A Finite Blocklength Approach for Wireless Hierarchical Federated Learning in the Presence of Physical Layer Security

Haonan Zhang*§, Chuanchuan Yang‡§, Bin Dai*§

* School of Information Science and Technology, Southwest Jiaotong University, Chengdu, 610031, China.

‡ Department of Electronics, Peking University, Beijing, 100871, China.

§ Peng Cheng Laboratory, Shenzhen, 518055, China.

zhanghaonan@my.swjtu.edu.cn, yangchuanchuan@pku.edu.cn, daibin@home.swjtu.edu.cn.

Abstract—The wireless hierarchical federated learning (HFL) in the presence of physical layer security (PLS) issue is revisited. Though a framework of this problem has been established in the previous work, practical secure finite blocklength (FBL) coding scheme remains unknown. In this paper, we extend the already existing FBL coding scheme for the white Gaussian channel with noisy feedback to the wireless HFL with quasi-static fading duplex channel, and derive achievable rate and upper bound on the eavesdropper's uncertainty of the extended scheme. The results of this paper are further explained via simulation results.

Index Terms—Finite blocklength coding, physical layer security, privacy-utility trade-off, wireless federated learning

I. INTRODUCTION

The wireless federated learning has been extensively studied in the literature [1]-[4]. Recently, with the development of edge computing, the client-edge-cloud hierarchical federated learning (HFL) systems receive much attention [5]-[6]. However, due to the broadcast nature of wireless communications, the wireless FL is susceptible to eavesdropping. In this paper, we study the wireless HFL in the presence of eavesdropping, see Figure 1. In Figure 1, users, edge servers and the cloud server cooperate with each other to jointly train a learning model, and in the meanwhile, the malicious cloud server may infer the presence of an individual data sample from a learnt model by various attacks. Differential privacy (DP) has been proved to be an effective way to protect the individual data against such attacks, and hence before aggregation of all users' gradients to the edge servers, the Gaussian noise which is used as local differential privacy (LDP) mechanism [7] is added to the gradient of each user. Moreover, each edge server communicates with the cloud server via a duplex fading channel, and due to the broadcast nature of wireless communication, this channel is eavesdropped by an external eavesdropper. The main object of Figure 1 is to *minimize the information leakage to the malicious cloud server subject to a certain mount of utility of the polluted data gradients (added by the Gaussian noises), and protect the transmitted data in wireless channels from eavesdropping*. In [8], the fundamental limit in the utility-privacy-physical layer security (PLS) trade-off was established. However, note that the secrecy capacity in [8] cannot be approached by finite blocklength (FBL) coding scheme since it is proved by using random binning coding scheme [9]. Then it is natural to ask: can we design a constructive FBL coding scheme for the edge server, and confuse the eavesdropper as much as possible?

In this paper, first, we extend an already existing FBL coding scheme for the white Gaussian channel with noisy feedback [10] to the model of Figure 1, then we derive achievable rate and upper bound on the eavesdropper's uncertainty of the extended scheme. Finally, we show the relationship between utility, privacy, PLS and other parameters via simulation examples.

Figure 1: The wireless HFL in the presence of eavesdroppers

II. PRELIMINARY, MODEL FORMULATION AND MAIN RESULTS

A. Preliminary: learning protocol

In Figure 1, there are a cloud server, L edge servers indexed by ℓ, and K users indexed by k and ℓ. $\{\mathcal{C}_\ell\}_{\ell=1}^{L}$ represents the disjoint user sets and $|\mathcal{C}_\ell|$ is the number of users in edge domain ℓ, $\{\mathcal{S}_{\ell,k}\}_{k=1}^{|\mathcal{C}_\ell|}$ represents the distributed datasets and $S_{\ell,k} = |\mathcal{S}_{\ell,k}|$ is the cardinality of $\mathcal{S}_{\ell,k}$, where $\mathcal{S}_{\ell,k} = \{(\mathbf{u}_{k,j}, v_{k,j})\}_{j=1}^{|\mathcal{S}_{\ell,k}|}$, $\mathbf{u}_{k,j} \in \mathbb{R}^q$ is the j-th vector of covariates with q features and $v_{k,j} \in \mathbb{R}$ is the corresponding associated label at user k. Denote the aggregated dataset in edge ℓ domain by $\mathcal{S}_\ell$, and each edge server aggregates gradients from its users. The global loss function $F(\mathbf{m})$ is given by

$$F(\mathbf{m}) = \frac{1}{S} \sum_{\ell=1}^{L} \sum_{k=1}^{|\mathcal{C}_\ell|} S_{\ell,k} F_{\ell,k}(\mathbf{m}), \qquad (2.1)$$

where $\mathbf{m} \in \mathbb{R}^q$ is the model vector and $S = \sum_\ell \sum_k S_{\ell,k}$. $F_{\ell,k}(\cdot)$ is the local loss function for user k, where

$$F_{\ell,k}(\mathbf{m}) = \frac{1}{S_{\ell,k}} \sum_{(\mathbf{u}_{k,j}, v_{k,j}) \in \mathcal{S}_{\ell,k}} f(\mathbf{m}; \mathbf{u}_{k,j}, v_{k,j}) + \lambda R(\mathbf{m}), \quad (2.2)$$

and $f(\mathbf{m}; \mathbf{u}_{k,j}, v_{k,j})$ is the sample-wise loss function. $R(\mathbf{m})$ is a strongly convex regularization function and $\lambda \geq 0$. The model training by minimizing the global loss function as

$$\mathbf{m}^\star = \arg\min_{\mathbf{m}} F(\mathbf{m}). \quad (2.3)$$

To minimize $F(\mathbf{m})$, we use a distributed gradient descent iterative algorithm. Specifically, *in the t-th ($t \in \{1, 2, ..., T\}$) communication round (the overall communication round is T)*, each user k computes its own local gradient $\nabla F_{\ell,k}(\mathbf{m}_t)$ and the users send the corrupted local gradients (added by Gaussian noises for LDP) to the edge servers. Then, the edge server ℓ computes its estimation $\widehat{\nabla F_\ell}(\mathbf{m}_t)$ of the partial gradient and $\nabla F_\ell(\mathbf{m}_t) = \frac{1}{S_\ell} \sum_{k \in \mathcal{C}^\ell} S_{\ell,k} \nabla F_{\ell,k}(\mathbf{m}_t)$, where $S_\ell = |\mathcal{S}_\ell|$ is the total number of $\mathcal{S}_\ell$. The cloud server's estimation $\widehat{\nabla F}(\mathbf{m}_t)$ of the global gradient is given by $\nabla F(\mathbf{m}_t) = \frac{1}{S} \sum_{\ell=1}^L S_\ell \nabla F_\ell(\mathbf{m}_t)$. The global model $\mathbf{m}_{t+1}$ updated by the cloud server is given by

$$\mathbf{m}_{t+1} = \mathbf{m}_t - \mu \widehat{\nabla F}(\mathbf{m}_t), \quad (2.4)$$

where μ is the learning rate. For convenience, in the t-th communication round, we denote $\mathbf{W}_{t,k} = S_{\ell,k} \nabla F_{\ell,k}(\mathbf{m}_t)$.

