PNAS
www.pnas.org

Supplementary Information for

Communication Efficient Federated Learning

Mingzhe Chen, Nir Shlezinger, H. Vincent Poor, Yonina C. Eldar, and Shuguang Cui

H. Vincent Poor
E-mail: poor@princeton.edu

This PDF file includes:

Supporting Information Text

1. Architecture of Wireless Federated Learning

Next, we introduce a practical scenario for the implementation of an federated learning (FL) algorithm, as shown in Fig. S1. In this figure, U devices and one central controller (base station) cooperatively participate in the FL training process. During the training process, each device first uses its collected data to train its local FL model. Then each device transmits it trained local FL parameters to the base station (BS). The BS will aggregate the received local FL parameters to generate the global FL model and broadcast it back to all devices. Since local FL parameters and the global FL model are transmitted over wireless links, the FL training performance will be affected by wireless network performance. In particular, the number of resource blocks (RBs) is limited and hence, the number of devices that can participate in FL is limited. Meanwhile, from Fig. S1, we can see that the devices in other service areas may use the same RBs to transmit data thus affecting the FL parameter transmission delay and FL convergence time.

2. Universal FL Model Parameter Compression

A. Rationale. Compression can be modeled as an encoding-decoding system. To faithfully represent the FL setup, we design our quantization strategy in light of the following requirements and assumptions:

A1 All devices share the same encoding function. This requirement significantly simplifies FL implementation.

A2 No *a-priori* knowledge or distribution of the model updates is assumed.

A3 The devices and the central controller (CC) share a source of common randomness. This is achieved by, e.g., letting the CC share with each device a random seed along with the weights. Once a different seed is conveyed to each device, it can be used to obtain a dedicated source of common randomness shared by the CC and each of the devices for the entire FL procedure. These seeds can be conveyed along with the weight.

Requirement *A2* gives rise to the need for a *universal quantization* approach, namely, a scheme which operates reliably regardless of the distribution of the FL parameter updates and without its prior knowledge.

B. Model Compression Algorithm. Here, we present the encoding and decoding functions. Following requirement *A1*, we utilize universal vector quantization, i.e., a quantization scheme which maps each set of continuous-amplitude values into a discrete representation in a manner which is ignorant of the underlying distribution. The source of common randomness assumed in *A3* implies that the CC and the devices can generate the same realizations of a dither signal. We thus use a compression based on dithered vector quantization, and particularly, on lattice quantization, detailed in the following.

Let L be a fixed positive integer, referred to henceforth as the lattice dimension, and let $\boldsymbol{G}$ be a non-singular $L \times L$ matrix, which denotes the lattice generator matrix. For simplicity, we assume that $M \triangleq \frac{m}{L}$ is an integer, where m is the number of elements in the weight matrices of the FL parameters used to represent an FL model, although the scheme can also be applied when this does not hold by replacing M with $\lceil M \rceil$. Next, we use $\mathcal{L}$ to denote the lattice, which is the set of points in $\mathbb{R}^L$ that can be written as an integer linear combination of the columns of $\boldsymbol{G}$, i.e.,

$$\mathcal{L} \triangleq \{\boldsymbol{x} = \boldsymbol{G}\boldsymbol{l} : \boldsymbol{l} \in \mathbb{Z}^L\}. \quad [1]$$

A lattice quantizer $Q_{\mathcal{L}}(\cdot)$ maps each $\boldsymbol{x} \in \mathbb{R}^L$ to its nearest lattice point, i.e., $Q_{\mathcal{L}}(\boldsymbol{x}) = \boldsymbol{l}_{\boldsymbol{x}}$ where $\boldsymbol{l}_{\boldsymbol{x}} \in \mathcal{L}$ if $\|\boldsymbol{x} - \boldsymbol{l}_{\boldsymbol{x}}\| \leq \|\boldsymbol{x} - \boldsymbol{l}\|$ for every $\boldsymbol{l} \in \mathcal{L}$. Finally, let $\mathcal{P}_0$ be the basic lattice cell, i.e., the set of points in $\mathbb{R}^L$ which are closer to $\mathbf{0}$ than to any other lattice point:

$$\mathcal{P}_0 \triangleq \{\boldsymbol{x} \in \mathbb{R}^L : \|\boldsymbol{x}\| < \|\boldsymbol{x} - \boldsymbol{p}\|, \forall \boldsymbol{p} \in \mathcal{L}/\{\mathbf{0}\}\}. \quad [2]$$

For example, when $\boldsymbol{G} = \Delta \cdot \boldsymbol{I}_L$ for some $\Delta > 0$, then $\mathcal{L}$ is the square lattice, for which $\mathcal{P}_0$ is the set of vectors $\boldsymbol{x} \in \mathbb{R}^L$ whose ℓ_∞ norm is not larger than $\frac{\Delta}{2}$. For this setting, $Q_{\mathcal{L}}(\cdot)$ implements entry-wise scalar uniform quantization with spacing Δ.

Define the model update of the kth device at iteration $t+1$ as $\boldsymbol{h}_{t+1}^{(k)} \triangleq \boldsymbol{o}_{k,t+1}^{\tau} - \boldsymbol{b}_t$. Using the above definitions in lattice quantization, the encoding and decoding procedures, which are based on subtractive dithered lattice quantization, consist of the following steps:

Encoder: The encoding function includes the following steps:

E1 **Normalize and partition:** The kth device scales $\boldsymbol{h}_{t+1}^{(k)}$ by $\zeta\|\boldsymbol{h}_{t+1}^{(k)}\|$ for some $\zeta > 0$, and divides the result into M distinct $L \times 1$ vectors, denoted $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$. The scalar quantity $\zeta\|\boldsymbol{h}_{t+1}^{(k)}\|$ is quantized separately from $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$ using some fine-resolution quantizer.

E2 **Dithering:** The encoder utilizes the source of common randomness, e.g., a shared seed, to generate the set of $L \times 1$ dither vectors $\{\boldsymbol{z}_i^{(k)}\}_{i=1}^M$, which are randomized in an i.i.d. fashion, independently of $\boldsymbol{h}_{t+1}^{(k)}$, from a uniform distribution over $\mathcal{P}_0$.

E3 **Quantization:** The vectors $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$ are discretized by adding the dither vectors and applying lattice quantization, i.e., by computing $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)})\}$.

