Federated Inference with Reliable Uncertainty Quantification over Wireless Channels via Conformal Prediction

Meiyi Zhu, Matteo Zecchin Student Member, IEEE, Sangwoo Park Member, IEEE, Caili Guo Senior Member, IEEE, Chunyan Feng Senior Member, IEEE, Osvaldo Simeone Fellow, IEEE

Abstract

Consider a setting in which devices and a server share a pre-trained model. The server wishes to make an inference on a new input given the model. Devices have access to data, previously not used for training, and can communicate to the server over a common wireless channel. If the devices have no access to the new input, can communication from devices to the server enhance the quality of the inference decision at the server? Recent work has introduced federated conformal prediction (CP), which leverages devices-to-server communication to improve the reliability of the server's decision. With federated CP, devices communicate to the server information about the loss accrued by the shared pre-trained model on the local data, and the server leverages this information to calibrate a decision interval, or set, so that it is guaranteed to contain the correct answer with a pre-defined target reliability level. Previous work assumed noise-free communication, whereby devices can communicate a single real number to the server.

The work of M. Zhu, C. Guo and C. Feng was supported by the Fundamental Research Funds for the Central Universities (No.2021XD-A01-1), and by the Beijing Natural Science Foundation (L222043). The work of M. Zhu was also supported by the BUPT Excellent Ph.D. Students Foundation. The work of M. Zecchin and O. Simeone was supported by the European Union's Horizon Europe project CENTRIC (101096379). The work of O. Simeone was also supported by an Open Fellowship of the EPSRC (EP/W024101/1), and by Project REASON, a UK Government funded project under the Future Open Networks Research Challenge (FONRC) sponsored by the Department of Science Innovation and Technology (DSIT).

Meiyi Zhu, Caili Guo and Chunyan Feng are with the Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: lia@bupt.edu.cn; guocaili@bupt.edu.cn; cyfeng@bupt.edu.cn).

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone are with the King's Communications, Learning & Information Processing (KCLIP) lab, Department of Engineering, King's College London, London WC2R 2LS, U.K. (e-mail: matteo.1.zecchin@kcl.ac.uk; sangwoo.park@kcl.ac.uk; osvaldo.simeone@kcl.ac.uk).

In this paper, we study for the first time federated CP in a wireless setting. We introduce a novel protocol, termed wireless federated conformal prediction (WFCP), which builds on type-based multiple access (TBMA) and on a novel quantile correction strategy. WFCP is proved to provide formal reliability guarantees in terms of coverage of the predicted set produced by the server. Using numerical results, we demonstrate the significant advantages of WFCP against digital implementations of existing federated CP schemes, especially in regimes with limited communication resources and/or large number of devices.

Index Terms

Conformal prediction, federated inference, wireless communications, type-based multiple access.

I. Introduction

Federation is a data processing paradigm whereby distributed devices with local, possibly private, data sets collaborate for the purpose of carrying out a shared information processing task without the direct exchange of the local data sets. The main exemplar of federated data processing is federated learning, which addresses the task of training a machine learning model. Federated learning has been widely studied in recent years, with research activities ranging from theoretical analyses [?], [?] to the design of communication protocols [?], [?] and to testbeds [?], [?]. This paper focuses on a different federated data processing task, namely federated inference, with the goal of leveraging collaboration across devices to ensure reliable decision-making.

A. Federated Reliable Inference

As illustrated in Fig. ??, we study a setting in which devices and a server share a pre-trained machine learning model. The model may have been obtained through a previous phase of federated learning, or it may have been downloaded from a repository of existing models trained in any other arbitrary manner. The server wishes to make an inference on a new input given the model. Specifically, given an input, it wishes to produce an interval, or set, of possible output values that is guaranteed to contain the correct answer with a pre-defined target reliability level. Devices have access to data, previously not used for training, and can communicate to the server over wireless channels. The devices do not have access to the new input.

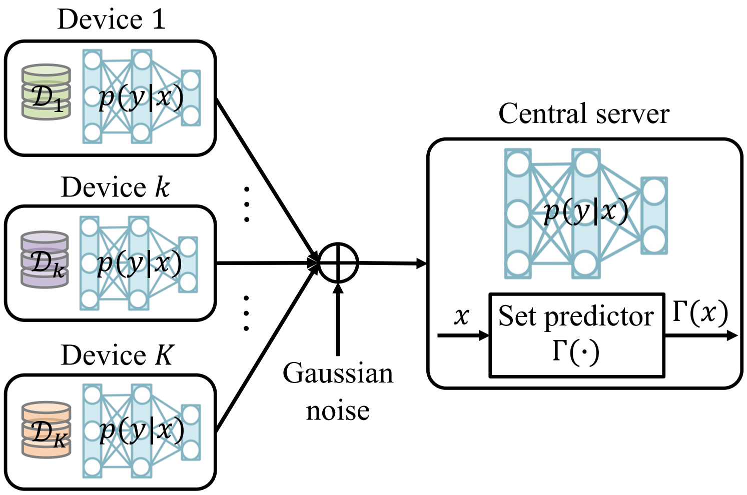

Fig. 1. Illustration of the wireless reliable federated inference problem under study: A pre-trained machine learning model $p(y|x)$ is available at devices and a server. The server wishes to make a reliable prediction on a test input x, which is not available at the devices. Following the CP framework, the prediction takes the form of a subset $\Gamma(x)$ of the label space $\mathcal{Y}$. The goal is to ensure that the predicted set $\Gamma(x)$ contains the true label with probability no smaller than a target reliability level $1-\alpha$ (see (??)). To this end, each device k communicates information about the local data set $\mathcal{D}_k$ to the server over a noisy shared channel. This information is then used at the server not to update the model $p(y|x)$ but rather to calibrate the prediction $\Gamma(x)$, ensuring the reliability condition (??).

For this setting, recent work has introduced federated conformal prediction (CP), which leverages devices-to-server communication to support reliable decision-making at the server [?]. With federated CP, devices communicate to the server information about the performance accrued by the shared pre-trained model on the local data. Intuitively, this information provides a yardstick with which the server can gauge the plausibility of each value of the output variable for the given input. For instance, if the model obtains a loss no larger than some value ℓ on 90% of the data points at the devices, then the server may safely exclude from the predicted interval/set all output values to which the model assigns a loss larger than ℓ, as long as it wishes to guarantee a 90% reliability level. In other words, the server leverages information received from the devices to calibrate its decision interval/set.

Previous work [?] assumed noise-free communication, whereby devices can communicate a single real number to the server. Specifically, reference [?] proposed a quantile-of-quantile (QQ) scheme, referred to as FedCP-QQ, whereby each device computes and communicates a pre-determined quantile of the local losses. In this paper, we study for the first time

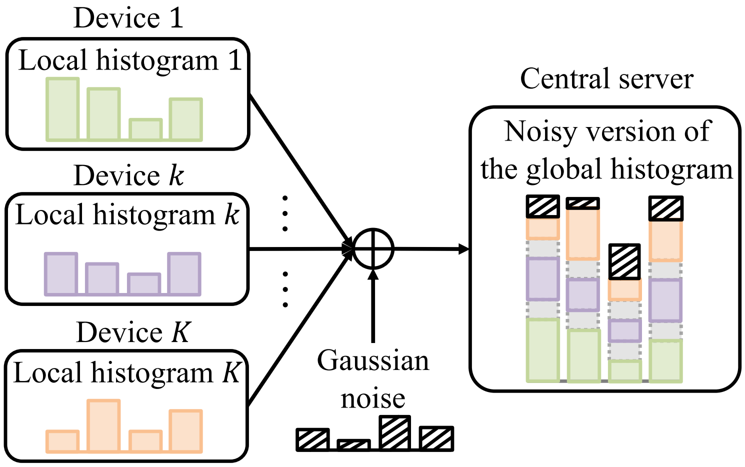

Fig. 2. TBMA enables the estimate of the global histogram of discrete scalar data available across all devices. To this end, orthogonal codewords are assigned to each histogram bin. All devices transmit simultaneously their individual local histograms over a shared wireless channel by allocating to each codeword a power proportional to the corresponding bin probability. This way, the server can obtain a noisy estimate of the global histogram thanks to the superposition of the signals received for each orthogonal codeword.

federated CP in a wireless setting.

B. Wireless Federated Conformal Prediction

Even with a perfect transmission of local quantiles, the performance of FedCP-QQ is inherently limited. In fact, for a target reliability level of, say, 90%, ideally, the server would need to know the 90-percentile of the losses obtained by the pre-trained model across all devices. However, the quantile-of-quantiles targeted by FedCP-QQ provides a generally inaccurate estimate of the overall quantile, particularly when the number of devices is large. Furthermore, a direct implementation of FedCP-QQ [?] on a wireless channel would require the transmission of quantized local quantiles, requiring a bandwidth that increases proportionally to the number of available devices.

In this paper, we introduce a novel protocol, termed wireless federated conformal prediction (WFCP), which addresses these shortcomings by building on type-based multiple access (TBMA) [?], [?] and on a novel quantile correction scheme. TBMA is a multiple access scheme that aims at recovering aggregated statistics, rather than individual messages. In particular, it can be used to support the estimate of the histogram of data available across the devices at the server. To explain it, assume that each device has scalar, quantized, data with a given, generally different, histogram. As illustrated in Fig. ??, the goal of the server is to estimate the average histogram across all devices, i.e., the histogram of the data available at all devices, without having to separately estimate the histograms of all devices.

To accomplish this objective, in TBMA, each histogram bin is assigned an orthogonal codeword. Devices divide their transmission energy across a number of orthogonal codewords in such a way that more energy is allocated to codewords corresponding to histogram bins with a larger number of data points. Allowing for all devices to transmit simultaneously, by collecting the energy received in each bin, the server can estimate the global histogram thanks to the superposition property of wireless communications.

By adopting TBMA as the communication protocol, the proposed WFCP scheme allows the server to estimate the histogram of the losses accrued by the pre-trained model across all the devices. This estimate is then used to estimate the desired global quantile. Importantly, the bandwidth required by TBMA scales only with the resolution of the quantization, i.e., with the number of bins, and not with the number of devices.

