One-Shot Federated Conformal Prediction

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, Sylvain Arlot

One-Shot Federated Conformal Prediction

Pierre Humbert [1] **Batiste Le Bars** [2] **Aurélien Bellet** [2] **Sylvain Arlot** [1]

Abstract

In this paper, we introduce a conformal prediction method to construct prediction sets in a one-shot federated learning setting. More specifically, we define a quantile-of-quantiles estimator and prove that for any distribution, it is possible to output prediction sets with desired coverage in only one round of communication. To mitigate privacy issues, we also describe a locally differentially private version of our estimator. Finally, over a wide range of experiments, we show that our method returns prediction sets with coverage and length very similar to those obtained in a centralized setting. Overall, these results demonstrate that our method is particularly well-suited to perform conformal predictions in a one-shot federated learning setting.

1. Introduction

Federated Learning (FL) is a recent paradigm that allows to learn from decentralized data sets stored locally by multiple agents (Kairouz et al., 2021). FL is particularly appealing when data are highly sensitive and cannot be centralized for privacy or security reasons. So far, the design of FL algorithms has mainly focused on the training phase of machine learning: the goal is to fit models on decentralized data sets while minimizing the amount of communication or optimizing the privacy-utility trade-off (see e.g. McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020; Geyer et al., 2017; Noble et al., 2022). However, FL poses further challenges regarding model evaluation, as this step must also be done without access to centralized data. In particular, with the increasing popularity of black-box methods, deploying machine learning models in real-world applications often requires to appropriately *quantify the uncertainty* of their predictions. Unfortunately, models trained with the above supervised FL algorithms only provide point predictions (e.g., class labels or regression targets). This is not sufficient in high-stakes applications like medicine (Begoli et al., 2019), where decisions may impact human lives.

In this work, we investigate the task of outputting a prediction set rather than a single point prediction in a FL setting. Formally, given some data stored by multiple agents and an additional test point (X, Y), we want to construct a *marginally valid* set which is likely to contain the unknown response Y. In other words, we want a set $\widehat{\mathcal{C}}(X)$ such that

$$\mathbb{P}\big(Y \in \widehat{\mathcal{C}}(X)\big) \geq 1 - \alpha \,, \tag{1}$$

where $\alpha \in (0, 1)$ is a desired miscoverage rate. Although there exist several methods to construct such a set (Vovk et al., 2005; Papadopoulos et al., 2002; Romano et al., 2019), they require access to a *centralized* data set. They are thus incompatible with the constraints of FL, in which agents process their data locally and only interact with a central server by sharing some aggregate statistics. Constructing a valid prediction set is even more challenging in the *one-shot FL* (Zhang et al., 2012; Guha et al., 2019; Yurochkin et al., 2019; Li et al., 2021; Dennis et al., 2021; Salehkaleybar et al., 2021) that we consider in this work, where the communication between the agents and the server is further restricted to a *single round*. One-shot FL is motivated by the fact that the number of communication rounds is often the main bottleneck in FL (Kairouz et al., 2021).

Contributions. In this paper, we present an intuitive one-shot FL method based on Conformal Prediction (CP) (Vovk et al., 2005) to construct distribution-free prediction sets satisfying (1). The key step of CP methods is the ordering of *scores* computed for each calibration data point. In the FL setting, this ordering step is not possible without exchanging the local data sets or performing many agent-server communication rounds. To circumvent this problem, we define a *quantile-of-quantiles* estimator: each agent sends to the server a local empirical quantile and the server aggregates them by computing a quantile of these quantiles. We describe how to choose the order of the quantiles (depending on the number of agents and the size of their local data sets) to obtain a prediction set that satisfies (1). We also prove that property (1) can be verified *conditionally to the observed data* with a slight modification of the selected quantiles. While the previous results rely on certain data homogeneity assumptions, we further quantify the impact

[1]Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France. [2]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189, CRIStAL, F-59000 Lille.

Preprint.

of heterogeneous (non-identically distributed) data on the performance of our algorithm. To address use cases with strong privacy constraints, we derive a version of our approach that satisfies differential privacy (Dwork et al., 2014), in which agents run the exponential mechanism to privately select their local quantile. Finally, we empirically evaluate the performance of our method on standard CP benchmarks and show that it produces prediction sets that are very close to the ones obtained when data are centralized.

2. Background and Related Work

2.1. Split Conformal Prediction

Conformal Prediction (CP) is a framework introduced by Vovk et al. (2005) to construct distribution-free prediction sets satisfying (1). One of the most popular methods to perform CP in a centralized setting is the *split conformal* (Papadopoulos et al., 2002) which is at the core of our main contribution described in Section 3.

To use the split conformal method (split CP), we first need to choose a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which measures the magnitude of a predictor error for a given point. Whether we are in the regression or classification setting, many different score functions exist in the literature (see e.g. Angelopoulos & Bates, 2021). In regression, for instance, a common choice is the fitted absolute residual $S_i \triangleq s(X_i, Y_i) = |Y_i - \widehat{f}(X_i)|$ where $\widehat{f}$ is some predictor learned on a training data set. Note that our approach does not assume a particular choice of score function, so throughout the paper, we will keep the function s abstract. Then, we split the data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ into two disjoint subsets: a *calibration* set $\mathcal{D}^{cal}_{n_1} = \{(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1})\}$ and a *training* set $\mathcal{D}^{tr}_{n_0} = \{(X_{n_1+1}, Y_{n_1+1}), \ldots, (X_n, Y_n)\}$ with $n_1 = n - n_0$ and $n_0 < n$. The predictor $\widehat{f}$ is fitted on $\mathcal{D}^{tr}_{n_0}$ and conformity scores $\mathcal{S}^{cal}_{n_1} \triangleq \{S_1, \ldots, S_{n_1}\}$ are calculated on $\mathcal{D}^{cal}_{n_1}$ via the previously chosen score function s. Finally, given a test point X and $\alpha \in (0, 1)$, we construct the conformal set

$$\widehat{\mathcal{C}}(X) = \left\{ y \in \mathbb{R} : s(X, y) \leq \widehat{Q}_{(\lceil (n_1+1)(1-\alpha) \rceil)}(\mathcal{S}^{cal}_{n_1}) \right\} ,$$

where $\widehat{Q}_{(\cdot)}(\cdot)$ is defined by

$$\widehat{Q}_{(k)}(\mathcal{S}') = \widehat{Q}_{(k)} \triangleq \begin{cases} S'_{(k)} & \text{if } k \leq |\mathcal{S}'| \\ \infty & \text{otherwise} , \end{cases} \tag{2}$$

with $|\mathcal{S}'|$ the size of the sample $\mathcal{S}'$, and $S'_{(1)} \leq \ldots \leq S'_{(|\mathcal{S}'|)}$ the order statistics of the scores $S'_1, \ldots, S'_{|\mathcal{S}'|}$ in $\mathcal{S}'$. In other words, $\widehat{Q}_{(k)}$ outputs the k-th smallest value in a given set of scores. The following theorem proves that the set returned by the split CP method satisfies (1) under very mild assumptions.

Theorem 2.1 (Vovk et al., 2005; Lei & Wasserman, 2014)**.** *Let consider n i.i.d. (or only exchangeable) random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathcal{X} \times \mathcal{Y}$ and an additional test point (X, Y). For any score function and any $\alpha \in (0, 1)$, the set returned by the split CP method satisfies*

$$\mathbb{P}\left(Y \in \widehat{\mathcal{C}}(X)\right) \geq 1 - \alpha .$$

Furthermore, if $S_1, \ldots S_{n_1}$ are almost surely distinct, this probability is upper bounded by $1 - \alpha + 1/(n_1 + 1)$.

Although the first CP methods were the *split* and the related *full* methods (Vovk et al., 2005; Papadopoulos et al., 2002), many extensions based upon them have been proposed recently. In regression, Lei & Wasserman (2014) present a method called locally weighted CP and provide theoretical insights for conformal inference. More recently, Romano et al. (2019) have developed a variant of the split CP called Conformal Quantile Regression (CQR). Other recent alternatives have been proposed (Kivaranovic et al., 2020; Sesia & Romano, 2021; Gupta et al., 2022; Ndiaye, 2022). We refer to Vovk et al. (2005), Angelopoulos & Bates (2021) and Fontana et al. (2023) for in-depth presentations of CP.

2.2. Related Work in Federated Learning

As already mentioned, FL methods are today mostly focused on the training part of the learning process (i.e. fitting $\widehat{f}$ to the data). Nevertheless, a few recent works have considered other types of FL problems that can be related to our work. The closest related work is the one of Lu & Kalpathy-Cramer (2021) which, to the best of our knowledge, is the only paper claiming to perform conformal prediction in the FL setting. Their idea is to locally calculate the quantiles $\widehat{Q}_{(\lceil (n_1+1)(1-\alpha) \rceil)}$ for all agents and to average them in the central server. Unfortunately, they do not prove that their prediction set has valid coverage. Furthermore, their method is non-robust, especially when the size of local data sets is small, and their experiments (and ours, in Section 5) suggest that this set is generally too large. We will see that by considering a quantile of quantiles instead of an average of quantiles, the method we propose addresses these limitations. Gauraha & Spjuth (2021) propose an ensemble-based CP approach that can be performed in a distributed setting. However, they assume that a shared calibration set is available on the central server, which is unrealistic in FL. Finally, we can also mention recent work on federated evaluation of classifiers (Cormode & Markov, 2022), federated quantile computation (Andrew et al., 2021; Pillutla et al., 2022), and on uncertainty quantification with Bayesian FL (El Mekkaoui et al., 2021; Kotelevskii et al., 2022) which, although related to our work, do not study CP and do not allow to obtain coverage guarantees.

3. Quantile-of-Quantiles for Federated CP

In this section, we present a method to perform conformal prediction in a *one-shot* FL setting (Zhang et al., 2012; Guha et al., 2019), where only one round of communication from the agents to the central server is allowed.

3.1. Setup and Objective

Consider a set of $m \in \mathbb{N}^*$ agents, with their own local data, that seek to collaborate in order to compute a valid prediction set. For simplicity, we suppose that each agent has exactly n calibration data points, and refer to Appendix A.1 for a discussion of the case where agents have calibration sets of different sizes. We also assume that the predictor $\widehat{f}$ is given in advance: for instance, it could be learned on a separate set of data points using standard FL algorithms such as FedAvg (McMahan et al., 2017). We therefore only focus on the calibration of the prediction set and not on the training step. As a consequence, in the following, all theoretical statements are made conditionally on $\widehat{f}$ (often implicitly).

Formally, each agent $j \in \{1, \dots, m\}$ holds a local calibration data set $\mathcal{S}^{(j)} \triangleq (S_1^{(j)}, \dots, S_n^{(j)})$ composed of n scores, where $S_i^{(j)} = s(X_i^{(j)}, Y_i^{(j)})$ is the score associated to the i-th calibration data point of agent j and we want to find a particular value $\widehat{q}$ such that for a test point (X, Y), the set $\widehat{\mathcal{C}}(X) = \{y \in \mathbb{R} : s(X, y) \leq \widehat{q}\}$ contains the unknown response Y with probability at least $1-\alpha$. In the centralized case, the split CP method presented in Section 2.1 requires to order all the scores and to choose $\widehat{q}$ as the $\lceil (mn+1)(1-\alpha) \rceil$ smallest score. In one-shot FL, this global ordering step is only possible if the agents send their whole list of local scores to the server. This naive implementation of the split method is impractical, due to both privacy concerns and unacceptable communication costs, requiring us to design another strategy. As a single round of communication is allowed, the main difficulty is to choose what should be sent from the agents to the server, and what kind of aggregation should be done by the server to yield the desired coverage.

3.2. Main contribution: FedCP-QQ

Our method is based on the idea that each agent j should return a quantile of its local scores $\mathcal{S}^{(j)}$, in the same way as for the split method described in Section 2.1. The main questions that then arise are (i) which quantile of the scores the agents should send, and (ii) how to aggregate them at the central server level. Lu & Kalpathy-Cramer (2021) propose to use an empirical average, but this aggregation strategy is not satisfactory. This is obvious in the extreme case where $n = 1$ (a single data point per agent): it amounts to calculating the average of the local scores, which will typically fail to provide the desired coverage (1). Instead, we propose to select a quantile of the locally computed quantiles. This *quantile-of-quantiles* estimator is defined below.

Definition 3.1 (Quantile-of-quantiles)**.** *For any (ℓ, k) in $[\![1, n]\!] \times [\![1, m]\!]$, the Quantile-of-Quantiles (QQ) estimator is defined by*

$$\widehat{Q}_{(\ell,k)} \triangleq \widehat{Q}_{(k)} \left(\widehat{Q}_{(\ell)}(\mathcal{S}^{(1)}), \dots, \widehat{Q}_{(\ell)}(\mathcal{S}^{(m)}) \right) , \qquad (3)$$

where $\widehat{Q}_{(\cdot)}(\cdot)$ is given by Equation (2).

In words, QQ takes for each agent the ℓ-th smallest local score and then takes the k-th smallest value of these scores. This requires a single round of communication and thus fits the constraints of one-shot FL. The associated plug-in prediction set is

$$\widehat{\mathcal{C}}_{\ell,k}(X) = \left\{ y \in \mathbb{R} : s(X, y) \leq \widehat{Q}_{(\ell,k)} \right\} . \qquad (4)$$

Our objective is now to find (ℓ, k) such that $\mathbb{P}(Y \in \widehat{\mathcal{C}}_{\ell,k}(X))$ is closest possible to $1 - \alpha$ while being guaranteed to be above. To this aim, we derive the following result.

Theorem 3.2. *Let $\{(X_i^{(j)}, Y_i^{(j)})\}_{i,j=1}^{m,n}$ and (X, Y) be i.i.d. random variables (given $\widehat{f}$). For any $(\ell, k) \in [\![1, n]\!] \times [\![1, m]\!]$ we have:*

$$\mathbb{P}\left(Y \in \widehat{\mathcal{C}}_{\ell,k}(X) \right) \geq M_{\ell,k} \triangleq$$
$$1 - \frac{1}{mn+1} \sum_{j=k}^{m} \binom{m}{j} \sum_{I_{1,j}=\ell}^{n} \sum_{I_{1,j}^c=0}^{\ell-1} \frac{\binom{n}{i_1} \cdots \binom{n}{i_m}}{\binom{mn}{i_1+\cdots+i_m}} , \qquad (5)$$

where $I_{1,j} = \{i_1, \dots, i_j\}$ and $I_{1,j}^c = \{i_{j+1}, \dots, i_m\}$. Moreover, when the associated scores $\{S_i^{(j)}\}_{i,j=1}^{n,m}$ and $S \triangleq s(X, Y)$ have continuous c.d.f, (5) is an equality.

