Capacity of Remote Classification Over Wireless Channels

Lan, Qiao; Du, Yuqing; Popovski, Petar; Huang, Kaibin

Published in:
I E E E Transactions on Communications

DOI (link to publication from Publisher):
10.1109/TCOMM.2021.3072735

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Capacity of Remote Classification Over Wireless Channels

Qiao Lan, *Graduate Student Member, IEEE,* Yuqing Du, *Member, IEEE,* Petar Popovski, *Fellow, IEEE,* and Kaibin Huang, *Follow, IEEE*

***Abstract*—Remote classification involves offloading complex object-recognition tasks from mobile devices to servers at the network edge. It brings to the mobile device the capability of discerning hundreds of object classes by using the computational and storage capabilities of the infrastructure. Remote classification is challenged by the finite and variable data rate of the wireless channel, which affects the capability to transfer high-dimensional features and thus limits the classification resolution. We introduce a set of metrics under the name of *classification capacity* that are defined as the maximum number of classes that can be discerned over a given communication channel while meeting a target probability for classification error. We treat both the cases of a channel where the instantaneous rate is known and unknown. The objective is to choose a subset of classes from a class library that offers satisfactory performance over a given channel. We treat two different cases of subset selection. *First,* a device can select the subset by pruning the class library until arriving at a subset that meets the targeted error probability while maximizing the classification capacity. Adopting a subspace data model, we prove the equivalence of classification capacity maximization to the problem of packing on the Grassmann manifold. The results show that the classification capacity grows *exponentially* with the instantaneous communication rate, and *super-exponentially* with the dimensions of each data cluster. This also holds for ergodic and outage capacities with fading if the instantaneous rate is replaced with an average rate and a fixed rate, respectively. In the *second* case, a device has a unique preference of class subset for every communication rate, which is modeled as an instance of uniformly sampling the library. Without class selection, the classification capacity and its ergodic and outage counterparts are proved to scale *linearly* with their corresponding communication rates instead of the exponential growth in the last case.**

I. Introduction

There is an emerging trend of deploying various *Artificial Intelligent* (AI) algorithms at the edge, away from the central cloud, to provide a context-aware and low-latency platform for supporting a wide range of applications such as Internet search (e.g., Google Lens), digital payment (e.g., Alipay's Smile to Pay), and inter-connected vehicles in 5G [1], [2]. The ubiquitous wireless connectivity results in a new paradigm merging communication and inference, called *edge inference*, referring to the broad set of techniques for deploying trained AI models at edge servers to remotely execute inference tasks posed by mobile users, such as object recognition or speech interpretation.

A large class of edge inference services can be reduced to the model in which a mobile user wirelessly uploads a multimedia data sample (photo, video or speech clip) over a wireless link, the edge server recognizes an object embedded in the sample and feeds back the object label. We term this operation *remote classification* and it is the main theme of this work. A large-scale remote classifier in the edge/central cloud is capable of classifying many object classes, more than 700 image classes for Google Cloud and 200 text classes for Tencent Cloud[1]. Maximizing the classification accuracy requires a user to upload high-dimensional features, or even large-size raw data. Two challenges emerge in this context. First, the wireless link varies due to fading and interference. Second, the latency requirements are stringent in real time and/or high-mobility applications. To address this issue, an existing remote classification service achieves the required versatility by deploying a system of classifiers with diversified capacities, which are switched according to the application requirements, input data quality, or dimensionality of input feature vectors [3], [4]. Motivated by this, we study the capacity of remote classification as a function of the communication rate offered by the wireless link.

A. Classification, Channel Coding, and Source Coding

The essence of the remote classification problem can be better understood by relating it to two classic problems in information theory: source coding and channel coding. In source coding (or compression), a transmitter represents source information using codewords that can be sent over a limited-rate channel and enable the receiver to accurately reconstruct the information [5]. In channel coding, the transmitter selects a set of codewords to which the messages are mapped and the receiver should be capable of differentiating the codewords even in the presence of channel noise, thereby decoding transmitted messages [6]. Source coding can be seen as a process of remote estimation, while channel coding as a process of remote classification in which codewords are subject to design. Remote classification can be related to coding based on the following interpretation. We can view class as "codewords" chosen by the nature and the objects as noisy instances of the

The work of Kaibin Huang was supported by Guang-dong Basic and Applied Basic Research Foundation under Grant 2019B1515130003, Hong Kong Research Grants Council under Grants 17208319 and 17209917, and Innovation and Technology Fund under Grant GHP/016/18GD.

Q. Lan, Y. Du, and K. Huang are with the Dept. of Electrical and Electronic Engineering at The University of Hong Kong, Pok Fu Lam, Hong Kong (Email: qlan@eee.hku.hk; yqdu@eee.hku.hk; huangkb@eee.hku.hk). Petar Popovski is with the Dept. of Electronic Systems at the Aalborg University, Aalborg, Denmark (Email: petarp@es.aau.dk). Corresponding author: K. Huang.

[1]See Google Cloud web (https://cloud.google.com/vision/automl/docs/resources) and Tencent Cloud web (https://cloud.tencent.com/document/product/271/36459).

classes [7]. Then the transmitter sends a description (features) of a noisy instance over a limited-rate channel, such that the receiver is able to "decode" (infer) the "codeword" (the covert class or the label of the instance). Therefore, the name of "remote classification" as used in this paper, refers to a particular remote classification process in which the classes ("codewords") are not subject to design.

Despite the similarity, there exist several fundamental differences between remote classification and source/channel coding. First, the "codewords" (classes) in the former are chosen by the nature and not subject to design as in the latter. As a result, a typical multimedia classifier cannot be derived theoretically. Instead, it is usually computed using a supervised machine learning technique, which includes choosing a suitable model [e.g., *support vector machine* (SVM) or *convolutional neural network* (CNN)] and training the model using a large labeled dataset [7]. Second, in source/channel coding, it is the transmitter that has the ground-truth information while the receiver gets an imperfect version of this information. In contrast, in remote classification, the receiver is the one responsible for inferring the ground-truth information (in the form of a label); the transmitter does not have the information and acquires it via a feedback channel. Finally, the general problem of multimedia classification can have different mathematical characteristics from those of coding such as data spaces (e.g., a feature space versus a Galois field).

Despite the differences, relating remote classification to source/channel coding creates the possibility of exploiting analytical tools from the rich literature on the latter to study the former, which benefits this work. An early work in this direction is [8] where the rate of a stand-alone classifier is found to be mathematically equivalent to the capacity of a MIMO channel with space-time modulation. In this work, we adopt a similar approach to investigate the performance of a different system of remote classification featuring a pair of separated classifier and data source that are connected using a wireless channel.

B. Edge Computing and Inference

Remote classification and edge inference at large are services supported on the edge computing architecture [1]. The current work shares the same spirit as that on computation offloading, a main theme of edge computing research, where mobile devices use unreliable wireless links to offload computation to edge servers. Remote classification can be seen as an example of computation offloading. Edge computing augments the capabilities of mobile devices while preserving their energy efficiency [9]. To reduce the devices' energy consumption, a key approach for energy efficient computation offloading is to jointly optimize radio resource allocation to multiple users and their offloaded computation loads [10]–[12]. Stochastic optimization tools, such as Lyapunov optimization, are applied to adapting offloading decisions [12] and servers' CPU frequencies [13] to the dynamics in computation tasks and channels in order to reduce both latency and power consumption. More complex techniques for accelerating offloaded computation include replicated computation at multiple servers [14], adding the new dimension of caching to the joint communication-and-computation control [15], and scheduling of computation tasks [16]. Without considering a specific application, the prior work is based on generic computation models, such that the load is measured by the number of bits and the speed by the number of bits computed per second.

Attempts on materializing the vision of edge AI has led to the emergence of edge learning (see, e.g., [17], [18], for an overview) and edge inference, which is the theme of this work. Edge learning focuses on the efficient training of AI models at the network edge [18]. Subsequently, edge inference involves the application of trained models to performing inference tasks such as object recognition and speech interpretation. Research in edge inference has resulted in several interesting design approaches. Building on the mentioned idea of replicated computation in [14], it is proposed in [19] that the association between servers (base stations) and devices can be optimized together with beamforming to reduce the total energy consumption of the devices. Several research groups have developed techniques to implement device-edge cooperative inference, where a learning task is partitioned and executed partially on device and partially offloaded to the servers [20]–[22]. To address the issue of limited computation capacity of a device, a CNN model can be pruned before partitioning, and the idea can be implemented using the techniques in [20]. There also exist techniques for channel adaptive model partitioning and coding [21]. Furthermore, the model partitioning can be adjusted according to the allocated bandwidth and the requirements on latency and inference accuracy, which is the approach advocated in [22]. In addition, data compression for communication-efficient edge inference has also been investigated. For example, a relevant architecture is proposed in [23]. A *deep neural network* (DNN) encoder is deployed at a transmitter to compress raw data. After being received, the compressed data is decoded by the server using a DNN decoder before feeding the output into another DNN model for inference. In view of prior works, they are focused on technique design and rely on experiments for performance evaluation. There exist few results on the fundamental limits of edge inference systems under the constraint of wireless channels connecting severs and devices, which motivates the current work.

C. Contributions and Organization

The objective of this work is to make the first attempt on quantifying the performance of a remote classification system under a *communication channel constraint*, referring to the finite and time-varying rate of a wireless communication channel. To this end, we consider a system in which a mobile device sends a feature vector over a wireless channel to an edge server, which performs classification and sends the result to the mobile device. The server supports classification of an arbitrary subset of a class library based on a mainstream architecture of large-scale classification (see e.g., [24]). On the one hand, even if the communication rate is sufficiently large for transmitting all features of each sample, classification errors can still occur as an inherent effect of *data noise*,

which is caused by the natural factors in sensing (e.g., pose, perspective, lighting, and background). On the other hand, as the rate varies and so does the received number of features per sample, if the classification error probability should be constrained, the maximum number of object classes that are chosen to be discerned by the remote classifier has to be adapted to the rate in a similar way as the maximum constellation order of adaptive modulation. This gives rise to a performance metric called *ϵ-classification capacity* defined as the maximum number of classes that can be discriminated under the channel constraint and for a given target classification error probability[2]. Furthermore, two derivative metrics, called *ergodic* and *outage classification capacities*, are defined to account for the effect of fading, which correspond to adaptive and fixed coding rates, respectively. Using these metrics, the system performance is analyzed for two cases.

(1) **Class-selection case**: This case corresponds to practical applications that adapt the classification resolution to the feature transmission rate. Consider the example of recognition of objects on the road. At a low-transmission rate, the remote classifier is required to discriminate fewer object classes with high differentiability such as vehicles and pedestrians. As the rate increases, the classifier is capable of discriminating more object classes with low differentiability such as (vehicles) cars and trucks, as well as (pedestrians) adults and children.

(2) **Random-class case**: This case corresponds to practical applications that have no flexibility in choosing object classes for recognition. For a given communication rate, the user makes a uniformly random choice of class subset from the library subject to the rate constraint. For instance, a remote classifier for intelligent transportation may receive a task in type of vehicle classification, traffic sign classification or obstacle classification from a user at a time slot.

The scope of contributions made by this work is described as follows. For tractability, we follow the relevant work in [8] to adopt a statistical data model from the area of linear regression and a matching subspace *maximum-likelihood* (ML) classifier. Though an alternative classifier model, namely a neural network, is also considered in experiments, tractable analysis of its classification capacity remains an open problem. For the current analysis, it is sufficient to use a generic model of wireless channel characterized by a time varying rate. Specific physical-layer techniques such as MIMO, OFDM and NOMA for supporting the rate are not explicitly considered. As a remark, there are several differences between this work and [8] as follows. First of all, [8] considers linear classification with compression while compression has not been included in this paper. Moreover, the derivation pathways of two papers are different. Finally, the classifier in [8] is stand-alone while, in this work, the data source is connected to a remote (separated) classifier via a wireless channel. Consequently, revealing impact of wireless factors on remote classification *quality of experience* (QoE) should be regarded as a major improvement.

