Data-Importance Aware User Scheduling for Communication-Efficient Edge Machine Learning

Dongzhu Liu, Guangxu Zhu, Jun Zhang, and Kaibin Huang

Abstract*—With the prevalence of intelligent mobile applications, edge learning is emerging as a promising technology for powering fast intelligence acquisition for edge devices from distributed data generated at the network edge. One critical task of edge learning is to efficiently utilize the limited radio resource to acquire data samples for model training at an edge server. In this paper, we develop a novel user scheduling algorithm for data acquisition in edge learning, called *(data) importance-aware scheduling*. A key feature of this scheduling algorithm is that it takes into account the informativeness of data samples, besides communication reliability. Specifically, the scheduling decision is based on a *data importance indicator* (DII), elegantly incorporating two "important" metrics from communication and learning perspectives, i.e., the *signal-to-noise ratio* (SNR) and *data uncertainty*. We first derive an explicit expression for this indicator targeting the classic classifier of *support vector machine* (SVM), where the uncertainty of a data sample is measured by its distance to the decision boundary. Then, the result is extended to *convolutional neural networks* (CNN) by replacing the distance based uncertainty measure with the entropy. As demonstrated via experiments using real datasets, the proposed importance-aware scheduling can exploit the two-fold multi-user diversity, namely the diversity in both the multiuser channels and the distributed data samples. This leads to faster model convergence than the conventional scheduling schemes that exploit only a single type of diversity.

***Index Terms*—Scheduling, Resource management, Image classification, Multiuser channels, Data acquisition**

I. Introduction

The proliferation of smart devices and the booming of *artificial intelligence* (AI) ushered in a new era of ambient intelligence. Materializing the vision motivates the deployment of machine learning algorithms at the network edge, named *edge learning* [1]–[5], to enable intelligent mobile applications. Edge learning aims at fast AI model training by exploiting computing resources at edge servers, and low-latency access to distributed data at edge devices. In return, downloading the trained models to the devices will equip them with human-like intelligence to cope with real-time inference and decision making. Edge learning sits at the intersection of two areas: wireless communications and machine learning. The emergence of the new area gives rise to many inter-disciplinary research opportunities that require joint designs interweaving the two said areas towards an ultimate goal of fast and efficient intelligence acquisition.

With rapidly growing data-processing speeds, the bottleneck of fast edge learning is more on the communication aspect. Specifically, wirelessly uploading high-dimensional data from a large number of edge devices can congest the air-interface due to the limited radio resource [6]. To overcome this bottleneck, it calls for innovations on highly efficient wireless data acquisition tailored for edge learning systems. The conventional wireless technologies focus on *rate maximization* or *Quality-of-Service (QoS)*, which implicitly assume that transmitted data bits are equally important. However, the assumption is improper for machine learning applications since some data samples are more effective than the others for improving a learning model [7]. This fact motivates a novel design principle of importance aware *radio resource management* (RRM) that the radio resource should be allocated to edge devices not only based on channel states but also accounting for the importance of their data for model training. In this work, we apply this principle to revamp user scheduling by exploiting multi-user diversity in both the channel and data domains for efficient wireless data acquisition.

A. Related Work and Motivation

1) Radio Resource Management for Edge Learning: The mission of conventional wireless communications is to reliably transmit data bits at a rate as high as possible, regardless of the data content and its usefulness. Therefore, directly applying such a communication-learning separation principle to edge learning will lead to inefficient transmission [8]. This has triggered a lot of research interests recently on redesigning communication techniques for edge learning [1], covering key topics such as RRM [9]–[15], multiple access [16]–[18], and signal encoding [19]–[21]. The new idea in RRM, the theme of this paper, is to allocate resources to edge devices for transmitting learning relevant data by considering the learning task. Prior work on this topic can be separated for two learning paradigms. The first paradigm is federated

D. Liu was with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. She is now with the Department of Engineering, King's College London, London, United Kingdom. (e-mail: dongzhu.liu@kcl.ac.uk).

G. Zhu was with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. He is now with the Shenzhen Research Institute of Big Data, Shenzhen 518000, China (e-mail: gxzhu@sribd.cn). The work of G. Zhu was supported by National Key R&D Program of China (No. 2018YFB1800800), Guangdong Province Key Area R&D Program (No. 2018B030338001), Leading Talents of Guangdong Province Program (No. 00201501), and Shenzhen Peacock Plan (No. KQTD2015033114415450)

J. Zhang is with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong (email: jun-eie.zhang@polyu.edu.hk). The work of J. Zhang was supported in part by a start-up fund of The Hong Kong Polytechnic University under Project ID P0013883.

K. Huang is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (email: haungkb@eee.hku.hk). The work of K. Huang was supported by Hong Kong Research Grants Council under Grants 17208319 and 17209917, Innovation and Technology Fund under Grant GHP/016/18GD, and Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515130003.

Part of this work has been presented in IEEE ICC workshop 2020.

edge learning that preserves privacy by avoiding direct data uploading. In this paradigm, a model is distributively trained at edge devices, and the purpose of uplink transmission is to upload local models which are aggregated at an edge server to cooperatively improve a global model [9], [10]. In this paradigm, a resource allocation method integrating computing and communication is proposed to improve the learning efficiency [11]. Specifically, the training batch size is adapted to the wireless channel condition for attaining higher learning accuracy without compromising the latency. A similar idea has also been investigated in the other paradigm, centralized edge learning, where edge devices directly upload data to the server for training the global model. In this paradigm, given the communication overhead, the offloading data size in each communication round is optimized to acquire sufficient data samples for reducing the learning bias, and avoids insufficient learning due to exceeding the computing capacity, thereby improving the learning performance [12]. The efficiency of wireless data acquisition can be further enhanced by differentiating the usefulness of training data samples [13], [14]. Specifically, a novel retransmission scheme is designed for adapting the reliability requirement of a received data sample according to its importance for model improvement [15]. By intelligently allocating the constrained transmission budget, the scheme allows more important data to be received with a guaranteed reliability compared to the conventional channel-aware design, thereby improving learning accuracy. This idea of importance-aware retransmission motivates us to propose the new principle of importance-aware user scheduling to explore the new dimension of multiuser diversity in both channels and data to improve the communication efficiency of edge learning.

2) Multiuser Diversity: Multiuser channel diversity is an intrinsic characteristic of wireless networks arising from independent fading in multiuser channels. To increase the network throughput, multiuser diversity can be exploited by scheduling the user with the *best channel* at any given time [22], [23]. The diversity gain tends to increase for channels with large dynamic ranges, e.g., rich scattering and fast fading, as well as the large number of users. On the other hand, the scheme targeting throughput maximization is biased towards users with favourable channels and unfair for others [24]. To address this issue, one solution is proportional fair scheduling where a scheduling metric being the radio between the instantaneous and average rates of each user is adopted to strike a balance between rate maximization and fairness [25]–[27]. In the existing work, data importance is assumed homogeneous. However, in the context of edge learning, data samples differ in their importance for learning, called *data diversity*. Then the distribution of data at multiple devices gives rise to a new type of multiuser diversity, namely, *multiuser data diversity*. In this work, we make the first attempt to exploit both types of multiuser diversity in scheduling so as to improve the communication efficiency of edge learning.

Data diversity is not new but a fundamental concept in the area of *active learning* [7]. It concerns a scenario where abundant unlabelled data are available and manual labelling is costly. Data diversity can be exploited by selecting the *most informative data samples* to be labeled (by querying an oracle), such that a model can be accurately trained using fewer labelled samples, thereby reducing the labelling cost. Generally, the informative data samples are those highly uncertain to be predicted under the current model. Their use in training can significantly improve the accuracy of the classifier model. There are several commonly adopted metrics for measuring data uncertainty including *entropy* [28], *expected model change* [29], and *expected error reduction* [30]. For active learning, all data are assumed to be located at a server and hence wireless transmission is irrelevant. Nevertheless, the data-uncertainty measures developed in the area are found in this work to be a useful tool for designing importance aware scheduling for wireless data acquisition in edge learning.

B. Contributions and Organization

In this work, we propose importance-aware scheduling for communication efficient edge learning. To this end, consider a centralized edge learning system where a classifier is trained at the edge server by utilizing the data distributed at multiple edge devices. To accelerate the model training, the edge server schedules a device for wireless data uploading under the criterion of maximum improvement on the classifier's accuracy. The proposed scheduling scheme exploits channel diversity and data diversity simultaneously, to ensure received data are both important and reliable in the presence of channel fading and noise. As a result, the model convergence is accelerated and channel use is reduced. To the authors' best knowledge, this work represents the first attempt on exploiting both the channel and data diversity to improve the communication efficiency of an edge learning system.

The main contributions of this work are summarized as follows.

- **Importance-aware scheduling for SVM:** We first consider the classic classifier model of *support vector machine* (SVM), and develop the basic principle of importance-aware scheduling. The core element of the scheme is a new scheduling metric, named *data importance indicator* (DII), that is proposed to be the expected uncertainty of a received data sample in the presence of channel fading and noise. For SVM, the DII is suitably defined as the expected negative distance from a received data sample to the decision boundary of the classifier. The theoretical contribution of the DII design lies in that its derived closed form elegantly combines the received *signal-to-noise ratio* (SNR) from the communication perspective and *data uncertainty* from the learning perspective in a simple addition form. This allows the DII to measure the effective importance of a received data sample for model training given fading and noise. Consequently, scheduling under the criterion of maximum DII, yielding the proposed importance-aware scheduling, is capable of exploiting both multiuser channel-and-data diversity to accelerate model convergence while effectively coping with channel hostility.
- **Extension to general classifiers:** The principle of importance-aware scheduling developed for SVM is extended to general classifier models. The generalization

essentially replaces the distance-based uncertainty measure in the previous DII design for SVM with a general measure. It can be specified as one of available measures (e.g., entropy or expected model change) depending on the design choice. For illustration, a case study for the modern *convolutional neural networks* (CNN) classifier is presented.