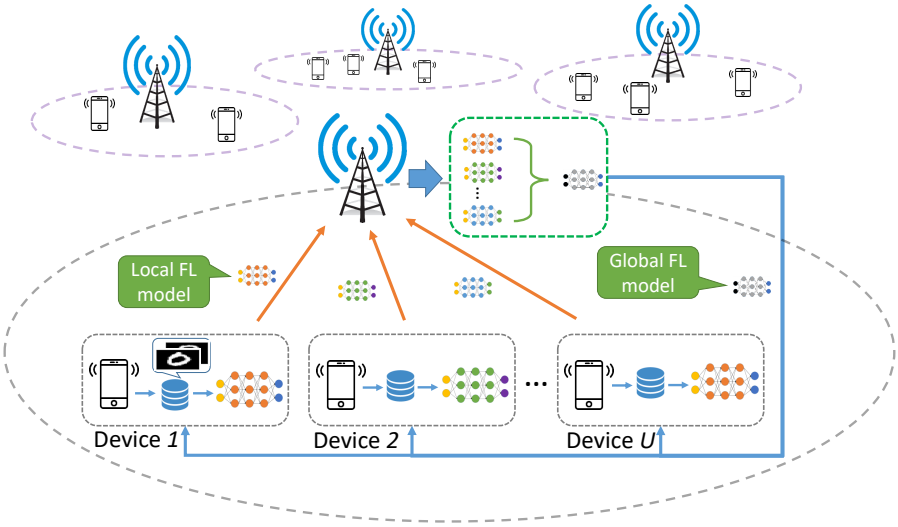

Fig. S1. The architecture of an FL algorithm that is implemented over a wireless network that consists of multiple devices and one base station.

E4 **Entropy coding:** The discrete values $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})\}$ are encoded into a digital codeword $u_{t+1}^{(k)}$ in a lossless manner.

In order to utilize entropy coding in step *E4*, the discretized $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})\}$ must take values on a *finite set*. This is achieved by the normalization in Step *E1*, which guarantees that $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$ all reside inside the L-dimensional ball with radius ζ^{-1}, in which the number of lattice points is not larger than $\frac{\pi^{L/2}}{\zeta^L\Gamma(1+L/2)\det(\boldsymbol{G})}$ (1, Ch. 2), where $\Gamma(\cdot)$ is the Gamma function. The overhead in accurately quantizing the single scalar quantity $\zeta\|\boldsymbol{h}^{(k)}\|$ is typically negligible compared to the number of bits required to convey the set of vectors $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$, hardly affecting the overall quantization rate.

Decoder: The decoding mapping implements the following:

D1 **Entropy decoding:** The CC first decodes each digital codeword $u_{t+1}^{(k)}$ into the discrete value $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})\}$. Since the encoding is carried out using a lossless source code, the discrete values are recovered without any errors.

D2 **Dither subtraction:** Using the source of common randomness, the CC generates the dither vectors $\{\boldsymbol{z}_i^{(k)}\}$, which can be carried out rapidly and at low complexity using random number generators as the dither vectors obey a uniform distribution. The CC then subtracts the corresponding vector from each lattice point, i.e., compute $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})-\boldsymbol{z}_i^{(k)}\}$. An illustration of the subtractive dithered lattice quantization procedure is illustrated in Fig. S2.

D3 **Collecting and scaling:** The values $\{Q_{\mathcal{L}}(\boldsymbol{h}_i^{(k)}+\boldsymbol{z}_i^{(k)})-\boldsymbol{z}_i^{(k)}\}$ are collected into an $m\times 1$ vector $\hat{\boldsymbol{h}}_{t+1}^{(k)}$ using the inverse operation of the partitioning and normalization in Step *E1*.

D4 **Model recovery:** The recovered matrices are combined into an updated model. Namely,

$$\boldsymbol{b}_{t+1}=\frac{1}{\sum_k N_k}\sum_k N_k\hat{\boldsymbol{h}}_{t+1}^{(k)}+\boldsymbol{b}_t. \tag{3}$$

The usage of subtractive dithered lattice quantization in Steps *E2*-*E3* and *D2* allow obtaining a digital representation which is relatively close to the true quantity, as illustrated in Fig. S2, without relying on prior knowledge of its distribution. The joint decoding aspect of the proposed scheme is introduced in the final model recovery Step *D4*. The remaining encoding-decoding procedure, i.e., Steps *E1*-*D3* is carried out independently for each device.

3. Optimization of Resource Block Allocation

Next, we explain how to solve the RB allocation optimization problem. We first define the data rate of device i transmitting its compressed local FL model to the CC as $c_i^{\mathrm{U}}\left(\boldsymbol{\chi}_{i,t}\right)=\sum_{r=1}^{R}\chi_{i,t}^r B\log_2\left(1+\frac{Ph_i}{I_r+BN_0}\right)$ where $\chi_{i,t}^r\in\{0,1\}$ is the RB allocation index with $\chi_{i,t}^r=1$ implying that RB r is allocated to device i at iteration t, otherwise, we have $\chi_{i,t}^r=0$, and $\boldsymbol{\chi}_{i,t}=\left[\chi_{i,t}^1,\ldots,\chi_{i,t}^R\right]$. Then, the transmission delay at each FL iteration is given by $\max_{i\in\mathcal{U}_{\boldsymbol{p}_t}}\frac{Z}{c_i^{\mathrm{U}}(\boldsymbol{\chi}_{i,t})}$, where $\mathcal{U}_{\boldsymbol{p}_t}$ is the subset of devices that transmit their compressed local FL model parameters to the CC at iteration t and Z is the data size of FL parameters. Given these definitions, we rewrite the optimization problem as follows:

$$\min_{\boldsymbol{\chi}_t}\max_{i\in\mathcal{U}_{\boldsymbol{p}_t}}\frac{Z}{c_i^{\mathrm{U}}\left(\boldsymbol{\chi}_{i,t}\right)} \tag{4}$$

$$\text{s. t. } \chi_{i,t}^r\in\{0,1\},\quad \forall i\in\mathcal{U}_{\boldsymbol{p}_t},r\in\mathcal{R}, \tag{4a}$$

$$\sum_{i\in\mathcal{U}_{\boldsymbol{p}_t}}\chi_{i,t}^r=1,\forall r\in\mathcal{R}, \tag{4b}$$

$$\sum_{r=1}^{R}\chi_{i,t}^r=1,\forall i\in\mathcal{U}_{\boldsymbol{p}_t}, \tag{4c}$$

where $\mathcal{R}$ is the set of RBs that can be allocated to the devices and $\boldsymbol{\chi}_t=\left[\boldsymbol{\chi}_{1,t},\ldots,\boldsymbol{\chi}_{|\mathcal{U}_{\boldsymbol{p}_t}|,t}\right]$ with $|\mathcal{U}_{\boldsymbol{p}_t}|$ being the number of devices in $\mathcal{U}_{\boldsymbol{p}_t}$. We assume that a variable q exists such that $\frac{Z}{c_i^{\mathrm{U}}(\boldsymbol{\chi}_{i,t})}\le q,\forall i\in\mathcal{U}_{\boldsymbol{p}_t}$. Hence, we can rewrite the problem in Eq. (4) as