B. Model formulation

In this paper, we assume that each edge server communicates with the cloud server without interference from other edge servers. Besides, we assume that the downlink communication is perfect, which is similar to [2], and eavesdropper only shows interest in the data transmitted in the uplink communication between the edge servers and the cloud server. Hence we only focus on **the T rounds uplink communication** between one of the edge servers and the cloud server. An information-theoretic approach of Figure 1 is illustrated in Figure 2. For simplification, we make the following assumptions:

- Similar to [2]-[3], we assume that the channel coefficients stay constants during the transmission (quasi-static fading channel).
- Similar to [3]-[4], we assume that the cloud server and the edge server have perfect channel state information (CSI) of the feedforward channel and feedback channel.
- From similar arguments in [11], we assume that eavesdropper is an active user but it is un-trusted by the cloud server, which indicates that the perfect CSI of eavesdropper's channel is known by the eavesdropper and the edge server. Moreover, we assume that the eavesdropper also knows the perfect CSI of the edge server-cloud server's channels.

Information source: In Figure 2(a), we assume that $\mathbf{W}_{t,k} \in \mathbb{R}^q$ is the k-th ($k \in \{1, 2, ..., K\}$) user's overall local gradient vector in t-th ($t \in \{1, 2, ..., T\}$) communication round, where $\mathbf{W}_{t,k} = (W_{t,k,1}, ..., W_{t,k,q})^\mathcal{T}$. Similar to [12], the elements of $\mathbf{W}_{t,k}$ are independent and identically distributed (i.i.d.) and $\mathbf{W}_{t,k} \sim \mathcal{N}(0, S_{\ell,k}\sigma_{w,t}^2 \mathbf{I})$. Let $\boldsymbol{\eta}_{t,k} = (\eta_{t,k,1}, ..., \eta_{t,k,q})^\mathcal{T}$ be local artificial Gaussian noise i.i.d. according to distribution $\mathcal{N}(0, \sigma^2 \mathbf{I})$. The corrupted local gradient $\mathbf{W}'_{t,k} = $

(a) An information-theoretic approach of Figure 1: encoding

(b) An information-theoretic approach of Figure 1: decoding

Figure 2: An information-theoretic approach of Figure 1

$(W'_{t,k,1}, ..., W'_{t,k,q})^\mathcal{T}$ that is aggregated by the edge server is given by

$$\mathbf{W}'_{t,k} = \mathbf{W}_{t,k} + \boldsymbol{\eta}_{t,k}, \quad (2.5)$$

where $\mathbf{W}'_{t,k} \sim \mathcal{N}(0, (S_{\ell,k}\sigma_{w,t}^2 + \sigma^2)\mathbf{I})$ for $k \in \{1, 2, ..., K\}$. The overall local gradients and the overall noises are defined as $\mathbf{W}_t = (W_{t,1}, ..., W_{t,q})^\mathcal{T}$ and $\boldsymbol{\eta}_t = (\eta_{t,1}, ..., \eta_{t,q})^\mathcal{T}$, respectively, where $W_{t,i} = \sum_{k=1}^K W_{t,k,i}$, $\eta_{t,i} = \sum_{k=1}^K \eta_{t,k,i}$ ($i \in \{1, 2, ..., q\}$). According to (2.5), we define the overall corrupted local gradients sent to the edge server as $\mathbf{W}'_t = (W'_{t,1}, ..., W'_{t,q})^\mathcal{T}$, where $W'_{t,i} = \sum_{k=1}^K W'_{t,k,i}$ ($i \in \{1, 2, ..., q\}$). Here note that since $\mathbf{W}_{t,k}$ and $\boldsymbol{\eta}_{t,k}$ are i.i.d. generated, $\mathbf{W}'_t$ is also composed of i.i.d. components, where $\mathbf{W}'_t \sim \mathcal{N}(0, (S_\ell \sigma_{w,t}^2 + K\sigma^2)\mathbf{I})$.

Definition 1 (Privacy by mutual information [13]): If the mutual information between $\mathbf{W}_t$ and $\mathbf{W}'_t$ satisfies $\frac{1}{qT} \sum_{t=1}^T I(\mathbf{W}_t; \mathbf{W}'_t) \leq \epsilon$, we say the LDP mechanism satisfies ϵ-mutual-information privacy for some $\epsilon > 0$.

Definition 2 (Utility by quadratic distortion [14]): The utility of $\mathbf{W}'_t$ is characterized by $d(\mathbf{W}_t, \mathbf{W}'_t) = ||\mathbf{W}'_t - \mathbf{W}_t||^2$, where $||\mathbf{X}||$ represents the l_2-norm of the vector $\mathbf{X}$. If $\frac{1}{qT} \sum_{t=1}^T E(d(\mathbf{W}_t, \mathbf{W}'_t)) \leq U$, we say the utility of $\mathbf{W}'_t$ is up to U.