- **Practical issues in implementation:** Several practical implementation issues of the proposed scheme are discussed and addressed by suitable design modifications.
  - **Exploiting data-label information:** The design involves revising the DII as the expectation of model update, which is derived to combine the SNR and label dependent hinge loss (see e.g., [31]) in a product form. Based on the revised DII with label information, the scheduling scheme can achieve faster convergence rate than the previous design.
  - **Model compression:** The second issue is the high computational complexity of data uncertainty evaluation at edge devices. This can be addressed by using a compressed model that prunes the model parameters with small values.
  - **Data deficiency:** In practice, the scheduling may face the data deficiency due to limited available devices in the system. That may degrade the performance of the proposed scheme since it is highly dependent on the global data size (or higher data diversity). To cope with this issue, several practical solutions are discussed for increasing the data size by increasing the number of devices, increasing buffer sizes, updating the local buffer with a higher frequency, or utilizing user mobility.
- **Experiments:** We evaluate the performance of the proposed importance-aware scheduling via extensive experiments using real datasets. The results demonstrate that the proposed method is able to exploit the two types of multiuser diversity, and as a result achieves better learning performance than the two baseline schemes that exploit only a single type of diversity. Moreover, the performance can be further improved by increasing the data size using several proposed methods. By exploiting the label information, the importance-aware scheduling could attain a faster convergence rate. Last, the computational complexity can be reduced by implementing a compressed evaluation model without significantly compromising the learning performance.

The remainder of the paper is organized as follows. The communication and learning models are introduced in Section II. In Section III, the principle of importance-aware scheduling is proposed. Several practical issues and solutions are discussed in Section IV. Section V provides experimental results, followed by concluding remarks in Section VI.

II. Communication and Learning Models

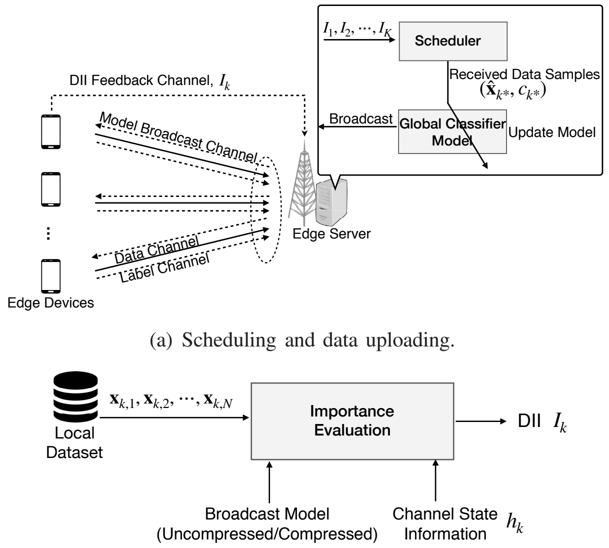
In this section, we first introduce the communication model, including multiuser scheduling and the data channel model. Then the learning models are introduced, followed by the data importance measures.

A. Communication Model

We consider an edge learning system in a single-cell wireless network as shown in Fig. 1, which comprises a single edge server and multiple edge devices, each equipped with a single antenna. A machine learning model is to be trained at the edge sever by utilizing the labeled data samples distributed over the K edge devices. The devices are coordinated by a scheduler to share the wireless channel in a time division manner, and they take turn to upload a data sample in each time slot. Each device is equipped a local buffer with the size of N samples. Note that both the buffer updating frequency and device mobility affect the learning performance which are discussed in Section IV-C. Denote the n-th data sample at the k-th device as $\mathbf{x}_{k,n} \in \mathbb{R}^p$, and its label $c_k \in \{1, 2, \cdots, C\}$ is acquired after the data sample is selected for transmission. Note that a label has a much smaller size than a data sample (e.g., a $0-9$ integer versus a vector of a million real coefficients), thus a low-rate noiseless channel for label transmission is assumed for simplicity.

1) Multiuser Scheduling: Time is divided into symbol durations, called *slots*. Slot synchronization among devices are assumed. Transmission of a data sample requires p slots, called a *symbol block*, which occupies a fixed duration of τ seconds. Thus, the slot during is $\tau_s = \frac{\tau}{p}$. Each sample coefficient is modulated into one symbol using linear analog modulation which will be discussed in sequel. The scheduled device is allocated a frequency flat fading channel for transmission and requires the passband bandwidth of $B = \frac{1+\alpha}{\tau_s}$, where $\alpha \in [0, 1]$ is the parameter of the raised-cosine pulse-shaping filter. The fading channel is assumed to be fixed within one sample duration and vary over multiple durations. In other words, a sample duration spans one channel coherence time and thus is referred to as one channel use in the sequel. At the beginning of each symbol block, the edge server broadcasts the global model for the devices to evaluate the importance of their data samples, measured by the DII and denoted as I_k at the k-th device, based on which, a device is selected for data uploading. The main purpose of this work is to design the DII. The model broadcast for data importance evaluation can be the current global model under training or a compressed one for low-complexity computation, as discussed in Section IV-B. Assuming a noiseless broadcast model and perfect *channel state information* (CSI), the data importance measure is evaluated at the devices and the results fed back to the server for scheduling. Upon receiving the DIIs, the server selects one of the devices for data-sample transmission.

2) Data Channel Model: Data channels are assumed to follow block-fading, where the channel coefficients remain static within a symbol block and are *independent and identically distributed* (i.i.d.) over different users. In a series of recent studies, analog modulation is seeing its revival to be a promising solution for multimedia transmission and found to outperform its digital counterpart in terms of edge learning performance [32], in the presence of Gaussian noise [33], in compression efficiency [34] and power consumption [35] for video transmission, and in alleviating the noise effect on video quality [36]. For these advantages as well as simplification



(a) Scheduling and data uploading.

(b) Data importance evaluation at an edge device.

Figure 1. An edge learning system with importance-aware scheduling.

of the design, analog modulation is adopted for transmitting training data which are typically images or videos. Specifically, during an arbitrary symbol block, the scheduled k-th device sends the data sample $\mathbf{x}$ using linear analog modulation, yielding the received signal given by

$$\mathbf{y} = \sqrt{P} h_k \mathbf{x} + \mathbf{z}_k, \tag{1}$$

where P is the transmit power, the Rayleigh fading coefficient h_k is a complex Gaussian random variable (r.v.), i.e., $h_k \sim \mathcal{CN}(0,1)$, and $\mathbf{z}_k$ is the *additive white Gaussian noise* (AWGN) vector with the entries following the i.i.d. $\mathcal{CN}(0,\sigma^2)$ distributions. The average power constraint is

$$\frac{P}{q}\mathsf{E}\left[\|\mathbf{x}\|^2\right] = \frac{P}{q}\sum_{i=1}^{q}\mathsf{E}\left[|X_i|^2\right] \le P_0 \tag{2}$$

where the expectation $\mathsf{E}[\cdot]$ is taken over the coefficients in the whole dataset. Since many data samples are transmitted and each has a large number of coefficients, the average transmission power approaches that at the left-hand side of (2). Therefore, P can be set as $P = \frac{qP_0}{\sum_{i=1}^{q}\mathsf{E}[|X_i|^2]}$ under the power constraint for sustaining the same SNR during the whole data transmission. Analog uncoded transmission is assumed here to allow fast data transmission [17] and for a higher energy efficiency (compared with the digital counterpart) as pointed out by [37]. We assume that perfect CSI is available at the edge server [1]. This allows the server to compute the instantaneous SNR and decode the received sample $\hat{\mathbf{x}}$ as follows:

$$\hat{\mathbf{x}} = \frac{1}{\sqrt{P}}\Re\left(\frac{{h_k}^*\mathbf{y}}{\|h_k\|^2}\right), \tag{2}$$

where $\mathbf{y}$ is given in (1). In (2), we extract the real part of the combined signal for further processing since the data for machine learning are real-valued in general (e.g., photos, voice clips or video clips). As a result, the receive SNR for sample $\hat{\mathbf{x}}$ is given as

$$\mathsf{SNR}_k = \frac{2P}{\sigma^2}|h_k|^2, \tag{3}$$

where the coefficient 2 at the right hand side arises from the fact that only the noise in the real dimension with variance $\frac{\sigma^2}{2}$ affects the received data. The SNR expression in (3) measures the reliability of a received data sample and serves as one of two performance metrics to be accounted for in scheduling as discussed in Section III.

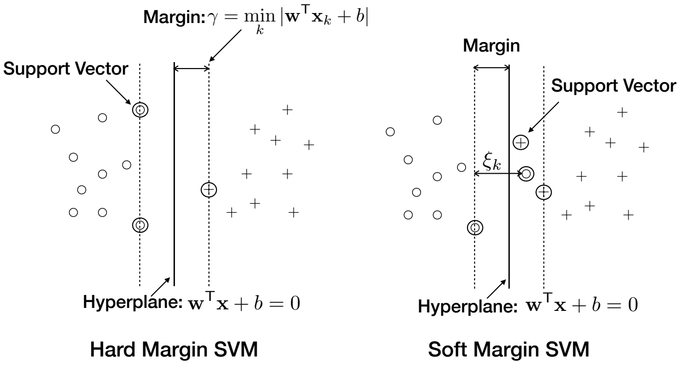

Margin: $\gamma = \min_k |\mathbf{w}^\mathsf{T}\mathbf{x}_k + b|$
Support Vector
Hyperplane: $\mathbf{w}^\mathsf{T}\mathbf{x} + b = 0$
Hard Margin SVM
Margin
Support Vector
ξ_k
Hyperplane: $\mathbf{w}^\mathsf{T}\mathbf{x} + b = 0$
Soft Margin SVM

Figure 2. Comparison of hard-margin SVM and soft-margin SVM for binary classification.