The proof is given in Appendix C.1. This theorem shows that we can lower bound the probability of coverage of our quantile-of-quantiles prediction set by a quantity $M_{\ell,k}$ that does not depend on the data distribution but only on m, n, ℓ and k. Furthermore, the lower bound becomes an *equality* when scores have a continuous c.d.f. This is the case for instance with the fitted absolute residual when the conditional distribution of Y given X has a continuous c.d.f., i.e. when the noise distribution is atomless. Note that although the theorem requires the data points to be i.i.d., in fact only the scores need to satisfy this hypothesis (conditionally to $\widehat{f}$). This is interesting since there are situations where the scores are i.i.d. even though data distributions are different across agents. In Section 3.5, we further discuss the impact of data heterogeneity across agents, an important aspect of many FL applications.

Based on Theorem 3.2, our algorithm returns $\widehat{Q}_{(\ell^*,k^*)}$ with

$$(\ell^*, k^*) = \arg\min_{\ell,k} \{ M_{\ell,k} : M_{\ell,k} \geq 1 - \alpha \} . \qquad (6)$$

Algorithm 1 FedCP-QQ

Input: Local scores $\{\mathcal{S}^{(j)}\}_{j=1}^m$, α, M (see Equation (5))

$(\ell^*, k^*) \longleftarrow \arg\min_{\ell,k} \{M_{\ell,k} : M_{\ell,k} \geq 1-\alpha\}$

for $j = 1, \ldots, m$ **do**
  Agent j sends $\widehat{Q}_{(\ell^*)}(\mathcal{S}^{(j)})$ to the central server
end for
Central server returns $\widehat{Q}_{(k^*)}\left(\widehat{Q}_{(\ell^*)}(\mathcal{S}^{(1)}), \ldots, \widehat{Q}_{(\ell^*)}(\mathcal{S}^{(m)})\right)$

By construction, the associated set (4) is *valid*, in the sense that it satisfies the desired coverage (1). The full procedure, called *Federated Conformal Prediction with Quantile-of-Quantiles* (FedCP-QQ), is summarized in Algorithm 1.

Particular cases. To gain more intuition on the FedCP-QQ procedure, let us consider the two extreme cases $n = 1$ and $n \to \infty$. When $n = 1$, each agent sends its unique score to the server. Thus, it suffices for the server to compute the k-th smallest score with $k = \lceil (m+1)(1-\alpha) \rceil$ to obtain a valid set. In the other extreme case where $n \to \infty$, if the agents send their ℓ-th smallest score with $\ell = \lceil (n+1)(1-\alpha) \rceil$, each agent has in fact sent the true quantile of order $(1-\alpha)$. The server can therefore choose any of the values and obtain a valid set. We see that in both cases, if both the agents and the server compute appropriate quantiles, we can obtain a valid set. Our method extends this idea to any values of m and n using Theorem 3.2 and Equation (6). In Appendix A.3, we study another interesting specific case where each machine sends its maximum value, i.e., $\ell = n$.

Computational optimizations. The brute-force computation of $M_{\ell,k}$ in Equation (5) for all (ℓ, k) can be quite costly in practice. To accelerate this step, we describe in Appendix A.2 a more efficient way to compute $M_{\ell,k}$, based on the calculation of rectangular probabilities of a multivariate hypergeometric distribution.

We also note that $M = (M_{\ell,k})_{(\ell,k) \in [\![1,n]\!] \times [\![1,m]\!]}$ can be precomputed and reused across multiple executions of FedCP-QQ. Indeed, as the probability of coverage is independent of the distribution of the data (Theorem 3.2), they do not change as long as m (the number of agents) and n (the size of local data sets) remain fixed. This is the case for instance when computing prediction sets for multiple score functions s, predictors $\widehat{f}$, and miscoverage rates α on the same data.

3.3. Upper Bounding the Probability of Coverage

While by construction our probability of coverage is necessarily lower bounded by $1-\alpha$, it is also interesting to have an upper bound, guaranteeing that the coverage of our prediction set is not too large. In the centralized case, Lei & Wasserman (2014) have shown that, if the scores have continuous c.d.f, the split conformal method with a calibration set of size mn gives $\mathbb{P}(Y \in \widehat{\mathcal{C}}(X)) \leq 1-\alpha+1/(mn+1)$ (Theorem 2.1). This means that when there is only one agent (or when agents do not collaborate), this probability is upper bounded by $1-\alpha+1/(n+1)$.

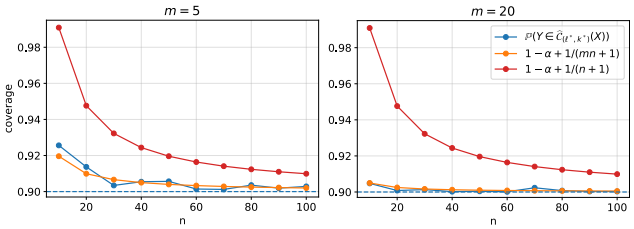

Figure 1. Comparison of the exact value of $\mathbb{P}(Y \in \widehat{\mathcal{C}}_{\ell^*,k^*}(X))$ (blue) with the upper bound when: data are centralized (orange), there is only one agent (red). Parameters are $\alpha = 0.1, m = \{5, 20\}$, and $n = \{10, \ldots, 100\}$.

Assuming that the scores have continuous c.d.f., in Figure 1 we compare the two upper bounds with the value of $M_{\ell^*,k^*} = \mathbb{P}(Y \in \widehat{\mathcal{C}}_{\ell^*,k^*}(X))$ returned by FedCP-QQ. Recall that, by Theorem 3.2, when the scores have a continuous c.d.f., the value of M_{ℓ^*,k^*} returned by FedCP-QQ is equal to the exact coverage of $\widehat{\mathcal{C}}_{\ell^*,k^*}(X)$. The figure shows (in blue) that FedCP-QQ returns prediction sets with coverage comparable to the (tight) upper bound of the centralized case with mn calibration points (in orange). We also see that the coverage is much larger if we consider the data of a single agent (in red), which illustrates the advantage of our method and the need for collaboration between the agents.

The form of our quantile-of-quantiles estimator does not allow us to extend the proof techniques of the centralized framework and obtain a theoretical upper bound similar to the one of Theorem 2.1. Nevertheless, the results obtained in Figure 1 make us conjecture that an upper bound could be of the same order as in the centralized framework, i.e., in $1-\alpha+\mathcal{O}(1/(mn+1))$.

3.4. Conditional Coverage Guarantee

In practice, we are interested in the coverage rate for test points when the data set is fixed. However, the guarantee in (1) does not address this as the randomness is also taken over the data. In other words, it bounds the miscoverage rate on average over all possible calibration data points (and over a training set if $\widehat{f}$ is learned). Let us define the conditional miscoverage rate as a function of the data:

$$\alpha_P(\mathcal{D}_{mn}) = \mathbb{P}\left(Y \notin \widehat{\mathcal{C}}_{\ell,k}(X) \mid \mathcal{D}_{mn}\right), \qquad (7)$$

with $\mathcal{D}_{mn}$ the full calibration set without the test point (Y, X). While, by construction of $\widehat{\mathcal{C}}_{\ell,k}(X)$, the expectation of $\alpha_P(\mathcal{D}_{mn})$ is smaller than α, the random variable

$\alpha_P(\mathcal{D}_{mn})$ may have a high variance. In particular, Bian & Barber (2022) show that we can construct a scenario where $\mathbb{P}\left(\alpha_P(\mathcal{D}_{mn}) = 1\right) = \alpha$ and $\mathbb{P}\left(\alpha_P(\mathcal{D}_{mn}) = 0\right) = 1 - \alpha$.

Here, we have $\mathbb{E}[\alpha_P(\mathcal{D}_{mn})] = \alpha$ but a non-negligible proportion of calibration data sets might result in a poor conditional coverage even though the average coverage is still $1 - \alpha$. In practice, we want to have $\alpha_P(\mathcal{D}_{mn}) \approx \alpha$ with a probability close to 1 to avoid this unfavorable scenario.

In the following theorem, we show that it is possible to control the conditional miscoverage of FedCP-QQ.

Theorem 3.3. *In the framework of Theorem 3.2, if $\delta \in (0, 0.5]$ and $\ell \cdot k \geq (1 - \alpha) \cdot mn$, then the conditional miscoverage rate —defined by Eq. (7)— is controlled as follows:*

$$\mathbb{P}\left(\alpha_P(\mathcal{D}_{mn}) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2mn}}\right) \geq 1 - \delta\,. \quad (8)$$

Theorem 3.3 is proved in Appendix C.2. It states that the probability that a particular data set results in a conditional miscoverage rate much higher than α vanishes with the number of data points used for calibration. A similar bound is obtained in the centralized setting (Vovk, 2012; Bian & Barber, 2022) for the split method. However, note that our theorem holds only for couples (ℓ, k) verifying a condition not necessarily verified by the couple (ℓ^*, k^*) used by FedCP-QQ. Nevertheless, our experiments suggest that this could still be true for (ℓ^*, k^*), up to a slight modification of the bound. However, similarly to the upper bound on the probability of coverage (see Section 3.3), the proof of this statement is difficult because it requires to study the rank of $\widehat{Q}_{(\ell,k)}$ in the full data set which, contrary to the centralized case, is a random variable. In the proof of Theorem 3.3, we rely on an almost sure lower bound for this rank, which is conservative and negatively impacts the final result. In the centralized case, the rank is almost surely fixed and this greatly simplifies the theoretical analysis.

3.5. Impact of Heterogeneous Data

An important challenge in FL is to deal with data heterogeneity across agents (Kairouz et al., 2021; Li et al., 2020; Le Bars et al., 2023). This heterogeneity can yield different distributions of scores across agents and thus affects the coverage of the set returned by CP methods. To better understand these effects, we no longer assume that all the variables are drawn from the same distribution. Instead, we only suppose that the local data points of agent j are drawn i.i.d. from an *agent-specific distribution* with a test point also drawn from a potentially different distribution.

As we do not have any information on the underlying distributions of the scores, we study how data heterogeneity affects the coverage of the set returned by FedCP-QQ, i.e., we quantify how much we lose in coverage if we apply the same strategy as in the i.i.d. case. Intuitively, the more the distributions of the scores $\{S^{(j)}\}_j$ are similar and close to the one of S, the less we lose in coverage. This is made precise in the following result.

Proposition 3.4. *Assume that the calibration data $\{(X_i^{(j)}, Y_i^{(j)})\}_{i,j=1}^{m,n}$ and the test point (X, Y) are such that, given $\widehat{f}$, the corresponding scores $\{S_i^{(j)}\}_{i,j=1}^{n,m}, S$ are independent, and that for every $j \in [\![1, m]\!]$, $\{S_i^{(j)}\}_{i=1}^{n}$ are i.i.d. Let $\{\widetilde{S}_i^{(j)}\}_{i,j=1}^{n,m}, \widetilde{S}$ be i.i.d. random variables (given $\widehat{f}$). Define, for every $j \in [\![1, m]\!]$, $p_j^*(S) = \mathbb{P}(S_{(\ell^\star)}^{(j)} \leq S | S)$ and $\tilde{p}^*(\widetilde{S}) = \mathbb{P}(\widetilde{S}_{(\ell^\star)}^{(1)} \leq \widetilde{S} | \widetilde{S})$. Then, we have*

$$\mathbb{P}\big(Y \in \widehat{\mathcal{C}}_{\ell^\star,k^\star}(X)\big) \geq 1 - \alpha - \mathbb{E}\left[d_{\mathrm{TV}}\Big(\mathrm{PoisBin}\big(p^*(S)\big)\,,\, \mathrm{Bin}\big(m, \tilde{p}^*(\widetilde{S})\big)\Big)\right],$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ is the total-variation (TV) distance, PoisBin *the Poisson-Binomial distribution and* Bin *the binomial distribution.*

Proposition 3.4 is proved in Appendix C.3. The general idea of this result is that when variables are i.i.d., probabilities on order statistics only depend on the c.d.f. of a certain binomial distribution, whereas when the variables are independent but with different distributions, the binomial needs to be replaced by a Poisson-Binomial distribution. The inequality indicates that, in the heterogeneous case, the coverage is reduced by the TV distance between the two distributions. We can note that this distance can be upper bounded (see Appendix C.3) and that it is equal to 0 when all the data are i.i.d and $S = \widetilde{S}$. We leave to future work the precise characterization of cases where the TV distance is negligible in front of $1 - \alpha$.

4. Differentially Private FedCP-QQ

While FL methods are often informally claimed to mitigate privacy issues, they still leak information about the local data sets during the execution of the algorithm. In the case of FedCP-QQ, it is easy to see how revealing a particular quantile of the local score distribution may leak sensitive information. In this section, we propose a privacy-preserving version of FedCP-QQ based on Differential Privacy (DP) (Dwork et al., 2014), a mathematical notion of privacy that essentially requires that the output distribution of a randomized algorithm is not too sensitive to a small modification of the input data set. In particular, we consider the strong *Local DP* (LDP) model where agents do not trust the central server and must locally privatize the messages they send.

Formally, a randomized algorithm $\mathcal{A}$ is said to be ε-LDP if for any two local data sets $\mathcal{S}$ and $\mathcal{S}'$ that differ in a sin-

Algorithm 2 Differentially Private Quantile

Input: Scores $(S_1, \dots, S_n) \in \mathbb{R}^n$, quantile $q \in (0,1)$, privacy level $\varepsilon > 0$, bins $\{I_1, \dots, I_B\}$
for $i = 1, \dots, n$ **do**
  Compute the discretized score $\bar{S}_i = e_b$ such that $S_i \in I_b$
end for
for $b = 1, \dots, B$ **do**
  Compute the weight $w_b = \max\left\{\frac{|\{i:\bar{S}_i < e_b\}|}{q}, \frac{|\{i:\bar{S}_i > e_b\}|}{1-q}\right\}$
end for
$\Delta_q \longleftarrow \max\{\frac{1}{q}, \frac{1}{1-q}\}$
Output: Bin e_b with probability $e^{-\frac{\varepsilon w_b}{2\Delta_q}} / \sum_{b'=1}^{B} e^{-\frac{\varepsilon w_{b'}}{2\Delta_q}}$

Algorithm 3 FedCP^2-QQ

Input: Local scores $\{\mathcal{S}^{(j)}\}_{j=1}^m$, miscoverage level α, privacy level $\varepsilon > 0$, bins $\{I_1, \dots, I_B\}$, $\gamma \in (0,1)$

The server finds (ℓ_γ, k_γ) as in FedCP-QQ (Algorithm 1) with coverage level $\frac{1-\alpha}{1-\gamma\alpha}$
$q \longleftarrow \max\left\{\frac{\ell_\gamma + \ell_{\text{cor}}}{n}, \frac{1}{2}\right\}$ with ℓ_{cor} from Eq. (10)
for $j = 1, \dots, m$ **do**
  Agent j sends $\widehat{Q}_j^\varepsilon$, the output of Alg. 2 with $\mathcal{S}^{(j)}$, to the server.
end for
Output: The server orders $\widehat{Q}_1^\varepsilon, \dots, \widehat{Q}_m^\varepsilon$ and outputs the k_γ-th value denoted $\widehat{Q}^\varepsilon_{(k_\gamma)}$.

gle data point (we call them *neighboring*), and any set of possible outputs O, we have:

$$\mathbb{P}\big(\mathcal{A}(\mathcal{S}) \in O\big) \leq \exp(\varepsilon)\mathbb{P}\big(\mathcal{A}(\mathcal{S}') \in O\big) \,. \tag{9}$$

A smaller ε therefore yields a better privacy. In our specific framework, $\mathcal{S}$ and $\mathcal{S}'$ correspond to two neighboring calibration data sets of an agent j and $\mathcal{A}(\mathcal{S})$ to the information sent by j to the central server.