Channels: At time instant i ($i \in \{1, 2, ..., N_t\}$ of t-th communication round, channel inputs and outputs are given by

$$Y_i(t) = hX_i(t) + \eta_{1,i}(t), \quad i = 1, 2, ..., N_t, \quad (2.6)$$

$$\widetilde{Y}_i(t) = \widetilde{h}\widetilde{X}_i(t) + \eta_{2,i}(t), \quad i = 1, 2, ..., N_t - 1, \quad (2.7)$$

$$Z_i(t) = gX_i(t) + \widetilde{g}\widetilde{X}_i(t) + \eta_{e,i}(t), \quad i = 1, 2, ..., N_t, \quad (2.8)$$

where $X_i(t)$ and $\widetilde{X}_i(t)$ respectively are the feedforward and feedback channel inputs, which satisfy the average power constraints $\frac{1}{N_t} \sum_{i=1}^{N_t} E[X_i(t)X_i(t)^\mathcal{H}] \leq P$ and $\frac{1}{N_t-1} \sum_{i=1}^{N_t-1} E[\widetilde{X}_i(t)\widetilde{X}_i(t)^\mathcal{H}] \leq \widetilde{P}$. $h, \widetilde{h}, g, \widetilde{g} \in \mathbb{C}$ are the CSI of the feedforward and feedback channels of the cloud server,

the feedforward and feedback channels of the eavesdropping channel, respectively, here note that $|h|$, $|\widetilde{h}|$, $|g|$ and $|\widetilde{g}|$ represent the modulus of h, $\widetilde{h}$, g and $\widetilde{g}$. $Y_i(t)$, $\widetilde{Y}_i(t)$ and $Z_i(t)$ respectively are the channel outputs of the cloud server, the edge server and the eavesdropper, $\eta_{1,i}(t)$, $\eta_{2,i}(t)$ and $\eta_{e,i}(t)$ are i.i.d. as $\mathcal{CN}(0, \sigma_1^2)$, $\mathcal{CN}(0, \sigma_2^2)$ and $\mathcal{CN}(0, \sigma_e^2)$, respectively. The signal-to-noise ratios of the feedforward and feedback channels are denoted by $\text{SNR} = \frac{P}{\sigma_1^2}$ and $\widetilde{\text{SNR}} = \frac{\widetilde{P}}{\sigma_2^2}$.

Source coding: We consider a lossy Gaussian source coding with quadratic distortion measure $d(\mathbf{W}'_t, \hat{\mathbf{W}}'_t) = ||\mathbf{W}'_t - \hat{\mathbf{W}}'_t||^2$, where $\hat{\mathbf{W}}'_t$ is the estimation of source decoder. According to [14, Chapter 3.8, pp. 64-65], there exists a source encoder mapping $\mathbf{W}'_t \rightarrow \{1, 2, ..., 2^{qR_t(D)}\}$, it compresses $\mathbf{W}'_t$ into an index W''_t which is uniformly distributed in $\mathcal{W}''_t = \{1, 2, ..., 2^{qR_t(D)}\}$, and the rate-distortion function $R_t(D)$ is given by

$$R_t(D) = \begin{cases} \frac{1}{2}\log\frac{S_\ell\sigma_{w,t}^2 + K\sigma^2}{D} & 0 \leq D < S_\ell\sigma_{w,t}^2 + K\sigma^2 \\ 0 & D \geq S_\ell\sigma_{w,t}^2 + K\sigma^2 \end{cases} \quad (2.9)$$

where $\frac{1}{qT}\sum_{t=1}^{T} E(d(\mathbf{W}'_t, \hat{\mathbf{W}}'_t)) \leq D$. For the source decoder, a source decoding mapping maps $\{1, 2, ..., 2^{qR_t(D)}\}$ to $\hat{\mathbf{W}}'_t$.

Channel encoder: At time i ($i \in \{1, 2, ..., N_t\}$) in t-th ($t \in \{1, 2, ..., T\}$) communication round, the transmitted codeword $X_i(t)$ is a stochastic function of the W''_t, h, $\widetilde{h}$ and $\widetilde{Y}_1^{i-1}(t) = (\widetilde{Y}_1(t), ..., \widetilde{Y}_{i-1}(t))$, i.e., $X_i(t) = f_{t,i}(W''_t, h, \widetilde{h}, \widetilde{Y}_1^{i-1}(t))$.

Channel decoder: The channel decoder's estimation $\hat{w}''_t = \varphi(h, \widetilde{h}, Y^{N_t})$, where φ is the channel decoder's decoding function. The channel decoder with outputs $\widetilde{X}_i(t) = \widetilde{f}_{t,i}(h, \widetilde{h}, Y_1^i(t))$, where $\widetilde{f}_{t,i}(\cdot)$ is a stochastic function. The average decoding error probability of message w''_t is given by

$$P_{e,t} = \frac{1}{|\mathcal{W}''_t|}\sum_{w''_t \in \mathcal{W}''_t} Pr\{\varphi(h, \widetilde{h}, Y^{N_t}) \neq w''_t | w''_t \text{ sent}\}. \quad (2.10)$$

Definition 3 The uncertainty of the eavesdropper (also called the secrecy level, which is first adopted in [15]) is defined as

$$\Delta = \frac{H(W''_1, ..., W''_T | Z^{N_1}, ..., Z^{N_T}, h, \widetilde{h}, g, \widetilde{g})}{H(W''_1, ..., W''_T)}, \quad 0 \leq \Delta \leq 1. \quad (2.11)$$

For fixed source encoding-decoding procedure, a rate R is said to be $(N, \tau, U, \epsilon, \delta, D)$ achievable, if for given decoding error probability τ, blocklength N, secrecy leve δ, $\frac{1}{qT}\sum_{t=1}^{T} I(\mathbf{W}_t; \mathbf{W}'_t) \leq \epsilon$ and $\frac{1}{qT}\sum_{t=1}^{T} E(d(\mathbf{W}_t, \mathbf{W}'_t)) \leq U$, there exists a channel code described above such that

$$\frac{H(W''_1, ..., W''_T)}{N} = R, \quad \frac{1}{T}\sum_{t=1}^{T} P_{e,t} \leq \tau, \quad \Delta \geq \delta, \quad (2.12)$$

where $N = \sum_{t=1}^{T} N_t$. The secrecy capacity $\mathcal{C}(N, \tau, U, \epsilon, \delta, D)$ is composed of all the secrecy achievable rates $R(N, \tau, U, \epsilon, \delta, D)$ defined above. Here note that $\delta \in [0, 1]$, and $\delta = 1$ corresponds to perfect secrecy.

