence with a 0 value after the longest run of zeros. Also note that Gold sequences with multiple of 4 degrees are not supported. Fig. 3 provides an example of the LFSR used to generate Gold codes of length 31.

Although Gold codes provide orthogonality, they have constraints on the size of  $C_i$  and  $\alpha$  value, which can only be 50% since most Gold sequences are balanced (i.e., Hamming weight  $2^{n_{LFSR}-1}$ ). Please refer to [22] for details on constructions and properties of Gold codes and LFSR-generated sequences.

**Constant weight codes** Another metric is the Hamming distance between two codewords  $u^1$  and  $u^2$ , which is the number of ones in  $u^1 \oplus u^2$ :

$$d_h(u^1, u^2) = \sum_{j=1}^{L_u} u_j^1 \oplus u_j^2 \quad (11)$$

We provide a method to create a matrix  $C_i$  with size  $M \times N_i$  where, in order to improve orthogonality, the minimum Hamming distance between rows is maximized. CWC are a family of non-linear codes where each sequence has a fixed Hamming weight (i.e., number of ones). CWC are flexible: they can provide sets of codewords with any cardinality, sequence length, and Hamming weight. We devise a variant of the

**Table 2.** Model summary. Total number of parameters is 6,603,710).

Layer Type	Output Shape	Param #	Activation	Hyper-parameters
Convolutional 2D	(-, 28, 28, 32)	832	ReLU	Num filters: 32 Kernel size: (5,5) Padding: Same
MaxPooling 2D	(-, 14, 14, 32)	0	-	Pool size: (2,2) Padding: Valid
Convolutional 2D	(-, 14, 14, 64)	51264	ReLU	Num filters: 64 Kernel size: (5,5) Padding: Same
MaxPooling 2D	(-, 7, 7, 64)	0	-	Pool size: (2,2) Padding: Valid
Flatten	(-, 3136)	0	-	
Dense	(-, 2048)	6424576	ReLU	Num units: 2048
Dense	(-, 62)	127038	-	Num units: 62

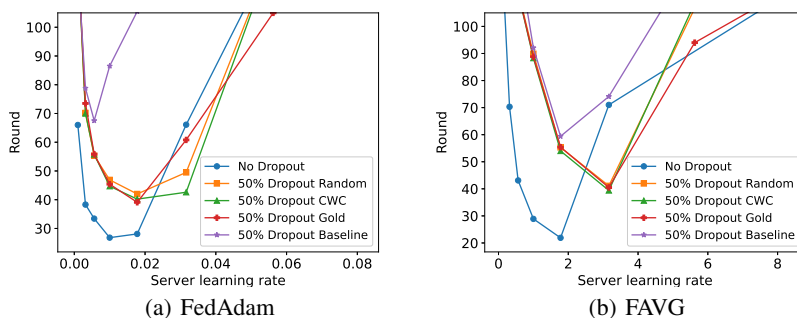
algorithm from [19] to generate  $M$  codewords with length  $N_i$  and weight  $(1 - \alpha) \cdot N_i$  by maximizing the minimum distance. Starting from a random codeword, we iteratively select new sequences with a fixed distance from the current set. The sequences are added in lexicographic order. If the final set is not complete (i.e., a set of  $M$  sequences with weight  $(1 - \alpha) \cdot N_i$  and length  $N_i$  does not exist for the given minimum distance), the required minimum distance is decremented and the selection procedure is performed again. The algorithm is described in 2, where we identify  $C_i$  as a set of codewords instead of a matrix.