[2]In the following text, it will be always implicitly assumed that there is a target classification probability that needs to be met.

The main contributions of the work are summarized as follows.

- **Classification capacity with class selection:** Consider the mentioned class-selection case. For a large library, the problem of maximizing the ϵ-classification capacity by class selection is shown to be equivalent to the mathematical problem of packing on a Grassmann manifold. The relation allows the application of packing results together with error probability analysis of space-time modulation to derive bounds on the maximum capacity. The results reveal the *exponential growth* of the capacity with the communication rate and *super-exponential growth* with the dimensions of each data cluster. Based on the results and considering Rayleigh fading, the scaling laws of ergodic and outage classification capacities are investigated. They are proved to scale as stated above if the communication rate is replaced by its ergodic counterpart or the maximum rate under an outage constraint.
- **Classification capacity with random classes:** Consider the other case of random classes. The expected classification error probability is related to the isotropic distribution on a Grassmann manifold. Applying relevant results allow the derivation of a lower bound on the classification capacity, which increases *linearly* with the communication rate. Lower bounds on ergodic and outage classification capacities with Rayleigh fading are also derived and shown to follow the same scaling law.
- **Extension to fast fading:** The preceding results based slow fading are extended to the case with fast fading, resulting in a *random* number of features used for remote classification of each data sample. It is found that fast fading does not change the classification-capacity scaling laws except for adding to the communication rates the multiplicative factor equal to some packet-success probability.
- **Experiment results:** Experiments based on both the statistical data model and real-world datasets (e.g., MNIST) are conducted to demonstrate the effects of wireless channel on the capacities of remote classification and the classification capacity gains of the class-selection case *with respect to* (w.r.t.) the random-class case.

Organization: The remainder of the paper is organized as follows. The models and performance metrics are introduced in Section II. Section III presents the analysis on classification capacities with class selection while that for the random-class case is investigated in Section IV. The derived results are further extended to fast fading channels in Section V. Section VI provides the experimental results, followed by concluding remarks in Section VII.

II. Models and Metrics

Consider the remote classification system in Fig. 1, where an edge device transmits feature vectors, extracted from data samples, to an edge server for classification using a trained model and receives from the server the inferred labels. The specific models and performance metrics are described as follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3072735, IEEE Transactions on Communications

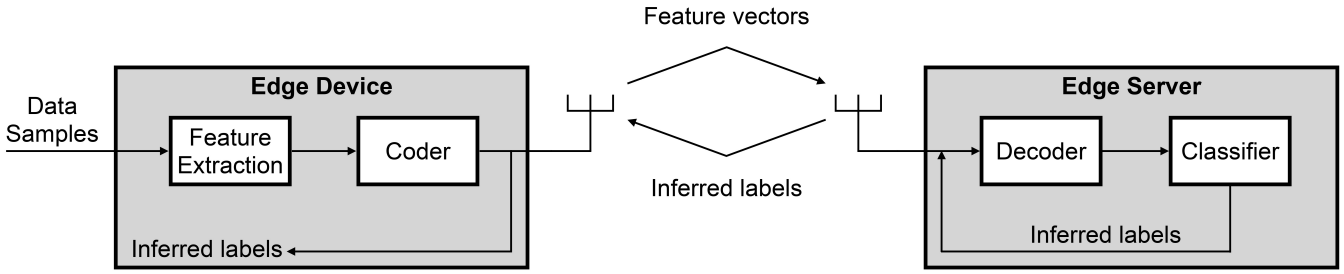


Figure 1. Remote classification system.

A. Classification Model

As in [8], we consider the classic statistical problem of classifying linear subspaces. The statistical data model and ML classifier are described as follows.

1) Statistical data model: Consider a clustered dataset comprising L separable classes, where the i-th class centroid is represented by a unitary matrix $\mathbf{U}_i \in \mathbb{R}^{N\times K}$ with $N \geq K$ and $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}_K$. An arbitrary data sample, denoted as $\tilde{\mathbf{x}}$, that belongs to the i-th class is modeled as [8]

$$\tilde{\mathbf{x}} = \Phi\mathbf{U}_i\mathbf{s} + \tilde{\mathbf{w}}, \tag{1}$$

where the unitary matrix $\Phi \in \mathbb{R}^{\tilde{N}\times N}$ represents the discriminant subspace embedded in the raw-data space, $\mathbf{s} \in \mathbb{R}^K$ results from the projection of the data sample into the class subspace $\mathbf{U}_i$, and $\tilde{\mathbf{w}} \in \mathbb{R}^{\tilde{N}}$ accounts for both the error in fitting the dataset distribution to the subspace model as well as the mentioned data noise. Note that $\tilde{\mathbf{w}}$ is the inherent cause of classification errors even in the absence of channel constraint. The random vector $\mathbf{s} \in \mathbb{R}^K$ is assumed to consist of *independent and identically distributed* (i.i.d.) $\mathcal{N}(0, \sigma_{\mathbf{s}}^2)$ elements, where $\mathcal{N}(0, \sigma_{\mathbf{s}}^2)$ denotes a zero-mean normal distribution with variance $\sigma_{\mathbf{s}}^2$. To compress the sample, an N-dimensional feature vector, denoted as $\mathbf{x}$, is extracted from $\tilde{\mathbf{x}}$ by projecting it onto the discriminant subspace:

$$\mathbf{x} = \Phi^T\tilde{\mathbf{x}} = \mathbf{U}_i\mathbf{s} + \mathbf{w}, \tag{2}$$

where $\mathbf{w} = \Phi^T\tilde{\mathbf{w}}$ comprises i.i.d. normal distribution $\mathcal{N}(0, \sigma_{\mathbf{w}}^2)$ elements and is hereafter simply referred to as data noise. The subspace Φ is assumed to be known to the server for calibrating the needed classifier; when Φ is determined by the sever, the operation is known in the literature as *feature selection*. Based on (2), the data model can be parameterized by the subspace set $\mathcal{U}_L = \{\mathbf{U}_\ell\}$.

Definition 1. *(Data SNR).* The *signal-to-noise ratio* (SNR) of the dataset is defined as the ratio between the variance of each cluster and that of data noise:

$$\text{Data SNR} = \frac{\sigma_{\mathbf{s}}^2}{\sigma_{\mathbf{w}}^2} = \sigma_{\mathbf{s}}^2, \tag{3}$$

where we set $\sigma_{\mathbf{w}}^2 = 1$ without loss of generality.

2) Maximum-likelihood remote classifier: Conditioned on $\mathbf{U}_i$, the *probability density function* (PDF) of $\mathbf{x}$ is given as

$$\begin{aligned} P(\mathbf{x}|\mathbf{U}_i) &= \frac{\exp\left(-\frac{1}{2}\mathbf{x}^T\left(\sigma_{\mathbf{s}}^2\mathbf{U}_i\mathbf{U}_i^T + \mathbf{I}_N\right)^{-1}\mathbf{x}\right)}{(2\pi)^{N/2}\det^{1/2}\left(\sigma_{\mathbf{s}}^2\mathbf{U}_i\mathbf{U}_i^T + \mathbf{I}_N\right)} \\ &= \frac{\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x} + \frac{\sigma_{\mathbf{s}}^2}{2(1+\sigma_{\mathbf{s}}^2)}\mathbf{x}^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{x}\right)}{(2\pi)^{N/2}\left(1+\sigma_{\mathbf{s}}^2\right)^{K/2}}. \end{aligned} \tag{4}$$

Given the knowledge of $\{\mathbf{U}_i\}_{i=1}^L$, the classifier estimates the class of a reliably received feature vector $\mathbf{x}$, say $\mathbf{U}_i$, (or equivalently the label i) by maximizing the above PDF:

$$\widehat{i} \triangleq \arg\max_{i\in\{1,2,\ldots,L\}} p(\mathbf{x}|\mathbf{U}_i) = \arg\max_{i\in\{1,2,\ldots,L\}} \mathbf{x}^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{x}, \tag{5}$$

which is a well-known ML classifier [26].

Remark 1. *(Geometric Interpretation).* The operation $\mathbf{U}_i\mathbf{U}_i^T\mathbf{x}$ projects the feature vector $\mathbf{x}$ onto the subspace, span$\{\mathbf{U}_i\}$. It gives the geometric interpretation that the ML classifier essentially aims at identifying the subspace forming the smallest angle with (or equivalently having the smallest subspace distance to) the feature vector $\mathbf{x}$.

B. Communication Model

Time is divided into slots, each of which has the duration of T seconds. Each feature is quantized into a sufficiently large number of bits, denoted as Q, such that distortion is negligible. The channel code is designed such that each quantized feature vector is encoded into a single codeword transmitted using one slot. The variation of the channel with bandwidth B is assumed to be slow w.r.t. the slot duration such that the channel remains constant within each slot but varies over slots. The extension to the scenario of fast channel variation is presented in Section V. Let R denote the communication rate (bit/s) of the channel. Both the cases of channel adaptive and fixed coding rates are considered as discussed in the sequel. As an example, given the transmit SNR (denoted as ρ) and without *channel state information at the transmitter* (CSIT), the rate for a *single-input-single-output* (SISO) channel is $R = B\log_2(1+\rho|h|^2)$, where h denotes the channel gain. As another example, the rate for a *multiple-input-multiple-output* (MIMO) channel is $R = B\log_2\det\left|\mathbf{I} + \frac{\rho}{N_t}\mathbf{H}\mathbf{H}^H\right|$ where $\mathbf{H}$ denotes the channel matrix and N_t the number of transmit antennas. The finite communication rate introduces a constraint on the feature-

space dimension (i.e., the number of features of each data sample) $N = \beta R$ where $\beta = \frac{T}{Q}$.

C. Performance Metrics

To facilitate defining the performance metrics, the notion of (object) *class library* is first formalized. In practical remote-classification, the server supports a large library of classes and can generate an active classifier for a user based on the chosen subset of classes (see e.g., [24]). The class library is represented by $\mathcal{F} = \{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_M\}$ where each element is a subspace matrix representing an available class. The L-class subset chosen by a user is specified by the subspace set $\mathcal{U}_L$ with $\mathcal{U}_L \subset \mathcal{F}$, which determines the dataset distribution.

1) Classification error probability: Labels inferred by the remote classifier can be erroneous as an inherent effect of data noise even though the channel is reliable and even if its rate is sufficiently large to transfer all features. A classification error is declared if the inferred label is different from the ground truth. Conditioned on the data distribution specified by $\mathcal{U}_L$ and the communication rate R, the classification error probability, denoted as P_{e}, can be written as

$$P_{\mathsf{e}}(R,\mathcal{U}_L) \triangleq \frac{1}{L}\sum_{\ell=1}^{L} \Pr\left(\mathcal{L}(\mathbf{x}) \neq \ell \mid y = \ell, \mathcal{U}_L, R\right), \quad (6)$$

where $\mathcal{L}$ denotes the classifier function mapping the input feature vector to the inferred label and y is the ground-truth label. Note the above definition assumes that the prior probability that the object $\mathbf{x}$ belongs to one of the L classes is uniform, as in [8] and R determines the length of $\mathbf{x}$ as described in the sequel. This classification error probability can be considered as a measure of the QoE of remote classification services. In addition, the future extension to the case with non-uniform probabilities requires modifying the classifier model by adding prior-dependent weights to the likelihoods of different labels.