C. Main result

Theorem 1. *For given N, τ, U, ϵ, δ and D, a lower bound on the secrecy capacity $\mathcal{C}(N, \tau, U, \epsilon, \delta, D)$ is given by*

$$\mathcal{C}(N, \tau, U, \epsilon, \delta, D) \geq R(N, \tau, U, \epsilon, \delta, D), \quad (2.13)$$

where

$$R(N, \tau, U, \epsilon, \delta, D) = \frac{\sum_{t=1}^{T} N_t R_t}{N}, \quad N = \sum_{t=1}^{T} N_t \quad (2.14)$$

$$R_t = \frac{1}{N_t}\log\left(\frac{3SNR|h|^2}{[Q^{-1}(\frac{\tau}{8})]^2}\left(1 + \frac{SNR|h|^2}{\Psi_1\Psi_2}\right)^{N_t - 1}\right), \quad (2.15)$$

$$\Psi_1 = 1 + L\frac{|h|^2 SNR}{|\widetilde{h}|^2 S\widetilde{N}R}, \quad \Psi_2 = \left(1 - \frac{L}{|\widetilde{h}|^2 S\widetilde{N}R}\right)^{-1}, \quad (2.16)$$

$$L = \frac{1}{3}\left[Q^{-1}\left(\frac{\tau}{8(N_t - 1)}\right)\right]^2, \quad (2.17)$$

$Q^{-1}(\cdot)$ *is the inverse function of the Gaussian Q-function* $Q(x) \stackrel{def}{=} \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-\frac{u^2}{2})du$, *and the uncertainty of the eavesdropper (the secrecy level) is upper bounded by*

$$\delta \leq \min_{t \in \{1, ..., T\}}[1 - \frac{\log\left(1 + \frac{|g|^2 P}{\sigma_e^2}\right)}{qR_t(D)}]^+, \quad [x]^+ = \max(x, 0), \quad (2.18)$$

where $R_t(D)$ is defined in (2.9), and

$$\max_{t \in \{1, ..., T\}}\left\{\frac{S_\ell\sigma_{w,t}^2}{K(2^{2\epsilon} - 1)}\right\} \leq \sigma^2 \leq \frac{U}{K}. \quad (2.19)$$

Proof: Theorem 1 is proved by a FBL approach, which will be explained in the next section. The formal proof is also given in the next section. ∎

III. A FBL APPROACH FOR WIRELESS HFL

Since the FBL approach of each communication round is similar, we only describe the FBL approach of t-th ($t \in \{1, 2, ..., T\}$) communication round. Therefore, for simplify the notation, we omit the index t of the signals and the noises in this section.

A. Channel re-presentation

At time i ($i \in \{1, 2, ..., N_t\}$) in t-th communication round, since the elements in (2.6)-(2.7) are complex numbers, (2.6)-(2.7) can be re-written as

$$Y_{R,i} + jY_{I,i} = (h_R + jh_I)(X_{R,i} + jX_{I,i}) + \eta_{R,1,i} + j\eta_{I,1,i},$$
$$\widetilde{Y}_{R,i} + j\widetilde{Y}_{I,i} = (\widetilde{h}_R + j\widetilde{h}_I)(\widetilde{X}_{R,i} + j\widetilde{X}_{I,i}) + \eta_{R,2,i} + j\eta_{I,2,i}, \quad (3.1)$$

where $j = \sqrt{-1}$, $Y_{R,i} = \text{Re}(Y_i)$, $Y_{I,i} = \text{Im}(Y_i)$, $h_R = \text{Re}(h)$, $h_I = \text{Im}(h)$, $X_{R,i} = \text{Re}(X_i)$, $X_{I,i} = \text{Im}(X_i)$, $\eta_{R,1,i} = \text{Re}(\eta_{1,i})$, $\eta_{I,1,i} = \text{Im}(\eta_{1,i})$, $\widetilde{Y}_{R,i} = \text{Re}(\widetilde{Y}_i)$, $\widetilde{Y}_{I,i} = \text{Im}(\widetilde{Y}_i)$, $\widetilde{h}_R = \text{Re}(\widetilde{h})$, $\widetilde{h}_I = \text{Im}(\widetilde{h})$, $\widetilde{X}_{R,i} = \text{Re}(\widetilde{X}_i)$, $\widetilde{X}_{I,i} = \text{Im}(\widetilde{X}_i)$, $\eta_{R,2,i} = \text{Re}(\eta_{2,i})$, $\eta_{I,2,i} = \text{Im}(\eta_{2,i})$, where $\text{Re}(\cdot)$ and $\text{Im}(\cdot)$ denote the real and imaginary parts of a complex element, respectively. Here note that $E(X_{R,i}^2) = P_R$, $E(X_{I,i}^2) = P_I$, $E(\widetilde{X}_{R,i}^2) = \widetilde{P}_R$ and $E(\widetilde{X}_{I,i}^2) = \widetilde{P}_I$, where $P_R = P_I = \frac{1}{2}P$ and $\widetilde{P}_R = \widetilde{P}_I = \frac{1}{2}\widetilde{P}$. From (3.1), we have

$$X_{R(I),i} = Y'_{R(I),i} - \eta'_{R(I),1,i}, \quad \widetilde{X}_{R(I),i} = \widetilde{Y}'_{R(I),i} - \eta'_{R(I),2,i}, \quad (3.2)$$

where $Y'_{R,i} = (h_R Y_{R,i} + h_I Y_{I,i})/|h|^2$, $Y'_{I,i} = (h_R Y_{I,i} - h_I Y_{R,i})/|h|^2$, $\widetilde{Y}'_{R,i} = (\widetilde{h}_R \widetilde{Y}_{R,i} + \widetilde{h}_I \widetilde{Y}_{I,i})/|\widetilde{h}|^2$, $\widetilde{Y}'_{I,i} = (\widetilde{h}_R \widetilde{Y}_{I,i} - \widetilde{h}_I \widetilde{Y}_{R,i})/|\widetilde{h}|^2$, $\eta'_{R,1,i} = (h_R \eta_{R,1,i} + h_I \eta_{I,1,i})/|h|^2$, $\eta'_{I,1,i} = (h_R \eta_{I,1,i} - h_I \eta_{R,1,i})/|h|^2$, $\eta'_{R,2,i} = (\widetilde{h}_R \eta_{R,2,i} + \widetilde{h}_I \eta_{I,2,i})/|\widetilde{h}|^2$ and $\eta'_{I,2,i} = (\widetilde{h}_R \eta_{I,2,i} - \widetilde{h}_I \eta_{R,2,i})/|\widetilde{h}|^2$. Hence, (3.2) is equivalent to (3.1), which indicates that the feedforward and feedback

channels are divide into the two sub-channels. In addition, we conclude that $\text{Var}(\eta'_{R,1,i}) = \text{Var}(\eta'_{I,1,i}) = \sigma_1^2/2|h|^2$ and $\text{Var}(\eta'_{R,2,i}) = \text{Var}(\eta'_{I,2,i}) = \sigma_2^2/2|\tilde{h}|^2$.