## 4 Evaluation

We run our code in vanilla TensorFlow [1] and Python 3, since the major frameworks for FL do not support FD. In particular, although TensorFlow Federated [26] allows FD with the same dropping masks, it does not allow broadcast of different models to the clients (which we require for CFD). The generation algorithm for CWC was implemented in MATLAB. Training was performed on the EMNIST dataset [9,6] for character and digits recognition. We normalize the pixels in each image in the  $[0, 1]$  interval and use a batch size of 10. We provide a description of the employed model in Table 2, which is *model C* from [23]. Each client optimizes the sparse categorical crossentropy loss with the SGD optimizer. We train the model for 500 rounds. We use four kinds of codes for CFD with dropout fraction  $\alpha = 0.5$ : random with same sub-model for each client (baseline FD), random with different sub-models, Gold and CWC. We experiment with two different server optimization methods (FAVG and FedAdam) and perform the fast server learning rate tuning for both of them. The FedAdam hyper-parameters  $\beta_1, \beta_2$ , and  $\tau$  do not need to be tuned since it has been demonstrated that default values are usually enough to achieve good convergence [21].

**Table 3.** Selected  $\eta$  values for 10 simulations and different codes and aggregation algorithms.

	FedAdam				FAVG	
Server Learning Rate (Log10)→	-2.25	-2	-1.75	-1.5	0,25	0,5
$\alpha$ :0.0 Fedadam No Dropout	0%	30%	60%	10%	100%	0%
$\alpha$ :0.5 Random Fedadam	0%	20%	30%	<b>50%</b>	0%	<b>100%</b>
$\alpha$ :0.5 CWC FedAdam	0%	10%	40%	<b>50%</b>	0%	<b>100%</b>
$\alpha$ :0.5 Gold FedAdam	0%	0%	<b>80%</b>	20%	0%	<b>100%</b>
$\alpha$ :0.5 Fedadam + Baseline FD	<b>50%</b>	50%	0%	0%	<b>70%</b>	30%

**Fig. 4.** The dropout approaches require a higher  $\eta$  in order to reach the minimal number of rounds for the target accuracy.

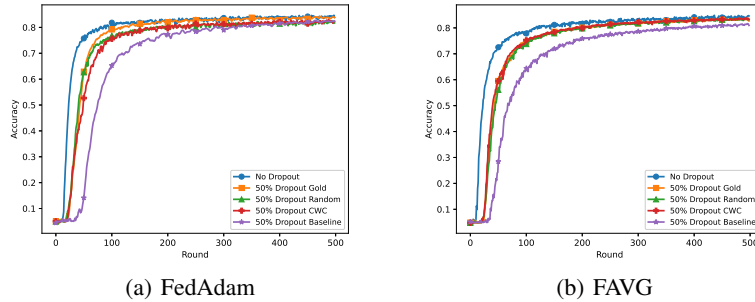
**The fast server learning rate tuning algorithm achieves consistent optimal  $\eta$  values across many simulations.** We run 10 training sessions for each coded approach with target accuracy  $\gamma^* = 20/62$  (which is 20 times the random accuracy) and showcase the results in Table 3 (highest probability is in bold). Whereas for FAVG the selected  $\eta$  is the same for all simulations (except for the baseline FD), FedAdam produces higher variability. However, the differences between the different selected  $\eta^*$  amounts to a maximum of 0.5 in log scale for each code.

**The optimal server learning rate in a traditional FL session is greater than that of an FD session, especially when a different sub-model per client is employed.** We experiment with 10 simulations with our tuning algorithm and keep track of the number of rounds to reach  $\gamma^*$  for each experimented  $\eta$ . Fig 4 shows the average number of rounds for each  $\eta$  for both FedAdam and FAVG and makes evident that the  $\eta$  producing the minimum number of rounds is greater for the coded approaches compared to the no dropout. Moreover, although the no dropout case is still the fastest one, the coded approaches achieve up to  $1.5\times$  speedup to reach  $\gamma^*$  compared to the dropout baseline, thus improving convergence time.