Our approach builds upon the (centralized) differentially private quantile mechanism recently introduced by Angelopoulos et al. (2022) and summarized in Algorithm 2. The main idea is to apply the exponential mechanism (McSherry & Talwar, 2007) to a discretization of the scores into bins and with an appropriate choice of utility function. It requires to fix a number of bins $B \in \mathbb{N}$, an upper bound on the scores $S_{\max}$ and a set of points $0 = e_0 < e_1 < \dots < e_{B-1} < e_B = S_{\max}$ defining the bins $I_b = (e_{b-1}, e_b]$. Algorithm 2 is ε-DP by a direct application of the exponential mechanism with utility function w_b and sensitivity Δ_q.

FedCP^2-QQ method. Our private algorithm, called *Federated Conformal Private Prediction* (FedCP^2)-QQ, is an extension of FedCP-QQ (Algorithm 1) with two key modifications: (i) exact local quantile computations are replaced by calls to DP Quantile (Algorithm 2), and (ii) the order of client and server-level quantiles is adjusted to guarantee the desired coverage. More precisely, if the central server asks for the ℓ-th smallest score of each agent, then the agents use Algorithm 2 to return a randomized bin around the true quantile $\widehat{Q}_{(\ell)}(\mathcal{S}^{(j)})$. To achieve the desired coverage $1-\alpha$ despite the randomness due to privacy, the server computes (ℓ_γ, k_γ) such that $\mathbb{P}(S \leq \widehat{Q}_{(\ell_\gamma, k_\gamma)})$ is above but close to $\frac{1-\alpha}{1-\gamma\alpha}$, where $\gamma \in (0,1)$ is a free parameter. Because the agents might return bins smaller than the one of the requested ℓ_γ-th score, the central server further compensates by asking agents for their $(\ell_\gamma + \ell_{\text{cor}})$-th smallest score with

$$\ell_{\text{cor}} = \frac{2}{\varepsilon} \log\left(\frac{B}{1 - (1-\gamma\alpha)^{\frac{1}{m}}}\right) . \tag{10}$$

Note that the smaller the privacy parameter ε (more privacy), the bigger the correction ℓ_{cor}. At first sight, one could think that B should be taken small to reduce the correction. In practice, it should also be taken sufficiently large to avoid aggressive rounding that could lead to a large final prediction set. We refer to Angelopoulos et al. (2022, Section 4.2) for an in-depth discussion on the selection of the number of bins B. The following theorem ensures that Algorithm 3 preserves privacy and allows to construct prediction sets that satisfy the desired coverage. The proof is given in Appendix C.4.

Theorem 4.1. *For any $\varepsilon > 0$, Algorithm 3 satisfies ε-LDP. Moreover, denoting $\widehat{\mathcal{C}}_\varepsilon(X) = \{y \in \mathbb{R} : s(X,y) \leq \widehat{Q}^\varepsilon\}$ with $\widehat{Q}^\varepsilon$ the output of the algorithm, we have*

$$\mathbb{P}(Y \in \widehat{\mathcal{C}}_\varepsilon(X)) \geq 1 - \alpha \,.$$

Choosing γ. Intuitively, in order to be equivalent to the non-private FedCP-QQ, γ should tend to 0 as the privacy parameter ε tends to infinity. To select γ automatically, we propose a grid-search strategy. We look for the γ that brings the smallest amount of correction, which we evaluate using the pre-computed table M. More precisely, for a given γ, we evaluate $M_{\ell_\gamma + \ell_{\text{cor}}, k_\gamma}$ which is the coverage obtained by the non-private FedCP-QQ estimator $\widehat{Q}_{(\ell_\gamma + \ell_{\text{cor}}, k_\gamma)}$. Note that this coverage is not the one of our private estimator since each agent might return a score smaller than the $(\ell_\gamma + \ell_{\text{cor}})$-th smallest. To find the best γ, we look at the one that brings the smaller coverage $M_{\ell_\gamma + \ell_{\text{cor}}, k_\gamma}$ over the grid. To gain more intuition on the degree of correction brought by the additional randomness of the private setting, we represent in Figure 2 the quantity $M_{\ell_\gamma + \ell_{\text{cor}}, k_\gamma}$ found for the best γ and for different values of n and ε. This plot shows how fast the correction is reduced as n and ε increase.

Privacy amplification by shuffling or aggregation. To achieve better privacy-utility trade-offs, it is common in FL to relax the LDP model and instead assume that the agents' messages are sent to a secure computation function whose output is received by the central server. This is sometimes referred to as *Distributed DP* (DDP) (Kairouz et al., 2021).

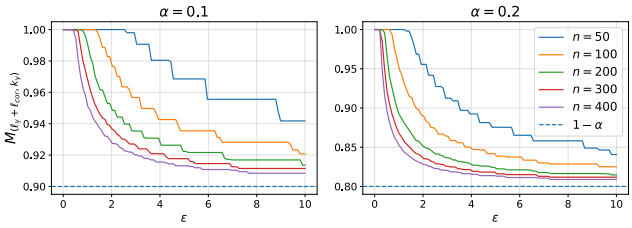

Figure 2. Degree of compensation $M_{\ell_\gamma+\ell_{\text{cor}},k_\gamma}$ for different values of α, n and ε when $m = 10$. We clearly observe that $M_{\ell_\gamma+\ell_{\text{cor}},k_\gamma}$ tends to the desired coverage $1-\alpha$ (dashed lines) as n and ε tends to $+\infty$, which means that the compensation vanishes.

Two standard secure computation primitives are compatible with FedCP^2-QQ: secure shuffling (Feldman et al., 2021) and secure aggregation (Bonawitz et al., 2017). Secure shuffling outputs a random permutation of the messages, which does not prevent the server from computing the desired quantile. For secure aggregation (which outputs the sum of the messages), each agent can encode its private quantile as a one-hot vector of size B indicating the corresponding bin. The sum of these vectors is sufficient for the server to find the bin corresponding to the k_γ-th smallest score. In both cases, ε is reduced by a factor of $\mathcal{O}(1/\sqrt{m})$. In other words, if one of the previous privacy amplification schemes is used, we can replace ε by $\varepsilon\sqrt{m}$ (up to a constant) and therefore reduce the correction ℓ_{cor} by a factor $\mathcal{O}(\sqrt{m})$, while still satisfying the same privacy guarantees. See (Feldman et al., 2021) and (McMillan et al., 2022) for detailed formulas.

Remark 4.2. *FedCP^2-QQ provides privacy guarantees with respect to the calibration data. To provide privacy guarantees with respect to the data used to train the model, one should train the model using locally differentially private algorithms (see e.g. Geyer et al., 2017; McMahan et al., 2018; Noble et al., 2022). Note that the training and calibration data sets are disjoint, and that FedCP^2-QQ only post-processes the private model to compute the calibration scores. Therefore, if model training satisfies ε_1-LDP and FedCP^2-QQ satisfies ε_2-LDP, the full pipeline satisfies $\max(\varepsilon_1, \varepsilon_2)$-LDP thanks to parallel composition.*

5. Experiments

In this section, we evaluate our method FedCP-QQ on synthetic and real regression data sets. Additional experiments on FedCP^2-QQ are presented in Appendix B.2. The code of our methods is available at `https://github.com/pierreHmbt/FedCP-QQ`.

Depending on the experiments, we use the split CP method presented in Section 2 or its popular variant Conformalized Quantile Regression (CQR) (Romano et al., 2019). For split CP, $\widehat{f}$ is a standard regressor, the score function s is $s(X,Y) = |Y - \widehat{f}(X)|$, and the prediction set is an interval of constant length $[\widehat{f}(X) \pm \hat{q}]$. In CQR, $\widehat{f}$ is replaced by a couple $(\widehat{f}_{\alpha/2}, \widehat{f}_{1-\alpha/2})$ where $\widehat{f}_\beta$ is a quantile regressor of order β (Koenker & Bassett Jr, 1978) and $s(X,Y) = \max(\widehat{f}_{\alpha/2}(X) - Y, Y - \widehat{f}_{1-\alpha/2}(X))$. In contrast to split CP, this method returns sets of the form $[\widehat{f}_{\alpha/2}(X) - \hat{q}, \widehat{f}_{1-\alpha/2}(X) + \hat{q}]$ which have a size adaptive to heteroscedasticity.

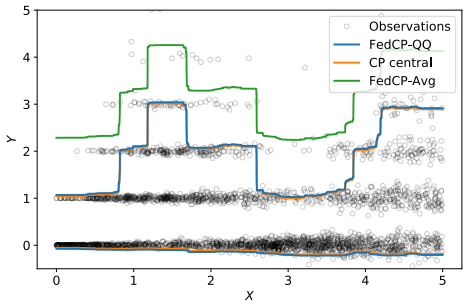

Figure 3. Prediction intervals on simulated data with FedCP-QQ (ours), centralized, and FedCP-Avg calibrations. The lower bound of the set returned by FedCP-Avg is beyond the figure.

For both split CP and CQR, we use FedCP-QQ to find the value of $\hat{q}$ (calibration step). We compare it with the centralized baseline (Equation 2) and FedCP-Avg (Lu & Kalpathy-Cramer, 2021). Recall that the latter simply averages the m quantiles of order $\lceil (n+1)(1-\alpha) \rceil / n$ sent by the agents (see Section 2.2).

5.1. Synthetic Data

Data set. We draw 2000 independent, univariate random variables X_i from a uniform distribution on $[1, 5]$. Following Romano et al. (2019), the response variable is sampled as

$$Y_i \,|\, X_i \sim \text{Pois}(\sin^2(X_i) + 0.1) + 0.03 \cdot X_i \varepsilon_{1,i} + 25 \cdot \mathbb{1}\{U_i < 0.01\} \varepsilon_{2,i} \,,$$

where $\text{Pois}(\lambda)$ is the Poisson distribution with mean λ, $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ are i.i.d. standard Gaussian variables, and U_i is uniform on the interval $[0, 1]$. Note that the last term of the equation can generate an outlier. Then, we split the data set into two disjoint subsets: one for training and one for calibration. To simulate a FL scenario, the calibration set is also divided into $m = 50$ disjoint subsets of size $n = 20$. Finally, we generate a test set of size 5000 with the same properties.

We construct the prediction sets using the CQR approach where the estimation of the (quantile) regression function is learned with quantile regression forests (Meinshausen & Ridgeway, 2006). The number of trees in the forest is set to 1000, the two parameters controlling the coverage rate on the training data are tuned using cross-validation and the remaining hyperparameters are set as in Romano et al. (2019).

Results. Figure 3 illustrates the performance of the different methods when $\alpha = 0.1$. We see that the set returned

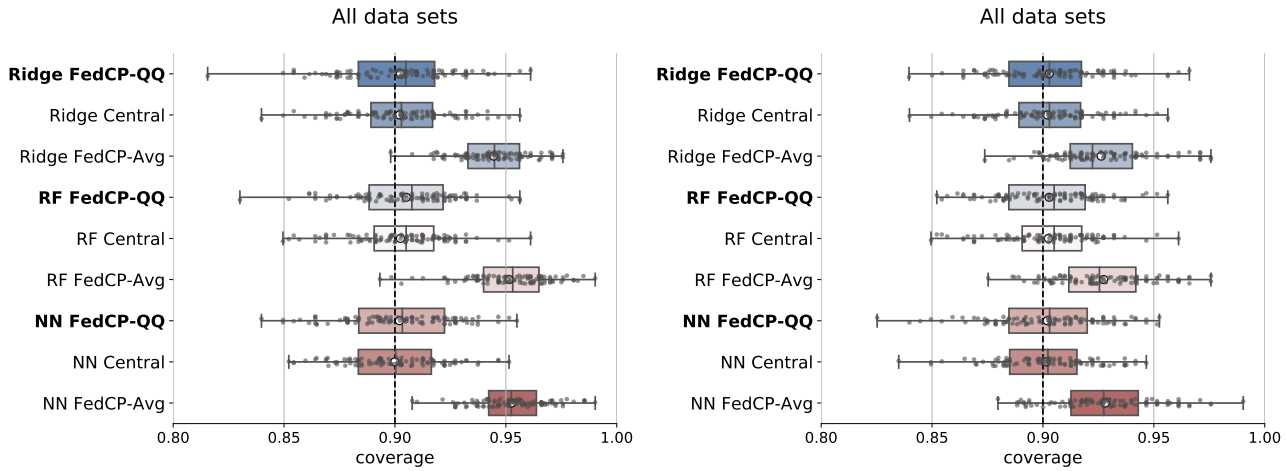

Figure 4. Empirical coverages of prediction intervals ($\alpha = 0.1$) constructed by various methods. On the left, when $m \gg n$. On the right, when $m \ll n$. Our method FedCP-QQ is shown in bold font. The white circle represents the mean.

by FedCP-QQ when the data are decentralized is almost identical to the one obtained when the data are centralized. This is not the case for FedCP-Avg which outputs a larger set. This may be due to the presence of outliers in the data and because the mean (the server aggregation strategy for FedCP-Avg) is not robust. On the contrary, by using a quantile function to aggregate the agents' quantiles, FedCP-QQ is robust to outliers and produces shorter valid sets. In the next section, we will see that the same behavior occurs on real data sets.