$$\min_{\boldsymbol{\chi}_t} q \tag{5}$$

$$\text{s. t. } \chi_{i,t}^r\in\{0,1\},\quad \forall i\in\mathcal{U}_{\boldsymbol{p}_t},r\in\mathcal{R}, \tag{5a}$$

$$\sum_{i\in\mathcal{U}_{\boldsymbol{p}_t}}\chi_{i,t}^r=1,\forall r\in\mathcal{R}, \tag{5b}$$

$$\sum_{r=1}^{R}\chi_{i,t}^r=1,\forall i\in\mathcal{U}_{\boldsymbol{p}_t}, \tag{5c}$$

$$q\ge\frac{Z}{c_i^{\mathrm{U}}(\boldsymbol{\chi}_{i,t})}. \tag{5d}$$

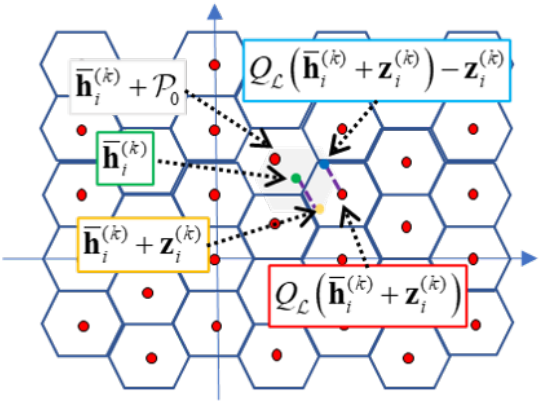

Fig. S2. Subtractive dithered lattice quantization illustration.

Here, $q \geq \frac{Z}{c_i^{\mathrm{U}}(\boldsymbol{\chi}_{i,t})}$ is nonconvex. Therefore, we first need to transform it to a convex equation. We assume that the data rate of device i using RB r to transmit local FL model parameters is $c_{ir,t}^{\mathrm{U}} = B\log_2\left(1+\frac{Ph_i}{I_r+BN_0}\right)$ where I_r is the interference over RB r. Then, we have $c_i^{\mathrm{U}}\left(\boldsymbol{\chi}_{i,t}\right) = \sum_{r=1}^{R} \chi_{i,t}^r c_{ir,t}^{\mathrm{U}}$. In consequence, the problem in Eq. (5) can be expressed by

$$\min_{\boldsymbol{\chi}_t} q \tag{6}$$

$$\text{s. t. } \chi_{i,t}^r \in \{0,1\}, \quad \forall i \in \mathcal{U}_{\boldsymbol{p}_t}, r \in \mathcal{R}, \tag{6a}$$

$$\sum_{i \in \mathcal{U}_{\boldsymbol{p}_t}} \chi_{i,t}^r = 1, \forall r \in \mathcal{R}, \tag{6b}$$

$$\sum_{r=1}^{R} \chi_{i,t}^r = 1, \forall i \in \mathcal{U}_{\boldsymbol{p}_t}, \tag{6c}$$

$$\sum_{r=1}^{R} \chi_{i,t}^r c_{ir,t}^{\mathrm{U}} \geq \frac{Z}{q}. \tag{6d}$$

The problem in Eq. (6) is an integer linear programming problem, which can be solved by using Matlab toolbox. In this paper, we use Matlab intlinprog function to solve the problem in Eq. (6).

4. Convergence Analysis

Here, we can analyze the convergence, proving Theorem 1. Our proof follows a similar outline to that used in (2, 3), with the introduction of additional arguments for handling the quantization constraints. The unique characteristics of the quantization error which arise from the dithered strategy allow us to rigorously incorporate its contribution into the overall flow of the proof. As a preliminary step, we first recall the assumptions in light of which our analysis is carried out:

AS1 The expected squared ℓ_2 norm of the random vector $\nabla f_k^i(\boldsymbol{o})$, representing the stochastic gradient evaluated at $\boldsymbol{b}$, is bounded by some $\xi^2 > 0$ for all $\boldsymbol{o} \in \mathbb{R}^m$.

AS2 The local objective functions $\{f_k(\cdot)\}$ are all ρ_s-smooth, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^m$ it holds that

$$f_k(\boldsymbol{v}_1) - f_k(\boldsymbol{v}_2) \leq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla f_k(\boldsymbol{v}_2) + \frac{1}{2}\rho_s \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2.$$

AS3 The local objective functions $\{f_k(\cdot)\}$ are all ρ_c-strongly convex, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^m$ it holds that

$$f_k(\boldsymbol{v}_1) - f_k(\boldsymbol{v}_2) \geq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla f_k(\boldsymbol{v}_2) + \frac{1}{2}\rho_c \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2.$$

AS4 The probabilistic device selection method selects the set $\mathcal{U}_{\boldsymbol{p}_t}$ of devices, such that $|\mathcal{U}_{\boldsymbol{p}_t}| = R$ and $\mathcal{U}_{\boldsymbol{p}_t}$ is uniformly distributed over all R-sized subsets of $\mathcal{U}$.

A. Recursive Bound on Weights Error. From (4), it follows that the effect of substractive dithered quantization can be modeled as additive noise, independent of the quantized value, whose distribution depends only on the properties of the lattice. In particular, it holds that the distortion induced in quantizing the model update $\boldsymbol{h}_t^{(k)}$, denoted $\boldsymbol{\epsilon}_t^{(k)}$, is an $m \times 1$ zero-mean additive noise vector *independent of* $\boldsymbol{h}_{t+1}^{(k)}$. Consequently, by defining the sequence $\boldsymbol{e}_t^{(k)}$ such that $\boldsymbol{e}_t^{(k)} = \boldsymbol{\epsilon}_t^{(k)}$ if t is an integer multiple of τ and $\boldsymbol{e}_t^{(k)} = \boldsymbol{0}$ otherwise, the instantaneous weights at the kth device, defined as $\tilde{\boldsymbol{o}}_t^{(k)} \triangleq \boldsymbol{o}_{k,\lfloor t/\tau \rfloor \tau}^{t-\lfloor t/\tau \rfloor \tau}$, can be written as

$$\tilde{\boldsymbol{o}}_{t+1}^{(k)} = \begin{cases} \tilde{\boldsymbol{o}}_t^{(k)} - \tilde{\lambda}_t \nabla f_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{o}}_t^{(k)}\right) + \boldsymbol{e}_{t+1}^{(k)} & t+1 \notin \mathcal{T}_\tau, \\ \sum_{k'=1}^{U} \alpha_{k'} \left(\tilde{\boldsymbol{o}}_t^{(k')} - \tilde{\lambda}_t \nabla f_k^{i_t^{(k')}}\left(\tilde{\boldsymbol{o}}_t^{(k')}\right) + \boldsymbol{e}_{t+1}^{(k')} \right) & t+1 \in \mathcal{T}_\tau, \end{cases} \tag{7}$$

where $\tilde{\lambda}_t \triangleq \lambda_{\lfloor t/\tau \rfloor \tau}^{t-\lfloor t/\tau \rfloor \tau}$, $\mathcal{T}_\tau$ is the set of integer multiples of τ, and f_k^i is the objective evaluated at the ith sample of the kth device.