**Our tuning mechanism saves communication resources compared to running full FL sessions with different values of  $\eta$ .** We compute the number of additional rounds as in Eq. 8 and report the results in Table 4. In FedAdam<sub>10</sub>,  $\gamma^*$  is 10 times the random accuracy (10/62) instead of 20. The overhead is directly dependent on the convergence

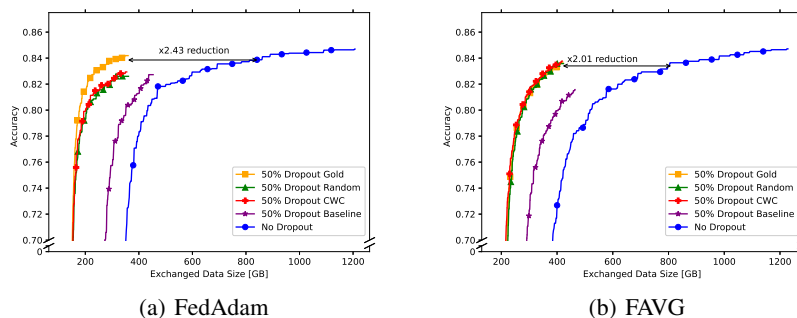
**Table 4.** Average number of additional rounds for different codes.

	No Drop	Rand	CWC	Gold	FD
<b>FedAdam<sub>20</sub></b>	154.9	262.5	248.6	259.3	456.3
<b>FAVG<sub>20</sub></b>	166.4	393.0	383.1	387.9	479.3
<b>FedAdam<sub>10</sub></b>	140.7	219.2	225.6	211.9	351.2

**Fig. 5.** Average test accuracy of FedAdam and FAVG for 5 simulations with the selected  $\eta^*$  from Table 3. Coded Federated Dropout with Gold codes and FedAdam improves convergence rate and final accuracy compared to random, CWC and baseline FD approaches.

speed of the model. Consequently, the no dropout case requires the least overhead and the baseline FD the most. Still, the additional number of rounds is much lower than running multiple full FL sessions. For both FedAdam and FAVG, our tuning algorithm tests 10  $\eta$  values. Therefore, running full sessions would require  $500 \cdot (10 - 1)$  additional rounds. We point out that decreasing the accuracy threshold  $\gamma^*$  to  $10/62$  notably reduces the additional number of rounds (FedAdam<sub>10</sub>). The best selected  $\eta$  is only slightly influenced by changing  $\gamma^*$ , thus demonstrating that tuning the threshold value is much easier than tuning  $\eta$  directly.

**Gold codes outperform other FD approaches for FedAdam and achieve 99.6% of the final accuracy of the no dropout case while saving  $> 2\times$  bandwidth.** Fig. 5 shows the average test accuracy of simulations run with the previously selected  $\eta^*$  values. While the benefits of using Gold codes or CWC is negligible in terms of final accuracy compared to random for FAVG, FedAdam plus Gold codes produces higher convergence speed. Moreover, Gold FedAdam reaches 99.6 % of the final accuracy of the no dropout case while saving almost  $1 - (1 - \alpha)^2 = 75\%$  of the bandwidth per round. The best result is achieved by FAVG without dropout (84.1% accuracy) when averaging over the last 100 rounds, while Gold codes achieve 83.8% for FedAdam, which is the 99.6%. Conversely, CWC does not perform well for FedAdam, achieving even worse performance than random codes. Regarding the overall bandwidth, we measure the size of the exchanged sub-models and compute the amount of gigabytes needed to reach a certain test accuracy. Fig.6 shows the reduction in overall bandwidth when CFD is employed with the selected  $\eta^*$  values. Gold codes plus FedAdam reduces the bandwidth



**Fig. 6.** Reachable test accuracy for FedAdam and FAVG given the amount of exchanged bytes in a FL session. Coded Federated Dropout with Gold codes and FedAdam reaches the same level of final accuracy as no dropout while reducing bandwidth usage by  $2.43\times$ .

needed to reach the maximum test accuracy by  $2.43\times$  compared to no dropout, while CFD plus FAVG by  $2.01\times$ .