The main technical challenge in the design of WFCP is how to ensure reliability – that is, the condition that the predicted interval/set at the server contains the true output with the desired reliability level. This challenge arises from the fact that the estimate of the global histogram of the losses, and hence of its quantile, is inherently noisy (see Fig. ??). WFCP addresses this problem by proposing a novel quantile correction method that is proved to guarantee reliability.

C. Related Work

We now provide a brief review of related work by focusing first on federated CP protocols and then on TBMA.

Federated CP. Prior to the introduction of the FedCP-QQ scheme [?] reviewed in Sec. ??, reference [?] initially applied CP in federated settings, aiming to provide distribution-free, set-valued predictions with reliability guarantees. In [?], each device calculates a quantile of its local losses, and the server aggregates these quantiles from all devices to form an average. However, applying CP with the averaged quantile does not guarantee reliability. To address these limitations, FedCP-QQ [?] was proposed whereby a QQ estimator is used in lieu of an average of quantiles, re-establishing formal reliability guarantees for federated CP.

Federated CP has been further generalized to address settings with statistical heterogeneity across the data available at the devices. In [?], the authors proposed an approach that ensures that the set predictor is well calibrated with respect to a specific mixture of distributions of the devices' local data. To reduce the communication overhead, they proposed to apply distributed quantile estimation methods [?], [?] to acquire an approximate quantile. Due to the imperfect estimate of the quantile, reliability guarantees are only proved under additional assumptions on the quality of the estimation error. In parallel, reference [?] studied a related setting with label distribution shifts among devices by generalizing the weighted quantile computation scheme proposed in [?] for centralized CP. In particular, by noting that a quantile can be obtained as the minimizer of the pinball loss [?], the authors applied a gradient-based approach to jointly estimate a quantile from distributed devices. Accordingly, unlike the setting studied in this paper and in previous work, which assumes one-shot, or embarrassingly parallel, protocols, the scheme in [?] is iterative, requiring multiple communication rounds.

All existing federated CP techniques do not consider the influence of noise on the communication channel. This paper aims to address this knowledge gap by investigating the problem of wireless federated CP and by focusing on the impact of channel noise on quantile estimation and, in turn, on model calibration. Unlike [?], [?], as in [?], we target formal reliability guarantees without additional assumptions on the quality of the quantile estimates. Furthermore, as in reference [?], we focus on statistically homogeneous data across devices.

TBMA. The pioneering papers [?], [?] introduced TBMA, whereby orthogonal codewords are assigned to different measurement values across multiple devices and a variant of the maximum likelihood estimator is devised to accomplish single parameter estimation. In reference [?], TBMA was applied to estimate multiple correlated parameters in a multi-cell set-up by leveraging in-cell orthogonal TBMA and inter-cell non-orthogonal frequency reuse strategy under centralized and decentralized decoding settings. Furthermore, papers [?], [?] developed a non-orthogonal variant of TBMA for multi-valued event detection in random access scenarios. Based on the assumption of sparse user activity, Bayesian approximate message passing estimators were designed for a single-cell [?] and a multi-cell fog-radio access network [?] respectively. Reference [?] proposed an end-to-end design of TBMA protocols, whereby the information bottleneck principle is adopted as the criterion to jointly optimize the codebook and neural network-based estimator under unknown source and channel statistics. None of these papers provide any insights into the key problem of using TBMA for reliable quantile estimation.

D. Contributions and Organization

In this paper, we introduce WFCP, the first wireless protocol for the implementation of federated CP. The main contributions of this paper are summarized as follows.

- We first review conventional centralized CP, which assumes that all data is available at the server [?]. Then, we introduce a digital communication framework in order to apply the state-of-the-art federated CP scheme, FedCP-QQ [?], to wireless systems, which will serve as a baseline scheme for WFCP. To this end, we assume that all the devices orthogonally share the available channel uses via time-division multiple access (TDMA), and that the server can detect and discard the received erroneous local quantile to implement the QQ estimator over the correctly received quantiles.
- We propose WFCP, a novel protocol based on TBMA that hinges on a carefully selected quantile threshold that accounts for the presence of channel noise.
- We provide a rigorous analysis of the reliability performance of WFCP, proving that it can achieve any target reliability level. The analysis also provides guidelines on the choice of important design parameters such as the number of quantization levels.
- Simulation results demonstrate the advantage of the proposed WFCP scheme over existing strategies, especially in the presence of limited communication resources and/or large number of devices.

The remainder of this paper is organized as follows. In Sec. ??, we describe the setting and define the problem. Sec. ?? presents the general framework of conventional CP, and introduces a quantized version for future reference. Sec. ?? reviews the FedCP-QQ scheme [?], which operates in an ideal noise-free scenario, and describes a digital wireless implementation of FedCP-QQ. In Sec. ??, we propose the WFCP scheme, also providing design guidelines and proof of reliability. Sec. ?? evaluates the performance of WFCP as compared to benchmarks via experiments, validating the effectiveness of WFCP. Sec. ?? summarizes this paper and points to directions for future work.

II. Setting and Problem Definition

A. Setting

We consider a wireless federated inference scenario in which a set of K devices and a central server communicate over a multiple access channel. A pre-trained machine learning model is available at both server and devices side. This model may have been previously trained using federated learning [?]. As in [?], we focus on the problem of reliable collaborative, or federated, inference using a fixed model along with communication between devices and server. In this setting, communication is not used to optimize the machine learning model, as in federated learning, but rather to ensure a higher level of reliability for an inference decision produced by the server on a new input available only at the server. As we will detail, thanks to communication, devices can help the server calibrate its decision, enhancing the server's estimate of the corresponding uncertainty.

To elaborate, we focus on a classification problem with C classes, which are labelled by elements in set $\mathcal{Y} = \{1, 2, \ldots, C\}$. Given any input $x \in \mathcal{X}$, the predictive model produces a conditional probability distribution $p(y|x)$ over the labels $y \in \mathcal{Y}$. Probability $p(y|x)$ is typically interpreted as a measure of the confidence that the model has in label y being the correct one. Conventionally, a decision y^* is obtained by selecting the label on which the model has maximum confidence, i.e.,

$$y^* = \arg\max_{y \in \mathcal{Y}} p(y|x). \tag{1}$$

The corresponding confidence level $p(y^*|x)$ produced by the model should ideally provide an indication of the true accuracy of the decision y^*. However, it is well known that machine learning models tend to be overconfident [?], [?], [?], and hence systems using the decision y^* produced by the model cannot trust the confidence level $p(y^*|x)$ to provide a reliable measure of the reliability of the decision.

In this paper, we are interested in producing decisions that provide trustworthy measures of uncertainty. To this end, following the CP framework [?], our goal is to ensure that, based on a generally unreliable model $p(y|x)$ and on communication with the devices, the server outputs set-valued decisions for any new input x with formal reliability guarantees.

To explain, we define a mapping from an input x to a subset of the label space as $\Gamma(x) \subseteq \mathcal{Y}$. Given a new input x available only at the server, as well as the model output distribution $\{p(y|x)\}_{y \in \mathcal{Y}}$, along with information received from the devices, the server aims at producing a set-valued decision $\Gamma(x)$ with reliability guarantees. For a target reliability level $1-\alpha$, with $\alpha \in [0,1]$, a set decision is said to be reliable if the set contains the true label y with probability at least $1-\alpha$, i.e., if the inequality

$$\Pr(y \in \Gamma(x)) \geq 1-\alpha \tag{2}$$

holds. The probability in (??) is evaluated with respect to the randomness of data generation and communications, as we discuss in the next subsections.

Before detailing the role of communications, we observe that the reliability requirement (??) can be trivially met by choosing as a set decision the set of all possible labels, i.e., $\Gamma(x) = \mathcal{Y}$, irrespective of the input x. This set predictor, while reliable, would be completely uninformative. It is hence also important to evaluate the performance of the set decision on the basis of the average size of its prediction. This is known as the inefficiency of the predictor, which is defined as

$$\mathbb{E}\left[|\Gamma(x)|\right], \tag{3}$$

where $|\cdot|$ represents the cardinality of the argument set. The expectation in (??) is evaluated with respect to the randomness of both data and communications, as in (??).

B. Data Model

As mentioned in the previous subsection, we assume that the model $p(y|x)$ is pre-trained using an arbitrary training technique and an arbitrary data set. Accordingly, we do not concern ourselves with the training data and with the training process in this paper. That said, the CP procedure requires a data set – distinct from training data – that is used to calibrate model $p(y|x)$ so as to obtain a reliable set-valued prediction in the sense of condition (??) for some input x. In conventional CP, such data set, known as calibration data set, is directly available at the decision-maker holding the model and the input x. In this paper, as in [?], we assume that, instead, calibration data sets are only present at the devices. By communicating information about such data to the server, the devices can facilitate the implementation of CP-based mechanisms to produce reliable estimates $\Gamma(x)$ that satisfy the inequality (??) for a desired reliability level $1-\alpha$. We will explain how CP works in the next section.

We assume that there are a total of N calibration data points, denoted as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, which are equally split across all K devices. Accordingly, each device k stores $N_d = N/K$ calibration data points, denoted as $\mathcal{D}_k = \{(x_{i,k}, y_{i,k})\}_{i=1}^{N_d}$. The union of all disjoint sets of data points $\mathcal{D}_k$ across all K devices recovers the overall calibration data set, i.e., $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$. Following the standard machine learning model, all calibration data points in data set $\mathcal{D}$ are assumed to be generated i.i.d. from some unknown distribution $p^*(x, y)$.

Furthermore, we denote a generic test data point as (x, y), which is also generated from distribution $p^*(x, y)$, independently from the calibration data. As explained, the input x of the test pair is known only to the server, while the corresponding label y is unknown and must be predicted by the server.

C. Communication Model

In prior work on federated inference via CP [?], [?], [?], [?], the communication channels between devices and server were assumed to be ideal. In contrast, in this work, we study the more challenging scenario in which devices are connected to the server via noisy channels. Accordingly, we will refer to this problem as wireless reliable federated inference.