5.2. Real Data

Data sets. We evaluate our method on five public-domain regression data sets also considered in Romano et al. (2019) and Sesia & Romano (2021): physicochemical properties of protein tertiary structure (bio) (Rana, 2013); bike sharing (bike) (Fanaee-T & Gama, 2013); communities and crimes (community) (Redmond, 2011); Tennessee's student teacher achievement ratio (star) (Achilles et al., 2008); and concrete compressive strength (concrete) (Yeh, 1998).

In this section, we use (i) split-CP with ridge regression —the regularization parameter is tuned by cross-validation; (ii) CQR with quantile Regression Forests (RF) —the hyper-parameters are the ones used in Section 5.1; and (iii) CQR with Neural Networks (NN) for quantile regression (Taylor, 2000) —the architecture and the parameters are those used in Romano et al. (2019).

The prediction sets, with a miscoverage rate fixed to $\alpha = 0.1$, are either calibrated with CP in the centralized setting or in a FL setting using FedCP-QQ and FedCP-Avg. For each experiment, we split the full data set into three parts: a training set (40%), a calibration set (40%), and a test set (20%). To simulate a FL scenario, we also split the calibration set in m disjoint subsets of equal size n. We consider scenarios where $m \gg n$, and $m \ll n$. Their exact values for each data set are given in Appendix B.1. All features are then standardized to have zero mean and unit variance. For each method, we compute the empirical coverage obtained on the test set and the average length of the conformal set. These two metrics are collected over 20 different training-calibration-test random splits.

Results. Figure 4 displays the boxplots of the empirical coverages obtained by each method over all the data sets and all the 20 different random splits (one point represents one random split of one data set). Results on individual data sets are presented in Appendix B.1, as well as boxplots of the lengths of the intervals obtained. The first observation we can make is that, on average (white circle), FedCP-QQ does return intervals whose coverage is greater than 0.90, without being too far from it. More importantly, our method returns prediction sets with coverage and length very similar to those returned by centralized calibration. In Figure 4 for instance, we see that the mean (white circle) and standard-deviation (size of the box) of the coverages obtained with FedCP-QQ and the centralized baseline have comparable values, with a slightly larger standard-deviation for FedCP-QQ. The same kind of observation can be made concerning the length of the prediction sets (see figures in Appendix B.1). Finally, it is interesting to note that, with FedCP-QQ, we obtain similar results for $m \gg n$ and $m \ll n$. This is in contrast to FedCP-Avg, which yields sets with higher coverages and lengths on all data sets and is therefore strictly inferior to our method. Note that Appendix B.2 provides additional results about our DP algorithm FedCP^2-QQ, showing how the coverage varies with the privacy parameter ε. Overall, these experiments support the fact that FedCP-QQ is a well-suited method to perform the calibration step of CP in a decentralized setting, placing it as the only one adapted to the context of (one-shot) FL.

6. Discussion

This paper introduces the method *Federated Conformal Prediction with Quantile-of-Quantiles* (FedCP-QQ) to output valid distribution-free prediction sets in a one-shot Federated Learning context. In addition to the analysis and discussion about the different properties of our method, we also introduce FedCP^2-QQ, a private version of FedCP-QQ based on Local Differential Privacy. Multiple experiments highlight that our method returns prediction sets with coverage and length close to those returned in a centralized setting, supporting the fact that FedCP-QQ is a well-suited method for (one-shot) FL scenarios.

This work brings many important future research directions. Among them, we expect that new proof techniques could lead to better theoretical guarantees, notably regarding conditional coverage and the private estimator. Our paper focuses on the calibration step, making it particularly suited for *split*-based conformal methods. However, it would be interesting to study how our FL approach could be extended to the full conformal or the nested conformal methods (Gupta et al., 2022), where the learning and the calibration steps or not fully split anymore. Finally, an interesting line of research is the derivation of estimators for the specific case where data are not identically distributed across agents, e.g., based on personalization or by allowing an additional round of communication.

Acknowledgements

This work was supported in part by grant ANR-20-CE23-0015 (Project PRIDE). Batiste Le Bars is supported by an Inria-EPFL fellowship. Sylvain Arlot is also supported by Institut Universitaire de France (IUF).

References

Achilles, C. M., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. Tennessee's student teacher achievement ratio (star) project. *Harvard Dataverse*, 1:2008, 2008.

Andrew, G., Thakkar, O., McMahan, B., and Ramaswamy, S. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34: 17455–17466, 2021.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Angelopoulos, A. N., Bates, S., Zrnic, T., and Jordan, M. I. Private prediction sets. *Harvard Data Science Review*, apr 2022.

Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Bian, M. and Barber, R. F. Training-conditional coverage for distribution-free predictive inference. *arXiv preprint arXiv:2205.03647*, 2022.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Cormode, G. and Markov, I. Federated calibration and evaluation of binary classifiers. *arXiv preprint arXiv:2210.12526*, 2022.

David, H. A. and Nagaraja, H. N. *Order statistics*. John Wiley & Sons, 2004.

Davis, P. J. Leonhard euler's integral: A historical profile of the gamma function: In memoriam: Milton abramowitz. *The American Mathematical Monthly*, 66(10):849–869, 1959.

Dennis, D. K., Li, T., and Smith, V. Heterogeneity for the win: One-shot federated clustering. In *ICML*, 2021.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Ehm, W. Binomial approximation to the poisson binomial distribution. *Statistics & Probability Letters*, 11(1):7–16, 1991.

El Mekkaoui, K., Mesquita, D., Blomstedt, P., and Kaski, S. Federated stochastic gradient langevin dynamics. In *Uncertainty in Artificial Intelligence*, pp. 1703–1712. PMLR, 2021.

Fanaee-T, H. and Gama, J. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pp. 1–15, 2013. ISSN 2192-6352.

Feldman, V., McMillan, A., and Talwar, K. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *FOCS*, 2021.

Fontana, M., Zeni, G., and Vantini, S. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Gauraha, N. and Spjuth, O. Synergy conformal prediction. In *Symposium on Conformal and Probabilistic Prediction and Applications*, pp. 91–110. PMLR, 2021.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Guha, N., Talwalkar, A., and Smith, V. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Learning. In *ICML*, 2020.

Kivaranovic, D., Johnson, K. D., and Leeb, H. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 4346–4356. PMLR, 2020.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Kotelevskii, N., Vono, M., Moulines, E., and Durmus, A. Fedpop: A bayesian approach for personalised federated learning. *arXiv preprint arXiv:2206.03611*, 2022.

Le Bars, B., Bellet, A., Tommasi, M., Lavoie, E., and Kermarrec, A.-M. Refined convergence and topology learning for decentralized sgd with heterogeneous data. In *AISTATS*, 2023.

Lebrun, R. Efficient time/space algorithm to compute rectangular probabilities of multinomial, multivariate hypergeometric and multivariate pólya distributions. *Statistics and Computing*, 23(5):615–623, 2013.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (1):71–96, 2014.

Li, Q., He, B., and Song, D. Practical one-shot federated learning for cross-silo setting. In *IJCAI*, 2021.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Lu, C. and Kalpathy-Cramer, J. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021.

Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

McMillan, A., Javidbakht, O., Talwar, K., Briggs, E., Chatzidakis, M., Chen, J., Duchi, J., Feldman, V., Goren, Y., Hesse, M., Jina, V., Katti, A., Liu, A., Lyford, C., Meyer, J., Palmer, A., Park, D., Park, W., Parsa, G., Pelzl, P., Rishi, R., Song, C., Wang, S., and Zhou, S. Private federated statistics in an interactive setting. *arXiv preprint arXiv:2211.10082*, 2022.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Meinshausen, N. and Ridgeway, G. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Ndiaye, E. Stable conformal prediction sets. In *International Conference on Machine Learning*, pp. 16462–16479. PMLR, 2022.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially Private Federated Learning on Heterogeneous Data. In *AISTATS*, 2022.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Pillutla, K., Laguel, Y., Malick, J., and Harchaoui, Z. Differentially private federated quantiles with the distributed discrete gaussian mechanism. In *International Workshop*

on Federated Learning: Recent Advances and New Challenges, 2022.

Rana, P. Physicochemical properties of protein tertiary structure data set. *UCI Machine Learning Repository*, 2013.

Redmond, M. Communities and crime unnormalized data set. *UCI Machine Learning Repository*, 2011.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Salehkaleybar, S., Sharif-Nassab, A., and Golestani, S. J. One-shot federated learning: Theoretical limits and algorithms to achieve them. *Journal of Machine Learning Research*, 22(189):1–47, 2021.

Sesia, M. and Romano, Y. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

Taylor, J. W. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Yeh, I.-C. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, T. N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In *ICML*, 2019.

Zhang, Y., Wainwright, M. J., and Duchi, J. C. Communication-efficient algorithms for statistical optimization. *Advances in neural information processing systems*, 25, 2012.

Appendix

A. Supplementary Discussions

A.1. FedCP-QQ with Different n_i

In the main article, we assumed for simplicity that all agents had the same amount of data n, but our method is in fact generalizable for $(n_1, \dots, n_m)$. In that case, the right-hand term of Equation (5) becomes

$$M_{\ell_1,\dots,\ell_m,k} = 1 - \frac{1}{n_1 + \cdots + n_m + 1} \sum_{j=k}^{m} \sum_{A \in \mathcal{P}_j} \sum_{i_1=\ell_{a_1}}^{n_{a_1}} \cdots \sum_{i_{|A|}=\ell_{a_{|A|}}}^{n_{a_{|A|}}} \sum_{i_{|A|+1}=0}^{\ell_{a_{|A|+1}}-1} \cdots \sum_{i_m=0}^{\ell_{a_m}-1} \frac{\binom{n_{a_1}}{i_1} \cdots \binom{n_{a_m}}{i_m}}{\binom{n_1+\cdots+n_m}{i_1+\cdots+i_m}} ,$$

where $\mathcal{P}_j$ is the set of all subsets of j integers that can be selected from $\{1, \dots, m\}$, $A = \{a_1, \dots, a_{|A|}\}$, and $A^c = \{a_{|A|+1}, \dots, a_m\}$. It can be computed in the same way as for the case where $n_j = n$. An important difference that appears if we want to apply the methodology of FedCP-QQ presented in the paper is that we now have to find different values for $\ell_1, \dots, \ell_m$ since the local sample sizes are different. Although possible, computing $M_{\ell_1,\dots,\ell_m,k}$ for all possible values of $(\ell_1, \dots, \ell_m, k)$ to find the smallest one above $1 - \alpha$, can be very time-consuming.

In practice, we propose to directly fix $\ell_j = \lceil (1-\alpha)(n_j+1) \rceil$ as it would be similarly done in the classical (centralized) split methodology. Hence, the previous probability function only needs to be computed for the different values of $k = 1, \dots, m$, reducing significantly the computation at the cost of being slightly less close to $1 - \alpha$. Note that this strategy can also be used in the context of the main paper, i.e., when $n_j = n$ and $\ell_j = \ell$.

A.2. Computation of Equation (5)

Let recall the right-hand side of Equation (5) is

$$M_{k,\ell} = 1 - \frac{1}{mn+1} \sum_{j=k}^{m} \binom{m}{j} \sum_{i_1=\ell}^{n} \cdots \sum_{i_j=\ell}^{n} \sum_{i_{j+1}=0}^{\ell-1} \cdots \sum_{i_m=0}^{\ell-1} \frac{\binom{n}{i_1} \cdots \binom{n}{i_m}}{\binom{mn}{i_1+\cdots+i_m}} = 1 - \frac{1}{mn+1} \sum_{j=k}^{m} \binom{m}{j} \sum_{I_{1,j}=\ell}^{n} \sum_{I_{1,j}^c=0}^{\ell-1} \frac{\binom{n}{i_1} \cdots \binom{n}{i_m}}{\binom{mn}{i_1+\cdots+i_m}} ,$$

where $I_{1,j} = \{i_1, \dots, i_j\}$ and $I_{1,j}^c = \{i_{j+1}, \dots, i_m\}$. The time complexity of its brute-force computation is too high. In this section, we provide an efficient algorithm to compute it. The first step is to rewrite the summations to bring out the mass function of a multivariate hypergeometric distribution:

$$\sum_{I_{1,j}=\ell}^{n} \sum_{I_{1,j}^c=0}^{\ell-1} \frac{\binom{n}{i_1} \cdots \binom{n}{i_m}}{\binom{mn}{i_1+\cdots+i_m}} = \sum_{r \in \tilde{R}_j} \sum_{I_{1,j}=\ell}^{n} \sum_{I_{1,j}^c=0}^{\ell-1} \frac{\binom{n}{i_1} \cdots \binom{n}{i_m}}{\binom{mn}{i_1+\cdots+i_m}} \mathbb{1}\{i_1 + \cdots + i_m = r\} , \tag{11}$$

with $\tilde{R}_j = \{j\ell, \dots, jn + (m-j)(\ell-1)\}$. The summation in $I_{1,j}$, and $I_{1,j}^c$ now writes

$$p_r(\mathbf{a}, \mathbf{b}) \triangleq \mathbb{P}(a_1 \le H_1 \le b_1, \cdots, a_m \le H_m \le b_m) ,$$

$$\text{where } (a_i, b_i) = \begin{cases} (\ell, n) & \text{if } i \in \{1, \dots, j\} \\ (0, \ell - 1) & \text{if } i \in \{j+1, \dots, m\} \end{cases} ,$$

and $(H_1, \dots, H_m)$ follows a multivariate hypergeometric distribution with parameters $(\{n, \dots, n\}, r)$. By a direct application of Bayes' theorem we obtain (Lebrun, 2013):

$$p_r(\mathbf{a}, \mathbf{b}) = \mathbb{P}\left(\sum_{i=1}^{m} T_i = r \right) \frac{\prod_{i=1}^{m} \mathbb{P}(a_i \le W_i \le b_i)}{\mathbb{P}(\sum_{i=1}^{m} W_i = r)} ,$$

where for any t in $(0, 1)$ and for all $1 \le i \le m$, the random variables W_i follow a binomial distribution $\mathcal{B}(n, t)$ and $T_i = (W_i \mid a_i \le W_i \le b_i)$ follows a truncated binomial distribution.

As there exists efficient algorithms to compute both $\mathbb{P}(a_i \le W_i \le b_i)$ and $\mathbb{P}(\sum_{i=1}^{m} W_i = r)$, the only difficulty remains the evaluation of $\mathbb{P}(\sum_{i=1}^{m} T_i = r)$. A straightforward approach is to multiply the generating probability functions of the T_i and then extract the coefficient of degree r. This algorithm has a time complexity of $\mathcal{O}(mr \log(r))$ if the multiplications are done using an FFT based algorithm. This strategy still remains costly for large values of m or r, and more advanced algorithms have been proposed by Lebrun (2013).

