Conformal Prediction for Federated Uncertainty Quantification Under Label Shift

Vincent Plassier [* 1 2]   Mehdi Makni [* 1]   Aleksandr Rubashevskii [3 4]   Eric Moulines [2 4]   Maxim Panov [5]

Abstract

Federated Learning (FL) is a machine learning framework where many clients collaboratively train models while keeping the training data decentralized. Despite recent advances in FL, the uncertainty quantification topic (UQ) remains partially addressed. Among UQ methods, conformal prediction (CP) approaches provides distribution-free guarantees under minimal assumptions. We develop a new federated conformal prediction method based on quantile regression and take into account privacy constraints. This method takes advantage of importance weighting to effectively address the label shift between agents and provides theoretical guarantees for both valid coverage of the prediction sets and differential privacy. Extensive experimental studies demonstrate that this method outperforms current competitors.

1. Introduction

Federated learning is an increasingly important framework for large-scale learning. FL allows many agents to train a model together under the coordination of a central server without ever transmitting the agents' data over the network, in an attempt to preserve privacy. There has been a considerable amount of FL work over the past 5 years, see e.g. (Bonawitz et al., 2019; Yang et al., 2019; Kairouz et al., 2021; Li et al., 2020). Compared to classical machine learning techniques, FL has two unique features. First, the networked agents are massively distributed, communication

bandwidth is limited, and agents are not always available (*system heterogeneity*). Second, the data distribution at different agents can vary greatly (*statistical heterogeneity*); see (Huang et al., 2022; Yoon et al., 2022). These features lead to serious challenges for both training and inference in federated systems. The focus of this work is on federated inference procedures that allow to build prediction sets for each agent with a confidence level that can be guaranteed.

Conformal Prediction, originally introduced in (Vovk et al., 1999; Shafer and Vovk, 2008; Balasubramanian et al., 2014), has recently gained popularity. It generates prediction sets with guaranteed error rates. Conformal algorithms are shown to be always valid: the actual confidence level is the nominal one, without requiring any specific assumption about the distribution of the data beyond exchangeability; see (Lei et al., 2013; Fontana et al., 2023) and references therein. With few exceptions, CP methods were developed for centralized environments.

We consider below a supervised learning problem with features x taking values in $\mathcal{X}$ and labels y taking values in $\mathcal{Y}$. Let $(X_k, Y_k)_{k=1}^{N_{\text{train}}+N}$ be an independent and identically distributed (i.i.d.) dataset. We divide the data into a *training* and a *calibration* dataset. Formally, let $\{\mathcal{K}_{\text{train}}, \mathcal{K}_{\text{cal}}\}$ be a partition of $\{1, \ldots, N_{\text{train}} + N\}$, and let $N = |\mathcal{K}_{\text{cal}}|$. Without loss of generality, we take $\mathcal{K}_{\text{cal}} = \{1, \ldots, N\}$. We learn a predictor $\hat{f} \colon \mathcal{X} \to \Delta_{|\mathcal{Y}|}$ on the training set $\mathcal{K}_{\text{train}}$, where $|\mathcal{Y}|$ is the number of classes and $\Delta_{|\mathcal{Y}|}$ is the $|\mathcal{Y}|$-dimensional probability simplex. For any covariate $x \in \mathcal{X}$ associated with a label $y \in \mathcal{Y}$, consider a classification score function $\mathcal{S} \colon \mathcal{Y} \times \Delta_{|\mathcal{Y}|} \to [0, 1]$, independent of other covariates and labels, which yields a non-conformity score given by $V(x, y) = \mathcal{S}(y, \hat{f}(x))$. This non-conformity measure estimates how unusual an example looks. Based on these non-conformity scores, standard CP procedure constructs, for each *significance level* $\alpha \in [0, 1]$, a (measurable) *set-valued* predictor $\mathcal{C}_\alpha(x)$ using $\{(X_k, Y_k)\}_{k=1}^N$ that satisfies the following conditions

$$\mathbb{P}\left(Y_{N+1} \in \mathcal{C}_\alpha(X_{N+1})\right) \geq 1 - \alpha, \tag{1}$$

where (X_{N+1}, Y_{N+1}) is a *test point* that is independent of $\mathcal{K}_{\text{train}}$ and $\mathcal{K}_{\text{cal}}$. The quantity $1 - \alpha$ is called the *confidence level*. The guarantee (1) is set up in a centralized environment – all data are available at a central node and usually