Specifically, the K devices communicate with the server over a shared multiple access channel using T channel uses, or symbol periods. Each channel use carries a real-valued symbol, and we focus on Gaussian-noise channels [?]. In practice, fading channels can also be approximated by Gaussian-noise channels via pre-equalization, as often done in studies involving federated learning [?], [?], [?], [?]. As we will detail below, we consider two types of protocols, namely orthogonal-access systems in which each device uses distinct subsets of channel uses, and non-orthogonal protocols in which devices are simultaneously active on all channel uses.

We assume an average per-symbol power constraint P for each device k. Accordingly, denoting the per-symbol power of the channel noise as N_0, we define the signal-to-noise ratio (SNR) as

$$\text{SNR} = \frac{P}{N_0}. \tag{4}$$

1) Orthogonal Multiple Access: Time-division multiple access (TDMA) is a conventional orthogonal multiple access scheme that assigns distinct subsets of the T channel uses to the K devices. In this paper, we focus on equal allocations whereby all devices are assigned $\lfloor T/K \rfloor$ channel uses. Accordingly, in symbol period t assigned to device k, the received signal at the central server can be expressed as

$$y_t = x_{k,t} + z_t, \tag{5}$$

where $x_{k,t}$ is the (real) symbol transmitted by device k at time t and $z_t \sim \mathcal{N}(0, N_0)$ is the (real) channel noise.

2) Non-Orthogonal Multiple Access: As we will discuss in Sec. ??, the proposed scheme relies on TBMA, which is a form of non-orthogonal multiple access protocol. In general, in non-orthogonal protocols, all devices transmit concurrently in each symbol period t. Accordingly, the signal received at the server in period t can be written as

$$y_t = \sum_{k=1}^{K} x_{k,t} + z_t, \tag{6}$$

with the same definition given above for orthogonal protocols.

III. Background on Conformal Prediction

In this section, we provide a brief primer on CP in order to set the necessary background required by benchmarks and proposed schemes for the problem of wireless reliable federated inference described in the previous section. The presentation also includes discussions about the impact of quantization on the performance of CP, which is not covered in standard references on CP. Unlike the federated setting of interest in this work, conventional CP applies to a centralized scenario with a server holding all the available calibration data, which will be assumed throughout this section.

A. Validation-Based Conformal Prediction

We focus on a practical variant of CP, known as split, inductive, or validation-based CP, that operates on a pre-trained model $p(y|x)$ [?], [?], [?]. Given a new input x, the goal of CP is to produce a set predictor $\Gamma(x) \in \mathcal{Y}$ with the property of satisfying the reliability condition (??) for some pre-determined target reliability level $1-\alpha$. To this end, the server is given access not only to the model $p(y|x)$ and to a test input x, but also to a calibration data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of N data points. As in the previous section, the N calibration data points and the test data pair (x, y) are assumed to be i.i.d. according to an unknown distribution $p^*(x, y)$. The probability in (??) is evaluated with respect to the joint distribution of calibration and test data.

In the centralized setting under study here, the server builds the set predictor $\Gamma(x)$ using the test input x, the calibration data, and the model $p(y|x)$. Note that the true label y is not known at the server, since it is the subject of the inference process.

To this end, we introduce the nonconformity (NC) score function

$$s(x,y) = 1 - p(y|x). \tag{7}$$

The NC score is a measure of the loss of the model $p(y|x)$ on the data point (x,y). In fact, a large value indicates that the model assigns a low probability to example (x,y). Other NC scores are also possible [?], [?], [?], [?], and the methodology developed in this paper applies more broadly to any scores, as long as they are non-negative and upper bounded, i.e.,

$$0 \leq s(x,y) \leq 1. \tag{8}$$

Note that the upper bound is set to 1 without loss of generality since one can always re-scale a bounded NC score to fit in the range (??).

CP includes in the predicted set $\Gamma(x)$ all labels $y \in \mathcal{Y}$ with a NC score smaller than a given threshold s_α, i.e.,

$$\Gamma_\alpha(x) = \{y \in \mathcal{Y} : s(x,y) \leq s_\alpha\}. \tag{9}$$

As we discuss next, the threshold s_α is determined based on the target reliability level $1-\alpha$ in the reliability constraint (??). Dependence on parameter α is accordingly added in the subscript of the set predictor $\Gamma_\alpha(x)$.

B. Evaluation of the Threshold

To evaluate the threshold s_α, the server computes the NC scores (??) for all the N calibration data points, obtaining the collection $\mathcal{S} \triangleq \{s(x_i, y_i)\}_{i=1}^N$ of NC scores. Note that multiple data points may have the same NC score, which is accordingly counted multiple times. Then, the server sets the threshold s_α to be approximately equal to the $\lceil(1-\alpha)N\rceil$-th smallest NC score in the set $\mathcal{S}$ (counting possible repetitions). Intuitively, as the reliability level $1-\alpha$ increases, so does the threshold s_α, ensuring that the predicted set (??) includes a larger number of labels.

To formalize the operation of CP, let us introduce a function that, given a set $\mathcal{S}$, produces the $\lceil(1-\alpha)(N+1)\rceil$-th smallest value in the set. Note that the smallest value is evaluated with respect to a set with cardinality $N+1$, and not N, as required by CP (see, e.g., [?]). For any given collection of real numbers $\mathcal{S} = \{s_1, ..., s_N\}$ with possible repetitions, we denote as $s_{(1)} \leq s_{(2)} \ldots \leq s_{(N)}$ the sorted values in ascending order. Ties are broken arbitrarily. Then, the desired function is defined as

$$Q_{1-\alpha}(\mathcal{S}) \triangleq \begin{cases} s_{(\lceil (1-\alpha)(N+1) \rceil)} & \text{if } \alpha \geq 1/(N+1), \\ 1 & \text{otherwise,} \end{cases} \tag{10}$$

where $\lceil \cdot \rceil$ denotes the ceiling operation. Accordingly, function $Q_{1-\alpha}(\mathcal{S})$ returns the $\lceil (1-\alpha)(N+1) \rceil$-th smallest value in the set as long as $\lceil (1-\alpha)(N+1) \rceil \leq N$, or equivalently $\alpha \geq 1/(N+1)$; while returning the maximum value 1 otherwise.

The value $Q_{1-\alpha}(\mathcal{S})$ can also be interpreted as the $(1-\alpha)(N+1)/N$-quantile of the empirical distribution of the entries of set $\mathcal{S}$. In fact, the $(1-\alpha)(N+1)/N$-quantile of the empirical distribution is, by definition, the smallest number in the set $\mathcal{S}$ that is at least as large as a fraction $(1-\alpha)(N+1)/N$ of the elements in $\mathcal{S}$.

With function $Q_{1-\alpha}(\cdot)$, the CP set predictor (??) can be succinctly expressed as

$$\Gamma_{\alpha}(x) = \left\{ y \in \mathcal{Y} : s(x,y) \leq s_{\alpha}^{\text{CP}} \triangleq Q_{1-\alpha}(\mathcal{S}) \right\}. \tag{11}$$

As mentioned, it can be proved that the prediction set $\Gamma_{\alpha}(x)$ in (??) satisfies the reliability condition (??), irrespective of the accuracy of the underlying model $p(y|x)$ and of the ground-truth distribution $p^{*}(x,y)$ of the data [?], [?].

C. Quantized Conformal Prediction

As discussed in the previous section, in this paper, we are concerned with decentralized settings in which calibration data is not available at the server. In such a setting, communication between devices holding the calibration data and the server is limited by the available transmission resources. As a step in the direction of accounting for limitations arising from finite communication capacity, in this subsection, we discuss a centralized CP setting in which NC scores used to evaluate the threshold s_{α}^{CP} in (??) are constrained to take a discrete finite set of values.

To this end, we adopt a uniform scalar quantizer in which the range $[0,1]$ of possible values for the NC score, by assumption (??), is divided into M equal intervals $[S_0, S_1], (S_1, S_2], \ldots (S_{M-1}, S_M]$ with $S_0 = 0$ and $S_M = 1$. Given an input NC score $x \in [0,1]$, the quantized output $q(x)$ equals the upper value S_m of the interval $(S_{m-1}, S_m]$ containing x. Accordingly, the quantization function is defined as

$$q\left(x\right) \triangleq \begin{cases} S_1 & x \in [S_0, S_1], \\ S_m & x \in (S_{m-1}, S_m] \text{ for } m = 2, \ldots, M. \end{cases} \tag{12}$$

Suppose now that the server has access to the set of quantized NC scores $\mathcal{S}^q \triangleq \{q(s(x_i, y_i))\}_{i=1}^N$. Following the CP procedure, we define the set predictor as

$$\Gamma_\alpha^q(x) = \{y \in \mathcal{Y} : q(s(x, y)) \leq s_\alpha^{q\text{−CP}} \triangleq Q_{1-\alpha}(\mathcal{S}^q)\}, \tag{13}$$

that is, as the set of labels $y \in \mathcal{Y}$ whose quantized NC scores $q(s(x, y))$ are no larger than the $\lceil(1 - \alpha)(N + 1)\rceil$-th smallest NC score, $Q_{1-\alpha}(\mathcal{S}^q)$, in the calibration set.

Since any function of the input-output pair (x, y) is a valid NC score, so is the quantized value $q(s(x, y))$. Therefore, the quantized predicted set $\Gamma_\alpha^q(x)$ satisfies the reliability condition (??). However, one should generally expect that, due to information loss caused by quantization, the size of the predicted set $\Gamma_\alpha^q(x)$ is generally larger than that of the predicted set $\Gamma_\alpha(x)$ obtained from the original NC score function $s(x, y)$.

D. Quantized Conformal Prediction via Empirical Quantiles

In this subsection, we make the observation that the threshold $s_\alpha^{q\text{−CP}} = Q_{1-\alpha}(\mathcal{S}^q)$ used in the set predictor (??) can be expressed in terms of the empirical distribution of the quantized NC scores in set $\mathcal{S}^q$. More precisely, it can be evaluated, approximately, as the $(1 - \alpha)$-quantile of the empirical distribution. This fact will be instrumental in the design of the proposed federated inference protocol in Sec. ??.