**Minimizing cross-correlation instead of maximizing minimal distance provides greater sub-model orthogonality.** Fig. 5 and 6 show that Gold codes always outperform CWC. Therefore, codes built by minimizing cross-correlation produce higher final accuracy and convergence rate than the ones obtained by maximizing the minimum distance. Nevertheless, optimizing any of the two metrics outperforms the baseline for federated dropout for both FedAdam and FAVG.

## 5 Conclusion and future works

We have presented a fast server learning rate tuning algorithm for Federated Dropout and shown considerable reduction on the number of rounds to assess the optimal  $\eta^*$ . Moreover, we have shown that convergence rate and final accuracy of models trained in a FL session are improved when using coding theory to carefully perform Federated Dropout. Specifically, CFD with Gold sequences paired with an optimization mechanism such as FedAdam can achieve up to the same accuracy of the no dropout case, with  $2.43\times$  bandwidth savings. However, Gold codes have specific lengths and Hamming weights, so they are not flexible enough, while CWC does not improve performance compared to random dropout. Hence, future work will investigate further sequences from coding theory for FD.

**Acknowledgements** This research work was conducted with funding awarded by the Swedish Research Council for the project "Scalable Federated Learning" with registration number 2021-04610. This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No. ORA-CRG2021-4699.

**Table 5.** Adopted notations and principal symbols.

Symbol	Description	Value
$N_i$	Number of units in model layer $i$	See Tab. 2
$T$	Number of total clients	3400
$M$	Number of clients per round	35
$\alpha$	Dropout fraction for FD	0.5
$\eta_l$	Client learning rate	0.035
$E$	Training epochs per client	1
$\eta$	Server learning rate	See Sec. 3.1
$\beta_1$	Momentum parameter for FedAdam	0.90
$\beta_2$	Momentum parameter for FedAdam	0.99
$\tau$	Adaptivity degree for FedAdam	0.001
$\mathbf{w}^{(t)}$	Server weights at round $t$	-
$\mathbf{w}_k^{(t)}$	Initial client $k$ weights at round $t$	-
$\hat{\mathbf{w}}_k^{(t)}$	Final client $k$ weights at round $t$	-
$D_k$	Dataset for client $k$	-
$\{\xi\}_k^{(t)}$	Set of batches for client $k$ at round $t$	-
$L(\cdot)$	Client loss function	-
$S(t)$	Set of clients selected at round $t$	-
$c_i^k$	Binary mask for client $k$ and layer $i$	-
$c_{ij}^k$	Component $j$ of $c_i^k$	-
$C_i$	Codes Matrix per layer $i$	-
$n_{\text{LFSR}}$	Size of the LFSR	See Tab.1
$R(u_1, u_2, l)$	Correlation between $u_1$ and $u_2$ for shift $l$	-
$d_h(u^1, u^2)$	Hamming distance between $u_1$ and $u_2$	-
$w_h(u)$	Hamming weight of $u$	-
$\gamma^*$	Target accuracy	$\frac{20}{62}$ or $\frac{10}{62}$
$\gamma_k^t$	Training accuracy for client $k$ at round $t$	-
$\bar{\gamma}^t$	Median training accuracy at round $t$	-
$\bar{\gamma}$	Average of $\bar{\gamma}^t$ at round $t$	-
$q$	Rounds number to compute $\bar{\gamma}$	-
$n_a$	Number of adaptation steps	3
$\eta^*$	Best server learning rate	See Tab.3
$\Delta\eta$	Log distance between tentative $\eta$ values	-
$\Delta\eta_0$	Initial $\Delta\eta$	1
$\mathbb{H}$	Set of tentative $\eta$ values	-
$r$	Current round number	-
$r^*$	Best round number to reach $\gamma^*$	-
$r_i^*$	Best round number to reach $\gamma^*$ at step $i$	-

