

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  $\ell$  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  $\ell$ , 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

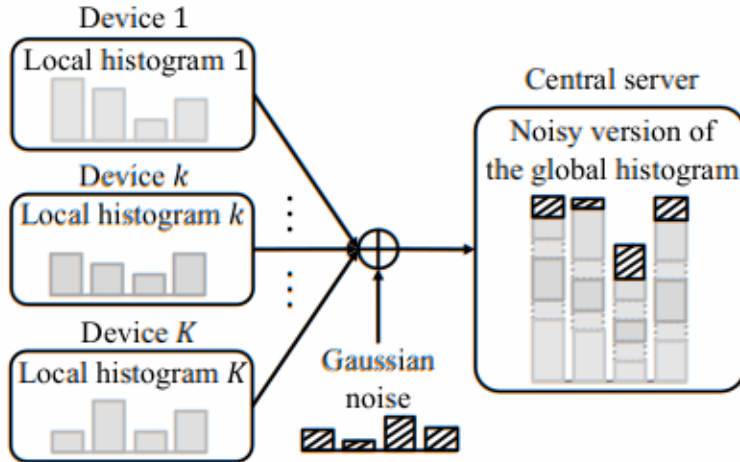


Fig. 2. TBMA enables the estimate of the global histogram of discrete scalar data available across all devices. To this end, orthogonal codewords are assigned to each histogram bin. All devices transmit simultaneously their individual local histograms over a shared wireless channel by allocating to each 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,

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

$$\Gamma_{\alpha_d, \alpha_s}^{\text{QQ}}(x) = \{y \in \mathcal{Y} : s(x, y) \leq s_{\alpha}^{\text{QQ}}\}. \quad (19)$$

The pair of miscoverage levels  $(\alpha_d, \alpha_s)$  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

$$\Pr(y \in \Gamma_{\alpha_d, \alpha_s}^{\text{QQ}}(x)) \geq 1 - \frac{1}{N+1} \sum_{j=k}^K \binom{K}{j} \sum_{I_{1,j}=n}^{N_d} \sum_{I_{1,j}^c=0}^{n-1} \frac{\binom{N_d}{i_1} \cdots \binom{N_d}{i_m}}{\binom{N}{i_1+\cdots+i_K}} \triangleq M_{\alpha_d, \alpha_s} \quad (20)$$

with  $n = \lceil (N_d+1)(1-\alpha_d) \rceil$ ;  $k = \lceil (K+1)(1-\alpha_s) \rceil$ ;  $I_{1,j} = \{i_1, \dots, i_j\}$ ;  $I_{1,j}^c = \{i_{j+1}, \dots, i_K\}$ ; 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  $(\alpha_d, \alpha_s)$  that minimizes the lower bound  $M_{\alpha_d, \alpha_s}$  while satisfying the target coverage rate  $1 - \alpha$ . The optimization objective can be formulated as

$$(\alpha_d^*, \alpha_s^*) \in \arg \min_{\alpha_d, \alpha_s} \{M_{\alpha_d, \alpha_s} : M_{\alpha_d, \alpha_s} \geq 1 - \alpha\}. \quad (21)$$

If the solution of (21) is not unique, it is suggested to find the pair with the largest value of  $\alpha_d$  and then choose among those the pair with the largest value of  $\alpha_s^*$ . 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]

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

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

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

and the channel dispersion  $V$  is defined as

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

Accordingly, with probability  $\epsilon$ , 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 TDMA and over-the-air computing to communicate the empirical distribution of the quantized NCrescores from the devices to the server via Viterbi-like 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 NCrescores as a closed-form Base-Based estimator, that 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