To elaborate, let us define as $p_m \in [0, 1]$ the fraction of quantized NC scores equal to S_m in the set of quantized NC scores $\mathcal{S}^q$, i.e.,

$$p_m = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{m_i = m\}, \tag{14}$$

where m_i is the index of the quantized i-th NC score, i.e., $S_{m_i} = q(s(x_i, y_i))$. We collect all M fractions into the vector

$$\boldsymbol{p} = [p_1, \ldots, p_M]^{\mathrm{T}}, \tag{15}$$

which satisfies the equality $\sum_{m=1}^{M} p_m = 1$. As we have discussed, CP relies on the evaluation of the $\lceil(1 - \alpha)(N + 1)\rceil$-th smallest element in the set $\mathcal{S}^q$, i.e., $Q_{1-\alpha}(\mathcal{S}^q)$. To evaluate this quantity, we modify the empirical distribution of the quantized NC scores by adding a fictitious $(N+1)$-th NC score equal to the maximum value. This yields the empirical distribution vector

$$\boldsymbol{p}^{+} = \frac{N}{N+1}\boldsymbol{p} + \left[0, \ldots, \frac{1}{N+1}\right]^{\mathrm{T}}. \tag{16}$$

With this definition, the $\lceil (1-\alpha)(N+1) \rceil$-th smallest element in set $\mathcal{S}^{q}$ can be obtained as the quantization level $S_{m_{\alpha}(\boldsymbol{p}^{+})}$, where the index $m_{\alpha}(\boldsymbol{p}^{+})$ is obtained by evaluating the $(1-\alpha)$-quantile of the empirical distribution $\boldsymbol{p}^{+}$, i.e.,

$$m_{\alpha}(\boldsymbol{p}^{+}) = \min\left\{ m \in \{1, \ldots, M\} : p_1^{+} + \cdots + p_m^{+} \geq 1-\alpha \right\}. \tag{17}$$

IV. Digital Wireless Federated Conformal Prediction

While the conventional CP scheme reviewed in the previous section assumes that predictive model and calibration data are both present at the server, in the wireless reliable federated inference setting as explained in Sec. ??, calibration data are only available at the devices. In this section, we first review the FedCP-QQ scheme proposed in [?], which addresses this problem by assuming noiseless links from devices to server that can support the noiseless transmission of a single real number from each device. Then, as a benchmark, we describe a direct digital wireless implementation of FedCP-QQ that accounts for the presence of noisy channels between devices and server.

A. Federated Conformal Prediction with Noiseless Communications

The FedCP-QQ scheme introduced in [?] is based on the quantile-of-quantiles (QQ) operation. Accordingly, as we detail next, it sets two probabilities α_d and α_s to identify target quantiles to be computed at devices and server, respectively.

Each device k has access to the local NC scores $\mathcal{S}_k = \{s(x_{i,k}, y_{i,k})\}_{i=1}^{N_d}$. Based on this collection of NC scores, it computes the $(1-\alpha_d)(N_d+1)/N_d$-quantile $Q_{1-\alpha_d}(\mathcal{S}_k)$. This real positive number is then communicated noiselessly to the server.

The server collects all the quantiles $\mathcal{Q}_{1-\alpha_d}^{1:K} = \{Q_{1-\alpha_d}(\mathcal{S}_1), \ldots, Q_{1-\alpha_d}(\mathcal{S}_K)\}$ from the K devices, and it evaluates the $(1-\alpha_d)(K+1)/K$-quantile of the K quantiles, i.e.,

$$s_{\alpha}^{\mathrm{QQ}} \triangleq Q_{1-\alpha_s}(\mathcal{Q}_{1-\alpha_d}^{1:K}). \tag{18}$$

The set predictor of the FedCP-QQ scheme is constructed using the obtained threshold as

$$\Gamma^{\mathrm{QQ}}_{\alpha_d,\alpha_s}(x) = \left\{ y \in \mathcal{Y} : s(x,y) \leq s^{\mathrm{QQ}}_{\alpha} \right\}. \tag{19}$$

The pair of miscoverage levels (α_d, α_s) must be selected in order to satisfy the coverage condition (??). To this end, reference [?] proved the following result.

Theorem 1 (Theorem 3.2 [?]). For any $(\alpha_d, \alpha_s) \in [1/(N_d+1), 1) \times [1/(K+1), 1)$, the coverage of the set predictor $\Gamma_{\alpha_d,\alpha_s}(x)$ is lower bounded as

$$\Pr\left(y \in \Gamma^{\mathrm{QQ}}_{\alpha_d,\alpha_s}(x)\right) \geq 1 - \frac{1}{N+1} \sum_{j=k}^{K} \binom{K}{j} \sum_{I_{1,j}=n}^{N_d} \sum_{I^c_{1,j}=0}^{n-1} \frac{\binom{N_d}{i_1} \cdots \binom{N_d}{i_m}}{\binom{N}{i_1+\cdots+i_K}} \triangleq M_{\alpha_d,\alpha_s} \tag{20}$$

with $n = \lceil (N_d+1)(1-\alpha_d) \rceil$; $k = \lceil (K+1)(1-\alpha_s) \rceil$; $I_{1,j} = \{i_1, \ldots, i_j\}$; $I^c_{1,j} = \{i_{j+1}, \ldots, i_K\}$; and the operation $\sum_{I_{1,j}=n}^{N}$ stands for the cascade of summations that takes into account for all elements in $I_{1,j}$ starting from n up to N, i.e., $\sum_{I_{1,j}=n}^{N} = \sum_{i_1=n}^{N} \sum_{i_2=n}^{N} \cdots \sum_{i_j=n}^{N}$.

With this result, one can find a pair of miscoverage levels (α_d, α_s) that minimizes the lower bound M_{α_d,α_s} while satisfying the target coverage rate $1-\alpha$. The optimization objective can be formulated as

$$(\alpha^*_d, \alpha^*_s) \in \arg\min_{\alpha_d,\alpha_s} \left\{ M_{\alpha_d,\alpha_s} : M_{\alpha_d,\alpha_s} \geq 1-\alpha \right\}. \tag{21}$$

If the solution of (??) is not unique, it is suggested to find the pairs with the largest value α^*_d and then choose among those the pair with the largest value α^*_s. Efficient ways to address this problem are discussed in [?], which also covers the more general case in which devices have different data set sizes.

B. Digital Transmission Benchmark

In this subsection, we propose a digital implementation of the FedCP-QQ scheme, which we refer to as Digital FedCP-QQ or DQQ for short. A direct implementation of the FedCP-QQ scheme requires every device k to quantize its local quantile $Q_{1-\alpha_d}(\mathcal{S}_k)$ in (??) before transmission in order to meet the capacity constraints on the shared noisy channel to the receiver. To this end, the device k applies the function $q(\cdot)$ defined in (??) to quantize the local quantile into one of M levels. Then, each device uses conventional digital communications to convey the quantized quantile to the server.

Specifically, to transmit the quantized local quantiles from K devices on the shared channel, we adopt a TDMA protocol whereby, as discussed in Sec. ??, the K devices are assigned $\lfloor T/K \rfloor$ channel uses each. Accordingly, the probability of error for each device can be closely approximated as [?, Theorem 54]

$$\epsilon = Q\left(\frac{\lfloor T/K \rfloor C - \log M}{\sqrt{\lfloor T/K \rfloor V}}\right), \tag{22}$$

in which the Q-function is complementary cumulative distribution function of a standard Gaussian variable; the capacity C is given by

$$C = \frac{1}{2}\log(1 + \mathrm{SNR}), \tag{23}$$

and the channel dispersion V is defined as

$$V = \frac{\mathrm{SNR}}{2}\frac{\mathrm{SNR}+2}{(\mathrm{SNR}+1)^2}\log^2 e. \tag{24}$$

Accordingly, with probability ϵ, the transmission is unsuccessful. Assuming that the server can detect errors, the QQ estimator (??) can be applied on the subset of quantiles that are received correctly. Note that the bound in (??) should now be evaluated by including only the correctly received quantiles from the devices.

While the resulting set predictor satisfies the reliability condition (??) by Theorem 1, the impact of lost quantiles due to channel errors is that of reducing the number of active devices, and hence the amount of calibration data effectively accessible by the server. This, in turn, generally increases the average predicted set size (??).

V. Wireless Federated Conformal Prediction

In this section, we introduce the proposed Wireless Federated Conformal Prediction (WFCP) scheme. WFCP implements a novel combination of TBMA and over-the-air computing to communicate the empirical distribution of the quantized NC scores from the devices to the server. Via over-the-air computing, thanks to the superposition property of the multiple access channel (??), the server obtains a noisy and unbiased estimate of the empirical distribution of the NC scores across all devices. Based on this estimate, the server computes an estimate of a global empirical quantile, which is judiciously selected to ensure the coverage property (??).

Unlike the existing FedCP-QQ scheme reviewed in the previous section, WFCP does not require devices to compute their local quantiles. This local computation, implemented by

FedCP-QQ to reduce bandwidth requirements, generally results in a performance loss, since the QQ estimator (??) used by FedCP-QQ cannot recover the global quantile required to implement CP on the overall calibration data set stored across all devices as reviewed in Sec. ??. This result could only be achieved by communicating separately all the quantized NC scores from devices to server. However, for a given transmission reliability level (??), this transmission would require a number T of channel uses that scale linearly with the number of calibration data points across all devices.

To mitigate this loss, WFCP enables a direct estimate of the global quantile at the server without imposing bandwidth requirements that scale linearly with the number of devices. Rather, the communication requirements of WFCP are only dictated by the precision with which the NC scores are represented for transmission to the server.

In the following, we first detail the transmission protocol based on TBMA adopted by WFCP. Then, we describe the set predictor implemented by WFCP on the basis of the received baseband signal. As it will be detailed, the main challenge in ensuring the reliability condition (??) is the determination of a suitable correction for the quantile estimated based on the noisy received signals. Finally, we provide some discussion on the trade-offs involved in the design choices, and we prove that condition (??) is guaranteed by WFCP.