---
[*]Equal contribution [1]Lagrange Mathematics and Computing Research Center, Paris, France [2]CMAP, Ecole Polytechnique, Paris, France [3]Skolkovo Institute of Science and Technology, Moscow, Russia [4]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, UAE [5]Technology Innovation Institute, Abu Dhabi, UAE. Correspondence to: Vincent Plassier <vincent.plassier@polytechnique.edu>, Maxim Panov <maxim.panov@tii.ae>.

assuming that the distributions of calibration and test data satisfy $P^{\mathrm{cal}} = P_Y^\star$. If there is a mismatch between the distributions P^{cal} and $P_Y^\star$, then corrections should be made to ensure an appropriate confidence level; see (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021; Barber et al., 2022) and references therein.

Setup. In this work, we consider a federated learning system with n agents. We assume that, instead of storing the entire dataset on a centralized node, each agent $i \in [n]$ owns a local calibration set $\mathcal{D}_i = \{(X_k^i, Y_k^i)_{k=1}^{N^i}\}$, where N^i is the number of calibration samples for the agent i. We further assume that the calibration data are i.i.d. and that the statistical heterogeneity is due to *label shifts*:

$$(X_k^i, Y_k^i) \sim P^i = P_{X|Y} \times P_Y^i,$$

where $P_{X|Y}$, the conditional distribution of the feature given the label, is assumed identical among agents but P_Y^i, the prior label distribution, may differ across agents. In federated learning, statistical heterogeneity is the rule rather than the exception, and it is essential to take into account the presence of label shift at the agent level. We assume that a predictive model $\hat{f}$ has been learned by federated learning. The results we present are agnostic to the learning procedure.

For an agent $\star \in [n]$, and each $\alpha \in (0, 1)$, we are willing to compute a set-valued predictor, $\mathcal{C}_\alpha$ with confidence level $1 - \alpha$, which depends on the calibration data of *all the agents*. The goal is to construct informative conformal prediction sets for each agent, even when its calibration set is limited in size, by using the calibration data of all the agents participating in the FL; we stress that the calibration data must always remain local to the networked agents. Most importantly, the resulting algorithm should attain both conformal and theoretical privacy guarantees – matched to the privacy guarantees that can be obtained in the FL training procedure.

Our main **contributions** to solving this challenging problem can be summarized as follows.

- We introduce a new method, `DP-FedCP`, to construct conformal prediction sets in a federated learning context that addresses label shift between agents; see Section 2. `DP-FedCP` is a federated learning algorithm based on federated computation of weighted quantiles of agent's non-conformity scores, where the weights reflect the label shift of each client with respect to the population. The quantiles are obtained by regularizing the pinball loss using Moreau-Yosida inf-convolution and a version of federated averaging procedure; see Section 3.
- We establish conformal prediction guarantees, ensuring the validity of the resulting prediction sets. Additionally, we provide differential private guarantees for `DP-FedCP`; see Section 4.

- We show that `DP-FedCP` provides valid confidence sets and outperforms standard approaches in a series of experiments on simulated data and image classification datasets; see Section 5.

Related Works. The construction of predictions sets with confidence guarantees has been the subject of much work, mostly in a centralized framework. The conformal framework, introduced in the pioneering works of (Vovk et al., 1999) is appealing in its simplicity/flexibility; see e.g. (Angelopoulos et al., 2021; Fontana et al., 2023) and the references therein. For exchangeable data, this framework provides a model-free methodology for constructing prediction sets that satisfy the desired coverage (Shafer and Vovk, 2008; Papadopoulos et al., 2002; Fannjiang et al., 2022; Angelopoulos et al., 2022b).