A.3. Reporting the Maximum of each Agent

Another particular case of interest is when $\ell = n$, i.e., agents send their maximum value. Using the fact that the c.d.f. of the maximal value of n i.i.d random variables with common c.d.f. F is F^n, it is possible to give a simpler formula for $M_{n,k}$.

Proposition A.1. *For every $m, n \geq 1$ and $k \in [\![1, m]\!]$, we have*

$$M_{n,k} = \frac{\Gamma(k + 1/n)}{\Gamma(k)} \cdot \frac{\Gamma(m+1)}{\Gamma(m + 1/n + 1)} \ ,$$

where $M_{n,k}$ is defined by Eq. (5) and for any complex number z such that $\Re(z) > 0$, $\Gamma : z \mapsto \int_0^{+\infty} t^{z-1} \mathrm{e}^{-t} \mathrm{d}t$ denotes the gamma function.

Furthermore, when $k = k_m \triangleq \lceil m(1-\alpha)^n \rceil$, we have

$$\lim_{m \to \infty} \mathbb{P}\left(Y \in \widehat{\mathcal{C}}_{n,k_m}(X)\right) \geq 1 - \alpha$$

Proposition A.1 is proved in Appendix C.5. It shows that when each agent sends the maximum to the central server, by taking the k_m-th smallest value of these maximums with $k_m \geq m(1-\alpha)^n$, the server obtains a valid coverage of $(1-\alpha)$. Note that for a fixed m, k_m decreases to 0 when n grows to infinity. This is expected since, intuitively, if the number of points per agent increases, the maximums also increase, and the server must compensate by taking a very small quantile of these values to obtain a coverage close to $(1-\alpha)$.

B. Additional Experimental Results

B.1. Results on Individual Data Sets

In this section, we present our results on individual data sets.

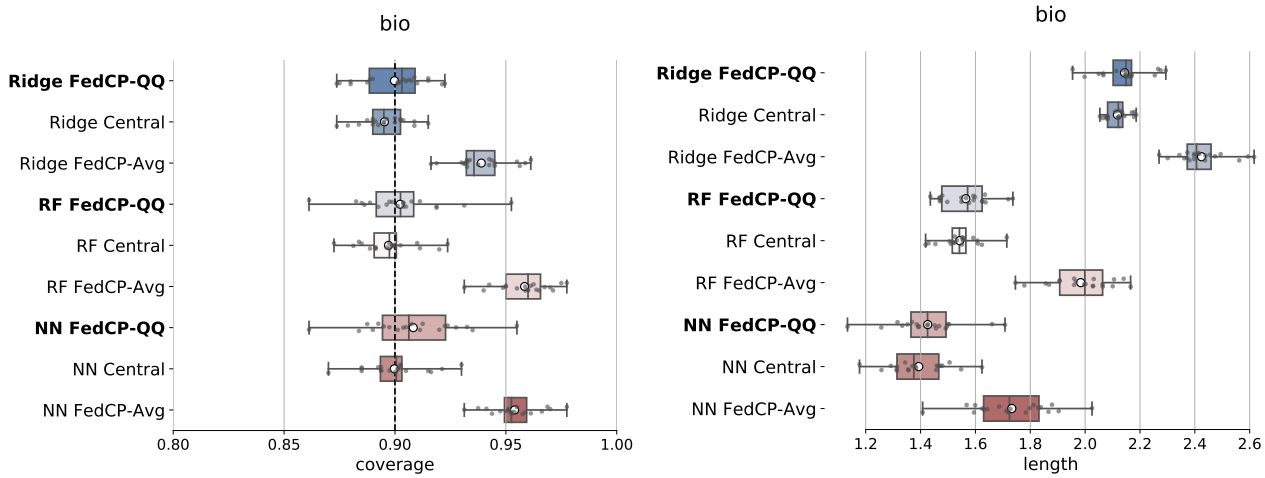

Figure 5. Coverage (left) and average length (right) of prediction intervals for 20 random training-calibration-test splits. The miscoverage is $\alpha = 0.1$, and the size of the calibration set is $m = 100$, and $n = 10$. The white circle represents the mean and the name of the data set is located at the top of each plot.

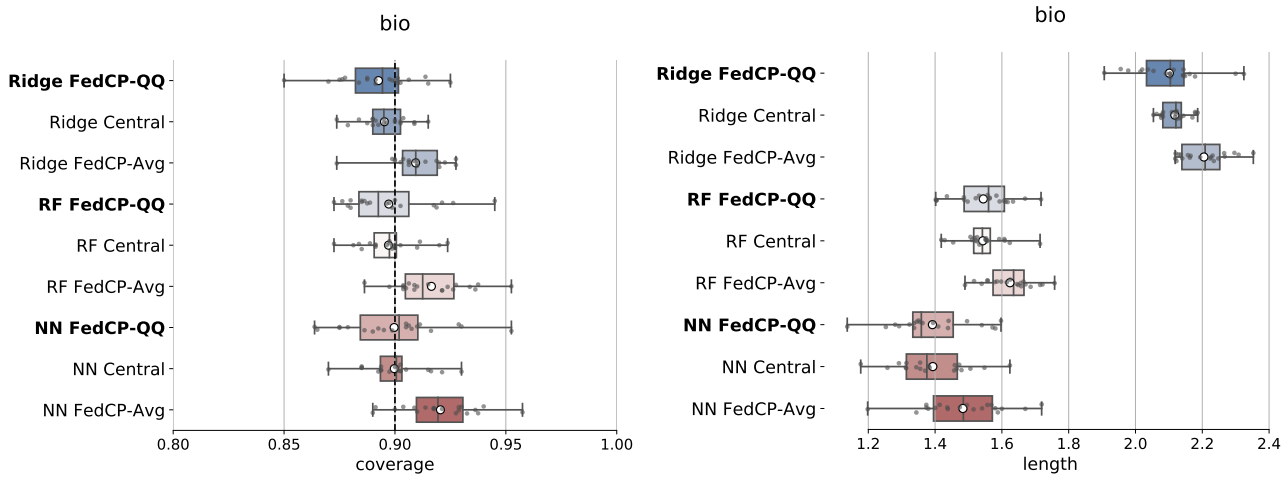

Figure 6. Coverage (left) and average length (right) of prediction intervals for 20 random training-calibration-test splits. The miscoverage is $\alpha = 0.1$, and the size of the calibration set is $m = 10$, and $n = 100$. The white circle represents the mean and the name of the data set is located at the top of each plot.

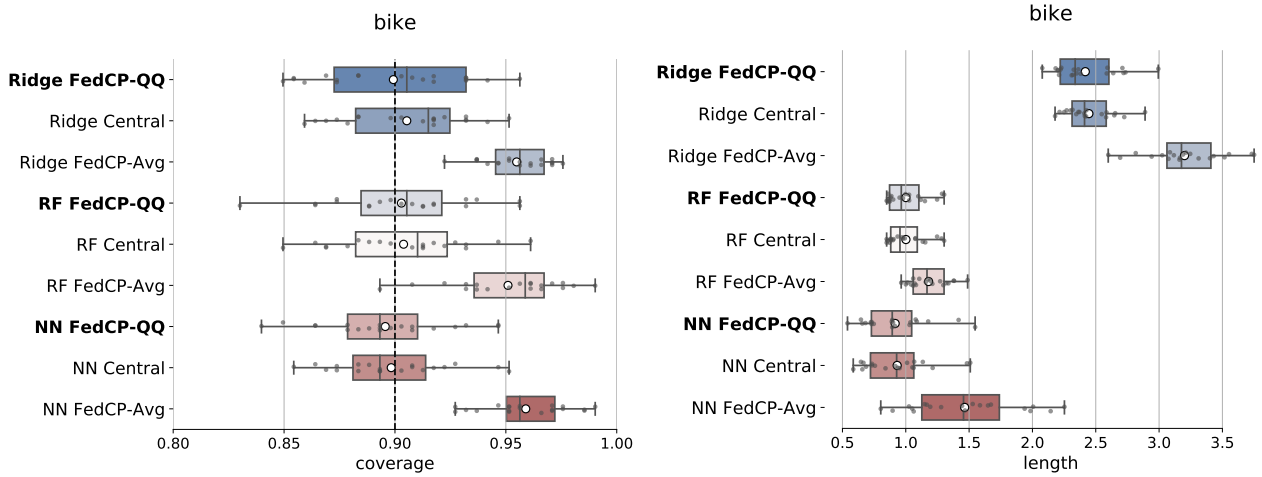

Figure 7. See caption of Figure 5. The size of the calibration set is $m = 100$ and $n = 10$.

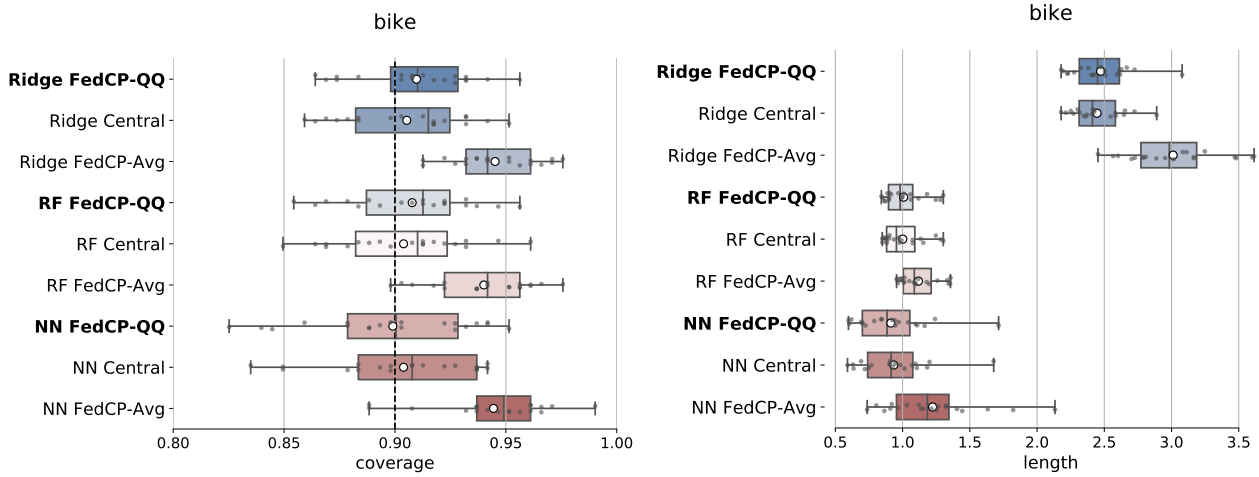

Figure 8. See caption of Figure 6. The size of the calibration set is $m = 10$ and $n = 100$.

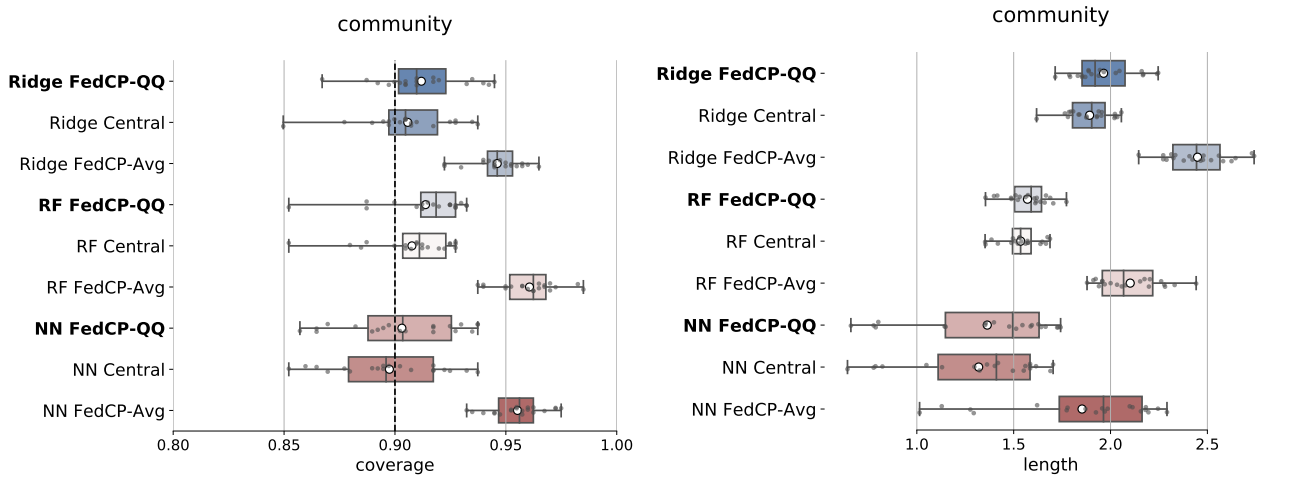

Figure 9. See caption of Figure 5. The size of the calibration set is $m = 80$ and $n = 10$.

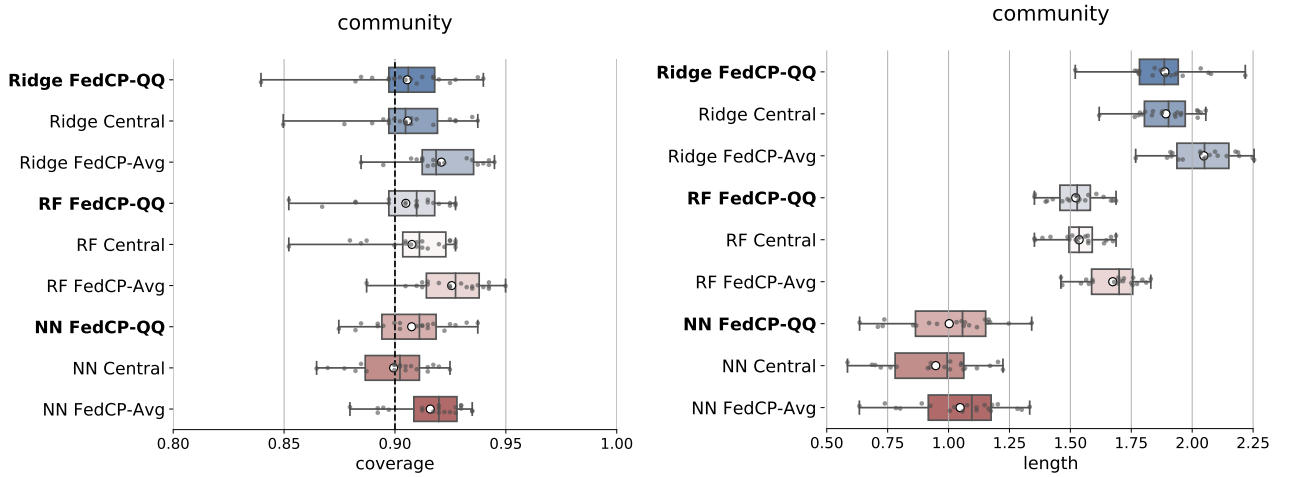

Figure 10. See caption of Figure 6. The size of the calibration set is $m = 10$ and $n = 80$.