A. TBMA-Based Communication Protocol

In WFCP, unlike FedCP-QQ, the devices do not first compute the empirical quantiles of their respective local NC scores. Rather, each device quantizes separately each of the N_d local NC scores using the uniform quantizer (??) with M levels. We denote as $m_{i,k} \in \{1, \ldots, M\}$ the index of the quantization value produced for the i-th NC score at device k, $s(x_{i,k}, y_{i,k})$, i.e.,

$$S_{m_{i,k}} = q(s(x_{i,k}, y_{i,k})). \tag{25}$$

Furthermore, in a manner that mirrors the presentation in Sec. ??, we introduce an M-dimensional probability vector that collects the quantized NC scores at device k as

$$\boldsymbol{p}_k = [p_{1,k}, \ldots, p_{M,k}]^{\mathrm{T}}, \tag{26}$$

where the probability

$$p_{m,k} = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathbb{1}\{m_{i,k} = m\} \tag{27}$$

corresponds to the fraction of NC scores associated to quantization level m at device k, such that the equality $\sum_{m=1}^{M} p_{m,k} = 1$ holds.

As an intermediate goal, WFCP obtains an unbiased estimate of the global empirical distribution $\boldsymbol{p}$ in (??) of the quantized NC scores across all devices. To this end, we first note that we have the equality

$$p_m = \frac{1}{K}\sum_{k=1}^{K} p_{m,k}, \tag{28}$$

and hence the fraction p_m of NC scores in the m-th quantization bin is equal to the corresponding fractions $p_{m,k}$ across all devices k. Once such an estimate is available, WFCP can apply the procedure discussed in Sec. ?? in order to mimic the operation of centralized quantized CP. As we will see, this requires a judicious adjustment of the threshold used in evaluating the predicted set (??).

To obtain an estimate of distribution $\boldsymbol{p}$, WFCP leverages TBMA and over-the-air computing. With TBMA, the devices share a codebook $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M] \in \mathbb{R}^{M\times M}$ of M orthogonal codewords, where each codeword $\boldsymbol{c}_m \in \mathbb{R}^{M\times 1}$ consists of M real symbols. The number M of channel uses should not be larger than the available number T of symbols. We will discuss the choice of the number of quantization points in Sec. ??. Assuming that each codeword is normalized to satisfy the energy constraint $\|\boldsymbol{c}_m\|^2 = 1$, by the orthogonality of the codewords, we have the equality $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} = \mathbf{I}_{M\times M}$. Each codeword $\boldsymbol{c}_m$ is assigned to the m-th quantization level S_m for $m \in \{1, \ldots, M\}$.

Accordingly, each device k transmits a superposition of the codewords $\boldsymbol{c}_{m_{i,k}}$ that correspond to the N_d quantized NC score $S_{m_{i,k}}$ for $i = 1, ..., N_d$. We use the simplified notation $\boldsymbol{c}_{i,k} = \boldsymbol{c}_{m_{i,k}}$. To express the transmitted signal mathematically, let us denote as $\boldsymbol{u}_{i,k} \in \mathbb{B}^{M\times 1}$ the one-hot vector with all zero entries except for a single 1 at the $m_{i,k}$-th position. The scaled superposition of the codewords transmitted by the device k can then be written as

$$\boldsymbol{x}_k = \gamma_k \sum_{i=1}^{N_d} \boldsymbol{c}_{i,k} = \gamma_k \boldsymbol{C} \sum_{i=1}^{N_d} \boldsymbol{u}_{i,k} = \gamma_k N_d \boldsymbol{C}\boldsymbol{p}_k, \tag{29}$$

where $\gamma_k > 0$ is a power control gain at device k, and we recall that $\boldsymbol{p}_k$ in (??) is the empirical probability vector of the quantized NC scores at device k.

All devices transmit simultaneously on the shared Gaussian-noise channel (??), Accordingly, thanks to the superposition property of the multiple access channel, the received signal at the server is given by

$$\boldsymbol{y}=\sum_{k=1}^{K}\boldsymbol{x}_k+\boldsymbol{z}=\boldsymbol{C}N_d\sum_{k=1}^{K}\gamma_k\boldsymbol{p}_k+\boldsymbol{z}, \tag{30}$$

where $\boldsymbol{z}\in\mathbb{R}^{M\times 1}\sim\mathcal{N}(\mathbf{0},N_0\mathbf{I})$ is an i.i.d. Gaussian noise vector with zero mean and variance N_0.

B. Estimate of the Global Empirical Distribution

The server wishes to extract an estimate of the global empirical distribution $\boldsymbol{p}$ of the quantized NC scores in (??) from the received signal (??). To this end, matched filtering is applied by left-multiplying the received signal by matrix $\boldsymbol{C}^{\mathrm{T}}$, yielding

$$\boldsymbol{w}=\boldsymbol{C}^{\mathrm{T}}\boldsymbol{y}=N_d\sum_{k=1}^{K}\gamma_k\boldsymbol{p}_k+\boldsymbol{C}^{\mathrm{T}}\boldsymbol{z}. \tag{31}$$

From (??), we note that the server obtains a weighted sum of the local empirical distribution vectors. In order to recover a noisy version of the global empirical distribution vector $\boldsymbol{p}$, the power control coefficients at the transmitters are set to be equal, i.e., $\gamma=\gamma_1=\cdots=\gamma_K$. The common scaling parameter γ must satisfy the average per-symbol transmit power constraint P. This constraint results in the inequalities

$$\|\boldsymbol{x}_k\|^2=\gamma^2N_d^2\|\boldsymbol{p}_k\|^2\leq MP, \quad \text{for } k=1,\ldots,K \tag{32}$$

or equivalently in the single inequality

$$\gamma\leq\frac{\sqrt{MP}}{N_d\max_k\|\boldsymbol{p}_k\|}=\frac{\sqrt{MP}}{N_d2^{-\min_k H_2(\boldsymbol{p}_k)}}, \tag{33}$$

where $H_2(\boldsymbol{p}_k)$ is the 2-Rényi entropy [?]

$$H_2(\boldsymbol{p}_k)=-\log_2\left(\sum_{m=1}^{M}p_{m,k}^2\right) \tag{34}$$

of the empirical probability vector $\boldsymbol{p}_k$. Inequality (??) reflects the fact that a more concentrated empirical distribution, with a smaller 2-Rényi entropy, yields a stricter restriction on the choice of the transmit power. Given that, in general, no prior information is available on the distribution of the NC scores, we set the power scaling factor γ by considering the worst-case situation, yielding the choice

$$\gamma=\frac{\sqrt{MP}}{N_d}. \tag{35}$$

With (??) in (??), the matched filtered received signal is given by

$$\boldsymbol{w} = \sqrt{MP}\sum_{k=1}^{K}\boldsymbol{p}_k + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{z} = \sqrt{MP}K\boldsymbol{p} + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{z}, \tag{36}$$

which is indeed a scaled and noisy version of the empirical probability distribution of all the $N = KN_d$ NC scores from K devices.

C. Set Predictor

Following the steps presented in the previous subsection, the server recovers the scaled and noisy version $\boldsymbol{w}$ of the empirical distribution $\boldsymbol{p}$ of all the calibration NC scores. In the ideal case of noiseless channels, i.e., with $N_0 = 0$, the server would have access to the empirical distribution $\boldsymbol{p}$ of all the NC scores, and the quantized CP procedure in Sec. ?? could be directly applied to obtain a reliable set predictor.

To address the availability of an estimate of vector $\boldsymbol{p}$, the proposed WFCP preprocesses the scaled and noisy version of the global empirical distribution (??), and then it computes the $(1-\alpha_c)(N+1)/N$-quantile of the distribution at a corrected unreliability level $\alpha_c < \alpha$, accounting for the presence of channel noise. We will demonstrate that, with a specific choice of the corrected unreliability level α_c, the proposed approach preserves the coverage property (??), with probability now taken also over the channel noise.

First, in order to facilitate the estimate of the global empirical distribution of all the quantized NC scores (??) from the estimate (??), the server carries out two steps: (i) it rescales the vector (??) by $N/(\sqrt{MP}K(N+1))$; and (ii) it adds $1/(N+1)$ to the last entry of the received signal vector. Step (ii) amounts to the same operation carried out in (??) within the centralized quantized CP scheme reviewed in Sec. ??.

This preprocessing yields the $M \times 1$ vector

$$\boldsymbol{v} = \frac{N}{\sqrt{MP}K(N+1)}\boldsymbol{w} + \left[0, \ldots, \frac{1}{N+1}\right]^{\mathrm{T}} = \boldsymbol{p}^{+} + \tilde{\boldsymbol{z}}, \tag{37}$$

where $\boldsymbol{p}^{+}$, as defined in (??), is the empirical distribution vector of the aggregated NC scores from the K devices as well as an additional NC score at the maximum quantization level S_M; and $\tilde{\boldsymbol{z}} \triangleq N/(\sqrt{MP}K(N+1))\boldsymbol{C}^{\mathrm{T}}\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is the effective noise vector with power

$$\sigma^2 = \frac{N^2}{MPK^2(N+1)^2}N_0 = \frac{N_d^2}{M\mathrm{SNR}(N+1)^2} \approx \frac{1}{M\mathrm{SNR}K^2}. \tag{38}$$

Note that for a fixed number N of NC scores, the effective noise power is inversely proportional to the square of the number K of devices. This can be interpreted as a form of coherent gain due to the use of TBMA [?]. Furthermore, for a fixed number K of transmitting devices, the smallest value of the effective noise power σ^2 is obtained when every device sends exactly one NC score, i.e., when $N_d = 1$ and $N = K$.