2) ϵ-classification capacity: Recall that the metric, denoted as C, is defined as the maximum number of classes that can be discriminated given an instantaneous communication rate, R, such that the classification error probability, P_{e}, is no larger than a given threshold $\epsilon \in (0,1)$. Conditioned on R and $\mathcal{U}_L$, the error probability can be written as the function $P_{\mathsf{e}}(R,\mathcal{U}_L)$. Using the notation, the *ϵ-classification capacity* for the class-selection case can be defined as:

$$C^{\mathsf{sel}}(R) = \sup_{\mathcal{U}_L \in \mathcal{F}, L} \left\{L \mid P_{\mathsf{e}}(R,\mathcal{U}_L) \leq \epsilon\right\}. \quad (7)$$

The counterpart for the random-class case is defined as

$$C^{\mathsf{rnd}}(R) = \sup_{L} \left\{L \mid \mathbb{E}_{\mathcal{U}_L}\left[P_{\mathsf{e}}(R,\mathcal{U}_L)\right] \leq \epsilon\right\}, \quad (8)$$

where the expectation is over the distribution of the classes, $\mathcal{U}_L$, given L.

3) Ergodic classification capacity: Consider the case where the device has CSIT and adapts the number of features per sample as well as coding rate to the channel state. Then we can define the *ergodic classification capacity* as:

$$\bar{C} = \begin{cases} \mathbb{E}_R\left[C^{\mathsf{sel}}(R)\right], & \text{class-selection case;} \\ \mathbb{E}_R\left[C^{\mathsf{rnd}}(R)\right], & \text{random-class case,} \end{cases} \quad (9)$$

where the expectations are over the distribution of communication rate R, $C^{\mathsf{sel}}(R)$ is defined in (7) and $C^{\mathsf{rnd}}(R)$ in (8).

4) Outage classification capacity: A different communication model is adopted where either the CSIT is unavailable or some form of channel inversion is used such that the channel cannot be inverted when its gain is below a given threshold. As a result, the device fixes the number of features per sample and coding rate, resulting in a required communication rate, r, for successful decoding of a received feature vector. Then a channel outage event is one that the channel capacity falls below a given threshold r, yielding the outage probability defined as

$$P_{\mathsf{out}}(r) \triangleq \Pr(R \leq r). \quad (10)$$

Under an outage constraint, $P_{\mathsf{out}} \leq \delta$, and a fixed transmit SNR, there exists a maximum rate of r. Then the outage classification capacity is defined as

$$C_{\mathsf{out}} = \begin{cases} \max\limits_{r} \left\{C^{\mathsf{sel}}(r) \mid P_{\mathsf{out}}(r) \leq \delta\right\}, \text{ class-selection case} \\ \max\limits_{r} \left\{C^{\mathsf{rnd}}(r) \mid P_{\mathsf{out}}(r) \leq \delta\right\}, \text{ random-class case,} \end{cases} \quad (11)$$

where $C^{\mathsf{sel}}(r)$ and $C^{\mathsf{rnd}}(r)$ are defined in (7) and (8), respectively.

Remark 2. *(Effective Classification Error Probability).* For remote classification, when the channel is in outage, the server receives zero features for a transmitted sample but may make a random guess on the sample's label with the error probability of $\frac{L-1}{L}$. If this is the case, the effective classification error probability is slightly larger than its constraint ϵ and should be given as $(1-\delta)\epsilon + \delta\frac{L-1}{L}$.

III. Classification Capacity with Class Selection

In this section, the ϵ-classification capacity and its ergodic and outage counterparts for the class-selection case are analyzed.

A. Classification Error Probability

To facilitate the derivation of classification capacities under a constraint on the classification error probability, we first analyze the probability as follows.

1) Pairwise classification error probability: Consider the classification of two specific classes, namely $\mathbf{U}_i$ and $\mathbf{U}_j$. The error probability of binary classification based on a similar data distribution model as the current one was studied in [29] in the context of space-time demodulation. Let "$i \to j$" denote the event that a sample of class i is assigned label j by the classifier. Then the *pairwise classification error probability* (PCEP) can be defined as $P(i \to j) = \Pr\left(\mathcal{L}(\mathbf{x}) = j \mid y = i, \mathcal{U}_L, R\right)$. A main result from [29] is given below.

Lemma 1. *(Exact Pairwise Classification Error Probability [29]).* The probability is given as

$$P(i\to j)=\frac{1}{4\pi}\int_{-\infty}^{\infty} dw \frac{1}{w^2+1/4} \cdot \prod_{\substack{k=1\\ \cos\theta_k^{(i,j)}<1}}^{K}\left[\frac{\left(1+\sigma_{\mathbf{s}}^2\right)\sigma_{\mathbf{s}}^{-4}}{\left(1-\cos^2\theta_k^{(i,j)}\right)(\omega^2+a_k^2)}\right]^{\frac{1}{2}} \quad (12)$$

where $\theta_k^{(i,j)}$ denotes the k-th principal angle between $\mathbf{U}_i$ and $\mathbf{U}_j$, and $a_k=\sqrt{\frac{1}{4}+\frac{\sigma_{\mathbf{s}}^2+1}{\sigma_{\mathbf{s}}^4\left(1-\cos^2\theta_k^{(i,j)}\right)}}$.

Note that the effect of the number of features per sample, N, (or the proportional communication rate, R) is not reflected in the above result given a fixed distance between $\mathbf{U}_i$ and $\mathbf{U}_j$, measured by $\{\cos^2\theta_k^{(i,j)}\}$. The effect of N (or R) lies in determining the dimensionality of the feature space and hence the number of classes that can be packed into the space as elaborated in Section III-B. To simplify analysis and gain insight, we further derive an upper and a lower bounds on the probability in the following lemma, proven in Appendix A.

Lemma 2. *(PCEP Bounds).* The PCEP can be bounded as

$$\frac{1}{3}\left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)d_{i,j}^2}\right)^K \le P(i\to j) \le \frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{\lfloor d_{i,j}^2\rfloor}{2}}, \quad (13)$$

where $g(\sigma_{\mathbf{s}}^2)=\frac{1}{4\left(\sigma_{\mathbf{s}}^{-4}+\sigma_{\mathbf{s}}^{-2}\right)}$ is a monotonically increasing function of the data SNR $\sigma_{\mathbf{s}}^2$ and $d_{i,j}=\sqrt{K-\mathrm{tr}\{\mathbf{U}_i\mathbf{U}_i^T\mathbf{U}_j\mathbf{U}_j^T\}}$ denotes the (chordal) subspace distance between the two classes $\mathbf{U}_i$ and $\mathbf{U}_j$.

Remark 3. *(Effects of Data SNR and Class Distance).* On the one hand, increasing the data SNR causes data clusters to shrink, improving their discernibility. For this reason, it is observed that both bounds on the PCEP in the above lemma decrease as the data SNR grows. On the other hand, the subspace distance between two classes determines their differentiability. Consequently, increasing the distance reduces the bounds on the PCEP. The improvement is known as the *discrimination gain* in the literature.

2) Classification error probability of L classes: Consider the error events $\{i\to j \mid i\neq j\}$ and the pairwise classification error probability analyzed in the preceding subsection. By the *union bound* and invoking (13), the probability can be bounded in terms of the pairwise counterpart as

$$P_{\mathsf{e}} = \frac{1}{L}\sum_{i=1}^{L}\Pr\left(\bigcup_{j\neq i}(i\to j)\right) \overset{(a)}{\le} \frac{1}{L}\sum_{i=1}^{L}\sum_{j\neq i}P(i\to j) \overset{(b)}{=} \frac{2}{L}\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}P(i\to j), \quad (14)$$

where (a) is due to the union bound and (b) due to the symmetry $P(i\to j)=P(j\to i)$. Define $d_{\mathsf{min}}=\min_{(i,j)}d_{i,j}$, the above bound can be further relaxed to give the upper bound in Lemma 3 in the sequel.

Next, P_{e} can be lower bounded as follows. Define $\mathcal{W}_{i,j^*}$ as an event that the ground-truth label is i while the inferred label is $j^*\neq i$ subject to $d_{i,j^*}=\min_{j\neq i}d_{i,j}\triangleq d_{\min}^{(i)}$. Then it follows from (6) that one lower bound of the classification error probability can be calculated as

$$P_{\mathsf{e}}\ge\frac{1}{L}\sum_{i=1}^{L}\Pr\left(\mathcal{W}_{i,j^*}\mid i,\ j^*=\arg\min_{j\neq i}d_{i,j}\right), \quad (15)$$

yielding the lower bound in Lemma 3.

Lemma 3. *(Classification Error Probability).* Given the class subspace set $\mathcal{U}_L$, the classification error probability can be bounded as

$$\frac{1}{3L}\sum_{i=1}^{L}\left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)\left(d_{\min}^{(i)}\right)^2}\right)^K \le P_{\mathsf{e}} \le \frac{L}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{\lfloor d_{\min}^2\rfloor}{2}}.$$

Remark 4. *(Effect of Number of Classes).* Apart from the effects of data SNR and class distance discussed earlier, one can further observe/infer from the above bounds that increasing the number of classes, L, makes the classification error probability grow. This is because that packing more classes into a fixed feature space reduces inter-class distances and thereby compromise their differentiability.

B. ϵ-Classification Capacity

The key step in deriving the ϵ-classification capacity is to establish the equivalence between the classification capacity maximization via class selection and the Grassmannian packing problem. To facilitate the analysis, we consider the scenario where the large-scale remote classifier at the server can support flexible classification as stated in the following assumption.

Assumption 1. *(Flexible Classification).* The server with a large class library supports classification of an arbitrary dataset (parameterized by a subspace set $\mathcal{U}_L$) with the classification error probability $P_{\mathsf{e}}(R,\mathcal{U}_L)$ in (7).

In practice, large-scale classification realizes flexible classification using a hierarchical architecture comprising a large number of component classifiers [24], [25].

1) Equivalence to Grassmannian packing: Given (N,K), a Grassmann manifold, $\mathcal{G}(N,K)$, refers to the space of K-dimensional subspaces embedded in the N-dimensional space, or equivalently the space of $N\times K$ unitary matrices. Based on the definition of ϵ-classification capacity in (7) and Assumption 1, the subspace set $\mathcal{U}^*$ that represents the class selection for capacity maximization can be found by solving the following optimization problem:

$$(\mathbf{P1})\quad \mathcal{U}^*=\arg\max_{\mathcal{U}\in\mathcal{G}} C(\mathcal{U}) \quad \text{s.t.} \quad P_{\mathsf{e}}(\mathcal{U})\le\epsilon,$$

where $\mathcal{G} = \mathcal{G}(N,K)$, $C(\mathcal{U}) = C(R,\mathcal{U})$ and $P_{\mathsf{e}}(\mathcal{U}) = P_{\mathsf{e}}(R,\mathcal{U})$ with N, K, and R in this subsection and omitted to simplify notation. Substituting the upper bound on P_{e} in Lemma 3 into $(\mathbf{P1})$, the problem can be recast as

$$(\mathbf{P2}) \quad \begin{aligned} \mathcal{U}^* = \arg \max_{|\mathcal{U}|=L,\ \mathcal{U}\in\mathcal{G}} \quad & L \\ \text{s.t.} \quad & d_{\min} \geq \beta_L, \end{aligned}$$

where $\beta_L = \sqrt{\frac{\log_2 \frac{L}{2\epsilon}}{\log_2(1+g(\sigma_{\mathbf{s}}^2))}}$. The solution of $(\mathbf{P2})$ lower bounds the maximum capacity from solving $(\mathbf{P1})$ and the approximation is accurate when the error probability is small. An intuitive interpretation of Problem $(\mathbf{P2})$ is to pack as many balls as possible (maximizing L), each centered at an element of $\mathcal{U}_L$ and with the radius $\frac{d_{\min}}{2}$, into the space $\mathcal{G}$, giving the name *Grassmannian packing* [27]. A standard approach for solving this class of mathematical problems is to convert them into equivalent problems of maximizing the minimum separation distance among L balls [27]:

$$\text{(Grassmannian Packing)} \quad \mathcal{U}^* = \arg \max_{\substack{\mathcal{U}\in\mathcal{G}, \\ |\mathcal{U}|=L}} d_{\min}. \tag{16}$$

Let $d^*_{\min}(L)$ denote the result from solving the above problem, called minimum class separation from packing. Then, L is increased to reach the maximum value under the constraint $d^*_{\min}(L) \geq \beta_L$, thereby solving the original Problem $(\mathbf{P1})$.