The equivalent model update representation in Eq. (7) allows us to model the effect of subtractive dithered quantization on the overall FL procedure as additional noise corrupting the computation of the stochastic gradients. Building upon this representation, we now follow the strategy proposed in (2) and adapted to heterogeneous data in (3). This is achieved by defining a virtual sequence $\{\boldsymbol{v}_t\}$ from $\{\tilde{\boldsymbol{o}}_t^{(k)}\}$ which can be shown to behave almost like mini-batch SGD with batch size τ, while being within a bounded distance of the FL model weights $\{\tilde{\boldsymbol{o}}_t^{(k)}\}$, by properly setting the step size $\tilde{\lambda}_t$. In particular, we define the virtual sequence $\{\boldsymbol{v}_t\}$ via

$$\bar{\boldsymbol{v}}_t \triangleq \begin{cases} \sum_{k=1}^{U} \alpha_k \tilde{\boldsymbol{o}}_t^{(k)} & t \notin \mathcal{T}_\tau, \\ \sum_{k \in \mathcal{U}_{\boldsymbol{p}_{t/\tau}}} \alpha_k \tilde{\boldsymbol{o}}_t^{(k)} & t \in \mathcal{T}_\tau, \end{cases} \tag{8}$$

which coincides with $\tilde{\boldsymbol{o}}_t^{(k)}$ when t is an integer multiple of τ. Also, let $\boldsymbol{v}_t \triangleq \sum_{k=1}^{U} \alpha_k \tilde{\boldsymbol{o}}_t^{(k)}$ be the virtual sequence representing the averaged model over all devices (both participating and non-participating) at each time instance. Further define the averaged noisy stochastic gradients and the averaged full gradients as

$$\tilde{\boldsymbol{g}}_t \triangleq \sum_{k=1}^{U} \alpha_k \left(\nabla f_k^{i_t^{(k)}} \left(\tilde{\boldsymbol{o}}_t^{(k)} \right) - \frac{1}{\tilde{\lambda}_t} \boldsymbol{e}_{t+1}^{(k)} \right), \tag{9a}$$

$$\boldsymbol{g}_t \triangleq \sum_{k=1}^{U} \alpha_k \nabla f_k \left(\tilde{\boldsymbol{o}}_t^{(k)} \right), \tag{9b}$$

respectively. Note that since the quantization error is zero-mean and each mini-batch consists of a single sample, whose indexes $\{i_t^{(k)}\}$ are independent and uniformly distributed, it holds that $\mathbb{E}\{\tilde{\boldsymbol{g}}_t\} = \boldsymbol{g}_t$. Additionally, the virtual sequence satisfies $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \tilde{\lambda}_t \tilde{\boldsymbol{g}}_t$. We use the following lemmas, proved in (3, Appendix B.4).

Lemma 1 *Under assumption AS4, $\bar{\boldsymbol{v}}_t$ is an unbiased estimation of $\boldsymbol{v}_t$, i.e. $\mathbb{E}_{\mathcal{U}_{\boldsymbol{p}_{t/\tau}}} \{\bar{\boldsymbol{v}}_t\} = \boldsymbol{v}_t$.*

Lemma 2 *The expected difference between $\boldsymbol{v}_t$ and $\bar{\boldsymbol{v}}_t$ is bounded by*

$$\mathbb{E}_{\mathcal{U}_{\boldsymbol{p}_{t/\tau}}} \left\{ \|\bar{\boldsymbol{v}}_t - \boldsymbol{v}_t\|^2 \right\} \leq \frac{4(U-R)}{(U-1)R} \eta_t^2 \tau^2 \xi^2. \tag{10}$$

We next use these lemmas to bound the distance between the FL model parameters and the optimal one, as

$$\begin{aligned} \|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{b}^*\|^2 &= \|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{v}_{t+1} + \boldsymbol{v}_{t+1} - \boldsymbol{b}^*\|^2 \\ &= \|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{v}_{t+1}\|^2 + \|\boldsymbol{v}_{t+1} - \boldsymbol{b}^*\|^2 + \underbrace{2 \left\langle \bar{\boldsymbol{v}}_{t+1} - \boldsymbol{v}_{t+1}, \boldsymbol{v}_{t+1} - \boldsymbol{b}^* \right\rangle}_{A}. \end{aligned} \tag{11}$$

The term $E_{\mathcal{U}_{\boldsymbol{p}_{t/\tau}}} \{A\} = 0$ since $\bar{\boldsymbol{v}}_t$ is unbiased by Lemma 1. Further, using Lemma 2, it follows from Eq. (11) that

$$\mathbb{E}\{\|\bar{\boldsymbol{v}}_{t+1} - \boldsymbol{b}^*\|^2\} \leq \mathbb{E} \left\{ \|\boldsymbol{v}_{t+1} - \boldsymbol{b}^*\|^2 \right\} + \frac{4(U-R)}{(U-1)R} \eta_t^2 \tau^2 \xi^2. \tag{12}$$

The resulting term which has to be bounded in Eq. (12) is thus equivalent to that used in (3), and as a result, by assumptions *AS2*-*AS3*, it follows from (3, Lemma 1) that if $\tilde{\lambda}_t \leq \frac{1}{4\rho_s}$ then

$$\mathbb{E} \left\{ \|\boldsymbol{v}_{t+1} - \boldsymbol{b}^*\|^2 \right\} \leq (1 - \tilde{\lambda}_t \rho_c) \mathbb{E} \left\{ \|\boldsymbol{v}_t - \boldsymbol{b}^*\|^2 \right\} + 6\rho_s \tilde{\lambda}_t^2 \psi + \tilde{\lambda}_t^2 \mathbb{E} \left\{ \|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^2 \right\} + 2\mathbb{E} \left\{ \sum_{k=1}^{U} \alpha_k \left\| \boldsymbol{v}_t - \tilde{\boldsymbol{o}}_t^{(k)} \right\|^2 \right\}. \tag{13}$$

Eq. (13) bounds the expected distance between the virtual sequence $\{\boldsymbol{v}_t\}$ and the optimal weights $\boldsymbol{b}^*$ in a recursive manner. We further bound the summands in Eq. (13), using the following lemmas:

Lemma 3 *If the step size $\tilde{\lambda}_t$ is non-increasing and satisfies $\tilde{\lambda}_t \leq 2\tilde{\lambda}_{t+\tau}$ for each $t \geq 0$, then, when assumption AS1 is satisfied, it holds that*

$$\tilde{\lambda}_t^2 \mathbb{E} \left\{ \|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^2 \right\} \leq \left(1 + 4M\zeta^2 \bar{\sigma}_{\mathcal{L}}^2 \tau^2 \right) \tilde{\lambda}_t^2 \xi^2 \sum_{k=1}^{U} \alpha_k^2. \tag{14}$$

Lemma 4 *If the step size $\tilde{\lambda}_t$ is non-increasing and satisfies $\tilde{\lambda}_t \leq 2\tilde{\lambda}_{t+\tau}$ for each $t \geq 0$, then, when assumption AS1 is satisfied, it holds that*

$$\mathbb{E} \left\{ \sum_{k=1}^{U} \alpha_k \left\| \boldsymbol{v}_t - \tilde{\boldsymbol{o}}_t^{(k)} \right\|^2 \right\} \leq 4(\tau - 1)^2 \tilde{\lambda}_t^2 \xi^2. \tag{15}$$

Next, we define $\delta_t \triangleq \mathbb{E} \left\{ \|\bar{\boldsymbol{v}}_t - \boldsymbol{b}^*\|^2 \right\}$. When $t \in \mathcal{T}_\tau$, the term δ_t represents the ℓ_2 norm of the error in the weights of the global model. Using Lemmas 3-4, while substituting Eq. (15) and Eq. (14) into Eq. (13) and Eq. (12), we obtain the following recursive relationship on the weights error:

$$\delta_{t+1} \leq (1 - \tilde{\lambda}_t \rho_c) \delta_t + \tilde{\lambda}_t^2 \varpi, \tag{16}$$

where

$$\varpi \triangleq \left(1 + 4M\zeta^2 \bar{\sigma}_{\mathcal{L}}^2 \tau^2 \right) \xi^2 \sum_{k=1}^{U} \alpha_k^2 + 6\rho_s \psi + 8(\tau - 1)^2 \xi^2 + \frac{4(U-R)}{(U-1)R} \tau^2 \xi^2.$$

The relationship in Eq. (16) is used in the sequel to prove the FL convergence bound stated in Theorem 1.

B. FL Convergence Bound. Here, we prove Theorem 1 based on the recursive relationship in Eq. (16). This is achieved by properly setting the step-size and the FL systems parameters in Eq. (16) to bound $\delta_t = \mathbb{E}\left\{\|\bar{\boldsymbol{v}}_t - \boldsymbol{b}^*\|^2\right\}$, and combining the resulting bound with the strong convexity of the objective *AS3* to prove the theorem.

In particular, we set the step size $\tilde{\lambda}_t$ to take the form $\tilde{\lambda}_t = \frac{\beta}{t+\gamma}$ for some $\beta > 0$ and $\gamma \geq \max\left(4\rho_s\beta, \tau\right)$, for which $\tilde{\lambda}_t \leq \frac{1}{4\rho_s}$ and $\tilde{\lambda}_t \leq 2\tilde{\lambda}_{t+\tau}$, implying that Eq. (13) and Eq. (15) hold.

Under such settings, we show that there exists a finite ν such that $\delta_t \leq \frac{\nu}{t+\gamma}$ for all integer $l \geq 0$. We prove this by induction, noting that setting $\nu \geq \gamma\delta_0$ guarantees that it holds for $t = 0$. Consequently, we next show that if $\delta_t \leq \frac{\nu}{t+\gamma}$, then $\delta_{t+1} \leq \frac{\nu}{t+1+\gamma}$. It follows from Eq. (16) that

$$\begin{aligned}\delta_{t+1} &\leq \left(1 - \frac{\beta}{t+\gamma}\rho_c\right)\frac{\nu}{t+\gamma} + \left(\frac{\beta}{t+\gamma}\right)^2 b \\ &= \frac{1}{t+\tau}\left(\left(1 - \frac{\beta}{t+\gamma}\rho_c\right)\nu + \frac{\beta^2}{t+\gamma}b\right).\end{aligned} \tag{17}$$

Consequently, $\delta_{t+1} \leq \frac{\nu}{t+1+\gamma}$ holds when

$$\frac{1}{t+\tau}\left(\left(1 - \frac{\beta}{t+\gamma}\rho_c\right)\nu + \frac{\beta^2}{t+\gamma}\varpi\right) \leq \frac{\nu}{t+1+\gamma},$$

or, equivalently,

$$\left(1 - \frac{\beta}{t+\gamma}\rho_c\right)\nu + \frac{\beta^2}{t+\gamma}\varpi \leq \frac{t+\gamma}{t+1+\gamma}\nu. \tag{18}$$

By setting $\nu \geq \frac{1+\beta^2\varpi}{\beta\rho_c}$, the left hand side of Eq. (18) satisfies

$$\begin{aligned}\left(1-\frac{\beta}{t+\gamma}\rho_c\right)\nu+\frac{\beta^2}{t+\gamma}\varpi &= \frac{t-1+\gamma}{t+\gamma}\nu+\left(\frac{1-\beta\rho_c}{t+\gamma}\nu+\frac{\beta^2}{t+\gamma}\varpi\right) \\ &= \frac{t-1+\gamma}{t+\gamma}\nu+\frac{1}{t+\gamma}\left(\left(1-\beta\rho_c\right)\nu+\beta^2\varpi\right) \\ &\overset{(a)}{\leq} \frac{t-1+\gamma}{t+\gamma}\nu,\end{aligned} \tag{19}$$

where (a) holds since $\nu \geq \frac{1+\beta^2\varpi}{\beta\rho_c}$. As the right hand side of Eq. (19) is not larger than that of Eq. (18), it follows that Eq. (18) holds for the current setting, which in turn proves that $\delta_{t+1} \leq \frac{\nu}{t+1+\gamma}$. Finally, the smoothness of the objective *AS2* implies that

$$\begin{aligned}\mathbb{E}\{f(\boldsymbol{b}_t)\} - f(\boldsymbol{b}^*) &= \mathbb{E}\{f(\bar{\boldsymbol{v}}_{t\tau})\} - f(\boldsymbol{b}^*) \\ &\leq \frac{\rho_s}{2}\delta_{t\tau} \leq \frac{\rho_s\nu}{2(t\tau+\gamma)},\end{aligned} \tag{20}$$

which, in light of the above setting, holds for $\nu \geq \max\left(\frac{1+\beta^2\varpi}{\beta\rho_c}, \gamma\delta_0\right)$, $\gamma \geq \max(\tau, 4\beta\rho_s)$, and $\beta > 0$. In particular, setting $\beta = \frac{\tau}{\rho_c}$ results in $\gamma \geq \tau\max(1, 4\rho_s/\rho_c)$ and $\nu \geq \max\left(\frac{\rho_c^2+\tau^2\varpi}{\tau\rho_c}, \gamma\delta_0\right)$, which, when substituted into Eq. (20), proves the theorem. ■

B.1. Deferred Proofs. Here we detail the proofs of the intermediate lemmas used for obtaining the recursion Eq. (16).