Then, WFCP computes the index $m_{\alpha_c}(\boldsymbol{v})$ corresponding to the $(1-\alpha_c)$-"quantile" of the noisy distribution $\boldsymbol{v}$. Note that, since the vector $\boldsymbol{v}$ is not normalized, and its entries may be even negative, the index $m_{\alpha_c}(\boldsymbol{v})$ does not correspond to a true quantile in general. To proceed, we define the set

$$\begin{aligned}\mathcal{M}_{\alpha_c}(\boldsymbol{v}) &= \Big\{m \in \{1,\dots,M\} : v_1 + \dots + v_m \geq 1 - \alpha_c\Big\} \\ &= \Big\{m \in \{1,\dots,M\} : p_1^+ + \dots + p_m^+ \geq 1 - \alpha_c - (\tilde{z}_1 + \dots + \tilde{z}_m)\Big\}. \end{aligned} \tag{39}$$

Noting that there may be no value $m \in \{1,\dots,M\}$ that satisfies the inequality in (??), we define the index $m_{\alpha_c}(\boldsymbol{v})$ as

$$m_{\alpha_c}(\boldsymbol{v}) = \begin{cases} M & \text{if } \mathcal{M}_{\alpha_c}(\boldsymbol{v}) = \emptyset, \\ \min \mathcal{M}_{\alpha_c}(\boldsymbol{v}) & \text{otherwise.} \end{cases} \tag{40}$$

In the absence of noise, i.e., with $\sigma^2 = 0$, the index $m_{\alpha_c}(\boldsymbol{v})$ corresponds to the index in (??) computed by the centralized quantized CP with $\alpha_c = \alpha$.

The index (??) is used to define the WFCP set predictor

$$\Gamma_{\alpha_c}^{\mathrm{WFCP}}(x|\boldsymbol{v}) = \{y \in \mathcal{Y} : q(s(x,y)) \leq s_{\alpha}^{\mathrm{WFCP}} \triangleq S_{m_{\alpha_c}(\boldsymbol{v})}\}. \tag{41}$$

The WFCP predicted set (??) coincides with the quantized CP predictor (??) when $\sigma^2 = 0$ with $\alpha_c = \alpha$.

D. Optimization of the Number of Quantization Levels

The number of quantization levels, M, causes an increase in the number of channel uses necessary for the server to recover vector $\boldsymbol{v}$ in (??). On the flip side, in the ideal case of noiseless transmission, a larger value of M generally yields a more informative set predictor thanks to the higher resolution of the NC scores.

As explained in Sec. ??, the system has access to T channel uses for transmission from devices to server. A possible choice of the number of quantization levels would be to set

$M = T$. As we argue next, this may not necessarily be the optimal design. In fact, choosing $M < T$ enables the application of a form of repetition coding, which, in turn, can reduce the effective noise power in (??) by increasing the SNR.

In fact, with $M < T$, a simple repetition coding strategy stipulates that the devices repeat their transmissions $R \triangleq \lfloor T/M \rfloor$ times, where R is the repetition rate. With this approach, the effective SNR, upon averaging the matched filter outputs (??), equals

$$\mathrm{SNR}_{\mathrm{rep}} = R \cdot \mathrm{SNR} = \lfloor T/M \rfloor \mathrm{SNR}. \tag{42}$$

Overall, the choice of the number M of quantization levels entails a tension between improving the resolution of the CP set predictor, which would require increasing the value of M, and decreasing the effective noise power (??), which calls for a decrease in the value of M.

E. Reliability Analysis

WFCP satisfies the following reliability guarantee.

Theorem 2. Select the corrected unreliability level as

$$\alpha_c = \alpha - \frac{\sigma^2 M}{4\alpha} = \alpha - \frac{N_d^2}{4\alpha \lfloor T/M \rfloor \mathrm{SNR}(N+1)^2}. \tag{43}$$

Then, the WFCP set predictor (??) satisfies the reliability guarantee

$$\Pr\left(y \in \Gamma_{\alpha_c}^{\mathrm{WFCP}}(x|\boldsymbol{v}) \right) \geq 1 - \alpha, \tag{44}$$

where the average is taken with respect to the joint distribution of calibration and test data, as well as over the channel noise.

Proof: The proof is detailed in Appendix ??.

Theorem ?? determines a correction for the threshold level $\alpha_c < \alpha$ in the presence of non-zero Gaussian-noise, in order to ensure the satisfaction of the reliability constraint (??). The correction term $\sigma^2 M/4\alpha$ increases with the effective channel noise power σ^2 and with the number M of quantization levels. Plugging in the definition (??) of the effective noise power with the SNR (??), the correction term can be expressed as

$$\frac{\sigma^2 M}{4\alpha} = \frac{N_d^2}{4\alpha \lfloor T/M \rfloor \mathrm{SNR}(N+1)^2} \approx \frac{M}{4\alpha T \mathrm{SNR} K^2}, \tag{45}$$

which is inversely proportional to the square of the number K of devices and to the number T of channel uses available for transmission. The dependence on K is particularly noteworthy: While conventional protocols like DQQ require communication resources in terms of channel uses T and/or SNR, which become increasingly more stringent as K increases, WFCP can benefit from the presence of multiple devices. In particular, as K grows, the correction term in (??) decreases as $1/K^2$, allowing WFCP to set a target reliability level $1-\alpha_c$ that becomes increasingly close to the true target $1-\alpha$.

VI. Experimental Settings and Results

In this section, we provide insights into the performance of the proposed WFCP via numerical results. We use as a benchmark the digital wireless implementation of FedCP-QQ [?], abbreviated as DQQ, reviewed in Section ??.

A. Setting

Following the experimental setting in reference [?], we use the CoronaHack data set, a public chest X-ray data set involving 5908 images classified using $C = 3$ labels, namely normal, viral pneumonia, and bacterial pneumonia. We use $N^{\text{tr}} = 5400$ data pairs for training the predictive model, while the remaining 508 data pairs are divided into $N = 300$ points for calibration and $N^{\text{te}} = 208$ points for testing. In the federated inference setup under study, each device holds N/K calibration data points, while only the server has access to the N^{te} test data points on which it wishes to generate reliable predictions.

The predictive model adopts the ResNet-18 architecture [?] with minor modifications. Specifically, the last layer is replaced by a linear layer with $C = 3$ output neurons, followed by a softmax layer that outputs the conditional probability distribution $p(y|x)$. We train the model using the standard federated gradient descent protocol [?]. To this end, we divide the N^{tr} training examples evenly across all devices. Following federated stochastic gradient descent, the server collects and averages the local gradients from a subset of the devices that are evaluated based on the respective local training data to update the model parameters via stochastic gradient descent. Specifically, we utilize cross-entropy as the loss function, while adopting the Adam optimizer with a learning rate of 0.001 over 50 epochs for the update procedure at the server. As depicted in Fig. ??, the trained predictive model $p(y|x)$ is deployed at both devices and server.

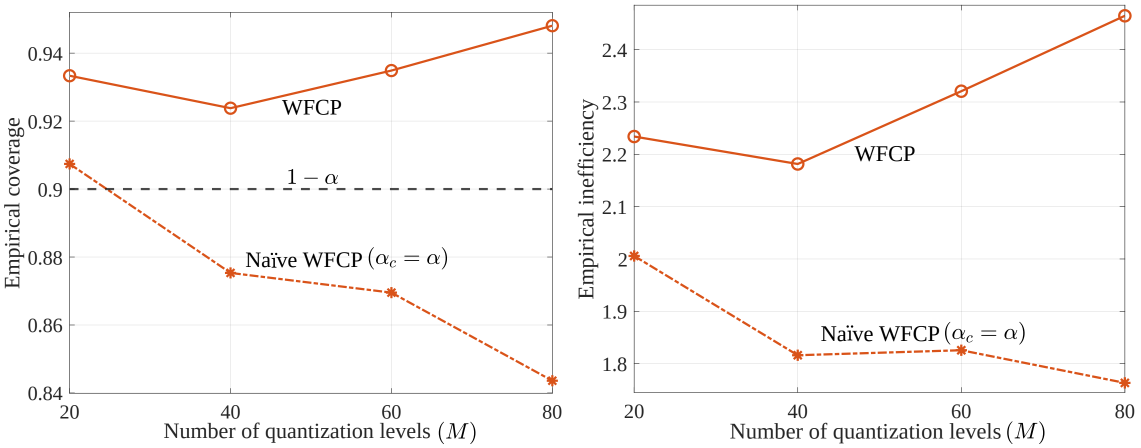

Fig. 3. Empirical coverage and empirical inefficiency of WFCP and naïve WFCP ($\alpha_c = \alpha$) versus the number M of quantization levels with target unreliability level $\alpha = 0.1$, number $T = 120$ of channel uses, number $K = 10$ of devices, and SNR $= -10$ dB.

Since training is done offline and since our focus is on the inference phase, we do not account for constraints on the communication links during training. Training techniques that operate on noisy channels, as in [?], [?], [?], [?], can be directly accommodated within the proposed federated inference framework.

We adopt as performance measures the empirical coverage and empirical inefficiency, which are defined respectively as

$$\text{Empirical coverage} = \frac{1}{N^{\text{te}}} \sum_{i=1}^{N^{\text{te}}} \mathbb{1}\left(y_i \in \Gamma(x_i)\right) \tag{46}$$

and

$$\text{Empirical inefficiency} = \frac{1}{N^{\text{te}}} \sum_{i=1}^{N^{\text{te}}} \left|\Gamma(x_i)\right|. \tag{47}$$

We run independent 400 experiments to evaluate the above criteria, and obtain an average. Each experiment involves a random split of the 508 data points not used for training into N calibration and N^{te} test pairs.