Though typically Grassmannian packing problems are intractable and usually solved numerically, there exists a rich literature on bounding the resultant minimum distance $d^*_{\min}(L)$. (see e.g., [27]). The following particular result is from [28].

Lemma 4. *(Packing Bounds [28]).* For large feature-space dimensions N, the minimum class separation distance from Grassmannian packing can be bounded as

$$KL^{-\frac{2}{NK}} \lesssim [d^*_{\min}(L)]^2 \lesssim 2K\left[1-\left(1-L^{-\frac{2}{NK}}\right)^2\right],\ N\to\infty. \tag{17}$$

2) Packing bounds on ϵ-classification capacity: Using Lemmas 3 and 4, the ϵ-classification capacity defined in (7) can be bounded as $C_{\mathsf{lb}} \leq C^{\mathsf{sel}}(R) \leq C_{\mathsf{ub}}$ with

$$C_{\mathsf{lb}} = \left\{ L : \frac{L}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{\lfloor d^2_{\mathsf{lb}} \rfloor}{2}} = \epsilon \right\}, \tag{18}$$

$$C_{\mathsf{ub}} = \left\{ L : \frac{1}{3}\left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)d^2_{\mathsf{ub}}}\right)^{K} = \epsilon \right\}, \tag{19}$$

where $d^2_{\mathsf{lb}} = KL^{-\frac{2}{NK}}$, $d^2_{\mathsf{ub}} = 2K\left(1-\left(1-L^{-\frac{2}{NK}}\right)^2\right)$ and (19) follows by substituting $\{d^{(i)}_{\mathsf{min}}\}$ in the lower bound of Lemma 3 with d_{ub} as $[d^{(i)}_{\mathsf{min}}]^2 \leq K \leq d^2_{\mathsf{ub}}, \forall i$. Solving the two equations (18) and (19) and substituting $N = \beta R$ yield the following theorem.

Theorem 1. *(ϵ-Classification Capacity with Class Selection).* Consider the class-selection case. For a high communication rate, the capacity can be asymptotically bounded as:

$$\begin{aligned} & 2^{\frac{\beta R}{2}(K\log_2 K + Kc_{\sigma_{\mathbf{s}}} - Kc_\epsilon)} \\ \lesssim\ & C^{\mathsf{sel}}(R) \\ \lesssim\ & 2^{\frac{\beta R}{2}\left(K\log_2 4K + K\log_2 \frac{4g(\sigma_{\mathbf{s}}^2)}{1-3\epsilon}\right)}, \quad R\to\infty, \end{aligned} \tag{20}$$

where $c_{\sigma_{\mathbf{s}}^2} = \log_2\log_2(1+g(\sigma_{\mathbf{s}}^2))$ and $c_\epsilon = \log_2\log_2\frac{1+g(\sigma_{\mathbf{s}}^2)}{2\epsilon}$. In particular, as $R, K \to \infty$, the capacity scales as

$$\lim_{R,K\to\infty} \frac{\log_2 C^{\mathsf{sel}}(R,K)}{RK\log_2 K} = \frac{\beta}{2}. \tag{21}$$

Proof: See Appendix B. □

Remark 5. *(Mathematical Intuition for Capacity Scaling Laws).* One can observe from the above theorem that the ϵ-classification capacity increases *exponentially* as the instantaneous communication rate R grows. The underpinning mathematical reason is that the volume of the Grassmann manifold containing the dataset classes is an *exponential function* of its dimensions N, which is proportional to R. Consequently, increasing R allows an exponentially growing number of "balls" (classes) to be packed into the manifold. One the other hand, the capacity scales *super-exponentially* with the dimensions of each data cluster (or each class), namely K. Note that increasing K improves the inter-class differentiability. One can infer from (18) and (19) that with the classification error probability fixed, the allowed number of "balls" (L) grows *exponentially* as the minimum pairwise distance of the "balls" (classes), $d^*_{\min}$, increases. Furthermore, $d^*_{\min}$ is a *super-linear* function of K as one can further observe from the definitions of d^2_{ub} and d^2_{lb} after (19). Combining the two relations gives the super-exponential capacity scaling w.r.t. K.

Remark 6. *(Effects of QoE Requirement and Data/Transmit SNR).* The dependence of the ϵ-classification capacity on the allowed maximum classification error probability ϵ (or QoE requirement), the data SNR $\sigma_{\mathbf{s}}^2$, and the transmit SNR ρ can be interpreted geometrically in terms of Grassmannian packing. Increasing ϵ, $\sigma_{\mathbf{s}}^2$ and ρ allows "balls" (classes) to get closer, shrinks "ball radiuses" (the variance of each data cluster), and increasing the Grassmannian volume (communication rate), respectively. They all contribute to packing more "balls" (larger capacity) though in different ways.

C. Ergodic and Outage Classification Capacities

Given a distribution function of the communication rate R, it is straightforward to use the results in Theorem 1 to analyze the ergodic and outage classification capacities based on their definitions in (9) and (11). In this section, we consider a Rayleigh fading channel and perform such analysis to provide concrete insight into the effect of channel fading on the performance of remote classification.

1) Ergodic classification capacity: Correspondently, the ergodic channel capacity is $\bar{R} = \mathbb{E}[B\log_2(1+\rho|h|^2)]$, where ρ is the transmit SNR and the channel gain $|h|^2 = \exp(1)$.

Proposition 1. *(Ergodic Classification Capacity for Rayleigh Fading).* Consider the class-selection case. The ergodic classification capacity defined in (9) can be bounded as

$$\sqrt{2\pi\gamma_{\mathsf{lb}}}\cdot\rho^{\gamma_{\mathsf{lb}}}\cdot e^{\gamma_{\mathsf{lb}}(\log\gamma_{\mathsf{lb}}-1)} \le \bar{C} \le \sqrt{2\pi\gamma_{\mathsf{ub}}}\cdot\rho^{\gamma_{\mathsf{ub}}}\cdot e^{\gamma_{\mathsf{ub}}(\log\gamma_{\mathsf{ub}}-1)}, \tag{22}$$

where $\gamma_{\mathsf{lb}} = \frac{\beta B}{2}\left(K\log_2 K + Kc_{\sigma_{\mathbf{s}}} - Kc_\epsilon\right)$ and $\gamma_{\mathsf{ub}} = \frac{\beta B}{2}\left(K\log_2 4K + K\log_2\frac{4g(\sigma_{\mathbf{s}}^2)}{1-3\epsilon}\right)$ with $c_{\sigma_{\mathbf{s}}}$ and c_ϵ defined in Theorem 1. In particular, for large $\bar{R}$ and K, the capacity scales as

$$\lim_{\bar{R},K\to\infty}\frac{\log_2\bar{C}}{\bar{R}K\log_2 K} = \frac{\beta}{2}, \tag{23}$$

where $\beta = \frac{T}{Q}$.

Proof: See Appendix C. □

Remark 7. *(Fading Does Not Affect Capacity Scaling).* The key observation from the above proposition is that both the scaling laws of the ergodic classification capacity are the same as those for ϵ-classification capacity in Theorem 1 except for replacing the instantaneous rate R with its ergodic counterpart $\bar{R}$. The remark also applies to outage classification capacity analyzed in the sequel if the communication rate is modified as the maximum rate under an outage constraint.

2) Outage classification capacity: To begin with, the maximum communication rate, denoted as R_δ, can be obtained from the active outage constraint $P_{\mathsf{out}} = \Pr(R \le R_\delta) \le \delta$ and the exponential distribution of the channel gain as

$$R_\delta = B\log_2\left(1+\rho\log\left(\frac{1}{1-\delta}\right)\right). \tag{24}$$

It is worth mentioning that R_δ is a monotonically increasing function of the outage probability δ. Moreover, note that the corresponding number of transmitted features per sample is now given as $N = \beta R_\delta$. The outage classification capacity is equal to the ϵ-classification capacity by replacing R with R_δ in (24), yielding the following proposition.

Proposition 2. *(Outage Classification Capacity for Rayleigh Fading)* Consider the class-selection case. The outage classification capacity defined in (11) can be bounded as

$$\left[1+\rho\log\left(\frac{1}{1-\delta}\right)\right]^{\gamma_{\mathsf{lb}}} \le C_{\mathsf{out}} \le \left[1+\rho\log\left(\frac{1}{1-\delta}\right)\right]^{\gamma_{\mathsf{ub}}}, \tag{25}$$

with $\rho \gg 1$, γ_{lb} and γ_{ub} defined in Proposition 1. In particular, as $R_\delta, K \to \infty$, the capacity scales as

$$\lim_{R_\delta,K\to\infty}\frac{\log_2 C_{\mathsf{out}}}{R_\delta K\log_2 K} = \frac{\beta}{2}. \tag{26}$$

Proof: See Appendix D. □

IV. Classification Capacity with Random Classes

In this section, the ϵ-classification capacity and its ergodic and outage counterparts are analyzed for the random-class case and compared with their counterparts in the class-selection case.

A. Expected Classification Error Probability

The expected classification error probability is analyzed in this subsection for a dataset with i.i.d. isotropic classes, $\{\mathbf{U}_\ell\}$, on the Grassmannian $\mathcal{G}(N,K)$.

1) Distribution of class separation: Let $\theta_{\max}$ denote the maximum principal angle between a pair of classes, $\mathbf{U}_i$ and $\mathbf{U}_j$.

Lemma 5. *(Class Separation Distribution [30]).* The PDF of $X = \sin^2\theta_{\max}$ is given as

$$f_X(x) = c_{N,K,\theta_{\max}} \cdot {}_2F_1\left(\frac{N-K-1}{2},\frac{1}{2};\frac{N+1}{2};\sin^2\theta_{\max}\mathbf{I}_{K-1}\right), \tag{27}$$

where $c_{N,K,\theta_{\max}} = K(N-K)\frac{\Gamma(\frac{K+1}{2})\Gamma(\frac{N-K+1}{2})}{\sqrt{\pi}\Gamma(\frac{N+1}{2})}(\sin\theta_{\max})^{K(N-K)-1}$ and ${}_2F_1(\cdot)$ denotes the Gaussian hypergeometric function with a matrix argument.

The squared chordal distance between $\mathbf{U}_i$ and $\mathbf{U}_j$ is defined as $d_c^2(\mathbf{U}_i,\mathbf{U}_j) = K - \mathrm{tr}\{\mathbf{U}_i\mathbf{U}_i^T\mathbf{U}_j\mathbf{U}_j^T\}$. Using Lemma 5, an upper bound on the *cumulative distribution function* (CDF) of the distance is derived as shown in the lemma below.