Proof of Lemma 3 To prove Eq. (14), we note that since the quantization noise and the stochastic gradients are mutually independent, it follows from the definition of the gradient vectors in Eq. (9) that

$$\begin{aligned}\tilde{\lambda}_t^2\mathbb{E}\left\{\|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^2\right\} &= \sum_{k=1}^{U}\alpha_k^2\mathbb{E}\left\{\left\|\boldsymbol{e}_{t+1}^{(k)}\right\|^2\right\} + \tilde{\lambda}_t^2\sum_{k=1}^{U}\alpha_k^2\mathbb{E}\left\{\left\|\nabla f_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{o}}_t^{(k)}\right) - \nabla f_k\left(\tilde{\boldsymbol{o}}_t^{(k)}\right)\right\|^2\right\} \\ &\overset{(a)}{\leq} \sum_{k=1}^{U}\alpha_k^2\mathbb{E}\left\{\left\|\boldsymbol{e}_{t+1}^{(k)}\right\|^2\right\} + \tilde{\lambda}_t^2\xi^2\sum_{k=1}^{U}\alpha_k^2,\end{aligned} \tag{21}$$

where (a) holds since the uniform distribution of the random index i_k implies that the expected value of the stochastic gradient is the full gradient, i.e., $\mathbb{E}\left\{\nabla f_k^{i_t^{(k)}}\left(\boldsymbol{b}\right)\right\} = \nabla f_k(\boldsymbol{b})$, and consequently, $\mathbb{E}\left\{\left\|\nabla f_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{o}}_t^{(k)}\right) - \nabla f_k\left(\tilde{\boldsymbol{o}}_t^{(k)}\right)\right\|^2\right\} \leq \mathbb{E}\left\{\left\|\nabla f_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{o}}_t^{(k)}\right)\right\|^2\right\} \leq \xi^2$

by assumption *AS1*. Furthermore, the definition of $\boldsymbol{e}_{t+1}^{(k)}$ implies that $\mathbb{E}\{\|\boldsymbol{e}_{t+1}^{(k)}\|^2\}=0$ for $t+1\notin\mathcal{T}$, while for $t+1\in\mathcal{T}$ it holds that $\mathbb{E}\{\|\boldsymbol{e}_{t+1}^{(k)}\|^2\}=\mathbb{E}\{\|\boldsymbol{\epsilon}_{t+1}^{(k)}\|^2\}=M\sigma_{\mathcal{L}}^2$. Now, the quantization error satisfies

$$\begin{aligned}\mathbb{E}\{\|\boldsymbol{e}_{t+1}^{(k)}\|^2\}&\leq M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\mathbb{E}\left\{\left\|\sum_{t'=t+1-\tau}^{t+1}\tilde{\lambda}_{t'}\nabla f_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{o}}_{t'}^{(k)}\right)\right\|^2\right\}\\&\overset{(a)}{\leq} M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau\sum_{t'=t+1-\tau}^{t+1}\tilde{\lambda}_{t'}^2\mathbb{E}\left\{\left\|\nabla f_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{o}}_{t'}^{(k)}\right)\right\|^2\right\}\\&\overset{(b)}{\leq} M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau^2\tilde{\lambda}_{t+1-\tau}^2\xi^2\overset{(c)}{\leq}4M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau^2\tilde{\lambda}_t^2\xi^2,\end{aligned}\tag{22}$$

where in (a) we used the inequality $\|\sum_t'=t+1-\tau^{t+1}\boldsymbol{r}_t\|^2\leq\tau\sum_{t'=t+1-\tau}^{t+1}\|\boldsymbol{r}_t\|^2$, which holds for any multivariate sequence $\{\boldsymbol{r}_t\}$; (b) is obtained from assumption *AS1*; and (c) follows since $\tilde{\lambda}_{t+1-\tau}\leq2\tilde{\lambda}_{t+1}\leq2\tilde{\lambda}_t$. Substituting Eq. (22) into Eq. (21) proves the lemma. ■

Proof of Lemma 4 Note that for $t_0=\lfloor t/\tau\rfloor\tau$, which is an integer multiple of τ, it holds that $\boldsymbol{v}_{t_0}=\tilde{\boldsymbol{o}}_{t_0}^{(k)}$. Since Eq. (15) trivially holds for $t=t_0$, we henceforth focus on the case where $t>t_0$. We now write

$$\begin{aligned}\mathbb{E}\left\{\sum_{k=1}^U\alpha_k\left\|\tilde{\boldsymbol{o}}_t^{(k)}-\boldsymbol{v}_t\right\|^2\right\}&=\mathbb{E}\left\{\sum_{k=1}^U\alpha_k\left\|\tilde{\boldsymbol{o}}_t^{(k)}-\tilde{\boldsymbol{o}}_{t_0}^{(k)}-(\boldsymbol{v}_t-\boldsymbol{v}_{t_0})\right\|^2\right\}\\&\overset{(a)}{\leq}\mathbb{E}\left\{\sum_{k=1}^U\alpha_k\left\|\tilde{\boldsymbol{o}}_t^{(k)}-\tilde{\boldsymbol{o}}_{t_0}^{(k)}\right\|^2\right\}\\&=\sum_{k=1}^U\alpha_k\mathbb{E}\left\{\left\|\tilde{\boldsymbol{o}}_t^{(k)}-\tilde{\boldsymbol{o}}_{t_0}^{(k)}\right\|^2\right\},\end{aligned}\tag{23}$$

where in (a) we used the fact that for every set $\{\boldsymbol{r}^{(k)}\}$, one can define a random vector $\boldsymbol{r}$ such that $\Pr(\boldsymbol{r}=\boldsymbol{r}^{(k)})=\alpha_k$, and thus

$$\begin{aligned}\sum_{k=1}^U\alpha_k\left\|\boldsymbol{r}^{(k)}-\sum_{l=1}^U\alpha_l\boldsymbol{r}^{(l)}\right\|^2&=\mathbb{E}\{\|\boldsymbol{r}-\mathbb{E}\{\boldsymbol{r}\}\|^2\}\\&\leq\mathbb{E}\{\|\boldsymbol{r}\|^2\}=\sum_{k=1}^U\alpha_k\|\boldsymbol{r}^{(k)}\|^2.\end{aligned}$$