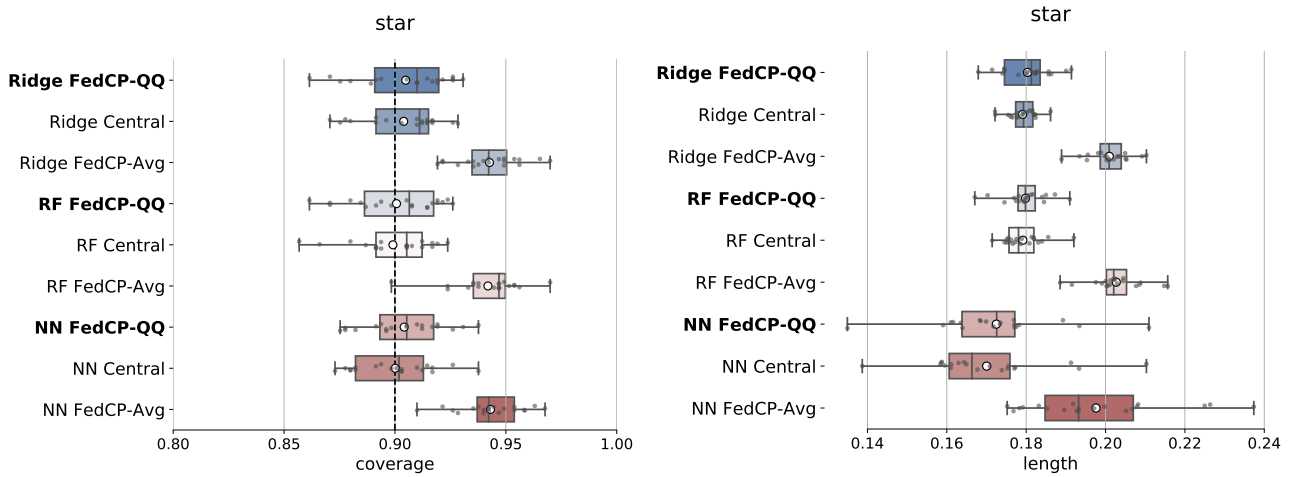

Figure 11. See caption of Figure 5. The size of the calibration set is $m = 80$ and $n = 10$.

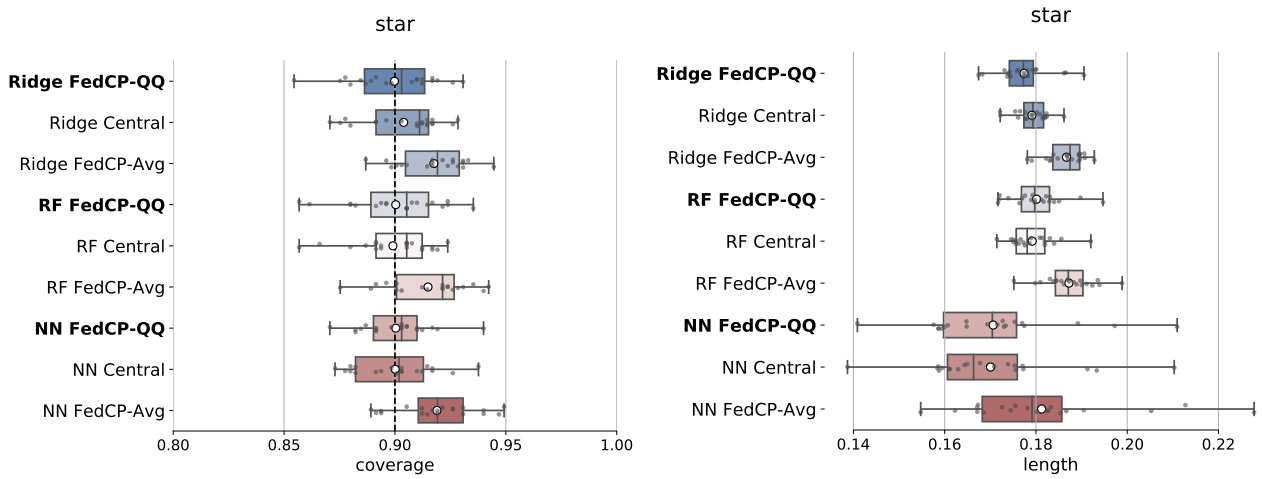

Figure 12. See caption of Figure 6. The size of the calibration set is $m = 10$ and $n = 80$.

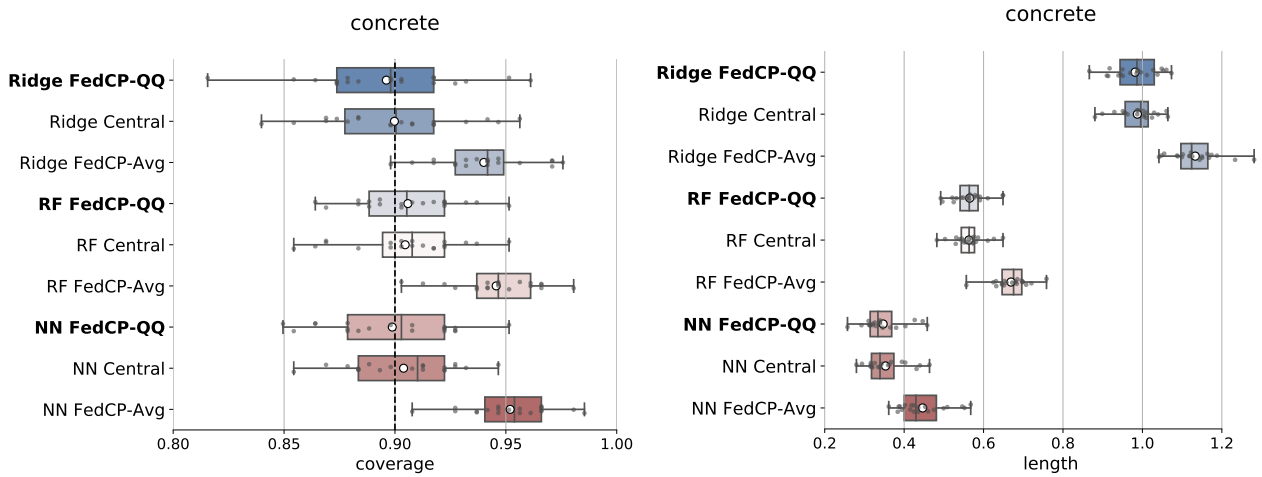

Figure 13. See caption of Figure 5. The size of the calibration set is $m = 40$ and $n = 10$.

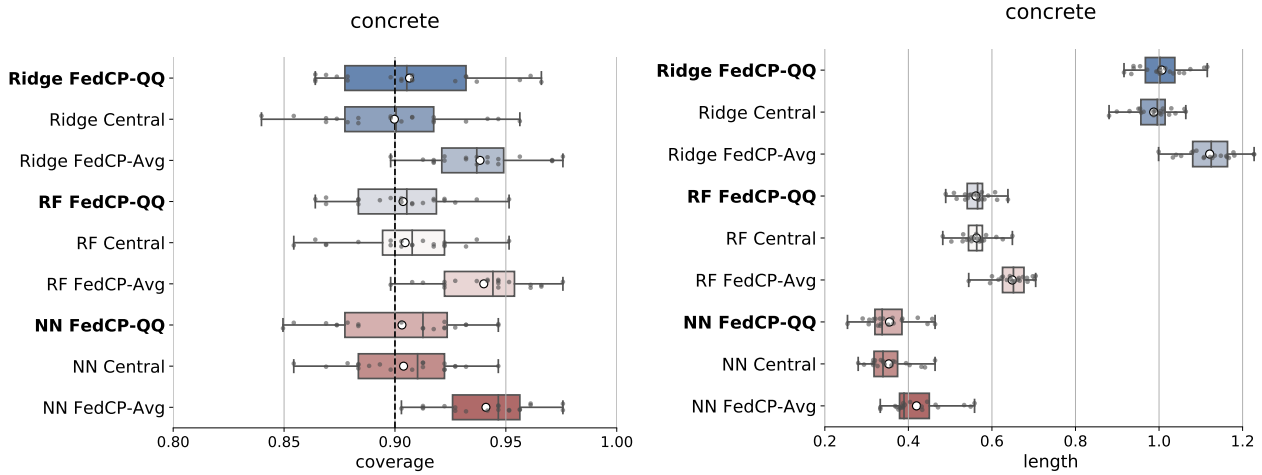

Figure 14. See caption of Figure 6. The size of the calibration set is $m = 10$ and $n = 40$.

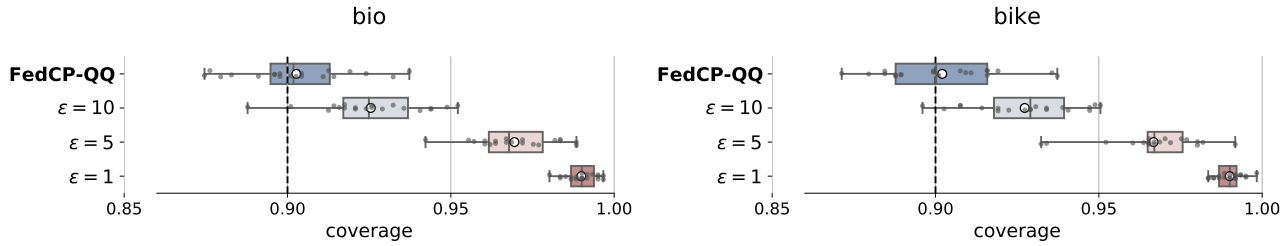

Figure 15. Empirical coverages of prediction intervals ($\alpha = 0.1$) constructed by FedCP-QQ and its private version FedCP^2-QQ for $\varepsilon = 10, 5, 1$. On top, coverages for the bio data set, and, on the bottom for the bike data sets. The white circle represents the mean.

B.2. Private Experiments

For the sake of completeness, we also evaluate the quality of our private algorithm FedCP^2-QQ described in Section 4. More specifically, we apply our private FedCP^2-QQ method on the bio and bike data sets with $m = 5$ and $n = 200$. The predictor is a quantile RF, the number of bins is set to $B = 100$, S_{max} is fixed to the largest score (no clipping), and $\varepsilon = 10, 5, 1$.

Figure 15 displays the empirical (test) coverages obtained over the 20 different random splits. As expected from Theorem 4.1, we observe that on average the desired coverage at 0.90 is well satisfied. However, we also see that the coverages become quickly conservative as the privacy parameter ε decreases. This suggests that the different corrections introduced to compensate for the extra randomness due to privacy may be overly strong. Last but not least, we note that these results would be significantly improved with the privacy amplification strategies discussed in Section 4.

C. Proofs

C.1. Proof of Theorem 3.2

The proof of our results heavily relies on order statistics. We refer to David & Nagaraja (2004) for an in-depth presentation. We begin by recalling the following important result.

Lemma C.1. *Let $X_1, \ldots, X_n$ be some i.i.d. sample drawn from a continuous distribution with c.d.f. F_X and density f_X. If we denote by $X_{(1)} \leq \cdots \leq X_{(n)}$ the corresponding ordered sample, for every $k \in [\![1, n]\!]$, the c.d.f. and density of $X_{(k)}$ are given by*

$$F_{X_{(k)}}(x) = \sum_{i=k}^{n} \binom{n}{i} F_X(x)^i \big[1 - F_X(x)\big]^{n-i} ,$$

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f_X(x) F_X(x)^{k-1} \big[1 - F_X(x)\big]^{n-k} .$$

We can now prove Theorem 3.2.

First, remark that if, conditionally to $\widehat{f}$, $(X_1^{(1)}, Y_1^{(1)}), \ldots, (X_n^{(m)}, Y_n^{(m)}), (X, Y)$ are i.i.d., then, conditionally to $\widehat{f}$, the associated scores $S_1^{(1)}, \ldots, S_n^{(m)}, S$ are i.i.d. We denote by F_S their c.d.f. (given $\widehat{f}$), and make the proof conditionally to $\widehat{f}$.

We know that F_S^{-1} is non-decreasing and that if $U \sim U_{[0,1]}$, $F_S^{-1}(U)$ has the same distribution as S (given $\widehat{f}$). Therefore, if $U_1^{(1)}, \ldots, U_n^{(m)}, U$ are independent with a uniform distribution over $[0, 1]$, and independent from the data, if

$$U_{(\ell,k)} \triangleq \widehat{Q}_{(k)} \left(\widehat{Q}_{(\ell)} \left(\{U_i^{(1)}\,,\, i = 1, \ldots, n\} \right), \ldots, \widehat{Q}_{(\ell)} \left(\{U_i^{(m)}\,,\, i = 1, \ldots, n\} \right) \right) ,$$

denotes the corresponding QQ estimator, then $F_S^{-1}(U_{(\ell,k)})$ has the same distribution as $\widehat{Q}_{(\ell,k)}$ (given $\widehat{f}$). We obtain that

$$\mathbb{P}\left(Y \in \widehat{\mathcal{C}}(X) \mid \widehat{f}\right) = \mathbb{P}\left(S \leq \widehat{Q}_{(\ell,k)} \mid \widehat{f}\right) = \mathbb{P}\left(F_S^{-1}(U) \leq F_S^{-1}(U_{(\ell,k)}) \mid \widehat{f}\right) \geq \mathbb{P}\left(U \leq U_{(\ell,k)} \mid \widehat{f}\right) . \tag{12}$$

Furthermore, if F_S is continuous, F_S^{-1} is increasing, and

$$\mathbb{P}\left(F_S^{-1}(U) \leq F_S^{-1}(U_{(\ell,k)}) \mid \widehat{f}\right) = \mathbb{P}\left(U \leq U_{(\ell,k)} \mid \widehat{f}\right) . \tag{13}$$