B. Message splitting

First, for given D, N_t, τ, ϵ and U, let

$$|\mathcal{W}''_t| = 2^{N_t R_t} = 2^{q R_t(D)}, \tag{3.3}$$

where $R_t = \frac{H(W''_t)}{N_t}$. Here note that when $R_t(D) = 0$, we do not transmit messages and choose $\hat{\mathbf{W}}'_t = \mathbf{0}$. Then, the message W''_t is divided into two independent parts $(W''_{t,R}, W''_{t,I})$, where $W''_{t,R}$ and $W''_{t,I}$ respectively take values in $\mathcal{W}''_{t,R} = \{1, 2, ..., 2^{N_t R_{t,R}}\}$ and $\mathcal{W}''_{t,I} = \{1, 2, ..., 2^{N_t R_{t,I}}\}$, and $R_{t,R} + R_{t,I} = R_t$. Divide the interval $[-\sqrt{3}, \sqrt{3}]$ into $2^{N_t R_{t,R}}(2^{N_t R_{t,I}})$ equally spaced sub-intervals, and the center of each sub-interval is mapped to a message value in $W''_{t,R}(W''_{t,I})$. Let $\theta_R(\theta_I)$ be the center of the sub-interval with respect to (w.r.t) the message $W''_{t,R}(W''_{t,I})$, and $E(\theta_R^2) = E(\theta_I^2) = 1$.

C. Coding scheme

For simplification, we only describe the coding scheme of message $W''_{t,R}$, and the coding scheme of message $W''_{t,I}$ is similar to the coding scheme of message $W''_{t,R}$.

Initialization: At time instant 1, the channel encoder maps the messages $W''_{t,R}$ to θ_R, and sends

$$X_{R,1} = \sqrt{P_R}\theta_R, \tag{3.4}$$

at the end of time 1, the channel decoder of cloud server receives Y_1. Then, the decoder obtains $Y'_{R,1}$ and computes the first estimation $\hat{\theta}_{R,1}$ of θ_R by

$$\hat{\theta}_{R,1} = \frac{Y'_{R,1}}{\sqrt{P_R}} = \theta_R + \frac{\eta'_{R,1,1}}{\sqrt{P_R}} = \theta_R + \varepsilon_{R,1}, \tag{3.5}$$

where $\varepsilon_{R,1} = \hat{\theta}_{R,1} - \theta_R = \frac{\eta'_{R,1,1}}{\sqrt{P_R}}$ is the decoding error of decoder at time instant 1. Define $\alpha_{R,1} = \text{Var}(\varepsilon_{R,1}) = \frac{\sigma_1^2}{2|h|^2 P_R}$.

Iteration: From the second time instant, we first introduce the dither signal sequence $V^{N_t-1} = (V_1, ..., V_{N_t-1})$, which follows from the extended SK-type feedback scheme in [10]. We assume that the i.i.d generated sequence V^{N_t-1} is perfectly known by the edge server and the cloud server, where $V_i \sim \text{Unif}[-\frac{d}{2}, \frac{d}{2}]$, $d = \sqrt{6\tilde{P}}$. The dither signals ensure that the encoded codeword of cloud server satisfies the power constraint.

At time instant i ($2 \le i \le N_t$), the channel decoder of cloud server computes and sends

$$\tilde{X}_{R,i-1} = \mathbb{M}_d[\gamma_{R,i-1}\hat{\theta}_{R,i-1} + V_{i-1}], \tag{3.6}$$

where $\gamma_{R,i-1}$ is the modulation coefficient of cloud server, $\mathbb{M}_d$ is the modulo-d function and it is defined in [10]. From property v of proposition 1 in [10], we have $E(\tilde{X}_{R,i-1}^2) = \frac{\tilde{P}}{2} = \tilde{P}_R$. After the channel encoder receives $\tilde{Y}_{i-1}$, the channel encoder obtains $\tilde{Y}'_{R,i-1}$ and computes the noisy version of decoding error $\varepsilon_{R,i-1} = \hat{\theta}_{R,i-1} - \theta_R$ by

$$\tilde{\varepsilon}_{R,i-1} = \frac{1}{\gamma_{R,i-1}}\mathbb{M}_d[\tilde{Y}'_{R,i-1} - \gamma_{R,i-1}\theta_R - V_{i-1}]$$
$$\overset{(a)}{=} \frac{1}{\gamma_{R,i-1}}\mathbb{M}_d[\gamma_{R,i-1}\varepsilon_{R,i-1} + \eta'_{R,2,i-1}], \tag{3.7}$$

where (a) is due to property ii of proposition 1 in [10]. The *modulo-aliasing errors* do not occur in channel encoder, if suitable $\gamma_{R,i-1}$ is chosen such that $\gamma_{R,i-1}\varepsilon_{R,i-1} + \eta'_{R,2,i-1} \in [-\frac{d}{2}, \frac{d}{2})$. Hence, the channel encoder obtains $\tilde{\varepsilon}_{R,i-1} = \varepsilon_{R,i-1} + \frac{\eta'_{R,2,i-1}}{\gamma_{R,i-1}}$. Then, the channel encoder sends

$$X_{R,i} = \lambda_{R,i-1}\gamma_{R,i-1}\tilde{\varepsilon}_{R,i-1}, \tag{3.8}$$

where $\lambda_{R,i-1}$ is chosen to satisfy the input power constraints P_R ($E[(X_{R,i})^2] = P_R$). Analogously, the channel decoder receives Y_i and computes $Y'_{R,i}$. Then, the channel decoder updates $\hat{\theta}_{R,i}$ by computing

$$\hat{\theta}_{R,i} = \hat{\theta}_{R,i-1} - \hat{\varepsilon}_{R,i-1} = \hat{\theta}_{R,i-1} - \beta_{R,i}Y'_{R,i}, \tag{3.9}$$

where $\hat{\varepsilon}_{R,i-1} = \beta_{R,i}Y'_{R,i}$, and $\beta_{R,i} = \frac{E(Y'_{R,i}\varepsilon_{R,i-1})}{E(Y'_{R,i})^2}$ is the Minimum Mean Square Error (MMSE) estimation coefficient, which ensures that $\varepsilon_{R,i-1}$ is correctly estimated from $Y'_{R,i}$. Define $\varepsilon_{R,i} = \hat{\theta}_{R,i} - \theta_R$, (3.9) yield that

$$\varepsilon_{R,i} = \varepsilon_{R,i-1} - \beta_{R,i}Y'_{R,i}, \tag{3.10}$$

and define $\alpha_{R,i} = \text{Var}(\varepsilon_{R,i})$.