Lemma 6. *(Upper Bound on Class Separation Distribution).* Consider a pair of independent and isotropic classes $\mathbf{U}_i$ and $\mathbf{U}_j$ on the Grassmannian $\mathcal{G}(N,K)$. The CDF of their squared chordal distance $d_c^2(\mathbf{U}_i,\mathbf{U}_j)$, denoted as $F_{d_c^2}(x)$, can be bounded as

$$F_{d_c^2}(x) \le \left(\frac{x}{K}\right)^{\frac{K(N-K)}{2}},\ x\in[0,K]. \tag{28}$$

Proof: See Appendix E. □

2) Expected classification error probability: Consider the ML classification of two random classes. Using Lemmas 2 and 6, the expected PCEP can be bounded as follows.

Lemma 7. *(Upper Bound on Expected PCEP).* For a pair of independent and isotropic random classes $\mathbf{U}_i$ and $\mathbf{U}_j$, the expected PCEP can be upper-bounded as

$$\mathbb{E}[P(i\to j)] \le \frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{K}{2}} + \frac{\log\left(1+g(\sigma_{\mathbf{s}}^2)\right)}{(1+g(\sigma_{\mathbf{s}}^2))}\frac{1}{N}. \tag{29}$$

Proof: See Appendix F. □

By applying the union bound and using Lemma 7, we obtain the following lemma.

Lemma 8. *(Expected Classification Error Probability).* For a dataset having L independent and isotropic classes $\mathcal{U}_L = \{\mathbf{U}_\ell\}$, the expected classification error probability can be upper-bounded as

$$P_{\mathsf{e}}^{\mathsf{rnd}}(L,R) = \mathbb{E}_{\mathcal{U}_L}[P_{\mathsf{e}}(\mathcal{U}_L,R)] \le \frac{L}{2}\left[\frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{K}{2}} + \frac{\log\left(1+g(\sigma_{\mathbf{s}}^2)\right)}{(1+g(\sigma_{\mathbf{s}}^2))\,N}\right], \tag{30}$$

where $N = \beta R$.

B. ϵ-Classification Capacity

The ϵ-classification capacity defined in (8) can be obtained by solving:

$$(\mathbf{P3}) \qquad \begin{aligned} C^{\mathsf{rnd}}(R) = \arg\max_{L} \quad & L \\ \text{s.t.} \quad & P_{\mathsf{e}}^{\mathsf{rnd}}(L,R) \leq \epsilon. \end{aligned}$$

By modifying the constraint using (30), the capacity can be lower-bounded as follows.

Theorem 2. *(ϵ-Classification Capacity with Random Classes).* For a large communication rate, the ϵ-classification capacity for the random-class case can be asymptotically bounded as:

$$C^{\mathsf{rnd}}(R) \gtrsim \frac{2\beta\epsilon\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}{\log\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}R, \qquad R\to\infty. \tag{31}$$

Remark 8. *(Mathematical Intuition for Capacity Scaling Laws).* As opposed to the *exponential capacity* scaling for the class-selection case, the ϵ-classification capacity is shown in Theorem 2 to scale *linearly* w.r.t. the communication rate. Unlike deterministic classes resulting from Grassmannian packing in the former case, the random classes in the current case do not have a guaranteed minimum separation distance and the randomness in their separations dramatically increases the classification error probability. As a result, the number of classes that can be contained in the Grassmannian has to be smaller so as to satisfy a constraint on the expected separation distances, which determines the expected classification error probability. This is the fundamental reason for much slower (linear) capacity scaling w.r.t. the communication rate that determines the Grassmannian volume. On the other hand, the data-cluster dimensions K does not appear in the scaling law as its effects on the classification error probability is negligible in the current case. This fact is reflected in the upper bound on the probability in Lemma 7 where the second term independent of K dominates the first that varnishes exponentially fast as K increases.

C. Ergodic and Outage Classification Capacities

The linear scaling of the ϵ-classification capacity w.r.t. to the communication rate R makes it straightforward to extend the result to ergodic and outage classification capacities by modifying R accordingly, giving the following proposition.

Proposition 3. *(Ergodic and Outage Classification Capacities with Random Classes).* The ergodic and outage classification capacities for the random-class case can be bounded as

$$\bar{C}^{\mathsf{rand}} \gtrsim \frac{2\beta\epsilon\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}{\log\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}\bar{R}, \quad \bar{R}\to\infty, \tag{32}$$

$$C_{\mathsf{out}}^{\mathsf{rand}} \gtrsim \frac{2\beta\epsilon\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}{\log\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}R_\delta, \quad R_\delta\to\infty, \tag{33}$$

where $\bar{R}$ is the expected communication rate and R_δ the maximum rate under the outage constraint.

A similar remark as Remark 7 can be made that fading affects the communication rate but does not change the capacity scaling laws w.r.t. to the rate, which are determined by the distribution of classes on the Grassmannian (see Remark 8).

V. Extension to Fast Fading

The preceding analysis assuming a static channel within each slot of transmitting a feature vector is extended to the case of channel variation within the slot due to fast fading. To this end, we modify the transmission and channel models as follows while other models and assumptions remain unchanged. To model fast fading, each slot is divided into sub-slots, over which the channel follows i.i.d. block fading. Considering an arbitrary slot, let N features to be transmitted over the slot be divided into S packets with $1\leq S\leq N$; each is transmitted using a sub-slot with a packet-loss probability (or equivalently outage probability) of $P_{\mathsf{out}}=\eta$. The features are extracted from the received packets and assembled as a single feature vector with missing features replaced by zeros, which is then used for classification. The variable S is suitably called *the fading speed*. Consider the class-selection case where classes are packed on a Grassmannian embedded in the feature space. If the fraction of lost feature dimension is small, the classes constituting packing in the original space remains approximately so in the reduced-dimension space. Assuming such a case, the ϵ-classification capacity is determined by the dimensionality of the latter space, or equivalently the number of successfully received features per sample, denoted as N_x. This also holds in the random-class case for a different reason that random erasures of some dimensions of the feature space does not change the isotropic distribution in the resultant space. The random variable N_x is determined by the number of successfully received packets, X, that follows the binomial distribution:

$$\Pr\left(X\right)=\binom{S}{n}(1-\eta)^n\eta^{S-n}. \tag{34}$$

Given the average number of successfully received packets, $(1-\eta)S$, fixed, for large S, the distribution can be approximated as Poisson:

$$\Pr(X=n)\approx\frac{[(1-\eta)S]^n e^{-(1-\eta)S}}{n!}, \qquad S\gg 1. \tag{35}$$

1) Class-selection case: Combining the approximate distribution function, $N_x=\frac{nN}{S}$ and Theorem 1, the ergodic classification capacity is derived as

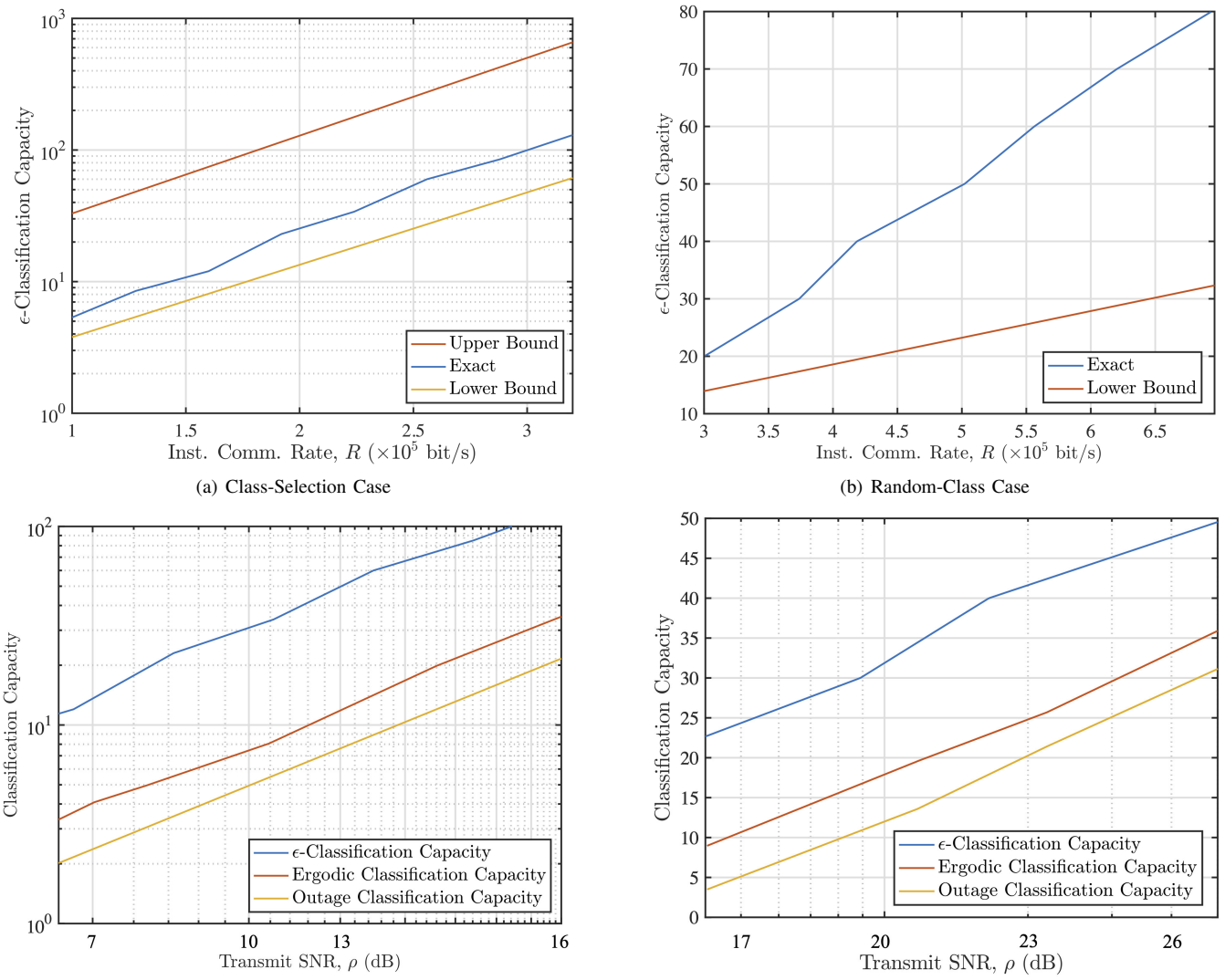
$$\begin{aligned} e^{(1-\eta)S\left(2^{\frac{\gamma_{\mathsf{lb}}N}{\beta BS}}-1\right)} &\leq \bar{C} \\ &\leq e^{(1-\eta)S\left(2^{\frac{\gamma_{\mathsf{ub}}N}{\beta BS}}-1\right)}, \ S\gg 1, \eta\ll 1. \end{aligned} \tag{36}$$

Define the ergodic communication rate $\bar{R}=\frac{(1-\eta)N}{\beta}$. For the maximum fading speed $S=N$, the ergodic classification capacity scales as

$$2^{\frac{\gamma_{\mathsf{lb}}}{\beta B}}-1\leq\lim_{\bar{R}\to\infty}\frac{\log\bar{C}}{\beta\bar{R}}\leq 2^{\frac{\gamma_{\mathsf{ub}}}{\beta B}}-1. \tag{37}$$

The above results suggest the following. First, as the number of packets S grows, both lower and upper bounds in (36) decrease, reflecting the effect of fast fading. Next, the capacity scaling law in (37) is exponential w.r.t. the ergodic communication rate as its slow-fading counterpart in Proposition 1. Therefore, the fading speed does not affect

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3072735, IEEE Transactions on Communications



(a) Class-Selection Case

(b) Random-Class Case

(c) Capacity Comparison for Class-Selection Case

(d) Capacity Comparison for Random-Class Case

Figure 2. Comparison of ϵ-classification capacity, ergodic and outage classification capacities in both the channel-selection and random-class cases at data SNR 17dB.

the classification-communication-rate relation, which is fundamentally attributed to class selection, except for scaling the communication rate by the packet-success probability $(1-\eta)$.