B. Learning Model

For the learning task, we consider supervised training of a classifier. Prior to wireless data acquisition, there is some initial data at the server. The data, denoted as $\mathcal{L}_0$, allow the construction of a coarse initial classifier, which is used for importance evaluation at the beginning. The classifier is refined progressively in the subsequent data acquisition (and training) process. In this work, we consider two widely used classifier models, i.e., the classic SVM classifier and the modern CNN classifier as introduced below.

1) SVM Model: As shown in Fig. 2, the original *hard margin SVM* is to seek the optimal hyperplane $\mathbf{w}^\mathsf{T}\mathbf{x} + b = 0$ as a decision boundary by maximizing its margin γ to data points, i.e., the minimum distance between the hyperplane to any data sample [38]. However, it works only for linearly separable datasets, which is hardly the case when the dataset is corrupted by channel noise in the current scenario. To enable the algorithm to cope with a potential outlier caused by noise, a variant of SVM called *soft-margin SVM* is adopted. Soft-margin SVM is widely used in practice to classify a noisy dataset that is not linearly separable by allowing misclassification, but with an additional penalty ξ_i for the non-separable sample $\mathbf{x}_i$. The comparison between hard margin SVM and soft margin SVM is graphically shown in Fig. 2. A convex formulation for the soft margin SVM problem is given by

$$\begin{aligned} \min_{\mathbf{w},b}\quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{s.t.}\quad & c_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \ge 1 - \xi_i, \\ & \xi_i \ge 0, \quad \forall i, \end{aligned} \tag{4}$$

where C is a parameter to control the tradeoff between maximizing the margin and minimizing the training error.

[1] The perfect CSI is assumed for simplifying the analysis. Nevertheless, it is straightforward to extend the current design to the imperfect CSI case. For instance, an additive noise can be introduced to account for the channel estimation error.

Small C tends to emphasize the margin maximization by allowing certain level of misclassification in the training data, and vice versa. In soft-margin SVM, as shown in Fig. 2, the support vector is defined to be the point lies either on or inside the margin [31]. Mathematically, a labelled training data sample $(\mathbf{x}, c)$ is a support vector if it satisfies the following equation:

$$(\textbf{Support Vector})\quad V(\mathbf{x}, c) = 1 - c(\mathbf{w}^{\mathsf{T}}\mathbf{x} + b) \geq 0. \tag{5}$$

After training, the learnt SVM model can be used for predicting the label of a new sample by computing its output score is given as

$$(\textbf{Output Score})\qquad s(\mathbf{x}) = (\mathbf{w}^{\mathsf{T}}\mathbf{x} + b)/\|\mathbf{w}\|, \tag{6}$$

where $\|\cdot\|$ represents the Euclidean norm.

2) CNN model: CNN is made up of neurons that have adjustable weights and biases to express a non-linear mapping from an input data sample to class scores as outputs [39]. Note that the weights and biases constitute the parameters of the CNN. Typical implementation of a CNN consists of multiple layers including convolutional layers, ReLu layers, pooling layers, fully connected layers and normalization layers. Without the explicitly defined decision boundaries as for SVM, CNN adjusts the parameters of the hidden layers to minimize the prediction error, calculated using the outputs of the softmax layer and the true labels of training data. The expression of output score is given as:

$$(\textbf{Output Score})\qquad s_{\hat{c}}(\mathbf{x}) = P_{\theta}(\hat{c}|\mathbf{x}), \tag{7}$$

indicating the posterior distribution of the predicted label of a data sample. After training, the learnt CNN model can then be used for predicting the label of a new sample by choosing one with the highest posterior probability.

C. Data Uncertainty Measures

The importance of a data sample for learning is usually measured by its *uncertainty*, as viewed by the model under training [7]. Two uncertainty measures targeting SVM and CNN respectively are introduced as follows.

1) Uncertainty Measure for SVM: For SVM, we adopt the distanced based uncertainty which is motivated by the fact that a classifier makes less confident inference on a data sample which is located near the decision boundary [15]. Given a data sample $\mathbf{x}$ and a binary classifier $\{\mathbf{w}, b\}$, the said distance can be readily computed by the absolute value of the output score as follows

$$d(\mathbf{x}) = |s(\mathbf{x})| = |\mathbf{w}^{\mathsf{T}}\mathbf{x} + b|/\|\mathbf{w}\|. \tag{8}$$

Then the distance based uncertainty measure is defined as

$$(\textbf{Distance Based Uncertainty})\quad \mathcal{U}_{\mathsf{d}}(\mathbf{x}) = -d^2(\mathbf{x}) = -\frac{(\mathbf{w}^{\mathsf{T}}\mathbf{x} + b)^2}{\|\mathbf{w}\|^2}. \tag{9}$$

2) Uncertainty Measure for CNN: For CNN, a suitable measure is *entropy*, an information theoretic notion, defined as follows [28]:

$$(\textbf{Entropy})\quad \mathcal{U}_{\mathsf{e}}(\mathbf{x}) = -\sum_{\hat{c}=1}^{C} P_{\theta}(\hat{c}|\mathbf{x}) \log P_{\theta}(\hat{c}|\mathbf{x}), \tag{10}$$

where $\hat{c}$ denotes a predicted class label and θ the set of model parameters to be learnt.

III. Principle of Importance-Aware Scheduling

In this section, we first consider the task of training a binary SVM classifier at the edge. To attain a more accurate model under the constrained transmission budget, it requires the edge server to schedule the device with the most useful data sample for transmission. The scheduling decision making is challenging as there lacks a selection metric to evaluate the importance of noisy data. The problem is tackled in this section by designing the DII, which combines two metrics in communication and learning to indicate the effective importance of a transmitted data sample for learning. Then, the importance-aware scheduling is proposed based on the indicator, so as to accelerate the model training at the edge server. Finally, the proposed scheme for SVM is extended to general classifiers.

A. Data Importance Indicator

The direct design of DII for optimizing the learning performance is difficult due to a lack of tractable mapping from noisy data importance to the learning speed and accuracy. Nevertheless, the following fact in active learning provides a potential connection between data uncertainty and model-convergence rate: a model can be trained using fewer labelled data samples if the highly uncertain data is selectively added into the training set. The fact suggests that data uncertainty should be incorporated into the design of DII to maximize the improvement on the classifier's accuracy. However, an uncertainty measure in active learning targets noiseless data selection, and thus cannot be directly used for edge learning, as the acquired training data is corrupted by channel fading and noise, thereby affecting the effective uncertainty. To address this issue, the expectation of received data uncertainty can serve as a reasonable measure of effective uncertainty, and thus is used for defining the DII.

Definition 1 (Data Importance Indicator)**.** Conditioned on the local dataset D_k at the k-th edge device and its associated channel, the corresponding DII is defined as:

$$I_k = \max_{n \in \mathcal{N}} \mathsf{E}_{\mathbf{z}_k}\left[\mathcal{U}_{\mathsf{d}}\left(\hat{\mathbf{x}}_{k,n}\right)\right], \tag{11}$$

where $\hat{\mathbf{x}}_{k,n}$ and $\mathcal{U}_{\mathsf{d}}(\cdot)$ are defined in (2) and (9) respectively, and $\mathcal{N} = \{1, 2, \cdots, N\}$ represents the sample index set.

The remainder of the sub-section focuses on deriving a closed-from expression for DII. To begin with, we first give the expression for calculating the distance from a received data sample to the decision boundary. The derivation of the

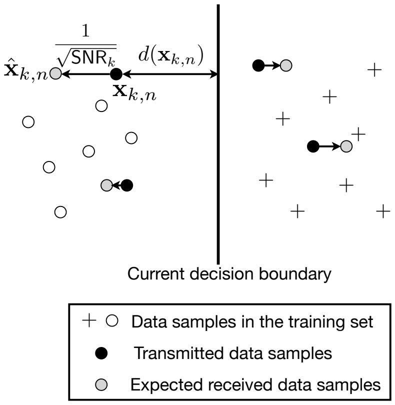

$\frac{1}{\sqrt{\mathsf{SNR}_k}}$
$d(\mathbf{x}_{k,n})$
$\hat{\mathbf{x}}_{k,n}$
$\mathbf{x}_{k,n}$
Current decision boundary
+ ○ Data samples in the training set
● Transmitted data samples
● Expected received data samples

Figure 3. Illustration of the derived DII and the effect of channel noise.

result involves utilizing the equivalence between the square of distance measure and that of the corresponding score.

Lemma 1. Conditioned on the parameters $\{\mathbf{w}, b\}$ of a binary SVM classifier, channel coefficient h_k, and channel noise $\mathbf{z}_k$, the distance from a received data sample $\hat{\mathbf{x}}$ to the decision boundary is given as:

$$d(\hat{\mathbf{x}}) = \sqrt{s^2(\mathbf{x}) + 2s(\mathbf{x}) \times \frac{\mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{z}}_k}{\|\mathbf{w}\|} + \left(\frac{\mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{z}}_k}{\|\mathbf{w}\|}\right)^2}, \qquad (12)$$

where $s(\cdot)$ is the output score given in (6), and $\widetilde{\mathbf{z}}_k = \frac{1}{\sqrt{P}}\Re\left(\frac{{h_k}^*}{\|h_k\|^2}\mathbf{z}_k\right)$ is the equivalently noise after decoding.

According to Lemma 1, the key step for deriving DII is to find the distribution of the projected channel noise $\frac{\mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{z}}_k}{\|\mathbf{w}\|}$. The derivation simply involves projecting the high-dimensional Gaussian distribution onto a particular direction specified by $\mathbf{w}$, which yields a univariate Gaussian distribution as elaborated below.