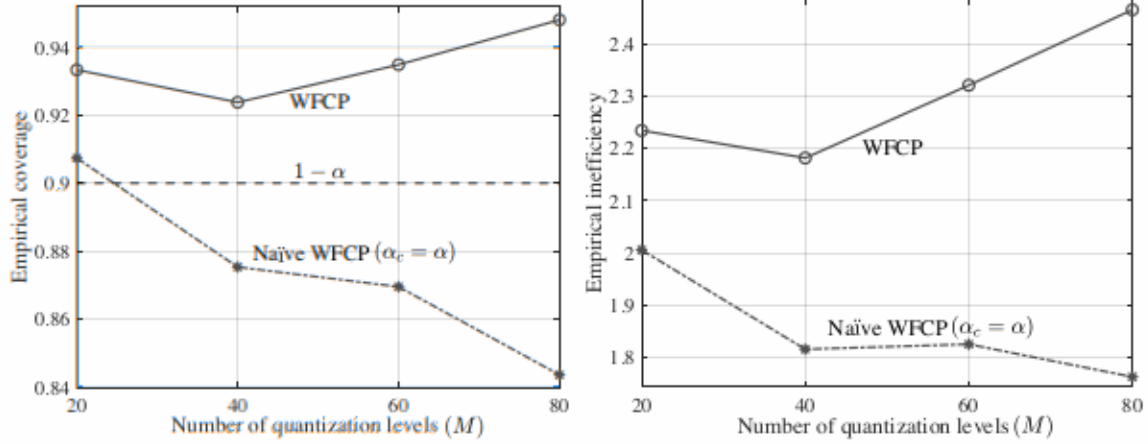


Fig. 3. Empirical coverage and empirical inefficiency of WFCP and naïve WFCP ( $\alpha_c = \alpha$ ) versus the number  $M$  of quantization levels with target unreliability level  $\alpha = 0.1$ , number  $T = 120$  of channel uses, number  $K = 10$  of devices, and SNR = -10 dB.

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

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

and

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

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  $N^{\text{te}}$  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



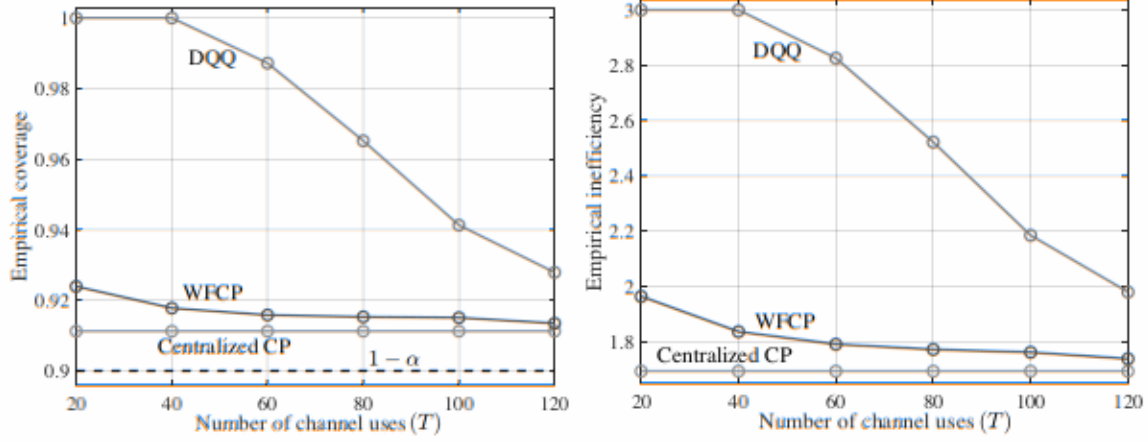


Fig. 4. Empirical coverage and empirical inefficiency of centralized CP, WFCP, and DQQ [?] versus the number  $T$  of channel uses available with target unreliability level  $\alpha = 0.1$ , number  $M = 20$  of quantization levels, number  $K = 20$  of devices, and SNR = 0 dB.