2) Random-class case: The ergodic classification capacity in this case can be easily modified from its slow-fading counterpart in Proposition 3 by redefining the ergodic communication rate for the current case:

$$\bar{C}^{\mathsf{rand}} \gtrsim \frac{2\beta\epsilon\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}{\log\left(1+g\left(\sigma_{\mathbf{s}}^2\right)\right)}\bar{R}, \quad \bar{R}\to\infty, \tag{38}$$

where $\bar{R}=\frac{(1-\eta)N}{\beta}$. As before, the effect of fast fading is to scale the ergodic classification capacity by the packet-success probability $(1-\eta)$.

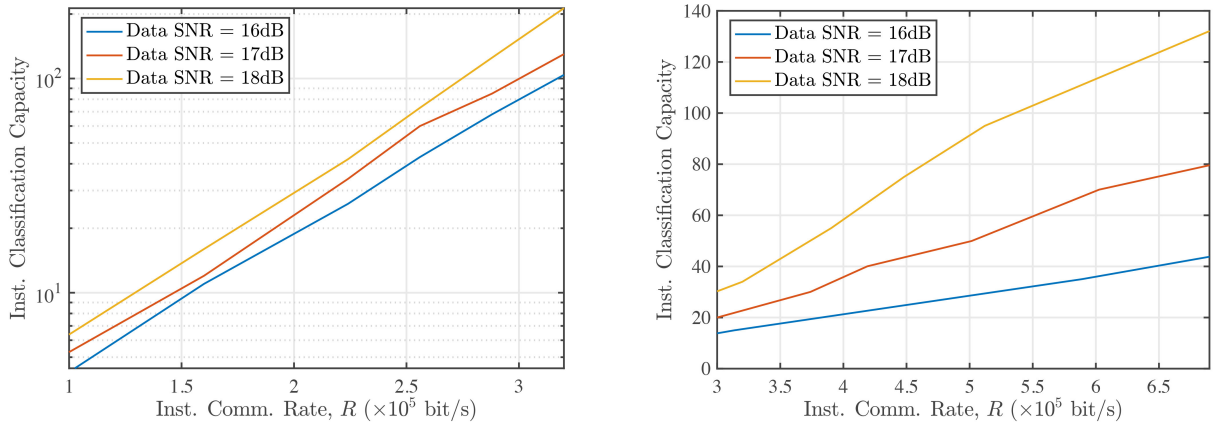
VI. Experimental Results

A. Experimental Settings

Two sets of experimental results are obtained based on the statistical data model used in the preceding analysis and a real dataset, respectively. Their corresponding experiment settings are as follows. For all experiments, fading is modeled as Rayleigh, the transmit SNR is set as 15 dB and channel bandwidth as 50 KHz.

- *Statistical data model:* The selected Grassmannian packing datasets were, in part, generated by Conway and Sloane [31]. The maximum classification error probability is 0.03 and 0.19 for the class-selection and random-class cases, respectively, and the maximum (channel) outage probability is 0.3.
- *MNIST dataset:* The well known MNIST dataset is used that comprises images of handwritten numbers. For inference, the popular neural network model, *multi-layer perceptron* (MLP), is adopted as the classifier and trained using the training dataset of MNIST. The maximum classification error probability is set as 0.03.
- *CIFAR-10 dataset:* Experiments are also conducted on another well-known dataset, CIFAR-10, containing 10 classes, e.g., airplanes, birds and horses. For inference, the popular *convolutional neural network* (CNN) is adopted and trained using the training dataset of CIFAR-10. The maximum classification error probability is set

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3072735, IEEE Transactions on Communications



(a) Class-Selection Case (b) Random-Class Case

Figure 3. Comparisons of ϵ-classification capacities at different data SNRs for both the channel-selection and random-class cases.

as 0.29.

B. Classification Capacities with Statistical Data Model

Fig. 2 shows the scalings of classification capacities of a remote-classification system, where the data SNR is set ad 17dB, as the communication rates grow and compares different capacity measures as well as the cases of class selection and random classes (with approximation when e.g., exact packing results not available). The exponential and linear scaling laws of the ϵ-classification capacities as presented in Theorems 1 and 2 are shown in Fig. 2(a) and (b) to hold even in a practical regime. Note that the small duration of the curves is caused by numerical computation of Grassmannian packing [31]. On the other hand, despite following the correct scaling laws, the bounds on the capacities are not tight due to the combined effect of the looseness of the union bounds on classification error probabilities and the distance bounds related to Grassmannian packing (in the class-selection case). Similar observations can be made on the bounds on ergodic and outage capacities with relevant curves omitted in Fig. 2 to keep the figures simple. Next, one can draw a conclusion from the comparisons in Fig. 2(c) and (d) that channel fading has a significant effect on the capacity of remote classification. For example, for a transmit SNR of 7 dB, the ergodic capacity (with fading and CSIT) and outage capacity (with fading but no CSIT) are 74% and 84% less than the ϵ-classification capacity (without fading), respectively, in the class-selection case; with a transmit SNR of 17 dB, the losses are 64% and 87% in the random-class case. Last, comparing Fig. 2(a) and (b) reveals a substantial capacity gain due to class selection such as 4-time increase in ϵ-classification capacity at the communication rate of 3×10^5 bit/s. The same conclusion holds for other capacity measures by comparing Fig. 2(c) and (d). In addition, the curves of capacity versus the instantaneous communication rate for a varying data SNR are plotted in Fig. 3. The capacity gain at high data SNRs increases as the communication rate grows in both the class-selection and the random-class cases.

Figure 4. Examples of 2-class subsets of the MNIST dataset. Consider classification based on $N = 7$ features per sample. Their corresponding classification error probabilities are different as specified.

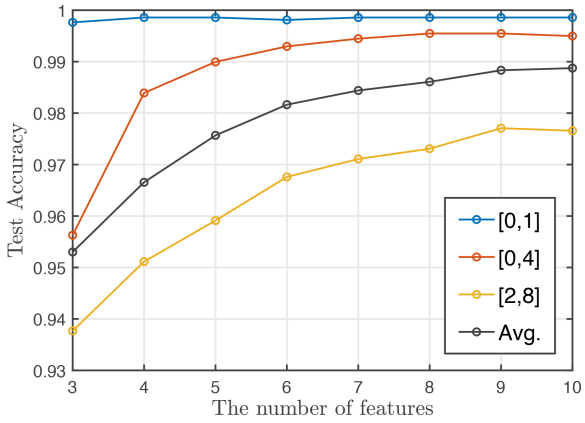
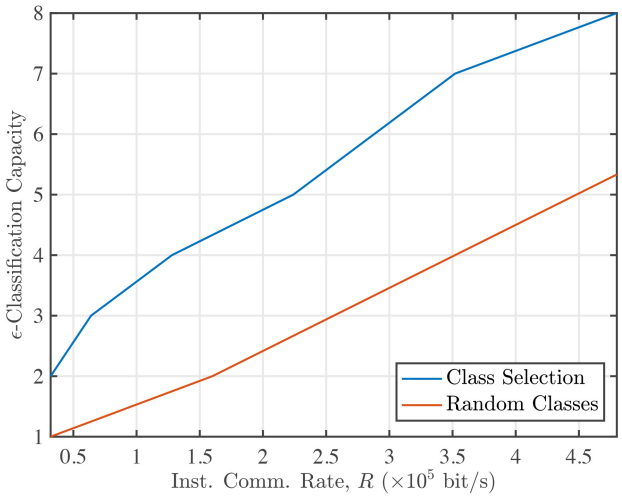


Figure 5. Accuracies of instances of 2-class subsets of the MNIST dataset.

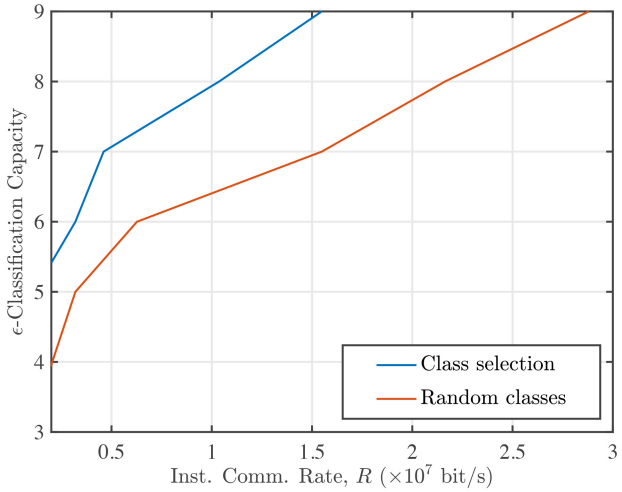
C. Classification Capacities with MNIST and CIFAR-10 Datasets

Available class subsets are generated as *all* combinations of classes in the MNIST dataset. The examples of three 2-class subsets are illustrated in Fig. 4. Next, the experimental settings are specified as follows. The MLP used in the experiments is designed and trained as follows. Its inputs are feature vectors extracted from raw images using a projection matrix obtained via PCA over the training dataset. The MLP has one hidden layer with 300 neurons and *ReLU* activations, where batch normalization is also applied. Finally, the last layer is a *softmax* output layer. During training and inference (test), the training set and test set are the standard sets specified by MNIST. The MLP is trained with the popular SGD+Momentum (SGDM) optimizer at a learning rate 0.01 and momentum 0.9 for 3 epochs. The test accuracies of instances of 2-class subsets are plotted in Fig. 5. Let $[x_i]_{i=1}^L$ denote a subset where class labels

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3072735, IEEE Transactions on Communications



(a) Capacities on MNIST



(b) Capacities on CIFAR-10

Figure 6. Classification capacity comparison between the cases of class-selection and random-class on real-world datasets.

$x_i,\ i = 1,\ ...,\ L$. The subset $[0, 1]$ corresponds to the class-selection case, which exhibits substantial accuracy gain beyond the average accuracy corresponding to the random-class case. In addition, the $[0, 4]$ and $[2, 8]$ subsets are observed to have medium and low accuracies, respectively.

The ϵ-classification capacities for the cases of class selection and random classes on MNIST are compared in Fig. 6(a). As the dataset is generated by the nature, the selected class subset is no longer generated by Grassmannian packing or the isotropic distribution as assumed in the theoretical analysis. However, we can still observe the capacity gain of class selection from Fig. 6(a), e.g., 75% capacity gain at the communication rate of 3.5×10^5 bit/s. Furthermore, the capacity with class selection scales with a growing communication rate at a rate faster than the random-class case. Both trends are aligned with the analytical results. Moreover, we have investigated the issue of whether the class-subset selection criteria for linear and MLP classifiers are aligned. Recall that the selection criterion is minimum class-pairwise separation distance for the linear classifier and classification error probability for the MLP. A class subset selected using the latter criterion as evaluated using the MLP is ranked among all subsets in terms of the former. In most cases, the rank is among the top 10%, demonstrating their alignment. For example, for the case of 7-class subsets, the subset selected using the classification error probability criterion as evaluated using the MLP is ranked 2nd in terms of the minimum class-pairwise separation distance.