These results can also be extended to non-exchangeable data. A method has been developed for dealing with covariate shift (Tibshirani et al., 2019). This method is based on evaluating the discrepancy between the distribution of the calibration data set P^{cal} and the test point distributed according to $P_Y^\star$. Using an estimate of the Radon-Nikodym derivative $\mathrm{d}P_Y^\star / \mathrm{d}P^{\mathrm{cal}}$, a valid prediction set can be obtained by weighting the non-conformity scores. The seminal work of (Tibshirani et al., 2019) led to several improvements, either to form valid prediction sets as long as the f-divergence of the discrepancy remains small (Cauchois et al., 2020), or to formulate hypothesis tests under covariate shifts (Hu and Lei, 2020). In addition, Gibbs and Candes (2021) examine the shift in an online environment; and Lei and Candès (2021) show the validity of the prediction sets even when the distributional shift is only approximated. Since many real-world data sets do not satisfy exchangeability, valid prediction sets are developed in (Barber et al., 2022) that put more mass around the point of interest.

Conformal methods adapted to label shift are considered in (Podkopaev and Ramdas, 2021; 2022) and have similar guarantees to those in (Tibshirani et al., 2019, Corollary 1). Methods for detecting and quantifying label shift have been proposed in (Lipton et al., 2018; Garg et al., 2020).

Differentially private quantiles can be derived based either on the exponential or Gaussian mechanisms (Gillenwater et al., 2021; Pillutla et al., 2022). Using the exponential mechanism, valid prediction sets are generated in (Angelopoulos et al., 2022a). However, quantile computation in a federated learning environment remains a challenge. A first federated approach based on quantile averaging was proposed in (Lu and Kalpathy-Cramer, 2021). However, this work does not provide theoretical guarantees, and the proposed method is vulnerable to distribution shifts. For federated deep learning, the differentially private versions are based on various techniques combination like gradient clipping and the addition of random noise (Triastcyn and

Faltings, 2019; Wei et al., 2020).

Notation. Denote by $[n]$ the set $\{1, \ldots, n\}$ and consider a finite number of labels, i.e., $|\mathcal{Y}| < \infty$. Each agent $i \in [n]$ has N_y^i calibration samples of label $y \in \mathcal{Y}$, and denote $N_y = \sum_{i=1}^n N_y^i$ their total number over all the calibration examples. Recall that N^i the total number of calibration samples on agent i, i.e., $N^i = \sum_{y \in \mathcal{Y}} N_y^i$. Define the total number of calibration data points $N := \sum_{i=1}^n \sum_{y \in \mathcal{Y}} N_y^i$. For a cumulative distribution function F and $\beta \in [0, 1]$, define by $Q_\beta(F) := \inf\{z \colon F(z) \geq \beta\}$ the β-quantile. Finally, for $v \in \mathbb{R}$ denote by δ_v the point-mass distribution.

2. Conformal Prediction for Federated Systems under Label Shift

Non-exchangeable data. In this section, we explain how to take advantage of calibration data to obtain a valid $(1-\alpha)$-prediction set. Consider the calibration dataset $\{(X_k^i, Y_k^i) \colon k \in [N^i]\}_{i \in [n]}$ with data distributed according to $\{P^i\}_{i \in [n]}$. For $\{\pi_i\}_{i \in [n]} \in \Delta_n$ we define the mixture distribution of labels given for $y \in \mathcal{Y}$ by

$$P_Y^{\mathrm{cal}}(y) = \sum_{i=1}^n \pi_i P_Y^i(y).$$

Our goal is to determine a set of likely outputs for a new data point $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star)$ drawn on agent $\star \in [n]$ from the distribution $P^\star$. The conformal approach relies on non-conformity scores $V_k^i = V(X_k^i, Y_k^i)$, $i \in [n]$, $k \in [N^i]$ to determine the prediction set – see (Shafer and Vovk, 2008). These non-conformity scores are uniformly weighted to generate the conventional prediction set

$$\mathcal{C}_{\alpha,\bar{\mu}}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu})\},$$
$$\bar{\mu} = (N+1)^{-1}\left(\sum_{i=1}^n \sum_{k=1}^{N^i} \delta_{V_k^i} + \delta_1\right). \quad (2)$$