## References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). pp. 265–283 (2016), <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>
2. Alistarh, D., Grubic, D., Li, J., Tomioka, R., Vojnovic, M.: Qsgd: Communication-efficient sgd via gradient quantization and encoding (2017)
3. Alistarh, D., Hoefler, T., Johansson, M., Khirirat, S., Konstantinov, N., Renggli, C.: The convergence of sparsified gradient methods (2018)
4. Bouacida, N., Hou, J., Zang, H., Liu, X.: Adaptive federated dropout: Improving communication efficiency and generalization for federated learning (2020)
5. Brouwer, A., Shearer, J., Sloane, N., Smith, W.: A new table of constant weight codes. *IEEE Transactions on Information Theory* **36**(6), 1334–1380 (1990). <https://doi.org/10.1109/18.59932>
6. Caldas, S., Duddu, S.M.K., Wu, P., Li, T., Konecny, J., McMahan, H.B., Smith, V., Talwalkar, A.: Leaf: A benchmark for federated settings (2019)
7. Caldas, S., Konecny, J., McMahan, H.B., Talwalkar, A.: Expanding the reach of federated learning by reducing client resource requirements. *CoRR* **abs/1812.07210** (2018), <http://arxiv.org/abs/1812.07210>
8. Chollet, F.: Keras. <https://github.com/fchollet/keras> (2015)
9. Cohen, G., Afshar, S., Tapson, J., van Schaik, A.: Emnist: an extension of mnist to handwritten letters (2017)
10. Ericsson: Mobile subscriptions shifting towards 5g. <https://www.ericsson.com/en/mobility-report/reports/june-2021> (2021)
11. Gold, R.: Optimal binary sequences for spread spectrum multiplexing (corresp.). *IEEE Transactions on Information Theory* **13**(4), 619–621 (1967). <https://doi.org/10.1109/TIT.1967.1054048>
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
13. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning (2021)
14. Kasami, T.: Weight distribution formula for some class of cyclic codes. *Coordinated Science Laboratory Report no. R-285* (1966)
15. Lee, W.: Overview of cellular cdma. *IEEE Transactions on Vehicular Technology* **40**(2), 291–302 (1991). <https://doi.org/10.1109/25.289410>
16. Li, T., Hu, S., Beirami, A., Smith, V.: Ditto: Fair and robust federated learning through personalization. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 139, pp. 6357–6368. PMLR (18–24 Jul 2021), <https://proceedings.mlr.press/v139/li21h.html>
17. Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of fedavg on non-iid data (2020)
18. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data (2017)
19. Montemanni, R., Smith, D.H.: Heuristic algorithms for constructing binary constant weight codes. *IEEE Transactions on Information Theory* **55**(10), 4651–4656 (2009). <https://doi.org/10.1109/TIT.2009.2027491>

20. Parliament, T.E., the Council of the European Union: Reform of eu data protection rules. <https://gdpr-info.eu/> (2018)
21. Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B.: Adaptive federated optimization (2020)
22. Sarwate, D., Pursley, M.: Crosscorrelation properties of pseudorandom and related sequences. *Proceedings of the IEEE* **68**(5), 593–619 (1980). <https://doi.org/10.1109/PROC.1980.11697>
23. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net (2015)
24. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (Jan 2014)
25. Taaghool, P., Evans, B., Buracchini, E., De Gaudinaro, G., Lee, J.H., Kang, C.G.: Satellite umts/imt2000 w-cdma air interfaces. *IEEE Communications Magazine* **37**(9), 116–126 (1999). <https://doi.org/10.1109/35.790970>
26. tensorflow.org: Tensorflow federated. <https://www.tensorflow.org/federated> (2017)
27. Wen, D., Jeon, K.J., Huang, K.: Federated dropout – a simple approach for enabling federated learning on resource constrained devices (2021)
28. Zeiler, M.D.: Adadelta: An adaptive learning rate method (2012)