As before, available subsets are generated as all combinations of classes in the CIFAR-10 dataset. Input features of different sizes are generated by bi-linear down-sampling of the original images. The CNN model and training strategy used in the experiments are adopted from an official tutorial on TensorFlow [3]. Fig. 6(b) shows the ϵ-classification capacity comparison between the class-selection and random-class cases. We can observe from Fig. 6(b) the capacity gain of the former case, e.g., 28% at the communication rate of 1.6×10^7 bit/s. In addition, the capacity with class selection grows faster with communication rate than the other case. The above results on real-world datasets demonstrate the practical relevance the proposed metric of classification capacity.

VII. Concluding Remarks

In this work, we have studied the performance of remote classification over wireless channels. The main contribution is the establishment of a relation between classification and communication by proposing various metrics of classification capacities and analyzing them using tools from differential geometry. This has led us to discover that the freedom of choosing object classes for classification under the channel constraint can attain an exponential scaling law of classification capacity w.r.t. the communication rate; without a deliberate selection, the scaling is linear.

The current study opens numerous directions for further investigation. Several of them are particularly interesting, including a realistic latency model, use of advanced wireless techniques (e.g., MIMO and OFDM) to increase the classification capacity, as well as the design of multiuser remote classification system that gives rise to new issues in terms of, e.g., resource allocation and cooperation.

Appendix

A. Proof of Lemma 2

First, we prove the upper bound on $P(i \to j)$. As $\omega^2 + a_k^2 \geq a_k^2, \forall \omega$, it follows from (12) that

$$P(i \to j) \leq \frac{1}{4\pi} \int_{-\infty}^{\infty} dw \frac{1}{w^2 + 1/4} \cdot \prod_{\substack{k=1 \\ \cos\theta_k^{(i,j)} < 1}}^{K} \left[\frac{\left(1 + \sigma_{\mathbf{s}}^2\right) \sigma_{\mathbf{s}}^{-4}}{a_k^2 \left(1 - \cos^2 \theta_k^{(i,j)}\right)} \right]^{\frac{1}{2}}$$

$$\leq \frac{1}{2} \prod_{k=1}^{K} \left[\frac{1}{1 + g(\sigma_{\mathbf{s}}^2) \sin^2 \theta_k^{(i,j)}} \right]^{\frac{1}{2}} \triangleq P_{\mathsf{ub}}. \qquad (39)$$

[3]See TensorFlow web (https://www.tensorflow.org/tutorials/images/cnn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3072735, IEEE Transactions on Communications

On the other hand, one can easily verify that

$$\frac{\partial P_{\mathsf{ub}}}{\partial \sin^2\theta_k^{(i,j)}} < 0 \quad \text{and} \quad \frac{\partial^2 P_{\mathsf{ub}}}{\partial \left(\sin^2\theta_k^{(i,j)}\right)^2} > 0. \tag{40}$$

The above results suggest that, given $d_{i,j}^2 = \sum_{k=1}^{K} \sin^2\theta_k^{(i,j)}$, P_{ub} is maximized when as many principal angles as possible are equal to zero. Consequently, one can further bound (39) as

$$P(i \to j) \leq \frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{\lfloor d_{i,j}^2 \rfloor}{2}}. \tag{41}$$

Next, we prove the lower bound on $P(i \to j)$. By Lemma 1,

$$P(i \to j) = \frac{1}{4\pi}\int_{-\infty}^{\infty} dw \frac{1}{w^2+1/4} \cdot \prod_{\substack{k=1 \\ \cos\theta_k^{(i,j)}<1}}^{K} \left[\frac{\left(1+\sigma_{\mathbf{s}}^2\right)\sigma_{\mathbf{s}}^{-4}}{\left(\omega^2+\frac{1}{4}\right)\sin^2\theta_k^{(i,j)}+1+\sigma_{\mathbf{s}}^2}\right]^{\frac{1}{2}} \tag{42}$$

Similarly, following the same argument as before with

$$\frac{\partial P(i \to j)}{\partial \sin^2\theta_k^{(i,j)}} < 0 \quad \text{and} \quad \frac{\partial^2 P(i \to j)}{\partial \left(\sin^2\theta_k^{(i,j)}\right)^2} > 0, \tag{43}$$

$P(i \to j)$ is minimized if all the principal angles have the same value given $d_{i,j}^2 = \sum_{k=1}^{K} \sin^2\theta_k^{(i,j)}$. This leads to:

$$\begin{aligned} P(i \to j) &\geq \frac{1}{4\pi}\int_{-\infty}^{\infty} dw \frac{1}{w^2+1/4} \cdot \left(\frac{1}{\frac{4}{K}g(\sigma_{\mathbf{s}}^2)\left(\omega^2+\frac{1}{4}\right)d_{i,j}^2+1}\right)^K, \quad (44) \\ &\geq \frac{1}{4\pi}\int_{\omega^2+\frac{1}{4}\leq 1} dw \frac{1}{w^2+1/4} \cdot \left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)\left(\omega^2+\frac{1}{4}\right)d_{i,j}^2}\right)^K, \\ &\geq \frac{1}{4\pi}\int_{\omega^2+\frac{1}{4}\leq 1} dw \frac{1}{w^2+1/4} \cdot \left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)d_{i,j}^2}\right)^K, \\ &= \frac{1}{\pi}\arctan(\sqrt{3})\left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)d_{i,j}^2}\right)^K. \quad (45) \end{aligned}$$

This completes the proof.

B. Proof of Theorem 1

First, we prove the lower bound on the ϵ-classification capacity. For large K, by (18),

$$KL^{-\frac{2}{NK}}\log_2(1+g(\sigma_{\mathbf{s}}^2))^{-1} = \log\frac{2\epsilon}{L(1+g(\sigma_{\mathbf{s}}^2))}. \tag{46}$$

For a high data SNR, it follows from the above equation that

$$L \gtrsim 2^{\frac{N}{2}\left(K\log_2 K + K\log_2\log_2(1+g(\sigma_{\mathbf{s}}^2)) - K\log_2\log_2\frac{1+g(\sigma_{\mathbf{s}}^2)}{2\epsilon}\right)}. \tag{47}$$

Next, we prove the upper bound on the ϵ-classification capacity. From (19),

$$\frac{1}{3}\left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)\delta_{\mathsf{ub}}^2}\right)^K = \epsilon. \tag{48}$$

As the direct approach is intractable, we find a lower bound on the right-hand side of (48). Using the fact that $4KL^{-\frac{2}{NK}} \geq \delta_{\mathsf{ub}}^2 > 1$ for large N and K,

$$\begin{aligned} &\frac{1}{3}\left(\frac{1}{1+\frac{4}{K}g(\sigma_{\mathbf{s}}^2)\delta_{\mathsf{ub}}^2}\right)^K \\ &\geq \frac{1}{3}\left(\frac{1}{1+16g(\sigma_{\mathbf{s}}^2)L^{-\frac{2}{NK}}}\right)^K, \quad N,K \to \infty. \end{aligned} \tag{49}$$

Then, (48) asymptotically reduces to $\frac{1}{3}\left(\frac{1}{1+16g(\sigma_{\mathbf{s}}^2)L^{-\frac{2}{NK}}}\right)^K \approx \epsilon$, as $N, K \to \infty$. This results in an asymptotic upper bound on the ϵ-classification capacity:

$$L \lesssim 2^{\frac{N}{2}\left(K\log_2 4K + K\log_2\frac{4g(\sigma_{\mathbf{s}}^2)}{1-3\epsilon}\right)}. \tag{50}$$

The substituting of $N = \beta R$ gives (20). Furthermore, as $R, K \to \infty$, the bounds on the ϵ-classification capacity scale in (21), which completes the proof.

C. Proof of Proposition 1

1) Bounds on ergodic classification capacity: The lower bound in Theorem 1 can be rewritten as $C^*(R) \gtrsim 2^{\frac{R}{B}\cdot\gamma_{\mathsf{lb}}}$, where $\gamma_{\mathsf{lb}} = \frac{\beta B}{2}\left(K\log_2 K + Kc_{\sigma_{\mathbf{s}}} - Kc_\epsilon\right)$. It follows that

$$\mathbb{E}[C^*(R)] \gtrsim \mathbb{E}\left[\left(1+\rho|h|^2\right)^{\gamma_{\mathsf{lb}}}\right]. \tag{51}$$

For a high transmit SNR,

$$\mathbb{E}\left[\left(1+\rho|h|^2\right)^{\gamma_{\mathsf{lb}}}\right] \approx \rho^{\gamma_{\mathsf{lb}}}\mathbb{E}[|h|^{2\gamma_{\mathsf{lb}}}] \overset{(a)}{=} \Gamma(\gamma_{\mathsf{lb}}+1), \quad \rho \to \infty, \tag{52}$$

where (a) uses $|h|^2 = \exp(1)$ and $\Gamma(\cdot)$ denotes the gamma function. Given large γ_{lb} and using the *stirling's approxximation*

$$\mathbb{E}[|h|^{2\gamma_{\mathsf{lb}}}] \approx \sqrt{2\pi\gamma_{\mathsf{lb}}}\cdot e^{\gamma_{\mathsf{lb}}(\log\gamma_{\mathsf{lb}}-1)}, \quad \gamma_{\mathsf{lb}} \gg 1. \tag{53}$$

Combining the above result with (51) and (52), (22) follows. Following the same procedure, the upper bound can be proved.

2) Scaling law: Consider the ergodic communication rate $\bar{R} = \mathbb{E}[B\log_2(1+\rho|h|^2)]$, $\bar{R} \to B\log_2\rho + B\mathbb{E}[\log_2|h|^2]$ and hence

$$\lim_{\rho\to\infty}\frac{\bar{R}}{\log_2\rho} = B, \quad \rho \to \infty, \tag{54}$$

implying $\bar{R} \to \infty$ as $\rho \to \infty$. Then, given sufficiently large ρ and furthermore letting $K \to \infty$, both the derived bounds in (22) scale as shown in (23) . This completes the proof.

D. Proof of Proposition 2

The bounds in (25) are straightforward by substituting R_δ in (24) into the outage classification capacity defined in (11). In the following, we prove the scaling law. To

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2021.3072735, IEEE Transactions on Communications

begin with, we show that as $\rho \to \infty$, $R_\delta \to \infty$. Given $R_\delta = \log_2\left(1+\rho\log\left(\frac{1}{1-\delta}\right)\right)$, at a high SNR, one can have

$$\begin{aligned} R_\delta &\approx \log_2\left(\rho\log\left(\frac{1}{1-\delta}\right)\right) \\ &= \log_2\rho + \log_2\log\frac{1}{1-\delta}, \quad \rho\to\infty. \end{aligned} \tag{55}$$

This implies that R_δ scales linearly with $\log_2\rho$ given δ. As a result, as $\rho\to\infty$, $R_\delta\to\infty$. Then, by letting $R_\delta, K\to\infty$, both bounds scale as shown in (26), which completes the proof.