Decoding: At time instant N_t, the channel decoder obtains the final estimation $\hat{\theta}_{R,N_t} = \theta_R + \varepsilon_{R,N_t}$. Then the channel decoder declares the center of sub-interval which $\hat{\theta}_{R,N_t}$ belongs to as the final estimation of the θ_R. The channel decoder successfully decodes the message $W''_{t,R}$ if $\hat{\theta}_{R,N_t}$ is in the sub-interval of θ_R, i.e., $\varepsilon_{R,N_t} \in [-\frac{\sqrt{3}}{2^{N_t R_{t,R}}}, \frac{\sqrt{3}}{2^{N_t R_{t,R}}})$.

The following algorithm 1 further explains the encoding-decoding scheme described above.

Algorithm 1 Encoding-decoding procedure of a sub-channel

Input: $W''_{t,R}, \tau, N_t, P_R, \tilde{P}_R, \sigma_1^2, \sigma_2^2, h, \tilde{h}, V^{N_t-1}, d = \sqrt{6\tilde{P}}$
Output: $X_{R,N_t}, \hat{\theta}_{R,N_t}$

    Initialization:
    Map $W''_{t,R} \to \theta_R$
    The edge server encodes $X_{R,1} = \sqrt{P_R}\theta_R$
    The cloud server computes $\hat{\theta}_{R,1}$ from (3.5)
    Iteration:
1: **for** $2 \le i \le N_t$ **do**
2:     Compute $\lambda_{R,i-1}, \gamma_{R,i-1}, \beta_{R,i}$ from (3.13)-(3.15)
3:     The cloud server encodes $\tilde{X}_{R,i-1}$ from (3.6)
4:     The edge server computes $\tilde{\varepsilon}_{R,i-1}$ from (3.7)
5:     The edge server encodes $X_{R,i}$ from (3.8)
6:     The cloud server computes $\hat{\theta}_{R,i}$ from (3.9)
7: **end for**

D. Performance analysis

1) Utility and privacy analysis: First, note that since $\mathbf{W}_t$, η_t and $\mathbf{W}'_t$ are i.i.d. generated, from Definition 1, we conclude that

$$\frac{1}{qT}\sum_{t=1}^{T} I(\mathbf{W}_t; \mathbf{W}'_t) \le \max_{t \in \{1,..,T\}} \frac{1}{2}\log\left(1 + \frac{S_\ell \sigma_{w,t}^2}{K\sigma^2}\right) \le \epsilon, \tag{3.11}$$

On the other hand, from Definition 2, we conclude that

$$\frac{1}{qT}\sum_{t=1}^{T} E(d(\mathbf{W}_t, \mathbf{W}'_t)) = \frac{1}{qT}\sum_{t=1}^{T} E(||\eta_t||^2) = K\sigma^2 \le U. \tag{3.12}$$

Combining (3.11) and (3.12), (2.19) in Theorem 1 is proved.

2) Parameter analysis: In our proposed scheme, we define the parameters $\lambda_{R,i}$, $\beta_{R,i}$, and $\gamma_{R,i}$ as follows, and the transmission performance of our scheme is determined by these parameters.

$$\lambda_{R,i} = \sqrt{L \cdot \frac{P}{\widetilde{P}}}, \quad \gamma_{R,i} = \sqrt{\frac{1}{\alpha_{R,i}}\left(\frac{\widetilde{P}}{2L} - \frac{\sigma_2^2}{2|\widetilde{h}|^2}\right)}, \qquad (3.13)$$

$$\beta_{R,i} = \frac{\sqrt{2\alpha_{R,i-1}}}{\sigma_1} \frac{\sqrt{\mathrm{SNR}(1 - L \cdot \mathrm{S\tilde{N}R}^{-1}|\widetilde{h}|^{-2})}}{\mathrm{SNR} + |h|^{-2}}, \qquad (3.14)$$

$$\alpha_{R,i} = |h|^{-2}\mathrm{SNR}^{-1}\left(1 + \frac{\mathrm{SNR}|h|^2}{\Psi_1\Psi_2}\right)^{1-i}, \qquad (3.15)$$

and Ψ_1, Ψ_2 and L are defined in (2.16)-(2.17). Combining the above definitions of parameters, and through the error probability analysis, the achievable rate of this FBL scheme in t-th communication round is given by

$$R_t = \frac{1}{N_t} \log\left(\frac{3\mathrm{SNR}|h|^2}{\left[Q^{-1}(\frac{\tau}{8})\right]^2}\left(1 + \frac{\mathrm{SNR}|h|^2}{\Psi_1\Psi_2}\right)^{N_t - 1}\right). \qquad (3.16)$$

The parameters derivation and error probability analysis of the FBL scheme are similar to those in [10], hence we omit it here.