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 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  $T$

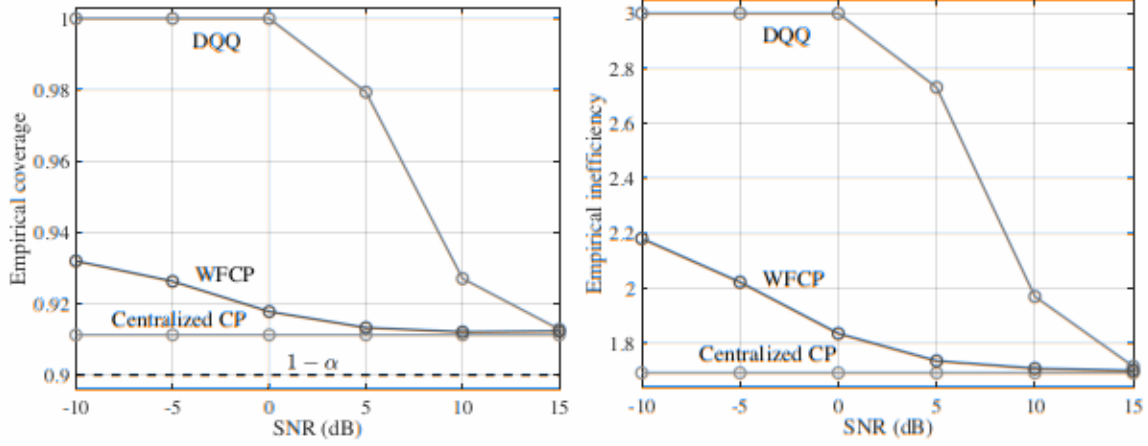


Fig. 5. Empirical coverage and inefficiency of centralized CP, WFCP, and DQQ [?] versus SNR with target unreliability level  $\alpha = 0.1$ , number  $M = 20$  of quantization levels, number  $T = 40$  of channel uses, and number  $K = 20$  of devices.

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  $\epsilon$  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  $\epsilon$  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

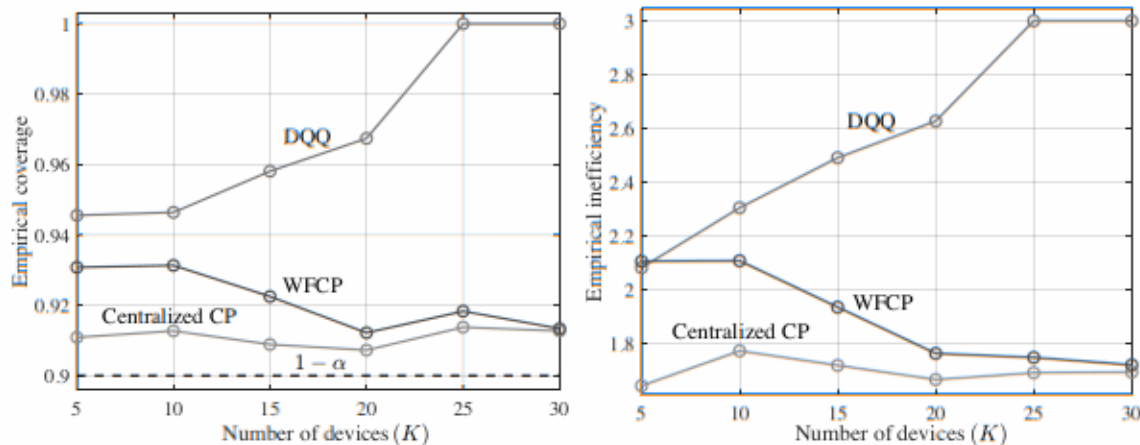


Fig. 6. Empirical coverage and inefficiency of centralized CP, WFCP, and DQQ [?] versus the number  $K$  of devices with  $N_d = 10$  per-device calibration data points, target unreliability level  $\alpha = 0.1$ , number  $M = 20$  of quantization levels, number  $T = 60$  of channel uses, and SNR =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