B. On the Choice of the Number of Quantization Levels

We start by focusing on the performance of the proposed WFCP scheme as a function of the number of quantization levels, M, for a fixed number $T = 120$ of channel uses. This study is meant to substantiate the discussion in Sec. ?? on the optimal choice of M as a

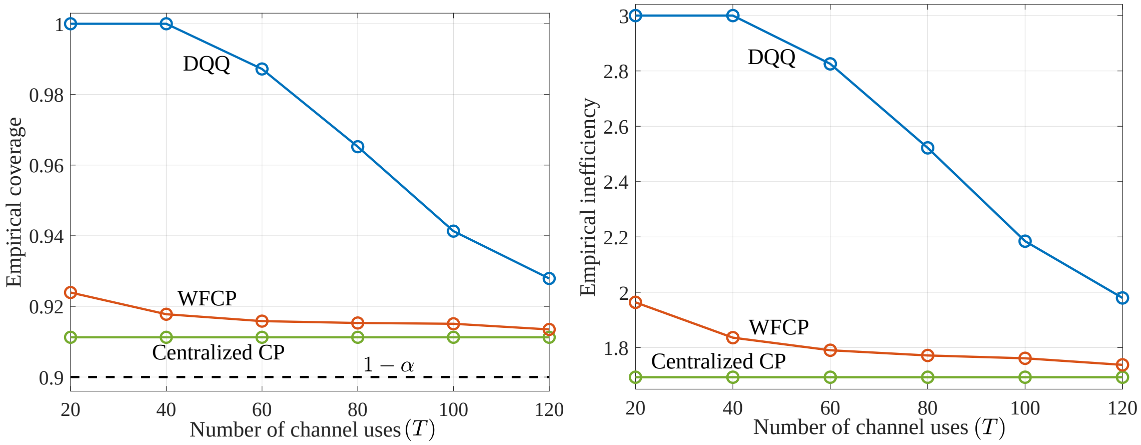

Fig. 4. Empirical coverage and empirical inefficiency of centralized CP, WFCP, and DQQ [?] versus the number T of channel uses available with target unreliability level $\alpha = 0.1$, number $M = 20$ of quantization levels, number $K = 20$ of devices, and $\text{SNR} = 0$ dB.

trade-off between improved effective SNR, requiring a smaller M, and a larger resolution, calling for a larger M. For reference, we also consider a naïve implementation of WFCP which simply sets the target reliability level $1 - \alpha_c$ in (??) to the true target $1 - \alpha$ without considering the impact of channel noise.

Fig. ?? shows empirical coverage and empirical inefficiency for $\alpha = 0.1$, $K = 10$ devices, and $\text{SNR} = -10$ dB as a function of M. As a first observation, confirming Theorem ??, WFCP achieves the target coverage reliability condition (??) for all quantization levels M. To obtain this goal, applying the corrected target reliability level $1 - \alpha_c$ in (??) is essential. In fact, as also seen in the figure, the naïve implementation of WFCP fails to meet the coverage requirements (??) as soon as M is sufficiently large, in which regime the performance is more sensitive to the presence of channel noise. For WFCP, the optimal value of M in terms of inefficiency is observed to be around $M = 40$, with smaller values causing a degraded performance due to an insufficient resolution and larger values being impaired by the smaller effective SNR.

C. Comparison between WFCP and DQQ

We now turn to comparing the performance of WFCP and DQQ (Sec. ??). We start by evaluating empirical coverage and empirical inefficiency as a function of the number T of channel uses for a fixed number $M = 20$ of quantization levels. As seen in Fig. ??, as T

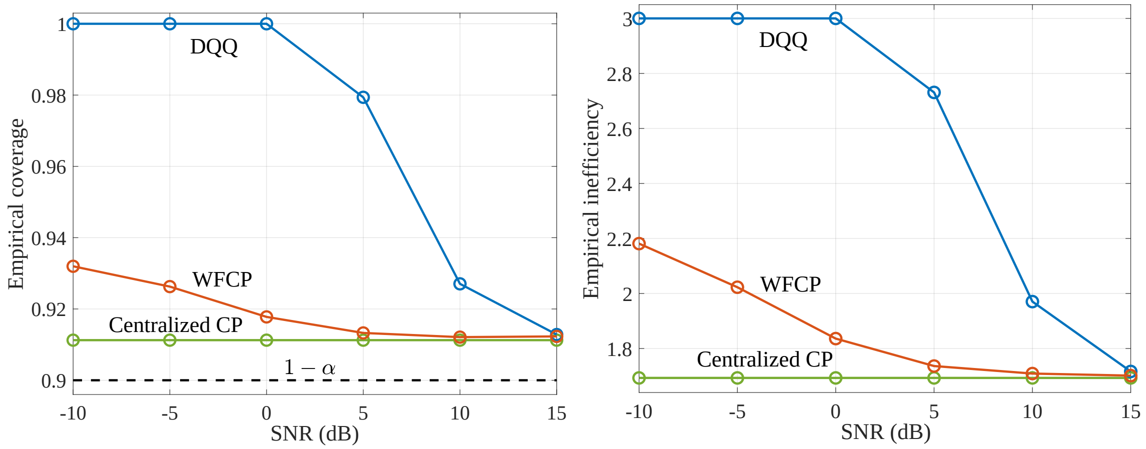

Fig. 5. Empirical coverage and inefficiency of centralized CP, WFCP, and DQQ [?] versus SNR with target unreliability level $\alpha = 0.1$, number $M = 20$ of quantization levels, number $T = 40$ of channel uses, and number $K = 20$ of devices.

increases, both methods maintain the target $(1-\alpha)$-coverage, while offering a decreasing inefficiency. This is because a larger T weakens the effect of channel noise by reducing the probability of error ϵ in (??) for DQQ, and by improving the effective SNR in (??) for WFCP. The proposed WFCP consistently outperforms DQQ, yielding highly informative prediction sets, with efficiency improvements being particularly evident in the regime of limited communication resources with low number T of channel uses. As T grows sufficiently large, the performance of both schemes approaches that of the centralized noiseless CP (Sec. ??).

The performance gains of WFCP in the presence of limited communication resources are further explored in Fig. ??, which evaluates the performance of WFCP and DQQ as a function of the SNR. As the SNR increases, the effective SNR in (??) improves along with a decrease in the correction term in (??), resulting in a more informative predicted set, which approaches the performance of the centralized CP. In a similar manner, as the SNR improves, the probability of error ϵ in (??) for DQQ decreases, thereby generating a smaller-sized predicted set, which approaches the performance of WFCP for SNR levels around 15 dB.

Fig. ?? evaluates the performance of WFCP and DQQ when varying the number of devices, K. Note that the number $N_d = 10$ of per-device calibration data points is kept fixed, so that, as K increases, the total number of calibration data points increases. For

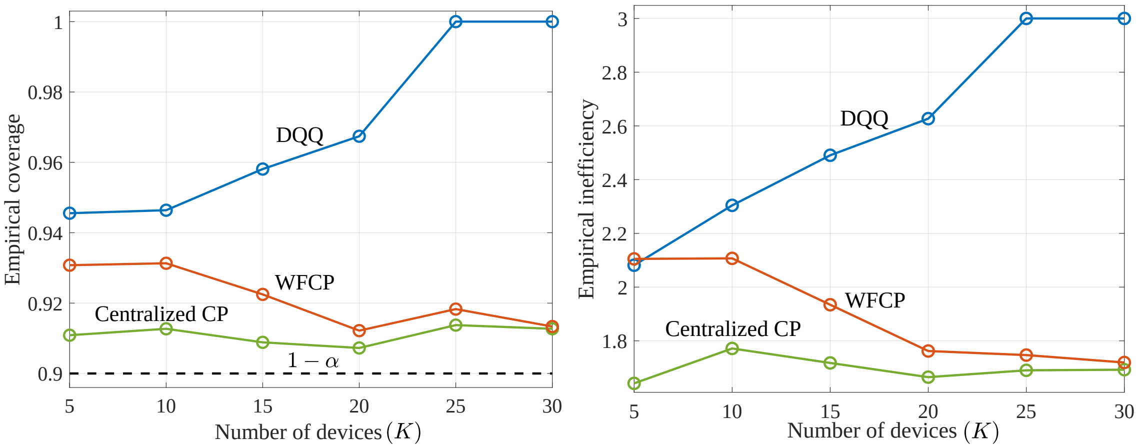

Fig. 6. Empirical coverage and inefficiency of centralized CP, WFCP, and DQQ [?] versus the number K of devices with $N_d = 10$ per-device calibration data points, target unreliability level $\alpha = 0.1$, number $M = 20$ of quantization levels, number $T = 60$ of channel uses, and $\mathrm{SNR} = 0$ dB.

DQQ, as the number of devices increases, the inefficiency tends to increase. In fact, an increase in the number of devices leads to a higher error probability ϵ in (??), which causes the average number of correctly received local quantiles, $K(1-\epsilon)$, to decrease.

In stark contrast, WFCP is observed to reduce the average predicted set size as the number K of devices increases. Intuitively, this is due to the adoption of the TBMA protocol, which allows the on-air combination of signals transmitted by all the devices. At a technical level, this result is aligned with (??), which shows that the correction term is approximately independent of the number of calibration data per device and that it is inversely proportional to the square of the number of devices, K. Accordingly, as K grows, the corrected target reliability level $1-\alpha_c$ approaches the true level $1-\alpha$, and the performance of WFCP approaches that of centralized CP.

VII. Conclusions

This paper has introduced wireless federated conformal prediction (WFCP), the first protocol for the deployment of federated inference via CP in shared noisy communication channels. Like conventional centralized CP and some of the existing federated extensions of CP for noiseless channels, WFCP provably provides formal guarantees of reliability, indicating that the predicted set produced at the server contains the true output with any target probability. WFCP builds on type-based multiple access (TBMA), a communication protocol that allows the estimate of a global histogram from distributed observations with a bandwidth that scales with the resolution of the histogram and not with the number of devices. The key technical challenge tackled by this paper is the definition of a novel quantile correction approach that ensures the reliability of the set predictor despite the presence of channel noise. The theoretical analysis of WFCP's reliability performance also offers valuable insight into the choice of critical design parameters, such as the number of quantization levels. Simulation results further substantiate the advantage of the proposed WFCP scheme over existing strategies, particularly under constraints of limited communication resources and/or large number of devices. All in all, the proposed WFCP protocol provides a promising framework for implementing federated CP in wireless communication scenarios, thereby establishing a robust foundation for future exploration in this domain.

For future work, we suggest broadening the application of WFCP to encompass a wider range of practical scenarios. This could involve exploring the impact of heterogeneous data distributions across the devices, as done in [?], [?] for noiseless channels, as well as the performance degradation arising from imperfect channel state estimation in fading channels. Another interesting direction for research is to devise a differentially private implementation of WFCP, potentially leveraging the idea of channel noise as a masking mechanism [?].