Next, we recall that $\boldsymbol{e}_{t'}=\boldsymbol{0}$ for each $t'=t_0+1,\ldots,t$. Consequently, similarly to the derivation in Eq. (22),

$$\begin{aligned}\mathbb{E}\left\{\left\|\tilde{\boldsymbol{o}}_t^{(k)}-\tilde{\boldsymbol{o}}_{t_0}^{(k)}\right\|^2\right\}&=\mathbb{E}\left\{\left\|\sum_{t'=t_0}^{t-1}\tilde{\lambda}_{t'}\nabla f_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{o}}_{t'}^{(k)}\right)\right\|^2\right\}\\&\overset{(a)}{\leq}(\tau-1)\sum_{t'=t_0}^{t-1}\tilde{\lambda}_{t'}^2\mathbb{E}\left\{\left\|\nabla f_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{o}}_{t'}^{(k)}\right)\right\|^2\right\}\\&\overset{(b)}{\leq}(\tau-1)^2\tilde{\lambda}_{t_0}^2\xi^2\overset{(c)}{\leq}4(\tau-1)^2\tilde{\lambda}_t^2\xi^2,\end{aligned}\tag{24}$$

where in (a) we used the inequality $\|\sum_{t'=t_0}^{t-1}\boldsymbol{r}_t\|^2\leq(t-1-t_0)\sum_{t'=t_0}^{t-1}\|\boldsymbol{r}_t\|^2\leq(\tau-1)\sum_{t'=t_0}^{t-1}\|\boldsymbol{r}_t\|^2$, which holds for any multivariate sequence $\{\boldsymbol{r}_t\}$; (b) is obtained from assumption *AS1*; and (c) follows since $\tilde{\lambda}_{t_0}\leq\tilde{\lambda}_{t-\tau}\leq2\tilde{\lambda}_t$. Substituting Eq. (24) into Eq. (23) proves the lemma. ■

5. Handwritten Digit Identification

Fig. S3 shows the architecture of the FL model used for handwritten digit identification. From Fig. S3, we can see that the FL model is a shallow fully-connected neural network that consists of 50 neurons. Since the size of one digit image is 28×28, the input size is 784. The purpose of training FL is to identify 10 class handwritten digits and hence, the output size is 10. The activation functions in the hidden and output layers are trasig and softmax functions.

6. Object Recognition in Images

Fig. S4 shows the architecture of the FL model used for object recognition in images. The FL model is a convolutional neural network (CNN) that consists of 16 layers. In this figure, Conv2D, MaxPooling, AvgPooling, and FullyConnected are short for 2D convolutional layer, maximum pooling layer, average pooling layer, and fully connected layer. The stride of each pooling or convolutional layer is 1 unless specified otherwise.

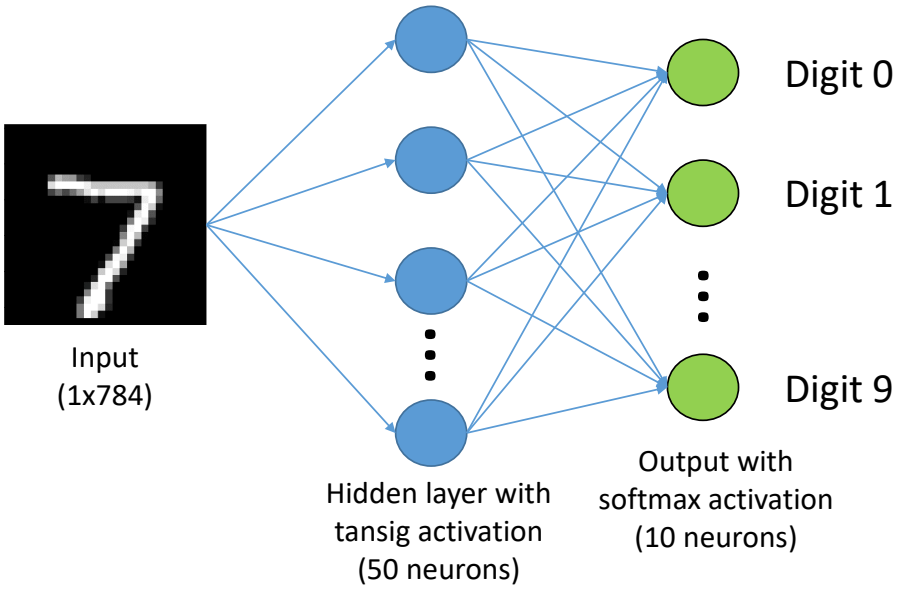

Fig. S3. The architecture of the FL model used for handwritten digit identification.

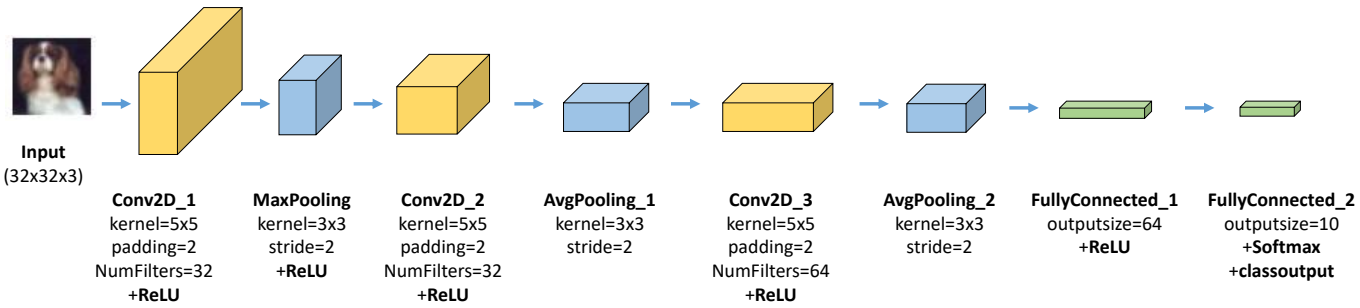

Fig. S4. The architecture of the FL model used for object recognition in images.

7. Finger Movement Detection

Fig. S5 shows the raw data collected from a hand close finger movement within five seconds. From Fig. S5, we can see that the raw data consists of 20000 signal samples. Hence, it is impossible to directly use the raw data as an input vector to train the FL model. A windowing method is used to split the raw data. We assume that the window size is 500 and the interval between two windows is 50. Therefore, 20000 signal samples can be divided into 391 windows. Fig. S6 shows the hand close finger movement data processed by the windowing method. From Fig. S6, we can see that the signal samples at each window are different which increases the complexity of training FL models. Fig. S7 shows the architecture of the FL model used for finger movement detection. The FL model is a CNN that consists of 19 layers. We consider the signal data in one window as an input vector of the CNN and hence, the size of a CNN input is 500×8. Since a CNN is used to detect 15 class finger movement, the size of a CNN output is 15×1. In Fig. S7, BatchNormalization is short for batch normalization layer. Meanwhile, the stride is a vector [2, 1] with 2 being the vertical stride and 1 being the horizontal stride.