3) Security analysis: First, we will perform a security analysis of our FBL scheme within the FBL regime, the eavesdropper's equivocation rate Δ can be re-written as

$$\Delta = \frac{H(W_1'', ..., W_T''|Z^{N_1}, ..., Z^{N_T}, h, \widetilde{h}, g, \widetilde{g})}{H(W_1'', ..., W_T'')}$$
$$\overset{(b)}{=} \frac{\sum_{t=1}^T H(W_t''|Z^{N_t}, h, \widetilde{h}, g, \widetilde{g})}{\sum_{t=1}^T H(W_t'')}, \qquad (3.17)$$

where (b) is due to the fact that $\mathbf{W}_t'$ is mapped into W_t'' at each communication round, which indicates that $(W_1'', ..., W_T'')$ and $(Z^{N_1}, ..., Z^{N_T})$ are independent of each other, and $H(W_t''|Z^{N_t}, h, \widetilde{h}, g, \widetilde{g})$ is given by

$$H(W_t''|Z^{N_t}, h, \widetilde{h}, g, \widetilde{g})$$
$$\overset{(c)}{\geq} H(W_t''|gX_1(t) + \widetilde{g}\widetilde{X}_1(t) + \eta_{e,1}(t), ..., gX_{N_t-1}(t) + \widetilde{g}\widetilde{X}_{N_t-1}(t) +$$
$$\eta_{e,N_t-1}(t), gX_{N_t}(t) + \eta_{e,N_t}(t), \eta_{1,1}(t), ..., \eta_{1,N_t}(t), \eta_{2,1}(t), ..., \eta_{2,N_t}(t),$$
$$\eta_{e,2}(t), ..., \eta_{e,N_t}(t), \widetilde{X}_1(t), ..., \widetilde{X}_{N_t-1}(t), h, \widetilde{h}, g, \widetilde{g})$$
$$\overset{(d)}{=} H(W_t''|gX_1(t) + \eta_{e,1}(t), \eta_{1,1}(t), ..., \eta_{1,N_t}(t), \eta_{2,1}(t), ..., \eta_{2,N_t}(t),$$
$$\eta_{e,2}(t), ..., \eta_{e,N_t}(t), \widetilde{X}_1(t), ..., \widetilde{X}_{N_t-1}(t), h, \widetilde{h}, g, \widetilde{g})$$
$$\overset{(e)}{=} H(W_t''|gX_1(t) + \eta_{e,1}(t))$$
$$\overset{(f)}{=} H(W_t'') + h(\eta_{e,1}(t)) - h(gX_1(t) + \eta_{e,1}(t))$$
$$\overset{(g)}{=} H(W_t'') - \log(1 + \frac{|g|^2 P}{\sigma_e^2}), \qquad (3.18)$$

where (c) follows from (2.8), (d) follows from $X_i(t) = X_{R,i}(t) + jX_{I,i}(t)$ $(i = 2, ..., N_t)$ is a function of $h, \widetilde{h}, \eta_{1,1}(t), ..., \eta_{1,i-1}(t), \eta_{2,1}(t), ..., \eta_{2,i-1}(t)$, (e) follows from the fact that $\widetilde{X}_i(t) = \widetilde{X}_{R,i}(t) + j\widetilde{X}_{I,i}(t)$ $(i = 1, ..., N_t - 1)$ is only related to V_i [16, Chapter 4.1, pp. 61-63], and $h, \widetilde{h}, g, \widetilde{g}, \eta_{1,1}(t), ..., \eta_{1,N_t}(t), \eta_{2,1}(t), ..., \eta_{2,N_t}(t), \eta_{e,2}(t), ..., \eta_{e,N_t}(t), V_1, ..., V_{N_t-1}$ are independent of W_t'', $X_1(t)$, $\eta_{e,1}(t)$, (f) is due to the fact that $X_1(t) = \sqrt{P_R}\theta_R + j\sqrt{P_I}\theta_I$, and $W_t'' = (W_{t,R}'', W_{t,I}'')$ are mapped into θ_R and θ_I, respectively, and (g) is follows from

$$h(gX_1(t) + \eta_{e,1}(t)) - h(\eta_{e,1}(t))$$
$$\leq \log\det\{\pi e[E(g(\sqrt{P_R}\theta_R + j\sqrt{P_I}\theta_I)(\sqrt{P_R}\theta_R + j\sqrt{P_I}\theta_I)^{\mathcal{H}}g^{\mathcal{H}})$$
$$+ E(\eta_{e,1}(t)\eta_{e,1}(t)^{\mathcal{H}})]\} - \log\det\{\pi eE(\eta_{e,1}(t)\eta_{e,1}(t)^{\mathcal{H}})\}$$
$$= \log(1 + \frac{|g|^2 P}{\sigma_e^2}). \qquad (3.19)$$

Substituting (3.18) into (3.17), we have

$$\Delta \geq \frac{\sum_{t=1}^T H(W_t'')(1 - \frac{\log(1 + \frac{|g|^2 P}{\sigma_e^2})}{H(W_t'')})}{\sum_{t=1}^T H(W_t')} \geq \min_{t \in \{1, ..., T\}}(1 - \frac{\log(1 + \frac{|g|^2 P}{\sigma_e^2})}{H(W_t'')}). \qquad (3.20)$$

From (3.3) and (3.20), $\Delta \geq \delta$ in (2.12) is guaranteed if

$$\delta \leq \min_{t \in \{1, ..., T\}}[1 - \frac{\log(1 + \frac{|g|^2 P}{\sigma_e^2})}{qR_t(D)}]^+. \qquad (3.21)$$

Then, the secrecy achievable rate $R(N, \tau, U, \epsilon, \delta, D)$ is given by

$$R(N, \tau, U, \epsilon, \delta, D) = \frac{H(W_1'', ..., W_T'')}{N} \overset{(h)}{=} \frac{\sum_{t=1}^T H(W_t'')}{N} = \frac{\sum_{t=1}^T N_t R_t}{N}, \qquad (3.22)$$

where (h) is similar to (b). Finally, combining the (3.11), (3.12), (3.16), (3.21) and (3.22), we complete the proof of Theorem 1.

IV. SIMULATION RESULTS

Similar to [1] and [4], we assume that the channel coefficients are i.i.d. as $\mathcal{CN}(0, 1)$, here note simulation results are based on an average of 1000 independent channel realizations. We consider a wireless HFL system with $K = 10$ users, a edge server and a cloud server, and the training samples are uniformly distributed across the 10 users. The regularization function $R(\mathbf{m}) = ||\mathbf{m}||^2$ with $\lambda = 5 \times 10^{-5}$. Before the channel coding, the edge server compresses the quantified data via Lempel Ziv Welch (LZW) source coding [17], and the total data amount to be transmitted is defined as M bits. The wireless data transmission latency for uploading information of edge server can be calculated by $T_c = M/R_{eg}$ [1], where R_{eg} represents the transmission rate of edge server. We train a neural network on the MNIST data set[1], and the neural network consists of 784 input nodes, a single hidden layer with 20 hidden nodes, and 10 output nodes. We use the cross entropy as the loss function, and the rectified linear unit (ReLU) and the softmax functions are the activation functions of the hidden and output layers, respectively. The total number of parameters in the neural network is $q = 15910$ and the learning rate be $\mu = 1$.