Therefore, it remains to treat the uniform case. By Lemma C.1, we have

$$\begin{aligned} F_{U_{(\ell,k)}}(t) &= \sum_{j=k}^{m} \binom{m}{j} F_{U_{(k)}}(t)^j (1 - F_{U_{(k)}}(t))^{m-j} \\ &= \sum_{j=k}^{m} \binom{m}{j} \left[\sum_{i=\ell}^{n} \binom{n}{i} t^i (1-t)^{n-i}\right]^j \left[1 - \sum_{i=\ell}^{n} \binom{n}{i} t^i (1-t)^{n-i}\right]^{m-j} \\ &= \sum_{j=k}^{m} \binom{m}{j} \left[\sum_{i=\ell}^{n} \binom{n}{i} t^i (1-t)^{n-i}\right]^j \left[\sum_{i=0}^{\ell-1} \binom{n}{i} t^i (1-t)^{n-i}\right]^{m-j} \end{aligned}$$

$$\text{since} \qquad 1 = \sum_{i=0}^{n} \binom{n}{i} t^i (1-t)^{n-i} = \sum_{i=0}^{\ell-1} \binom{n}{i} t^i (1-t)^{n-i} + \sum_{i=\ell}^{n} \binom{n}{i} t^i (1-t)^{n-i},$$

hence we get that

$$F_{U_{(\ell,k)}}(t) = \sum_{j=k}^{m} \binom{m}{j} \sum_{i_1=\ell}^{n} \cdots \sum_{i_j=\ell}^{n} \sum_{i_{j+1}=0}^{\ell-1} \cdots \sum_{i_m=0}^{\ell-1} \binom{n}{i_1} \cdots \binom{n}{i_m} t^{i_1+\cdots+i_m} (1-t)^{mn-(i_1+\cdots+i_m)}.$$

As a consequence, we obtain

$$\begin{aligned} \mathbb{P}\left(U_{(\ell,k)} \leq U\right) &= \mathbb{E}\left[F_{U_{(\ell,k)}}(U)\right] \\ &= \int_0^1 F_{U_{(\ell,k)}}(t) \mathrm{d}t \\ &= \int_0^1 \sum_{j=k}^{m} \binom{m}{j} \sum_{i_1=\ell}^{n} \cdots \sum_{i_j=\ell}^{n} \sum_{i_{j+1}=0}^{\ell-1} \cdots \sum_{i_m=0}^{\ell-1} \binom{n}{i_1} \cdots \binom{n}{i_m} t^{i_1+\cdots+i_m} (1-t)^{mn-(i_1+\cdots+i_m)} \mathrm{d}t \\ &= \sum_{j=k}^{m} \binom{m}{j} \sum_{i_1=\ell}^{n} \cdots \sum_{i_j=\ell}^{n} \sum_{i_{j+1}=0}^{\ell-1} \cdots \sum_{i_m=0}^{\ell-1} \binom{n}{i_1} \cdots \binom{n}{i_m} \mathrm{B}\left(i_1+\cdots+i_m+1, mn-(i_1+\cdots+i_m)+1\right), \end{aligned}$$

where

$$\mathrm{B} : (a,b) \in (0,+\infty)^2 \mapsto \int_0^1 t^{a-1}(1-t)^{b-1} \mathrm{d}t$$

denotes the Beta function. The identity $\binom{a}{b} = \dfrac{1}{(a+1)\mathrm{B}(b+1, a-b+1)}$, with $a = mn$ and $b = (i_1 + \cdots + i_m)$, implies that $\mathbb{P}\left(U_{(\ell,k)} \leq U\right) = 1 - M_{n,k}$, hence

$$\mathbb{P}\left(U \leq U_{(\ell,k)}\right) = M_{n,k}. \tag{14}$$

By Eq. (12), we obtain that

$$\mathbb{P}\left(Y \in \widehat{\mathcal{C}}_{\ell,k}(X) \,|\, \widehat{f}\right) \geq M_{n,k}$$

almost surely, hence Eq. (5) by integrating this inequality. When F_S is continuous, Eq. (13) and (14) show that

$$\mathbb{P}\left(Y \in \widehat{\mathcal{C}}_{\ell,k}(X) \,|\, \widehat{f}\right) = M_{n,k}$$

hence the result. □

C.2. Proof of Theorem 3.3

First, let us remark that $\sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}\left\{S_i^{(j)} \leq \widehat{Q}_{(\ell,k)}\right\}$ is almost surely greater or equal to $\ell \cdot k$ by definition of $\widehat{Q}_{(\ell,k)}$. Now, following the proof of Bian & Barber (2022, Theorem 1), by definition of the FedCP-QQ method, we have

$$\left\{Y \in \widehat{\mathcal{C}}_{k,\ell}(X)\right\} = \left\{S \leq \widehat{Q}_{(\ell,k)}\right\}$$

$$
\begin{aligned}
&\supseteq \left\{ \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}\left\{ S_i^{(j)} < S \right\} < \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}\left\{ S_i^{(j)} \leq \widehat{Q}_{(\ell,k)} \right\} \right\} \\
&\supseteq \left\{ \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}\left\{ S_i^{(j)} < S \right\} < \ell \cdot k \right\} \\
&= \left\{ \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{1}\left\{ S_i^{(j)} \geq S \right\} \geq mn - \ell \cdot k \right\} \\
&= \left\{ \bar{F}_{mn}(S) \geq \frac{mn - \ell \cdot k}{mn} \right\} ,
\end{aligned}
$$

where $\bar{F}_{mn}(S)$ is the right-tail empirical c.d.f of the $\{S_i^{(j)}\}_{i,j=1}^{n,m}$ at S. Note that this is a random variable in both the data set and S. We now have

$$
\begin{aligned}
\alpha_P(\mathcal{D}_{mn}) &= \mathbb{P}\left(Y \notin \widehat{\mathcal{C}}_{\ell,k}(X) \mid \mathcal{D}_{mn} \right) \leq \mathbb{P}\left(\bar{F}_{mn}(S) < 1 - (\ell \cdot k)/(mn) \mid \mathcal{D}_{mn} \right) \\
&= \mathbb{P}\left(\bar{F}_{mn}(S) + \bar{F}_S(S) - \bar{F}_S(S) < 1 - (\ell \cdot k)/(mn) \mid \mathcal{D}_{mn} \right) \\
&\leq \mathbb{P}\left(\bar{F}_S(S) \leq 1 - (\ell \cdot k)/(mn) + \sup_{s \in \mathbb{R}} \left\{ \bar{F}_S(s) - \bar{F}_{mn}(s) \right\} \mid \mathcal{D}_{mn} \right) .
\end{aligned}
$$

Fixing any $\Delta > 0$, let consider the event

$$
\left\{ \sup_{s \in \mathbb{R}} \left\{ \bar{F}_S(s) - \bar{F}_{mn}(s) \right\} \leq \Delta \right\} .
$$

Note that it depends of the data $\mathcal{D}_{mn}$. On this event, we have

$$
\alpha_P(\mathcal{D}_{mn}) \leq \mathbb{P}\left(\bar{F}_S(S) \leq 1 - (\ell \cdot k)/(mn) + \Delta \mid \mathcal{D}_{mn} \right) \leq 1 - (\ell \cdot k)/(mn) + \Delta
$$

since $\bar{F}_S(S)$ is a valid p-value (Bian & Barber, 2022, Lemma 1). As a consequence,

$$
\mathbb{P}\big(\alpha_P(\mathcal{D}_{mn}) > 1 - (\ell \cdot k)/(mn) + \Delta \big) \leq \mathbb{P}\left(\sup_{s \in \mathbb{R}} \left\{ \bar{F}_S(s) - \bar{F}_{mn}(s) \right\} > \Delta \right) .
$$

Applying the Dworetzky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990), the last term is bounded by $\delta \in (0, 0.5]$ when we choose $\Delta = \sqrt{\frac{\log(1/\delta)}{2mn}}$. Finally, for $\ell \cdot k \geq (1 - \alpha) \cdot mn$, we have

$$
\mathbb{P}\left(\alpha_P(\mathcal{D}_{mn}) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2mn}} \right) \geq \mathbb{P}\left(\alpha_P(\mathcal{D}_{mn}) \leq 1 - \frac{\ell \cdot k}{mn} + \sqrt{\frac{\log(1/\delta)}{2mn}} \right) \geq 1 - \delta .
$$

□

C.3. Proof of Proposition 3.4

All the proof is made conditionally to the predictor $\widehat{f}$, which means that we prove below that

$$
\mathbb{P}\big(Y \in \widehat{\mathcal{C}}_{\ell^\star,k^\star}(X) \,|\, \widehat{f} \big) \geq 1 - \alpha - \mathbb{E}\left[d_{\mathrm{TV}}\Big(\mathrm{PoisBin}\big(p^*(S)\big) \,,\, \mathrm{Bin}\big(m, \tilde{p}^*(\widetilde{S})\big) \Big) \,\Big|\, \widehat{f} \right] . \tag{15}
$$

The result follows by taking an expectation. In the remainder of the proof, for simplicity, we write $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ instead of $\mathbb{P}(\cdot \mid \widehat{f})$ and $\mathbb{E}[\cdot \mid \widehat{f}]$, respectively.

First, for every $k \in [\![1, m]\!]$ and $\ell \in [\![1, n]\!]$, by definition of $\widehat{\mathcal{C}}_{\ell,k}$, we have

$$
\mathbb{P}\big(Y \notin \widehat{\mathcal{C}}_{\ell,k}(X) \big) = \mathbb{P}\left(\widehat{Q}_{(\ell,k)} < S \right) = \mathbb{P}\Bigg(\underbrace{\sum_{j=1}^{m} \mathbb{1}\left\{ \widehat{Q}_{(\ell)}(\mathcal{S}^{(j)}) < S \right\}}_{\triangleq W} \geq k \Bigg) . \tag{16}
$$

Similarly,

$$\mathbb{P}\left(\widehat{Q}_{(\ell,k)}\left(\widetilde{\mathcal{S}}^{(1)},\dots,\widetilde{\mathcal{S}}^{(m)}\right) < \widetilde{S}\right) = \mathbb{P}\Bigg(\underbrace{\sum_{j=1}^{m} \mathbb{1}\left\{\widehat{Q}_{(\ell)}(\widetilde{\mathcal{S}}^{(j)}) < \widetilde{S}\right\}}_{\triangleq \widetilde{W}} \geq k\Bigg). \tag{17}$$

Given S (and $\widehat{f}$), the random variables $\mathbb{1}\{\widehat{Q}_{(\ell)}(\mathcal{S}^{(j)}) < S\}$, $j = 1,\dots,m$, are independent Bernoulli random variables with respective parameters $p_j(S,\ell) \triangleq \mathbb{P}(S^{(j)}_{(\ell)} \leq S|S)$, so their sum W follows the $\mathrm{PoisBin}(p(S,\ell))$ distribution, where $p(S,\ell) \triangleq (p_1(S,\ell),\dots,p_m(S,\ell))$. Given $\widetilde{S}$ (and $\widehat{f}$), the random variables $\mathbb{1}\{\widehat{Q}_{(\ell)}(\widetilde{\mathcal{S}}^{(j)}) < \widetilde{S}\}$, $j = 1,\dots,m$, are i.i.d. Bernoulli random variables with common parameter $\tilde{p}(\widetilde{S},\ell) \triangleq \mathbb{P}(\widetilde{S}^{(1)}_{(\ell)} \leq \widetilde{S}|\widetilde{S})$, so their sum $\widetilde{W}$ follows the $\mathrm{Bin}(m,\tilde{p}(S,\ell))$ distribution. As a consequence, we have

$$\begin{aligned}\mathbb{P}(W \geq k \,|\, S) - \mathbb{P}(\widetilde{W} \geq k \,|\, \widetilde{S}) &= \mathrm{PoisBin}\big(p(S,\ell)\big)\big([k,+\infty)\big) - \mathrm{Bin}\big(m,\tilde{p}(\widetilde{S},\ell)\big)\big([k,+\infty)\big) \\ &\leq d_{\mathrm{TV}}\Big(\mathrm{PoisBin}\big(p(S,\ell)\big)\,,\,\mathrm{Bin}\big(m,\tilde{p}(\widetilde{S},\ell)\big)\Big),\end{aligned}$$

by definition of the total-variation (TV) distance $d_{\mathrm{TV}}(\mu,\nu) = \sup_{A \text{ measurable}}\{\mu(A)-\nu(A)\}$ for any probability distributions μ and ν.
Taking an expectation and using Eq. (16) and (17), we get that

$$\begin{aligned}\mathbb{P}\big(Y \notin \widehat{\mathcal{C}}_{\ell,k}(X)\big) &= \mathbb{P}(\widetilde{W} \geq k) + \mathbb{P}(W \geq k) - \mathbb{P}(\widetilde{W} \geq k) \\ &\leq \mathbb{P}\left(\widehat{Q}_{(\ell,k)}\left(\widetilde{\mathcal{S}}^{(1)},\dots,\widetilde{\mathcal{S}}^{(m)}\right) < \widetilde{S}\right) + \mathbb{E}\Big[d_{\mathrm{TV}}\Big(\mathrm{PoisBin}\big(p(S,\ell)\big)\,,\,\mathrm{Bin}\big(m,\tilde{p}(\widetilde{S},\ell)\big)\Big)\Big] \\ &\leq 1 - M_{\ell,k} + \mathbb{E}\Big[d_{\mathrm{TV}}\Big(\mathrm{PoisBin}\big(p(S,\ell)\big)\,,\,\mathrm{Bin}\big(m,\tilde{p}(\widetilde{S},\ell)\big)\Big)\Big]\,,\end{aligned}$$

by Theorem 3.2, which applies here since $\widetilde{S}^{(1)}_1,\dots,\widetilde{S}^{(m)}_n,\widetilde{S}$ are i.i.d., conditionally to $\widehat{f}$. Therefore,

$$\mathbb{P}\big(Y \in \widehat{\mathcal{C}}_{\ell,k}(X)\big) \geq M_{\ell,k} - \mathbb{E}\Big[d_{\mathrm{TV}}\Big(\mathrm{PoisBin}\big(p(S,\ell)\big)\,,\,\mathrm{Bin}\big(m,\tilde{p}(\widetilde{S},\ell)\big)\Big)\Big]\,, \tag{18}$$

which implies the result by taking $(\ell,k) = (\ell^*,k^*)$ since $M_{\ell^*,k^*} \geq 1-\alpha$. □

Remark C.2. *In Proposition 3.4, let us emphasize that the auxiliary random variables $\{\widetilde{S}^{(j)}_i\}^{n,m}_{i,j=1}, \widetilde{S}$ can be dependent from the scores $\{S^{(j)}_i\}^{n,m}_{i,j=1}, S$, as long as they satisfy the only assumption required: $\{\widetilde{S}^{(j)}_i\}^{n,m}_{i,j=1}, \widetilde{S}$ must be i.i.d. given $\widehat{f}$. One also can choose the common distribution of the $\{\widetilde{S}^{(j)}_i\}^{n,m}_{i,j=1}, \widetilde{S}$. Here, the best choice is the one that maximizes the right-hand side of Eq. (15). We conjecture that a good choice is to take $\widetilde{S} = S$, and to define the $\widetilde{S}^{(j)}_i$ as independent copies of S (given $\widehat{f}$).*

Finally, let us recall a result from Ehm (1991, Theorem 1) which can be useful to control the right-hand side of Eq. (15).