Lemma 2. Given the specific direction $\mathbf{w}/\|\mathbf{w}\|$, the projected channel noise follows a Gaussian distribution:

$$\frac{\mathbf{w}^{\mathsf{T}}\widetilde{\mathbf{z}}_k}{\|\mathbf{w}\|} \sim \mathcal{N}\left(0, \frac{1}{\mathsf{SNR}_k}\right). \qquad (13)$$

Applying the expectation and variance of projected channel noise (see Lemma 2) into Lemma 1, the expected distance of a received data sample is presented in the following lemma.

Lemma 3. The expected distance from the received data sample $\hat{\mathbf{x}}_{k,n}$ to the decision boundary is

$$\mathsf{E}_{\mathbf{z}_k}\left[d^2(\hat{\mathbf{x}}_{k,n})\right] = d^2(\mathbf{x}_{k,n}) + \frac{1}{\mathsf{SNR}_k}. \qquad (14)$$

Lemma 3 suggests that the channel fading and noise tend to degrade the data importance. The effect of noise on the receive data importance can be further illustrated in Fig. 3, where the solid black dots represent the transmitted data samples. The corresponding received data sample is expected to be the grey dot, which is more likely to be pushed away from the decision boundary, i.e., it suffers importance reduction. The degradation is proportional to the power of channel noise (or the inverse of SNR). The result is aligned with the intuition that channel noise is harmful and can not be exploited for improving learning performance.

With Lemma 3, we are ready to derive a closed-form expression of DII, as shown in the following proposition.

Proposition 1. Consider the training of a binary SVM classifier at the edge and the model is broadcast to the edge devices for data uncertainty evaluation. Given the local dataset $D_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \cdots, \mathbf{x}_{k,N}\}$ of the k-th edge device and SNR_k, the DII is given as

$$I_k = -\frac{1}{\mathsf{SNR}_k} + \max_{n\in\mathcal{N}} \mathcal{U}_{\mathsf{d}}\left(\mathbf{x}_{k,n}\right), \qquad (15)$$

where $\mathcal{N} = \{1, 2, \cdots, N\}$ represents the sample index set, and $\mathcal{U}_{\mathsf{d}}(\cdot)$ is a distance-based uncertainty measure defined in (9).

Remark 1 (How does the local buffer size affect DII?)**.** With the increase of buffer size, the DII is dominated by the SNR due to convergence of the second term related to data uncertainty towards zero. A larger buffer size suggests potentially higher diversity of the dataset for selection. Specifically,

$$\lim_{N\to\infty} \max_{n\in\mathcal{N}_k} \mathcal{U}_{\mathsf{d}}\left(\mathbf{x}_{k,n}\right) = \lim_{N\to\infty} \min_{n\in\mathcal{N}_k} d^2\left(\mathbf{x}_{k,n}\right) \to 0, \ \forall k. \qquad (16)$$

As a result, the DII in the case of large buffer becomes:

$$\lim_{N\to\infty} I_k = -\frac{1}{\mathsf{SNR}_k}. \qquad (17)$$

The developed DII for binary SVM can be extended to multi-class SVM. Given a data sample $\mathbf{x}$, the distance based uncertainty is evaluated based on the predicted label as illustrated in the sequel. Specifically, a C-class SVM classifier is implemented by $L = C(C-1)/2$ *one-versus-one* binary component classifiers that each trained using the samples from the two concerned classes only [40]. To predict the label $\hat{c}$, the output L-dimension vector, denoted as $\mathbf{s} = [s_1(\mathbf{x}), s_2(\mathbf{x}), \cdots, s_L(\mathbf{x})]$ is compared with a *reference coding matrix* of size $C \times L$, denoted by $\mathbf{M}$. An example of the reference coding matrix with C = 4 is provided as follows:

$$\mathbf{M}= \begin{array}{c} \\ \text{class1} \\ \text{class2} \\ \text{class3} \\ \text{class4} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{binary1} & \text{binary2} & \text{binary3} & \text{binary4} & \text{binary5} & \text{binary6} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix} \end{array},$$

where each row gives the "reference output pattern" corresponding to the associated class. Given $\mathbf{M}$, the prediction of the class index of $\mathbf{s}$ involves simply comparing the Hamming distances between $\mathbf{s}$ and different rows in $\mathbf{M}$, and choosing the row index with the smallest distance as the predicted class index:

$$\hat{c} = \arg\min_{c} \sum_{\ell=1}^{L} |m_{c\ell}|[1 - \mathrm{sgn}(m_{c\ell}s_\ell(\mathbf{x}))]/2, \qquad (18)$$

where $m_{c\ell}$ denotes the ℓ-th element in vector $\mathbf{m}_c$, and $\mathrm{sgn}(\mathrm{x})$ denotes the sign function taking a value from $\{1, 0, -1\}$ corresponding to the cases $x > 0$, $x = 0$ and $x < 0$, respectively. Having obtained the predicted label $\hat{c}$, the distance based uncertainty is averaged over all the effective component classifiers of the predicted label and DII is defined below:

$$I_k = -\frac{1}{\mathsf{SNR}_k} + \max_{n\in\mathcal{N}} \Big\{ -\frac{1}{C-1} \sum_{\ell=1}^{L} |m_{\hat{c}\ell} s_\ell(\mathbf{x}_{k,n})|^2 \Big\}. \quad (19)$$

B. Importance-Aware Scheduling

In this section, the importance-aware scheduling is designed for binary SVM classification. Specifically, the edge sever schedules the device with highest value of DII. The design can be extended to multi-class SVM following the procedure described in the preceding section. Given the derived DII in Proposition 1, the resultant scheme is described in Scheme 1 below.

The summation form of DII in (20) elegantly incorporates both data uncertainty and channel quality in the design of scheduling criterion, which provides a simple mechanism for simultaneous exploitation of multiuser data-and-channel diversity. Any criterion purely exploits only a single type of diversity may compromise the learning performance degradation and lead to inefficient use of radio resources. Particularly, a scheduling criterion based on only SNR (only exploiting channel diversity, see Scheme 2) is prone to selecting a useless data sample which has little contribution to refining the decision boundary. On the other hand, the one based on data importance (only exploiting data diversity, see Scheme 3) may suffer from selecting highly noisy data, and thereby compromise the learning.

Scheme 1 (Importance-aware scheduling for binary SVM)**.** Consider the acquisition of a data sample from multiple edge devices in an edge learning system. The edge server schedules device k^* for data transmission if

$$k^* = \arg\max_{k} \left\{ -\frac{1}{\mathsf{SNR}_k} + \max_{n \in \mathcal{N}_k} \mathcal{U}_{\mathsf{d}}\left(\mathbf{x}_{k,n}\right) \right\}, \quad (20)$$

where $\mathcal{U}_{\mathsf{d}}(\cdot)$ is the distance based uncertainty defined in (9).

Remark 2 (How does a transmit SNR affect scheduling?)**.** The effect of the transmit SNR P/σ^2 on scheduling can be understood by rewriting the scheduling scheme using the definition of SNR given in (3):

$$k^* = \arg\max_{k} \left\{ -\frac{\sigma^2}{P} \times \frac{1}{2\|h_k\|^2} + \max_{n \in \mathcal{N}_k} \mathcal{U}_{\mathsf{d}}\left(\mathbf{x}_{k,n}\right) \right\}. \quad (21)$$

One can observe that the transmit SNR P/σ^2 is a weight factor to balance the influences of channel quality and data uncertainty on the scheduling decision. The scheduling schemes for low and high transmit SNR scenarios are discussed as follows.

- *Low transmit SNR:* For this case, wireless channels are unreliable. The channel diversity is more critical to be exploited for reliably receiving a data sample. Otherwise, received data samples are severely corrupted by noise and become useless regardless of their uncertainty (importance) prior to transmission. This fact causes the proposed scheme to enforce a large weight factor (low transmit SNR) for channel quality in the scheduling metric. Moreover, the scheme is reduced to channel-aware scheduling (see Scheme 2) when the transmit SNR approaches zero.
- *High transmit SNR:* On the contrary, when the transmit SNR is high, it is more critical to exploit the data diversity as all wireless data channels are reliable. In this case, acquiring data samples with high original uncertainty values accelerates the model training. This fact is translated into the small weight factor (high transmit SNR) for channel quality so as to make data uncertainty dominant in scheduling decision making. If the transmit SNR is sufficiently large, the scheduling scheme reduces to pure important data selection (named data-aware scheduling in Scheme 3) as the first term in (21) vanishes.

Last, the two mentioned conventional schemes that are special cases of importance-aware scheduling are presented as follows.

Scheme 2 (Channel-aware scheduling)**.** Consider the acquisition of a data sample from multiple edge devices in an edge learning system. The edge server schedules the k^* device for data transmission if

$$k^* = \arg\max_{k} \ \mathsf{SNR}_k, \quad (22)$$

where SNR_k is defined in (3), and the transmitted data sample is randomly selected from the scheduled edge device.

Scheme 3 (Data-aware scheduling)**.** Consider the acquisition of a data sample from multiple edge devices in an edge learning system. The edge server schedules the k^* device for data transmission if

$$k^* = \arg\max_{k} \ \max_{n \in \mathcal{N}_k} \mathcal{U}_{\mathsf{d}}\left(\mathbf{x}_{k,n}\right), \quad (23)$$

where $\mathcal{U}_{\mathsf{d}}(\cdot)$ is the distanced based uncertainty defined in (9).