E. Proof of Lemma 6

To derive the upper bound on the CDF, namely $F_{d_c^2}(x)$, we fist obtain an upper bound on the probability of $\Pr(\sin^2\theta_{\max}<x)$. Given (27),

$$\begin{aligned} \Pr(\theta_{\max}<x) = &\frac{\Gamma\left(\frac{K+1}{2}\right)\Gamma\left(\frac{N-K+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{N+1}{2}\right)}(\sin x)^{K(N-K)} \\ &\cdot {}_2F_1\left(\frac{N-K}{2},\frac{1}{2};\frac{N+1}{2};\sin^2 x\mathbf{I}_K\right). \end{aligned} \tag{56}$$

Due to the fact that $\Pr(\sin^2\theta_{\max}<x) = \Pr(\theta_{\max}<\arcsin\sqrt{x})$, one can have

$$\begin{aligned} \Pr(\sin^2\theta_{\max}<x) &= \frac{\Gamma\left(\frac{K+1}{2}\right)\Gamma\left(\frac{N-K+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{N+1}{2}\right)}x^{\frac{K(N-K)}{2}} \\ &\quad\cdot {}_2F_1\left(\frac{N-K}{2},\frac{1}{2};\frac{N+1}{2};x\mathbf{I}_K\right), \\ &\le \frac{\Gamma\left(\frac{K+1}{2}\right)\Gamma\left(\frac{N-K+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{N+1}{2}\right)}x^{\frac{K(N-K)}{2}} \\ &\quad\cdot {}_2F_1\left(\frac{N-K}{2},\frac{1}{2};\frac{N+1}{2};\mathbf{I}_K\right), \end{aligned} \tag{57}$$

where the inequality uses the fact that ${}_2F_1$ is a non-decreasing function in x. On the other hand, according to [32], one can have

$${}_2F_1\left(\frac{N-K}{2},\frac{1}{2};\frac{N+1}{2};\mathbf{I}_K\right) = \frac{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{K+1}{2}\right)\Gamma\left(\frac{N-K+1}{2}\right)}. \tag{58}$$

Then, by substituting (58) into (57),

$$\Pr(\sin^2\theta_{\max}<x) = \Pr(\theta_{\max}<\arcsin x) \le x^{\frac{K(N-K)}{2}}. \tag{59}$$

As two random subspaces of dimension K embedded in $\mathbb{R}^N$ are quasi-orthogonal, given large N, the squared chordal distance d_c^2 can be approximated as $K\sin^2\theta_{\max}$. Then, we can bound the said CDF, namely $F_{d_c^2}(x)$, as

$$\begin{aligned} F_{d_c^2}(x) &= \Pr(d_c^2<x) \\ &\approx \Pr(K\sin^2\theta_{\max}<x) \\ &= \Pr\left(\sin^2\theta_{\max}<\frac{x}{K}\right), \quad N\to\infty. \end{aligned} \tag{60}$$

By combining (59) and (60), the desired result follows.

F. Proof of Lemma 7

It follows from (13) that

$$\begin{aligned} \mathbb{E}\left[P(i\to j)\right] &\le \mathbb{E}_{d_c^2}\left[\frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{\lfloor d_{i,j}^2\rfloor}{2}}\right] \\ &= \int_0^K \frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{\lfloor x\rfloor}{2}}\mathrm{PDF}_{d_c^2}(x)dx, \\ &\le \frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{K}{2}} + \frac{\log(1+g(\sigma_{\mathbf{s}}^2))}{2} \\ &\quad\cdot\int_0^K F_{d_c^2}^{\mathsf{ub}}(x)\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{x}{2}-1}dx. \end{aligned} \tag{61}$$

Combining (28) and (61), gives

$$\begin{aligned} \mathbb{E}\left[P(i\to j)\right] &\le \frac{1}{2}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{K}{2}} + \frac{K\log(1+g(\sigma_{\mathbf{s}}^2))}{2} \\ &\quad\cdot\int_0^1 x^{\frac{K(N-K)}{2}}\frac{1}{\left[1+g(\sigma_{\mathbf{s}}^2)\right]^{\frac{Kx}{2}-1}}dx. \end{aligned} \tag{62}$$

We decompose the second term at the RHS of (62) as follows

$$\begin{aligned} &\frac{K\log(1+g(\sigma_{\mathbf{s}}^2))}{2}\left[\int_0^{\frac{2}{K}} x^{\frac{K(N-K)}{2}}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{Kx}{2}-1}dx\right. \\ &\qquad\left.+\int_{\frac{2}{K}}^1 x^{\frac{K(N-K)}{2}}\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)^{\frac{Kx}{2}-1}dx\right] \\ \le\ &\frac{K\log(1+g(\sigma_{\mathbf{s}}^2))}{2}\left[\left(1+g(\sigma_{\mathbf{s}}^2)\right)\int_0^{\frac{2}{K}} x^{\frac{K(N-K)}{2}}dx\right. \\ &\qquad\left.+\left(\frac{1}{1+g(\sigma_{\mathbf{s}}^2)}\right)\int_{\frac{2}{K}}^1 x^{\frac{K(N-K)}{2}}dx\right]. \end{aligned} \tag{63}$$

For large N, (63) can be asymptotically expressed as $\frac{\log(1+g(\sigma_{\mathbf{s}}^2))}{1+g(\sigma_{\mathbf{s}}^2)}\frac{1}{N}$. Substituting it into (62), (29) follows. This completes the proof.

REFERENCES

[1] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 869-904, 2020.

[2] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, and W. Xu, "Connected roads of the future: Use cases, requirements, and design considerations for vehicle-to-everything communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 110-123, Sep. 2018

[3] Y. Wang, J. Shen, T.-K. Hu, P. Xu, T. Nguyen, R. Baraniuk, Z. Wang and Y. Lin, "Dual dynamic inference: Enabling more efficient, adaptive, and controllable deep inference," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 4, pp. 623-633, May 2020.

[4] X. Huang and S. Zhou, "Dynamic compression ratio selection for edge inference systems with hard deadlines," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8800-8810, Sept. 2020.

[5] T. Berger, *Rate distortion theory: A mathematical basis for data compression*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[6] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: Wiley, 1991.

[7] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[8] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the Grassmann manifold: Fundamental limits of subspace classifiers," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2133-2147, Apr. 2015.

[9] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322-2358, 2017.

[10] C. You, K. Huang, H. Chae, and B. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397-1411, Mar. 2017.

[11] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. S. Quek, "Offloading in mobile edge computing: Task allocation and computational frequency scaling," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3571-3584, Aug. 2017.

[12] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 1991-1995, Jun. 2012.

[13] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994-6009, Sep. 2017.

[14] K. Li, M. Tao, and Z. Chen, "Exploiting computation replication for mobile edge computing: A fundamental computation-communication tradeoff study," to appear in *IEEE Trans. Wireless Commun.*.

[15] A. Ndikumana, N. H. Tran, T. M. Ho, Z. Han, W. Saad, D. Niyato and C. S. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1359-1374, 1 Jun. 2020.

[16] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, July 10-15, 2016.

[17] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *[online]. Available: https://arxiv.org/pdf/2007.01790.pdf*.

[18] H. T. Nguyen, N. C. Luong, J. Zhao, C. Yuen, and D. Niyato, "Resource allocation in mobility-aware federated learning networks: A deep reinforcement learning approach," *[online]. Available: https://arxiv.org/pdf/1910.09172.pdf*.

[19] X. Yang, S. Hua, Y. Shi, H. Wang, J. Zhang, and K. B. Letaief, "Sparse optimization for green edge AI inference," *J. Commun. Info. Netw.*,vol. 5, no. 1, pp. 1-15, Mar. 2020.

[20] W. Shi, Y. Hou, S. Zhou, Z. Niu, Y. Zhang, and L. Geng, "Improving device-edge cooperative inference of deep learning via 2-step pruning," *[online]. Available: https://arxiv.org/pdf/1903.03472.pdf*, 2019.

[21] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *[online]. Available: https://arxiv.org/pdf/2006.02166.pdf*, 2020.

[22] E. Li, Z. Zhou, and X. Che, "Edge intelligence: On-demand deep learning model co-inference with device-edge synergy," in *Proc. ACM Workshop Mobile Edge Commun. (MECOMM'18)*, Budapest, Hungary, Aug. 20-25, 2018.

[23] S. P. Chinchali, E. Cidon, E. Pergament, T. Chu, and S. Katti, "Neural networks meet physical networks: Distributed inference between edge devices and the cloud," in *Proc. ACM Workshop Hot Topics Netw. (HotNets'18)*, Redmond, Washington, Nov. 15-16, 2018.

[24] W. Yu, Z. Sun, H. Liu, Z. Li, and Z. Zheng, "Multi-level deep learning based e-Commerce product categorization," in *Proc. ACM SIGIR 2018 Workshop on eCommerce,* Ann Arbor, Michigan, July 8-12, 2018.

[25] M. Ceci and D. Malerba, "Classifying web documents in a hierarchy of categories: A comprehensive study", *J. Intell. Info. Sys.*, vol 28, no. 1, pp. 37-78, Jan. 2017.

[26] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. New York: Springer, 2001.

[27] J. H. Conway, N. J. A. Sloane, and E. Bannai, *Sphere packings, lattices, and groups*. New York: Springer, 1987.

[28] A. Barg and D. Y. Nogin, "Bounds on packings of spheres in the Grassmann manifold," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2450-2454, Sep. 2002.

[29] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 543-564, Mar. 2000

[30] P.-A. Absil, A. Edelman, and P. Koev, "On the largest principal angle between random subspaces," *Linear Algebra and Its Applications*, vol. 141, no. 1, pp. 288-294, Apr. 2006.

[31] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian spaces," *Experimental Mathematics*, vol. 5, no. 2, pp. 139-159, Feb. 1996.

[32] D. Richards, and Q. Zhang, "A reflection formula for the Gaussian hypergeometric function of matrix argument," *[online]. Available: https://arxiv.org/pdf/2002.05248.pdf*.

Qiao Lan (Graduate Student Member, IEEE) received the B.Eng. degree (with honor) from the Southern University of Science and Technology (SUSTech), Shenzhen, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong (HKU). His research interests include edge intelligence and B5G systems.

Yuqing Du (Member, IEEE) received the B.Eng. degree from Harbin Engineering University (HEU) in 2016 and the Ph.D. degree from The University of Hong Kong (HKU) in 2020, all in electrical engineering. His research interests include edge learning, distributed machine learning.

Petar Popovski (S'97–A'98–M'04–SM'10–F'16) is a Professor at Aalborg University, where he heads the section on Connectivity. He received his Dipl.-Ing and M. Sc. degrees in communication engineering from the University of Sts. Cyril and Methodius in Skopje and the Ph.D. degree from Aalborg University in 2005. He is a Fellow of the IEEE. He received an ERC Consolidator Grant (2015), the Danish Elite Researcher award (2016), IEEE Fred W. Ellersick prize (2016), IEEE Stephen O. Rice prize (2018), Technical Achievement Award from the IEEE Technical Committee on Smart Grid Communications (2019) and the Danish Telecommunication Prize (2020). He is a Member at Large at the Board of Governors in IEEE Communication Society, Vice-Chair of the IEEE Communication Theory Technical Committee and IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He is currently an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Prof. Popovski was the General Chair for IEEE SmartGridComm 2018 and IEEE Communication Theory Workshop 2019. His research interests are in the area of wireless communication and communication theory. He authored the book "Wireless Connectivity: An Intuitive and Fundamental Guide", published by Wiley in 2020.

Kaibin Huang (Fellow, 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 at 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 is an Associate Editor for IEEE Transactions on Wireless Communications and Journal on Selected Areas in Communications (JSAC), and an Area Editor for IEEE Transactions on Green Communications and Networking. Previously, he has also served on the editorial board of IEEE Wireless Communications Letters. Moreover, he has guest edited special issues for IEEE JSAC, IEEE Journal on Selected Topics on Signal Processing, and IEEE Communications Magazine. He is an IEEE Fellow and an IEEE Distinguished Lecturer of both the IEEE Communications and Vehicular Technology Societies. He has been named a Highly Cited Researcher by Clarivate Analytics in 2019 and 2020.