Theorem C.3. *Let $m \geq 1$ be an integer, $p_1,\dots,p_m \in [0,1]$ and $\tilde{p} = \frac{1}{m}\sum_{j=1}^m p_j$. Let $\mathrm{Bin}(m,\tilde{p})$ denote the binomial distribution and $\mathrm{PoisBin}(m,(p_1,\cdots,p_m))$ denote the Poisson-binomial distribution. The following inequalities hold true:*

$$\begin{aligned}C\big[1-\tilde{p}^{m+1}-(1-\tilde{p})^{m+1}\big]\cdot\left[1-\frac{\sum_{i=1}^m p_i(1-p_i)}{m\tilde{p}(1-\tilde{p})}\right] &\leq d_{\mathrm{TV}}\Big(\mathrm{PoisBin}(p_1,\cdots,p_m)\,,\,\mathrm{Bin}(m,\tilde{p})\Big) \\ &\leq \frac{m}{m+1}\big[1-\tilde{p}^{m+1}-(1-\tilde{p})^{m+1}\big]\cdot\left[1-\frac{\sum_{i=1}^m p_i(1-p_i)}{m\tilde{p}(1-\tilde{p})}\right],\end{aligned}$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ is the total-variation distance, and C is a universal constant.

C.4. Proof of Theorem 4.1

The privacy guarantee is a direct consequence of the fact that Algorithm 2 is ε-DP (exponential mechanism). Indeed, each agent calls this algorithm only one time during FedCP^2-QQ, making it ε-DP with respect to each local agent (ε-LDP).

It remains to prove that the desired coverage is achieved. To do so, recall the following utility lemma related to the output of Algorithm 2 (Angelopoulos et al., 2022).

Lemma C.4. (Utility of Algorithm 2). *For any $\delta \in (0,1)$ and $q \in [0.5, 1)$, the output of Algorithm 2, denoted $\hat{s}_q$, satisfies:*

$$\mathbb{P}\left(\frac{|\{i : \bar{S}_i \leq \hat{s}_q\}|}{n} \geq q - \frac{2\log(B/\delta)}{n\varepsilon} \right) \geq 1 - \delta \tag{19}$$

Proof. The proof, provided in Angelopoulos et al. (2022), Lemma 1, is a direct application of the utility guarantee of the general exponential mechanism (see Dwork et al., 2014, Corollary 3.12 therein). □

We can now prove our main result. Let first define the event $E = \{\widehat{Q}^{\varepsilon} \geq \widehat{Q}_{(\ell_\gamma, k_\gamma)}\}$, i.e., when the private estimator $\widehat{Q}^{\varepsilon}$ returned by FedCP^2-QQ is greater than the non-private estimator $\widehat{Q}_{(\ell_\gamma, k_\gamma)}$ that would be returned by FedCP-QQ (Algorithm 1) with coverage $\frac{1-\alpha}{1-\gamma\alpha}$. Then, we have:

$$\begin{aligned}
\mathbb{P}\Big(Y \in \widehat{\mathcal{C}}_{\varepsilon}(X)\Big) &= \mathbb{P}\Big(S \leq \widehat{Q}^{\varepsilon}\Big) \\
&= \mathbb{P}\Big(S \leq \widehat{Q}^{\varepsilon}, E\Big) + \underbrace{\mathbb{P}\Big(S \leq \widehat{Q}^{\varepsilon}, E^c\Big)}_{\geq 0} \\
&\geq \mathbb{P}\Big(S \leq \widehat{Q}^{\varepsilon} | E\Big) \cdot \mathbb{P}(E) \\
&\geq \mathbb{P}\Big(S \leq \widehat{Q}_{(\ell_\gamma, k_\gamma)}\Big) \cdot \mathbb{P}(E) \\
&\geq \frac{1-\alpha}{1-\gamma\alpha} \cdot \mathbb{P}(E) \ ,
\end{aligned}$$

where the last inequality comes from the fact that $\widehat{Q}_{(\ell_\gamma, k_\gamma)}$ is the output of FedCP-QQ (Algorithm 1) with coverage $\frac{1-\alpha}{1-\gamma\alpha}$. It remains to prove that $\mathbb{P}(E) \geq 1 - \gamma\alpha$.

Notice that a sufficient condition for the event E to be satisfied is if each agent j outputs a value $\widehat{Q}_j^{\varepsilon}$ greater that the ℓ_γ-th ordered score $S^{(j)}_{(\ell_\gamma)}$ of the local data set $S^{(j)}$. Indeed, in that case the k_γ-th ordered value of $\widehat{Q}_1^{\varepsilon}, \ldots, \widehat{Q}_m^{\varepsilon}$, i.e., $\widehat{Q}^{\varepsilon}$, would necessarily be bigger than the k_γ-th ordered value of $S^{(1)}_{(\ell_\gamma)}, \ldots, S^{(m)}_{(\ell_\gamma)}$, i.e., $\widehat{Q}_{(\ell_\gamma, k_\gamma)}$. In the end, we have $E \supset \bigcap_{j=1}^{m} \{\widehat{Q}_j^{\varepsilon} \geq S^{(j)}_{(\ell_\gamma)}\}$, which allows us to obtain a lower bound for $\mathbb{P}(E)$:

$$\mathbb{P}(E) \geq \mathbb{P}\Big(\bigcap_{j=1}^{m} \{\widehat{Q}_j^{\varepsilon} \geq S^{(j)}_{(\ell_\gamma)}\} \Big) = \mathbb{P}\Big(\widehat{Q}_1^{\varepsilon} \geq S^{(1)}_{(\ell_\gamma)}\Big)^m \geq \mathbb{P}\Big(\widehat{Q}_1^{\varepsilon} \geq \bar{S}^{(1)}_{(\ell_\gamma)}\Big)^m \ ,$$

where the equality comes from the fact that the events $\{\widehat{Q}_j^{\varepsilon} \geq S^{(j)}_{(\ell_\gamma)}\}$ are independent and have identical probability. The last inequality comes from the fact that the discretized score $\bar{S}^{(1)}_{(\ell_\gamma)}$ is larger than (or equal to) the non-discretized score $S^{(1)}_{(\ell_\gamma)}$. We can finally bound $\mathbb{P}\Big(\widehat{Q}_1^{\varepsilon} \geq \bar{S}^{(1)}_{(\ell_\gamma)}\Big)$ thanks to Lemma C.4:

$$\begin{aligned}
\mathbb{P}\Big(\widehat{Q}_1^{\varepsilon} \geq \bar{S}^{(1)}_{(\ell_\gamma)}\Big) &= \mathbb{P}\Big(|\{i : \bar{S}_i \leq \widehat{Q}_1^{\varepsilon}\}| \geq \ell_\gamma\Big) \\
&= \mathbb{P}\Big(|\{i : \bar{S}_i \leq \widehat{Q}_1^{\varepsilon}\}| \geq \ell_\gamma + \ell_{\text{cor}} - \ell_{\text{cor}}\Big) \\
&= \mathbb{P}\Big(\frac{|\{i : \bar{S}_i \leq \widehat{Q}_1^{\varepsilon}\}|}{n} \geq \frac{\ell_\gamma + \ell_{\text{cor}}}{n} - \frac{2}{n\varepsilon}\log\Big(\frac{B}{1-(1-\gamma\alpha)^{\frac{1}{m}}}\Big)\Big) \\
&\geq (1-\gamma\alpha)^{\frac{1}{m}} \ ,
\end{aligned}$$

where the last inequality in obtained by applying Lemma C.4 with $q = \frac{\ell_\gamma + \ell_{\text{cor}}}{n}$ and $\delta = 1 - (1-\gamma\alpha)^{\frac{1}{m}}$. Finally, we have $\mathbb{P}(E) \geq 1 - \gamma\alpha$, which concludes the proof. □

C.5. Proof of Proposition A.1

We start by proving the following lemma.

Lemma C.5. *The following equality holds true for every integer $k \geq 1$:*

$$\sum_{j=0}^{k-1} \frac{\Gamma(j+1/n)}{\Gamma(j+1)} = \frac{n \cdot k \cdot \Gamma(k+1/n)}{\Gamma(k+1)} \, .$$

Proof. Throughout the proof, we use that for any $x > 0$, $\Gamma(x) = \dfrac{\Gamma(x+1)}{x}$, according to Davis (1959).

We proceed by induction on k. First, for $k = 1$,

$$\sum_{j=0}^{k-1} \frac{\Gamma(j+1/n)}{\Gamma(j+1)} = \frac{\Gamma(1/n)}{\Gamma(1)} = \Gamma(1/n) = n \cdot \Gamma(1/n+1) = \frac{n\Gamma(1/n+1)}{\Gamma(2)} \, .$$

Then, assume that the result holds true for some $k \geq 1$, that is,

$$\sum_{j=0}^{k-1} \frac{\Gamma(j+1/n)}{\Gamma(j+1)} = \frac{n \cdot k \cdot \Gamma(k+1/n)}{\Gamma(k+1)} \, ,$$

and let us prove that it holds true for $k+1$:

$$\begin{aligned}
\sum_{j=0}^{k} \frac{\Gamma(j+1/n)}{\Gamma(j+1)} &= \sum_{j=0}^{k-1} \frac{\Gamma(j+1/n)}{\Gamma(j+1)} + \frac{\Gamma(k+1/n)}{\Gamma(k+1)} \\
&= \frac{n \cdot k \cdot \Gamma(k+1/n)}{\Gamma(k+1)} + \frac{\Gamma(k+1/n)}{\Gamma(k+1)} \\
&= \frac{(n \cdot k + 1)\Gamma(k+1/n)}{\Gamma(k+1)} \\
&= \frac{(k+1)}{\Gamma(k+2)} \cdot (n \cdot k + 1)\Gamma(k+1/n) \\
&= \frac{(k+1)}{\Gamma(k+2)} \cdot (n \cdot k + 1)\frac{\Gamma(k+1/n+1)}{k+1/n} \\
&= \frac{(k+1)}{\Gamma(k+2)} \cdot (n \cdot k + 1)\frac{\Gamma(k+1/n+1)}{(n \cdot k+1)/n} \\
&= \frac{n \cdot (k+1) \cdot \Gamma(k+1+1/n)}{\Gamma(k+2)} \, .
\end{aligned}$$

□

We can now prove Proposition A.1. Let us assume that $\{U_i^{(j)}\}_{i,j=1}^{m,n}, U$ are i.i.d. uniform on $[0,1]$ and use the notation of the proof of Theorem 3.2. We have $M_{n,k} = \mathbb{P}(U \leq U_{(n,k)})$ by Theorem 3.2 and by Lemma C.1, for every $t \in [0,1]$:

$$F_{U_{(n,k)}}(t) = \sum_{j=k}^{m} \binom{m}{j} F_{U_{(n)}}(t)^j \left[1 - F_{U_{(n)}}(t)\right]^{m-j} = \sum_{j=k}^{m} \binom{m}{j} \left[t^n\right]^j \left[1 - t^n\right]^{m-j} \, .$$

Therefore,

$$\begin{aligned}
1 - M_{n,k} &= \mathbb{P}\left(U_{(n,k)} < U\right) = \mathbb{P}\left(U_{(n,k)} \leq S\right) = \int_0^1 F_{U_{(n,k)}}(t)\mathrm{d}t \\
&= \int_0^1 \sum_{j=k}^{m} \binom{m}{j} (t^n)^j (1-t^n)^{m-j} \mathrm{d}t
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{j=k}^{m}\binom{m}{j}\int_0^1 u^j(1-u)^{m-j}u^{1/n-1}\mathrm{d}u \qquad \text{(change of variable } u=t^n\text{)}\\
&= \frac{1}{n}\sum_{j=k}^{m}\binom{m}{j}\int_0^1 u^{j+1/n-1}(1-u)^{m-j}\mathrm{d}u\\
&= \frac{1}{n}\sum_{j=k}^{m}\binom{m}{j}\mathrm{B}(j+1/n, m-j+1)\ ,
\end{aligned}
$$

where

$$
\mathrm{B}:(a,b)\in(0,+\infty)^2 \mapsto \int_0^1 u^{a-1}(1-u)^{b-1}\mathrm{d}u = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
$$

denotes the Beta function. We obtain that

$$
\begin{aligned}
1-M_{n,k} &= \frac{1}{n}\sum_{j=k}^{m}\frac{\Gamma(j+1/n)}{\Gamma(j+1)}\cdot\frac{\Gamma(m+1)}{\Gamma(m+1/n+1)}\\
&= \frac{1}{n}\cdot\frac{\Gamma(m+1)}{\Gamma(m+1/n+1)}\sum_{j=k}^{m}\frac{\Gamma(j+1/n)}{\Gamma(j+1)}\\
&= \frac{1}{n}\cdot\frac{\Gamma(m+1)}{\Gamma(m+1/n+1)}\left(\sum_{j=0}^{m}\frac{\Gamma(j+1/n)}{\Gamma(j+1)}-\sum_{j=0}^{k-1}\frac{\Gamma(j+1/n)}{\Gamma(j+1)}\right)\ .
\end{aligned}
$$

Using Lemma C.5, we get that

$$
\begin{aligned}
1-M_{n,k} &= \frac{1}{n}\cdot\frac{\Gamma(m+1)}{\Gamma(m+1/n+1)}\left(\frac{n(m+1)\Gamma(m+1/n+1)}{\Gamma(m+2)}-\frac{n\cdot k\cdot\Gamma(k+1/n)}{\Gamma(k+1)}\right)\\
&= \Gamma(m+1)\left(\frac{m+1}{\Gamma(m+2)}-\frac{k\cdot\Gamma(k+1/n)}{\Gamma(k+1)\Gamma(m+1/n+1)}\right)\\
&= 1-\frac{\Gamma(k+1/n)}{\Gamma(k)}\cdot\frac{\Gamma(m+1)}{\Gamma(m+1/n+1)}\ ,
\end{aligned}
$$

which proves the first formula.

Now, using Stirling's formula, when $k, m \to +\infty$, we have

$$
\frac{\Gamma(k+1/n)}{\Gamma(k)}\cdot\frac{\Gamma(m+1)}{\Gamma(m+1/n+1)} \sim \frac{\Gamma(k)k^{1/n}}{\Gamma(k)}\cdot\frac{\Gamma(m)m}{\Gamma(m)m^{1/n+1}} = \frac{k^{1/n}}{m^{1/n}}\ .
$$

By setting $k = k_m \geq m(1-\alpha)^n$, we obtain the second result. □