Appendix

A. Proof of Theorem 1

In this section, we denote quantized calibration NC scores as $s_i = q(s(x_i, y_i))$ for $i = 1, \ldots, N$ and quantized NC score for the true test pair (x, y) as $s_{n+1} = q(s(x, y))$. We also introduce two sets of $N+1$ NC scores. The first includes both calibration and test NC scores, i.e.,

$$\mathcal{S}^* = \{s_i\}_{i=1}^{N+1}, \tag{48}$$

while the second replaces the test NC score s_{N+1} with the maximum NC score value S_M, i.e.,

$$\mathcal{S}^{\max} = \{s_i\}_{i=1}^{N} \cup \{S_M\}. \tag{49}$$

For such a genie-aided set $\mathcal{S}^* = \{s_1, \ldots, s_{N+1}\}$ that has access to the test NC score, we use the bag notation $\wr \mathcal{S}^* \int$ to refer to a set of numerical values of the NC scores, which excludes the identity of the data point to which each NC score is assigned. Furthermore, we write as $\pi(\boldsymbol{\mathcal{S}}^*)$ the indices of the data points assigned to each element in the bag $\lbag\boldsymbol{\mathcal{S}}^*\rbag$. We use the same notation for $\boldsymbol{\mathcal{S}}^{\max}$. Based on these definitions, the set $\boldsymbol{\mathcal{S}}^*$ is unambiguously identified by the bag $\lbag\boldsymbol{\mathcal{S}}^*\rbag$ and by the assignment $\pi(\boldsymbol{\mathcal{S}}^*)$.

Finally, given the bag $\lbag\boldsymbol{\mathcal{S}}^*\rbag$, we introduce the $M \times 1$ vector as $\boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^*\rbag)$, in which each m-th entry represents the fraction of NC scores in $\lbag\boldsymbol{\mathcal{S}}^*\rbag$ equal to quantization level S_m. With this definition, the vector $\boldsymbol{p}^+$ in (??) can be equivalently defined as

$$\boldsymbol{p}^+ = \boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^{\max}\rbag), \tag{50}$$

and hence vector $\boldsymbol{v}$ in (??) as

$$\boldsymbol{v} = \boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^{\max}\rbag) + \tilde{\boldsymbol{z}}. \tag{51}$$

In (??) and (??), we have used the fact that histograms do not depend on the ordering of defined set. Recall that we are interested in finding a lower bound on the probability (??), which can be expressed as the expectation

$$\begin{aligned}
\Pr\left(y \in \Gamma_{\alpha_c}^{\mathrm{WFCP}}(x|\boldsymbol{v})\right) &= \mathbb{E}_{\boldsymbol{\mathcal{S}}^*,\tilde{\boldsymbol{z}}}\left[\mathbb{1}(y \in \Gamma_{\alpha_c}^{\mathrm{WFCP}}(x|\boldsymbol{v}))\right] \\
&= \mathbb{E}_{\boldsymbol{\mathcal{S}}^*,\tilde{\boldsymbol{z}}}\left[\mathbb{1}(s_{N+1} \leq S_{m_{\alpha_c}(\boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^{\max}\rbag)+\tilde{\boldsymbol{z}})})\right] \\
&\geq \mathbb{E}_{\boldsymbol{\mathcal{S}}^*,\tilde{\boldsymbol{z}}}\left[\mathbb{1}(s_{N+1} \leq S_{m_{\alpha_c}(\boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^*\rbag)+\tilde{\boldsymbol{z}})})\right] \\
&= \mathbb{E}_{\tilde{\boldsymbol{z}},\lbag\boldsymbol{\mathcal{S}}^*\rbag}\mathbb{E}_{\pi(\boldsymbol{\mathcal{S}}^*)|\tilde{\boldsymbol{z}},\lbag\boldsymbol{\mathcal{S}}^*\rbag}\left[\mathbb{1}(s_{N+1} \leq S_{m_{\alpha_c}(\boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^*\rbag)+\tilde{\boldsymbol{z}})})\right],
\end{aligned} \tag{52}$$

where in the second equality we have used (??) and (??), while for the third inequality, we have followed a standard trick of CP (see, e.g., Lemma 1 of [?]). This states that replacing any single value with the maximum value will never decrease the respective empirical quantile value. In the last equality, we have used the law of iterated expectations, which allows us to apply the expectations in the sequence as explained next.

First step: Bounding $\mathbb{E}_{\pi(\boldsymbol{\mathcal{S}}^*)|\tilde{\boldsymbol{z}},\lbag\boldsymbol{\mathcal{S}}^*\rbag}[\cdot]$

We begin by studying the inner expectation over the ordering $\pi(\boldsymbol{\mathcal{S}}^*)$ after conditioning on the bag $\lbag\boldsymbol{\mathcal{S}}^*\rbag$ and the noise vector $\tilde{\boldsymbol{z}}$. In the following, we write $\boldsymbol{p}^* = \boldsymbol{p}(\lbag\boldsymbol{\mathcal{S}}^*\rbag)$ to simplify the notation. Recall that $S_1, \ldots, S_M$ are the M quantization levels. From exchangeability of the data, we have the equality (see, e.g., [?])

$$\Pr\left[s_{N+1} = S_i|\tilde{\boldsymbol{z}}, \lbag\boldsymbol{\mathcal{S}}^*\ \rbag\right] = p_i^*. \tag{53}$$

It follows that we have the series of equalities

$$\begin{aligned}\mathbb{E}_{\pi(\mathcal{S}^*)|\tilde{\boldsymbol{z}},\wr\mathcal{S}^*\int}\left[\mathbb{1}(s_{N+1}\leq S_{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})})\right] &= \Pr\left[s_{N+1}\leq S_{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}|\tilde{\boldsymbol{z}},\wr\mathcal{S}^*\int\right]\\ &= \sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})} p_i^*\\ &\geq \min\left\{1, 1-\alpha_c-\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right\}, \qquad (54)\end{aligned}$$

where the inequality follows from the definition of $m_{\alpha_c}(\boldsymbol{p}(\wr\mathcal{S}^{\max}\int)+\tilde{\boldsymbol{z}})$ in (??) and the $\min\{\cdot\}$ operator is introduced to account for the case $m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})=M$.

Second step: Bounding ($\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}[\cdot]$)

We now marginalize over the noise vector $\tilde{\boldsymbol{z}}$ given the bag $\wr\mathcal{S}^*\int$. Given the bag $\wr\mathcal{S}^*\int$ we have

$$\begin{aligned}\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\mathbb{E}_{\pi(\mathcal{S}^*)|\tilde{\boldsymbol{z}},\wr\mathcal{S}^*\int}\left[\mathbb{1}\left(y\in\Gamma_{\alpha_c}^{\mathrm{WFCP}}(x|\boldsymbol{v})\right)\right] &\geq \mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\left[\min\left\{1,1-\alpha_c-\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right\}\right]\\ &= 1+\frac{1}{2}\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\left[-\alpha_c-\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i-\left|\alpha_c+\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right|\right],\end{aligned} \qquad (55)$$

in which we have used the identity $\min\{x,y\}=y+\frac{x-y-|x-y|}{2}$ to obtain the last equality.

We now note that the index $m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})$ depends on $\tilde{\boldsymbol{z}}$ and that, once conditioned on $\wr\mathcal{S}^*\int$ it depends only on the realization of the sequence $\tilde{z}_1,\ldots,\tilde{z}_{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}$. Therefore $m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})$ is a stopping time for the sequence $\tilde{z}_1,\tilde{z}_2,\ldots$, and we are allowed to invoke first Wald's identity [?] to obtain

$$\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\left[\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right]=\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\Big[m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})\Big]\mathbb{E}[\tilde{z}_1]=0, \qquad (56)$$

where the last equality follows from the noise being zero mean. Furthermore, from Wald's second identity [?], we have

$$\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\left[\left(\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right)^2\right]\leq\sigma^2\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\mathcal{S}^*\int}\left[m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})\right]\leq\sigma^2 M, \qquad (57)$$

which will be used next. Applying Jensen's inequality $\mathbb{E}[|x|]^2 \leq \mathbb{E}[x^2]$, we can further bound (??),

$$\begin{aligned}
& 1+\frac{1}{2}\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\boldsymbol{\mathcal{S}}^*\int}\left[-\alpha_c-\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i-\left|\alpha_c+\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right|\right] \\
& \geq 1-\frac{\alpha_c}{2}-\frac{1}{2}\sqrt{\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\boldsymbol{\mathcal{S}}^*\int}\left[\alpha_c^2+2\alpha_c\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i+\left(\sum_{i=1}^{m_{\alpha_c}(\boldsymbol{p}^*+\tilde{\boldsymbol{z}})}\tilde{z}_i\right)^2\right]} \\
& \geq 1-\frac{\alpha_c}{2}-\frac{1}{2}\sqrt{\alpha_c^2+\sigma^2 M}. \qquad (58)
\end{aligned}$$

Accordingly, we have

$$\mathbb{E}_{\tilde{\boldsymbol{z}}|\wr\boldsymbol{\mathcal{S}}^*\int}\mathbb{E}_{\pi(\boldsymbol{\mathcal{S}}^*)|\tilde{\boldsymbol{z}},\wr\boldsymbol{\mathcal{S}}^*\int}\left[\mathbb{1}\left(y\in\Gamma_{\alpha_c}^{\mathrm{WFCP}}\left(x|\boldsymbol{v}\right)\right)\right]\geq 1-\frac{\alpha_c}{2}-\frac{1}{2}\sqrt{\alpha_c^2+\sigma^2 M}. \qquad (59)$$

Final step: Bounding ($\mathbb{E}_{\wr\boldsymbol{\mathcal{S}}^*\int}[\cdot]$)

The final step follows directly from (??), in which the lower bound does not depend on the bag $\wr\boldsymbol{\mathcal{S}}^*\int$. This gives that

$$\Pr\left(y\in\Gamma_{\alpha_c}^{\mathrm{WFCP}}(x|\boldsymbol{v})\right)\geq 1-\frac{\alpha_c}{2}-\frac{1}{2}\sqrt{\alpha_c^2+\sigma^2 M}. \qquad (60)$$

Therefore, to satisfy the target coverage rate $1-\alpha$, we set the corrected unreliability level as

$$\alpha_c=\alpha-\frac{\sigma^2 M}{4\alpha}. \qquad (61)$$