C. Extension to General Classifier Models

In this section, the proposed importance-aware scheduling targeting for SVM classifier is extended to a general model. However, the derivation for SVM may not be directly applied to a generic classifier (e.g., CNN), due to the lack of an explicitly defined functional mapping from input noisy data to the output score. Nevertheless, the general form of the DII derived in the SVM setting is applicable to a generic model. This motivates the simple generalization of the importance-aware scheduling by replacing the uncertainty measure in (9) targeting SVM with a general measure, which can be properly chosen depending the specific learning model. The modified scheme is descried as follows.

Scheme 4 (Importance-aware scheduling for a generic classifier)**.** Consider the acquisition of a data sample from multiple edge devices in an edge learning system. The edge server schedules the k^* device for data transmission if

$$k^* = \arg\max_{k} \left\{ -\frac{1}{\mathsf{SNR}_k} + \max_{n \in \mathcal{N}_k} \mathcal{U}_{\mathsf{x}}\left(\mathbf{x}_{k,n}\right) \right\}, \quad (24)$$

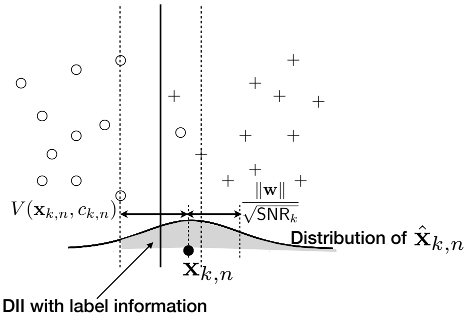

$V(\mathbf{x}_{k,n}, c_{k,n})$
$\frac{\|\mathbf{w}\|}{\sqrt{\mathsf{SNR}_k}}$
Distribution of $\hat{\mathbf{x}}_{k,n}$
$\mathbf{x}_{k,n}$
DII with label information

Figure 4. Illustration of the DII with label information.

where $\mathcal{U}_{\mathsf{x}}$ is a general uncertainty measure. In particular, it can be the entropy $\mathcal{U}_{\mathsf{e}}$ defined in (10) if CNN classifier model is adopted.

As a guideline, the selection of the uncertainty measure should allow easy computation using the model output. For example, for SVM, the output score evaluated using the linear decision boundaries allows easy evaluation of the distance-based uncertainty. On the other hand, for CNN, the softmax output, which gives the posterior probability for each predicted class, makes the entropy to be a more natural choice for measuring uncertainty.

IV. Implementation Issues and Solutions

A. Importance-Aware Scheduling With Label Information

In the preceding sections, the data importance (uncertainty) is evaluated and a scheduling decision made based on unlabelled data with a label generated after scheduling. This targets the scenario where labelling (e.g., by a human labeler) is costly. However, in some cases, data at the edge devices are generated together with labels. For example, training data for auto-driving (e.g., outputs of cameras and radar) are automatically labelled by sensing the driving decisions of a human driver. Then, the design of DII should exploit label information to further accelerate the learning speed.

The design of DII for the current case of all labelled data is based on the fact in the incremental learning of SVM that a newly added data sample to the training dataset can update the improve model if it is a support vector [41]. As a result, the DII is defined to be the expectation of model update. With the definition of support vector given in (5), the event of model update, denoted as $\mathcal{V}$, is defined as that both transmitted and received data samples are support vectors: $\{\mathcal{V}(\hat{\mathbf{x}}|\mathbf{x}, c)|\,(V(\mathbf{x}, c) \geq 0) \cap (V(\hat{\mathbf{x}}, c) \geq 0)\}$. The event ensures the model update is due to the important data sample instead of the channel noise. Mathematically, the DII with label information is defined as follows:

$$\begin{aligned} I_k &= \max_{n \in \mathcal{N}} \mathsf{E}_{\mathbf{z}_k} \left[\mathcal{V}(\hat{\mathbf{x}}_{k,n} | \mathbf{x}_{k,n}, c_{k,n}) \right] && (25) \\ &= \max_{n \in \mathcal{N}} \int_{-V(\mathbf{x}_{k,n}, c_{k,n})}^{\infty} \sqrt{\frac{\mathsf{SNR}_k}{2\pi \|\mathbf{w}\|^2}} \exp\left(-\frac{t^2}{2\|\mathbf{w}\|^2/\mathsf{SNR}_k} \right) dt \\ &= \max_{n \in \mathcal{N}} \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{V(\mathbf{x}_{k,n}, c_{k,n})}{\sqrt{2\|\mathbf{w}\|^2/\mathsf{SNR}_k}} \right) \right], \\ &\qquad \forall\ V(\mathbf{x}_{k,n}, c_{k,n}) \geq 0, && (26) \end{aligned}$$

where $V(\mathbf{x}, c)$ has been given in (5). The result is graphically shown in Fig. 4 as the shaded area. One can notice that DII is a probability that requires the variation of channel noise lies inside the margin boundary, which is determined by the ratio of $V(\mathbf{x}, c)$ and noise power $\frac{1}{\sqrt{\mathsf{SNR}_k}}$ as derived in (26). Then the importance-aware scheduling with label information is designed as follows based on the simplified DII.

Scheme 5 (Importance-aware scheduling with label information)**.** Consider the acquisition of a data sample from multiple edge devices in an edge learning system. The edge server schedules the k^* device for data transmission if

$$k^* = \arg\max_k \left\{ \sqrt{\mathsf{SNR}_k} \times \max_{n \in \mathcal{N}_k} \ \max\left[0, V(\mathbf{x}_{k,n}, c_{k,n})\right] \right\}, \qquad (27)$$

where $V(\mathbf{x}_{k,n}, c_{k,n})$ is defined in (5) and $\max\left[0, V(\mathbf{x}_{k,n}, c_{k,n})\right]$ is to pre-select the support vector $(V(\mathbf{x}_{k,n}, c_{k,n}) \geq 0)$ for transmission.

It is remarked that $\max\left[0, V(\mathbf{x}_{k,n}, c_{k,n})\right]$ is exactly the same as the definition of hinge loss [31], and thus DII with label information elegantly incorporates two metrics from communication and learning perspectives. Compared with the data selection by using the uncertainty measure (without label information), hinge loss is another way to exploit data diversity which has its pros and cons. For the data selection based on uncertainty, the new coming data sample near to the decision boundary helps to refine the optimal classifier (reduce the hypothesis space) in a binary search manner. With the label information, the selected data sample based on hinge loss could cross the decision boundary with a wrong predicted label, thereby achieve a faster rate than the binary search so as to accelerate learning speed. On the other hand, when the hypothesis space is small, the hinge loss may guide to select the outlier in non-separable dataset, and thus mislead the classifier to the opposite side.

B. Compressed Model for DII Evaluation

One practical issue in implementing importance-aware scheduling is high local computing complexity of DII evaluation. Specifically, the energy consumption and the requirement of computing resources for data uncertainty evaluation using the full model may be too costly for a resource constrained edge device. This motivates the use of compressed model for DII evaluation that can reduce local computing complexity without significantly compromising the scheduling performance. Underpinning the design is the fact that no performance loss will be incurred as long as the compressed model can provide sufficient differentiability amongst data in terms of DII values. To illustrate this point, as shown in Fig. 5, the scheduling based on the DII evaluated using the compressed model may make exactly the same decision as that based on DII evaluated using the full model. This fact indicates the existence of model redundancy to be reduced for efficient data uncertainty evaluation.

To avoid performance degradation, the compression ratio should be properly selected according to data distribution,

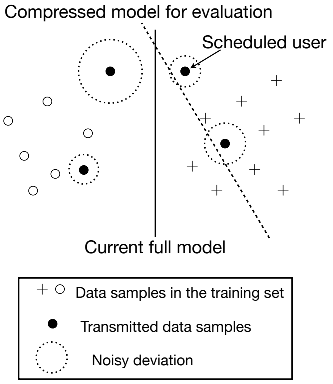

Compressed model for evaluation
Scheduled user
Current full model
+ ○ Data samples in the training set
● Transmitted data samples
Noisy deviation

Figure 5. Effect of the compressed model for unlabeled DII evaluation.

SNR and the number of users. For example, a highly separable dataset and sparse device population allows large space for model compression. On the other hand, for the low transmit SNR scenario, the scheduling scheme relies less on data importance, thus requires less accurate model to evaluate its value.

In the experiment, we vary the compression ratio $C_r \in (0,1)$, representing the ratio of parameters remained, and show its effect on learning performance. Given the compression ratio, we could further characterize the computational efficiency of data uncertainty evaluation at edge devices, which is determined by the number of computer operations for calculating the output score. In general, it is scaled by the total number of parameters W in the learning model. Specifically, for a linear classifier like SVM, the overall local computational cost is $\mathcal{O}\left(WC_rN\right)$, for calculating linear score functions of N data samples. On the other hand, computing an output score of CNN requires forward propagation in the compressed neural network, and thus the local computational cost is also $\mathcal{O}\left(WC_rN\right)$ as indicated in [31].

C. Several Methods for Performance Enhancement

As the importance-aware scheduling exploits data diversity, the resultant performance highly depends on total distributed data samples for selection, which is proportional to the number of users, local buffer size, the update frequency of local buffer and user mobility as characterized in the following. Assume the number of edge users is K in the network in each time slot, during T slots of the wireless data acquisition, the available number of data samples for scheduling is given as

$$\mathcal{D}(T) = KN + \sum_{t=2}^{T} \left[K - P_u(t)\right] P_d(t) + P_u(t)N, \qquad (28)$$

where $P_d(t) \in \{0,1,2,\cdots,N\}$ is the number of updated samples in each devices, and $P_u(t) \in \{0,1,2,\cdots,K\}$ is the number of users replaced in the coverage cell due to mobility. The result in (28) suggests that the performance of importance-aware scheduling can be improved in several possible ways such as increasing the number of users K and the buffer size N, updating the local buffered samples with a higher frequency, or utilizing user mobility. They are verified by simulation to be effective.