References

1. JH Conway, NJA Sloane, *Sphere Packings, Lattices and Groups.* (Springer Science & Business Media) Vol. 290, (2013).
2. SU Stich, Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767* (2018).
3. X Li, K Huang, W Yang, S Wang, Z Zhang, On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* (2019).
4. R Zamir, M Feder, On lattice quantization noise. *IEEE Transactions on Inf. Theory* **42**, 1152–1159 (1996).

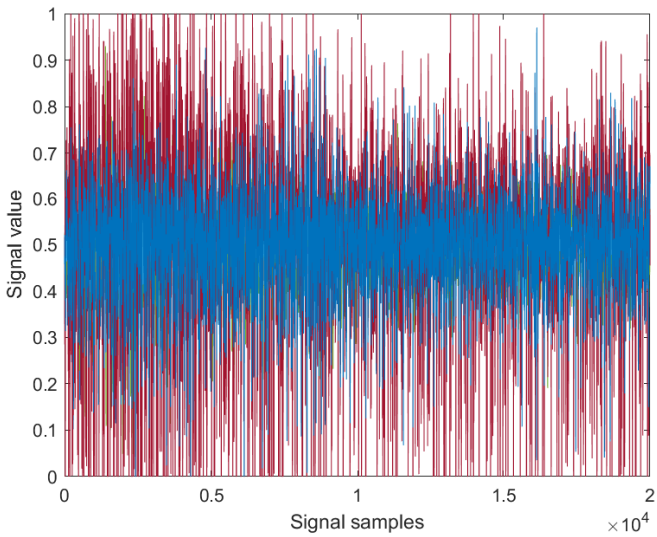

Fig. S5. Raw data collected from a close finger movement.

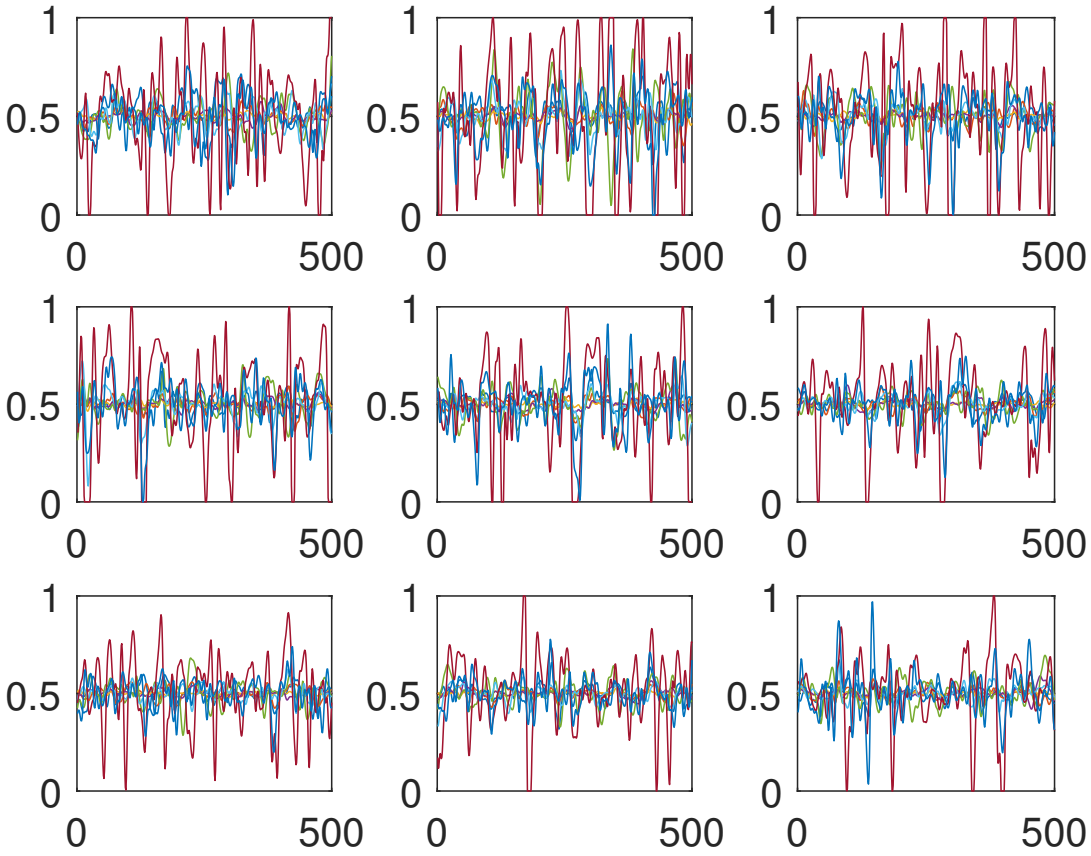

Fig. S6. The close finger movement data processed by the windowing method.

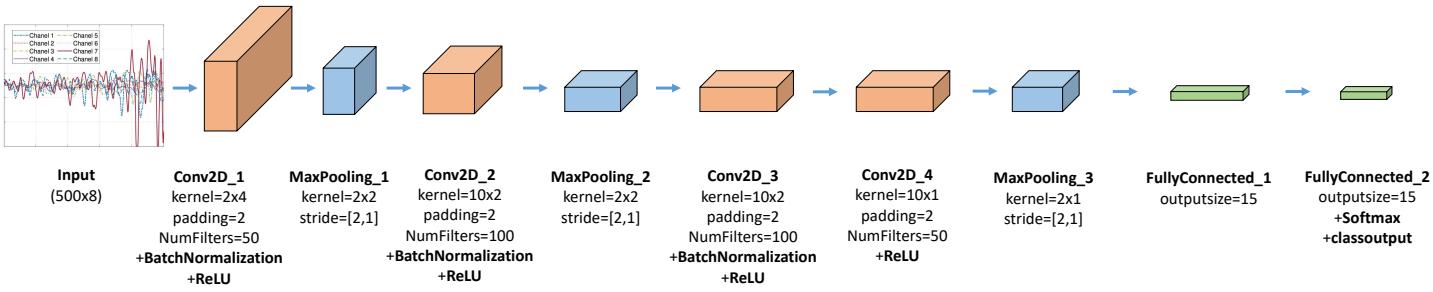

Fig. S7. The architecture of the FL model used for finger movement detection.