From Figure 3(a) and Figure 3(b), we see that the channel coding does not affect the learning performance of HFL, and the eavesdropper has a poor learning performance when using our proposed FBL scheme, which indicates that the PLS of the data is guaranteed by the proposed FBL scheme. In Figure 4, we conclude that the transmission latency of the proposed FBL scheme is significantly low compared with LDPC codes (10x less latency). Figure 5 plots the learning performance of our FBL scheme under different privacy-utility constraints, we conclude that more stringent privacy-utility constraints (smaller ϵ and larger U) lead to lower learning performance. The different privacy-utility constraints do not affect the achievable secrecy rates, as shown in Figure 6(a), because the achievable secrecy rate approaches a constant as blocklength increases. Figure 6(b) shows that the secrecy level increases as the privacy-utility constraints become more stringent. Moreover, as the

[1]http://yann.lecun.com/exdb/mnist/

communication round increases, the secrecy level decreases because the variance of the gradient decreases as the training continues [12].

(a) Test accuracy

(b) Training loss

Figure 3: Performance comparison between the different schemes ($U = 5$, $\epsilon = 0.1$, $D = 10^{-4}$, $\tilde{\text{SNR}} = 15$dB, $\tau = 10^{-6}$, $\sigma^2 = 0.5$, $\sigma_1^2 = \sigma_2^2 = \sigma_e^2 = 1$, $P = 10$, $S_\ell = 60000$)

Figure 4: Transmission latency with different schemes ($U = 5$, $\epsilon = 0.1$, $D = 10^{-4}$, $\tilde{\text{SNR}} = 15$dB, $\tau = 10^{-6}$, $\sigma^2 = 0.5$, $\sigma_1^2 = \sigma_2^2 = \sigma_e^2 = 1$, $T = 200$, $S_\ell = 60000$)

(a) Test accuracy

(b) Training loss

Figure 5: Performance comparison between the different U and ϵ of proposed FBL scheme ($\tilde{\text{SNR}} = 15$dB, $\tau = 10^{-6}$, $\sigma_1^2 = \sigma_2^2 = \sigma_e^2 = 1$, $P = 10$, $D = 10^{-4}$, $S_\ell = 60000$)

V. CONCLUSION AND FUTURE WORK

This paper proposes a practical FBL coding scheme for the wireless HFL in the presence of PLS, which almost achieves perfect secrecy without affecting learning performance. Besides this, simulation results show that the coding blocklength of our proposed scheme is significantly shorter than classical LDPC code. One possible future work is to study the case that imperfect CSI is obtained by all parties.

(a) Achievable secrecy rates

(b) Secrecy level

Figure 6: Performance of proposed FBL scheme ($\tilde{\text{SNR}} = 15$dB, $\tau = 10^{-6}$, $\sigma_1^2 = \sigma_2^2 = \sigma_e^2 = 1$, $P = 10$, $D = 10^{-4}$, $S_\ell = 60000$)

REFERENCES

[1] G. Zhu, Y. Wang and K. Huang, "Broadband Analog Aggregation for Low-Latency Federated Edge Learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491-506, Jan. 2020.

[2] D. Liu and O. Simeone, "Privacy for Free: Wireless Federated Learning via Uncoded Transmission With Adaptive Power Control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170-185, 2021.

[3] M. Seif, R. Tandon and M. Li, "Wireless Federated Learning with Local Differential Privacy," *2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, 2020, pp. 2604-2609.

[4] K. Yang, T. Jiang, Y. Shi and Z. Ding, "Federated Learning Based on Over-the-Air Computation," *2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, 2019, pp. 1-6.

[5] S. Liu, G. Yu, X. Chen and M. Bennis, "Joint User Association and Resource Allocation for Wireless Hierarchical Federated Learning with IID and Non-IID Data," *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 7852-7866, Oct. 2022.

[6] L. Liu, J. Zhang, S. H. Song and K. B. Letaief, "Client-Edge-Cloud Hierarchical Federated Learning," *2020 IEEE International Conference on Communications (ICC)*, Dublin, Ireland, 2020, pp. 1-6.

[7] M. Abadi et al., "Deep Learning with Differential Privacy," *Proceeding of ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, pp. 303-318, 2016.

[8] H. Zhang, C. Yang and B. Dai, "When Wireless Federated Learning Meets Physical Layer Security: The Fundamental Limits," *2022 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, New York, NY, USA, 2022, pp. 1-6.

[9] A. D. Wyner, "The Wire-Tap Channel," *Bell System Techical Journal*, vol. 54, no. 8, pp. 1355-1387, 1975.

[10] A. Ben-Yishai and O. Shayevitz, "Interactive Schemes for the AWGN Channel with Noisy Feedback," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2409-2427, 2017.

[11] A. Mukherjee, S. A. A. Fakoorian, J. Huang and A. L. Swindlehurst, "Principles of Physical Layer Security in Multiuser Wireless Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1550-1573, Third Quarter 2014.

[12] Z. J. Chen, E. E. Hernandez, Y. C. Huang and S. Rini, "DNN gradient lossless compression: Can GenNorm be the answer?," *2022 IEEE International Conference on Communications (ICC)*, Seoul, Korea, 2022, pp. 407-412.

[13] W. Wang, L. Ying and J. Zhang, "On the Relation Between Identifiability, Differential Privacy, and Mutual-Information Privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5018-5029, 2016.

[14] A. A. El Gamal and Y.-H. Kim, Network Information Theory. Cambridge, U.K.: Cambridge University Press, 2011.

[15] E. Tekin and A. Yener, "The Gaussian Multiple Access Wire-Tap Channel," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5747-5755, Dec. 2008.

[16] R. Zamir, Lattice Coding for Signals and Network. Cambridge, U.K.: Cambridge University Press, 2014.

[17] T. A. Welch, "A technique of high-performance data compression," *IEEE Computer*, vol. 17, no. 6, pp. 8-19, June 1984.