V. Experimental Results

A. Experiment Setup

The default experiment settings are as follows unless specified otherwise. The number of edge devices is $K = 10$. Each device is equipped with a local buffer, with the size $N = 10$, and updates one of outdated data samples with a new one, denoted as $P_d(t) = 1$, for an arbitrary slot t. Consider the static user case where the number of users replaced in the coverage cell is set as $P_u(t) = 0$ for all t. The maximum transmission budget T for the learning task is given as 100 and $1,000$ (channel uses) for binary and multi-class classifications, respectively. Under the transmission budget constraint, we consider the test accuracy as the performance metric. All results are averaged over 150 and 20 experiments for binary SVM and multi-class CNN, respectively.

1) Channel Model: We assume the classic Rayleigh fading channel with channel coefficients $\{h_k\}$ following i.i.d. complex Gaussian distribution $\mathcal{CN}(0,1)$. The average transmit SNR defined as $\bar{\rho} = P/\sigma^2$ is by default set as 15 dB.

2) Experimental Dataset: We consider the learning task of training a classifier using the well-known MNIST dataset of handwritten digits. The training and test sets consist of $60,000$ and $10,000$ samples, respectively. Each sample is a grey-valued image of 28×28 pixels that gives the sample dimensions $p = 784$. The experiments of multi-class classification are conducted by using the whole dataset. For binary classification, we choose the relatively less differentiable class pair of “3” and “5” (according to *t-distributed stochastic neighbor embedding* visualization) from the whole data set, including $11,552$ training samples and 1902 test samples. The training set used in experiments is partitioned as follows. At the edge server, the initially available training dataset $\mathcal{L}_0$ is constructed by randomly sampling 2 data samples for each class. The remaining training data are evenly and randomly partitioned for constructing the local datasets at edge devices, which are used for updating local buffers. The placement of data samples into a local buffer are randomly sampled from its local dataset, following the update rules as discussed at the beginning of this section.

3) Learning Model Implementation: The considered classifier models include binary SVM and CNN. For binary SVM, the soft-margin SVM is implemented with slack variable set as 1. *Iterative Single Data Algorithm* (ISDA) [42] is used for solving the SVM problem with maximum 10^6 iterations. For the implementation of CNN, we use a 6-layer CNN including two 3×3 convolution layers with batch normalization before ReLu activation (the first with 16 channels, the second with 32), the first one followed with a 2×2 max pooling layer and the second one followed with a fully connected layer, a softmax layer, and a final classification layer. The model is trained using stochastic gradient descent with momentum [43]. The mini-batch size is 2048, and the number of epochs is 120. To accelerate training, the CNN is updated in a batch mode with the incremental sample size set as 10. The broadcast model is uncompressed, i.e., $C_r = 1$, by default.

B. Learning Performance for SVM

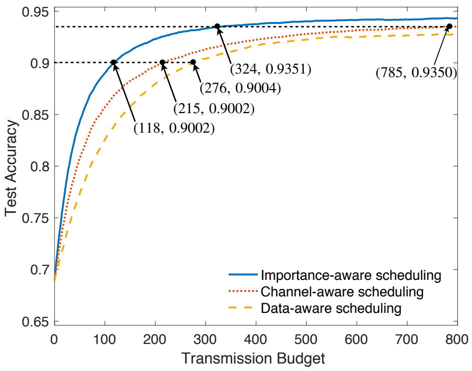


Figure 6. Test accuracy versus transmission budget.

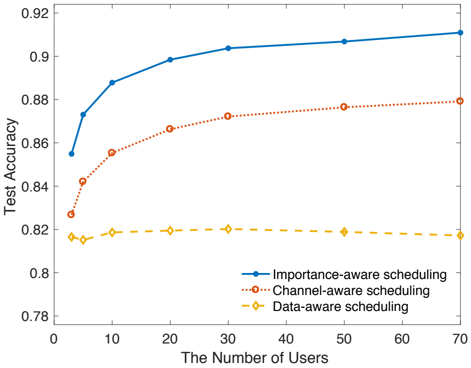


Figure 7. Test accuracy versus the numbers of users.

1) Convergence Rate: In Fig. 6, the learning performance of the proposed importance-aware scheduling is compared with two baseline schemes, namely the channel-aware scheduling and data-aware scheduling, corresponding to pure channel selection and important data selection respectively. The transmission budget varies from 0 to its maximum value which is set as 800 for ensuring model convergence of all schemes. It is observed that the proposed scheme outperforms the two benchmarks. Specifically, if the targeted accuracy is 0.9, the required budget is 118 for importance-aware scheduling while it is 215 and 276 for channel-aware scheduling and data-aware scheduling respectively. Thus, it saves more than half budget to achieve the targeted performance by using importance-aware scheduling. The comparison is more remarkable if the targeted accuracy is 0.935, where the budget requirements are 324 and 785 for the proposed scheme and conventional channel-aware scheme respectively. In contrast, data-aware scheduling can not achieve that targeted accuracy within the maximum transmission budget. This confirms the fast convergence by exploiting both data and channel diversities, and verifies the effectiveness of the proposed scheme for fast edge learning. In the following experiments, we fix the transmission budget of all schemes and compare their test accuracies instead, which is equivalently to reflect the difference in terms of convergence rate.

2) Multi-user Diversity: In Fig. 7, we investigate the gain of multi-user diversity by plotting test accuracy over the number of users. The performance of importance-aware scheduling consistently outperforms two benchmarks in varying number of users scenarios. This verifies the performance gain by intelligently allocating radio resources according to both data importance and channel condition. It is observed that the data-aware scheduling is hardly to exploit data diversity in the wireless edge learning scenario. As the scheme itself is unconscious of channel conditions, the important data samples are contaminated by large channel noise that impeding the performance improvement by data selection. The result indicates that reliable transmission is the principle requirement before exploiting data importance for learning. Second, by involving more users in the learning system, importance-aware scheduling achieves larger performance improvement than that of the channel-aware scheduling at the initial stage (e.g., $K < 10$). The reason is that small number of users incurs data deficiency that could be overcome by adding more users. In contrast, when the number of users is large (e.g., $K > 10$), the improvement rates of two schemes are comparable. In this case, the improvement is mainly due to channel diversity while the gain by exploiting data is saturated if more users are involved. The result reflects that multi-user diversity include two folds, data and channel, which should be jointly exploited for improving learning performance. The schemes only exploiting one aspect like the baseline schemes will cause a potential degradation in learning performance.

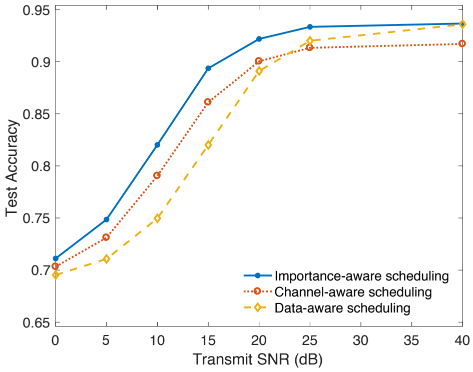
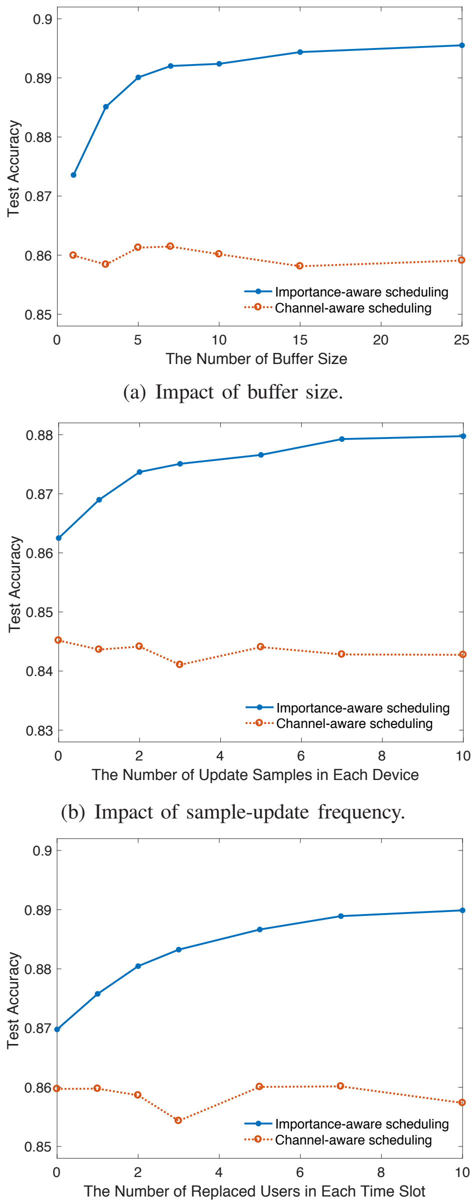


Figure 8. Test accuracy versus the average transmit SNR $\bar{\rho}$.

3) Transmit SNR: To demonstrate its robustness against the hostile channel conditions, the proposed importance-aware scheduling is tested under different values of transmit SNR and the results are shown in Fig. 8. One can notice that the test accuracy of proposed scheme is always better than that of the two baseline schemes. The results further substantiates the performance gain by jointly exploiting both data and channel diversities. To be specific, it is more essential to balance the tradeoff between data importance and data reliability in a moderate transmit SNR scenario (e.g., $\rho = 5 - 25$ dB), since the proposed scheme is shown to achieve a more remarkable performance gain than the two benchmarking schemes. In contrast, the proposed scheme reduces to channel-aware scheduling and data-aware scheduling in low transmit SNR (e.g., $\rho = 0$ dB) and high transmit SNR (e.g., $\rho = 25$ dB) scenarios respectively, which verifies the discussion in Remark 2. The comparison of three schemes reflects that data reliability is the most critical requirement, since all of schemes suffer severe performance degradations in low SNR scenario. Upon a certain guarantee on data reliability, then the performance can be further improved by exploiting data importance

4) Data Diversity: Fig. 9 demonstrates the performance improvement by increasing the number of available data



(a) Impact of buffer size.

(b) Impact of sample-update frequency.

(c) Impact of user mobility.

Figure 9. Impact of different approaches for data diversity enhancement.

samples for selection, which depends on the number of users, local buffer size, the update frequency of local buffer and user mobility, as discussed in Section IV-C. Since the effect of the number of users has been discussed, this part will focus on the other three, which purely exploit the diversity in distributed data. The relevant results are shown in subfigures respectively, where the performance of importance-aware scheduling is compared with the channel-aware scheduling. The three figures verify that the conventional scheme is unable to exploit the distributed data samples.

- *Local buffer size:* Fig. 9(a) presents the performance improvement by increasing the number of buffer size. The incremental buffers size leads to a remarkable performance improvement at the initial stage ($N < 5$), corresponding to a data deficiency scenario. Then the improvement will be saturated if continuously increase the buffer size ($N > 15$).

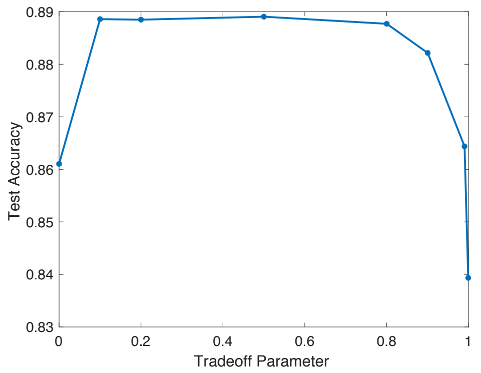


Figure 10. The effect of tradeoff factor α.

- *Update frequency of local buffer:* Fig. 9(b) displays the performance curves of test accuracy versus the number of update samples in each device. Note that the update data frequency of local buffer could affect the learning performance only in a data deficiency scenario and the number of users is set as $K = 5$ in this experiment. The result of importance-aware scheduling reveals that the data deficiency can be overcome by frequently updating the samples in local buffer.
- *User Mobility:* The user mobility could be specified as the number of replaced users $P_u(t)$ in each time slot, and high mobility corresponds to the large value of $P_u(t)$. In Fig. 9(c), the test accuracy is plotted over the number of replaced users. In this experiment, the buffer size is set as $N = 5$ for a data deficiency scenario, and $P_d(t)$ is set as 0 to reflect the unique performance improvement by exploiting user mobility, which is shown to be prominent.

5) Tradeoff Between SNR and Data Uncertainty: Fig. 10 shows the effect of different tradeoff between SNR and uncertainty on the learning performance. In this experiment, the DII is constructed as follows.

$$I_k = -\frac{1}{\mathsf{SNR}_k} \times (1-\alpha) + \max_{n\in\mathcal{N}} \mathcal{U}_{\mathsf{d}}\left(\mathbf{x}_{k,n}\right) \times \alpha \ , \qquad (29)$$

where the tradeoff factor α ranges from 0 to 1. The tradeoff factor specifies how the scheduling scheme weights the importance of SNR and uncertainty in the decision making, varying from purely SNR-based selection ($\alpha = 0$) to purely data-uncertainty-based selection ($\alpha = 1$). The experimental result shows that the best test accuracy is achieved at $\alpha = 0.5$, verifying the optimality of the equal treatment between SNR and data uncertainty derived in (15). It is interesting to find that, the accuracy barely decreases for a large range of α, e.g., $\alpha \in [0.1, 0.8]$. This suggests that as long as both the SNR and data-uncertainty can be taken into account to some extent in the scheduling decision, considerable performance gains can be achieved compared with the schemes based on either SNR or data-uncertainty.

6) Compressed Model for Importance Evaluation: In Fig. 11, the proposed scheme is tested under different importance-evaluation models, by varying the model compression ratio. Although a simple model as SVM, it is able to reduce half computing operations ($C_r = 1/2$) without incurring performance loss. The performance of importance-aware

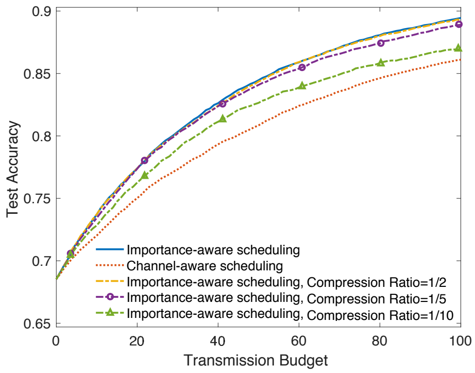


Figure 11. Test accuracy for different evaluation models by varying its compression ratio C_r.

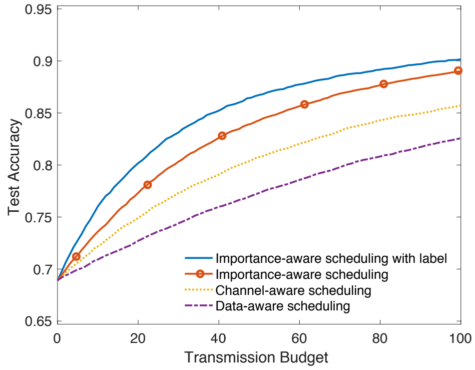


Figure 12. Learning performance evaluation for importance-aware scheduling with label information.

scheduling by using a compressed importance-evaluation model is shown to consistently outperform the channel-aware scheduling, even the number of model parameters is reduced by $10\times$ ($C_r = 1/10$). On the other hand, the performance loss due to model compression is related with the training stage: the precision of evaluation model should be increased as the learnt model becomes more accurate, that is, allowing fewer number of model parameters to be reduced. As shown by the curve of $C_r = 1/5$, at the initial stage, it achieves same performance as the one using uncompressed model, while the performance loss increases as the model being more accurate.

7) Scheduling With Label Information: In Fig 12, importance-aware scheduling with label information is compared with the unlabeled scheduling scheme and the two benchmarks. By exploiting additional label information, the model attains faster convergence rate than the unlabeled scheme. However, if the transmission budget is large, the learning accuracies of two schemes will be converged to a comparable level. Comparing with the benchmarks, the importance-aware scheduling with label information achieves remarkable improvement in terms of learning accuracy, corresponding to 5% for channel-aware scheduling and 8 % for data-aware scheduling.

C. Learning Performance for CNN

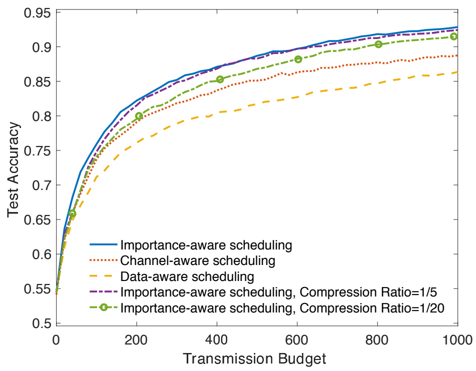


Figure 13. Learning performance evaluation for CNN classifier.

Our heuristic design for CNN is tested in the scenario of multi-class classification and the results are provided in Fig. 13. For the uncompressed evaluation model, the test accuracy consistently outperforms the two baseline schemes. Although the heuristic design can not guarantee an optimal tradeoff between data importance and data (channel) reliability, the performance gain is prominent. That confirms the benefit of exploiting both channel-and-data diversity in CNN. On the other hand, each training sample in CNN contributes to define the multiple decision boundaries, that is different from SVM where a single hyperplane determined by few support vectors. In another word, CNN is less robust with the existing of channel noise, and requires more training samples to construct multiple boundaries. Thereby it has a potential to achieve more remarkable performance gain if more users are involved in the system, since the achievable multiuser diversity for SVM may not be enough for the training of CNN classifier.

In general, a large number of CNN parameters are redundant that enables a highly compressed model for importance evaluation. This can be supported by the curve of $C_r = 1/5$, which achieves almost same performance as that of by using the uncompressed one. Furthermore, the performance loss will be less than 50% if the compression ratio is $1/20$. In contrast, the SVM classifier suffers a prominent performance loss for $C_r = 1/10$. That verifies the redundancy of CNN which is capable to utilize a compressed evaluation model to reduce computing complexity.

VI. Concluding Remarks

In this paper, we have proposed the novel scheduling scheme, namely importance-aware scheduling, for wireless data acquisition in edge learning systems. The scheme intelligently makes a joint channel-and-data selection for training data uploading so as to accelerate learning speed. Comprehensive experiments using real datasets substantiate the performance gain by exploiting two-fold multi-user diversity, namely multiuser data-and-channel diversity.

At a higher level, the work contributes the new principle of exploiting data importance to improve the efficiency of multiuser data acquisition for distributed edge learning. It is interesting to study the convergence rate and the resultant diversity gain in future work. Specifically, the convergence rate can be quantified as the speed of gradient norm vanishing with respect to the number of iterations. The analysis of the convergence rate involves applying the extreme value theory over two types of stochastic processes: data stochasticity and channel stochasticity, in a similar way as in [44]. The scaling of the convergence rate is expected to be monotonically increasing with the size of dataset and the number of individual channels, which represents the order of multiuser diversity gain. This work can be generalized to the more sophisticated batch mode training and more complex systems, such as broadband wireless systems with OFDMA and MIMO. Moreover, the scheduling algorithms can be further designed to ensure the global data diversity, where the feedback is required to account for both multiuser diversity and data diversity with respect to global dataset. Besides raw data acquisition, another interesting direction is the acquisition of learning relevant information in a federated learning framework, e.g., gradient updates and model updates. The relevant scheduler design can build on the current one by changing the data-importance measure to a suitable measure based on gradient divergence [45] or model variance [46].

References

[1] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, pp. 19–25, Jan. 2020.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *Proc. of IEEE Int. Conf. Comput. Commun. (INFOCOM)*, (Honolulu, USA), April 2018.

[3] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[4] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[5] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the internet of vehicles," *Proc. IEEE*, vol. 108, pp. 245–261, 2020.

[6] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, S. Mazzocchi, H. B. McMahan, *et al.*, "Towards federated learning at scale: System design," in *Conf. Sys. Machine Learning (SysML)*, (California, USA), March 2019.

[7] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[8] P. Popovski, O. Simeone, F. Boccardi, D. Gunduz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity," *[Online]. Available: https://arxiv.org/pdf/1907.02441.pdf*, 2019.

[9] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. of Conf. on Neural Info. Process. Sys. (NIPS) Workshop*, (Barcelona, Spain), Dec. 2016.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. of 20th Intl. Conf. Artif. Intell. Stat. (AISTATS)*, (Florida, USA), May 2017.

[11] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning system," *[Online]. Available: https://arxiv.org/pdf/1905.09712.pdf*, 2019.

[12] N. Skatchkovsky and O. Simeone, "Optimizing pipelined computation and communication for latency-constrained edge learning," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1542–1546, 2019.

[13] M. Chen, Y. Hao, K. Lin, Z. Yuan, and L. Hu, "Label-less learning for traffic control in an edge network," *IEEE Network*, vol. 32, pp. 8–14, Nov. 2018.

[14] J. Qian, S. Sengupta, and L. K. Hansen, "Active learning solution on distributed edge computing," *[Online]. Available: https://arxiv.org/pdf/1906.10718.pdf*, 2019.

[15] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Wireless data acquisition for edge learning: Importance aware retransmission," *[Online]. Available: https://arxiv.org/pdf/1812.02030.pdf*, 2019.

[16] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, pp. 491–506, Jan. 2020.

[17] M. M. Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *[Online]. Available: https://arxiv.org/pdf/1901.00844.pdf*, 2019.

[18] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *[Online]. Available: https://arxiv.org/pdf/1812.11750.pdf*, 2018.

[19] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient sgd via gradient quantization and encoding," in *Proc. of Conf. on Neural Info. Process. Sys. (NIPS)*, (Long Beach, USA), Dec. 2017.

[20] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik, "Stochastic distributed learning with gradient quantization and variance reduction," *[Online]. Available: https://arxiv.org/pdf/1904.05115.pdf*, 2019.

[21] Y. Du and K. Huang, "Fast analog transmission for high-mobility wireless data acquisition in edge learning," *IEEE Wireless Commun. Lett.*, vol. 8, pp. 468–471, Oct. 2018.

[22] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. of Intl. Conf. on Commun. (ICC)*, (Seattle, USA), June 1995.

[23] D. N. C. Tse and S. V. Hanly, "Multiaccess fading channels. i. polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, 1998.

[24] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2053–2064, 2001.

[25] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proc. of IEEE Int. Conf. Comput. Commun. (INFOCOM)*, (San Franciso, USA), April 2003.

[26] R. Kwan, C. Leung, and J. Zhang, "Proportional fair multiuser scheduling in LTE," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 461–464, 2009.

[27] H. J. Bang, T. Ekman, and D. Gesbert, "Channel predictive proportional fair scheduling," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 482–487, 2008.

[28] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, (Anchorage, USA), June 2008.

[29] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Proc. of Conf. on Neural Info. Process. Sys. (NIPS)*, (Vancouver and Whistler, Canada), Dec. 2008.

[30] N. Roy and A. McCallum, "Toward optimal active learning through monte carlo estimation of error reduction," in *Proc. of Intl. Conf. on Machine Learning (ICML)*, (Williamstown, USA), June 2001.

[31] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[32] E. Bourtsoulatze, D. B. Kurka, and G. Deniz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. on Cognitive Commun. and Networking*, vol. 5, pp. 567–579, March 2019.

[33] S. Jakubczak, J. Z. Sun, D. Katabi, and V. K. Goyal, "Analog transmission of degradable content over wireless channels," in *Proc. of 48th Annual Allerton Conf. on Commun., Control, and Comp.*, 2010.

[34] G. Lai, Y. Liu, and L. Zhang, "Distributed realcast: A channel-adaptive video broadcast delivery scheme," in *Proc. of IEEE 80th Veh. Tech. Conf.*, 2014.

[35] T. T. Nu, T. Fujihashi, and T. Watanabe, "Power-efficient video uploading for crowdsourced multi-view video streaming," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 9-13 Dec. 2018.

[36] S. Jakubczak and D. Katabi, "Softcast: Clean-slate scalable wireless video," in *Proc. of 48th Annual Allerton Conf. on Commun., Control, and Comp.*, 2010.

[37] S. Cui, J.-J. Xiao, A. J. Goldsmith, Z.-Q. Luo, and H. V. Poor, "Energy-efficient joint estimation in sensor networks: Analog vs. digital," in *Proc. of Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, (Philadelphia, USA), March 2005.

[38] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer, 2001.

[39] S. Haykin, *Neural Networks: A Comprehensive Foundation*, vol. 2. New York: Prentice hall, 1994.

[40] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Proc. of Conf. on Neural Info. Process. Sys. (NIPS)*, (Denver, USA), Dec. 2000.

[41] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Proc. of Conf. on Neural Info. Process. Sys. (NIPS)*, (Denver, USA), Dec. 2001.

[42] V. Kecman, T.-M. Huang, and M. Vogt, "Iterative single data algorithm for training kernel machines from huge data sets: Theory and perfor-

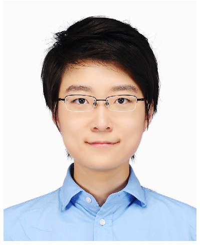
mance," in *Support Vector Machines: Theory and App.*, pp. 255–274, Springer, 2005.

[43] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. of Intl. Conf. on Machine Learning (ICML)*, (Atlanta, USA), June 2013.

[44] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, 2019.

[45] T. Chen, G. Giannakis, T. Sun, and W. Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Proc. of Conf. on Neural Info. Process. Sys. (NIPS)*, (Montreal, Canada), Dec. 2018.

[46] M. Kamp, L. Adilova, J. Sicking, F. Hüger, P. Schlicht, T. Wirtz, and S. Wrobel, "Efficient decentralized deep learning by dynamic model averaging," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 393–409, Springer, 2018.



Dongzhu Liu received the B.Eng. degree from the University of Electronic Science and Technology of China (UESTC) in 2015, and the Ph.D. degree from The University of Hong Kong in 2019. She is now a postdoc research associate in the Dept. of Engineering at King's College London. Her research interests include edge intelligence, federated learning, and wireless communications.

Guangxu Zhu received the B.Eng and M.Eng degrees from Zhejiang University, and the Ph.D. degree from The University of Hong Kong in 2019. He is now a research scientist with the Shenzhen Research Institute of Big Data. His research interests include edge intelligence, distributed machine learning, 5G technologies such as massive MIMO, mmWave communication, and wirelessly powered communications. He is a recipient of the Hong Kong Postgraduate Fellowship (HKPF) and a Best Paper Award from WCSP 2013.

Kaibin Huang (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the National University of Singapore, and the Ph.D. degree from The University of Texas at Austin, all in electrical engineering. Presently, he is an associate professor in the Dept. of Electrical and Electronic Engineering at The University of Hong Kong. He received the IEEE Communication Society?s 2019 Best Tutorial Paper Award, 2015 Asia Pacific Best Paper Award, and 2019 Asia Pacific Outstanding Paper Award as well as Best Paper Awards from IEEE GLOBECOM 2006 and IEEE/CIC ICCC 2018. Moreover, he received an Outstanding Teaching Award from Yonsei University in S. Korea in 2011. He has served as the lead chairs for the Wireless Comm. Symp. of IEEE Globecom 2017 and the Comm. Theory Symp. of IEEE GLOBECOM 2014 and the TPC Co-chairs for IEEE PIMRC 2017 and IEEE CTW 2013. He has edited special issues for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING, and IEEE Communications Magazine. He is/was an Associate Editor for several major journals in wireless including IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE WIRELESS COMMUNICATIONS LETTERS, and IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He is a Distinguished Lecturer of IEEE IEEE Vehicular Technology Society and an ISI Highly Cited Researcher.

Jun Zhang (Senior Member, IEEE) received the B.Eng. degree in Electronic Engineering from the University of Science and Technology of China in 2004, the M.Phil. degree in Information Engineering from the Chinese University of Hong Kong in 2006, and the Ph.D. degree in Electrical and Computer Engineering from the University of Texas at Austin in 2009. He is an Assistant Professor in the Department of Electronic and Information Engineering at the Hong Kong Polytechnic University. His research interests include wireless communications and networking, mobile edge computing and edge learning, distributed learning and optimization, and big data analytics.

Dr. Zhang co-authored the books Fundamentals of LTE (Prentice-Hall, 2010), and Stochastic Geometry Analysis of Multi-Antenna Wireless Networks (Springer, 2019). He is a co-recipient of the 2019 IEEE Communications Society & Information Theory Society Joint Paper Award, the 2016 Marconi Prize Paper Award in Wireless Communications, and the 2014 Best Paper Award for the EURASIP Journal on Advances in Signal Processing. He also received the 2016 IEEE ComSoc Asia-Pacific Best Young Researcher Award. He is an Editor of IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, and Journal of Communications and Information Networks.