However, this method can lead to significant under-coverage in the presence of label shift (Podkopaev and Ramdas, 2021). In fact, since the data $\{(X_k^i, Y_k^i) \colon k \in N^i\}_{i \in [n]}$ are often not exchangeable, it is required to correct the quantile to account for label shift to obtain valid prediction sets (Tibshirani et al., 2019). As proposed by Podkopaev and Ramdas (2021), we begin by assuming that, for all $i \in [n]$ and $y \in \mathcal{Y}$, we have access to the likelihood ratios:

$$w_y^i = P_Y^i(y)/P_Y^{\mathrm{cal}}(y).$$

Denote by $\mathcal{I} = \{(i, k) \colon i \in [n], k \in [N^i]\} \cup \{(\star, N^\star+1)\}$. Using the weights $\{W_k^i \colon (i, k) \in \mathcal{I}\}$ provided in (40), the non-exchangeability correction of Tibshirani et al. (2019) is given for any $\mathbf{y} \in \mathcal{Y}$ by

$$p_{Y_k^i, \mathbf{y}}^\star = \frac{W_k^i}{W_{N^\star+1}^\star + \sum_{j=1}^n \sum_{l=1}^{N^j} W_l^j}, \quad (3)$$

$$\mu_{\mathbf{y}}^\star = p_{\mathbf{y}, \mathbf{y}}^\star \delta_1 + \sum_{i=1}^n \sum_{k=1}^{N^i} p_{Y_k^i, \mathbf{y}}^\star \delta_{V_k^i}.$$

For any covariate $\mathbf{x} \in \mathcal{X}$, define the $(1-\alpha)$-prediction set with *oracle weights*

$$\mathcal{C}_{\alpha,\mu^\star}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\mu_{\mathbf{y}}^\star)\}.$$

In contrast to the exchangeable setting, the quantile is calculated based on a weighted empirical distribution depending on $\mathbf{y}$. The validity of the prediction set is based on the concept of weighted exchangeability, which was introduced in (Tibshirani et al., 2019, Definition 1); see also (Podkopaev and Ramdas, 2021, Theorem 2). In the following, we will suppose that the next assumption holds.

H1. The calibration data points $\{(X_k^i, Y_k^i) \colon (i, k) \in \mathcal{I}\}$ are pairwise independent, and there are no ties between $\{V_k^i \colon (i, k) \in \mathcal{I}\}$ almost surely.

Theorem 2.1. *If* **H1** *holds, then for any* $\alpha \in [0, 1)$, *we have*

$$1 - \alpha \leq \mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\mu^\star}(X_{N^\star+1}^\star)\right)$$

$$\leq 1 - \alpha + \mathbb{E}\left[\max_{(i,k)\in\mathcal{I}}\left\{p_{Y_k^i, Y_{N^\star+1}^\star}^\star\right\}\right], \quad (4)$$

where $p_{Y_k^i, Y_{N^\star+1}^\star}^\star$ *is defined in* (3).

This theorem is directly adapted from (Tibshirani et al., 2019, Corollary 1). For completeness, a formal proof is postponed to Appendix C.1. It is important to note that the lower bound in (4) holds even in the presence of ties between non-conformity scores. Although Theorem 2.1 guarantees the validity of $\mathcal{C}_{\alpha,\mu^\star}(X_{N^\star+1}^\star)$, this prediction set requires the challenging computation of the weights $p_{y,\mathbf{y}}^\star$. Indeed, the calculation of $W_{y,\mathbf{y}}$ requires the summation over $N!$ elements. The first key contribution of our work is given in Theorem 2.2, where we show that alternative weights, which are easier to compute, can lead to valid prediction sets. Specifically, the new weights $\bar{p}_{y,\mathbf{y}}^\star$ are computed on a smaller number of data points $\bar{N} \leq N$, which are randomly selected based on a multinomial random variable with parameter $(\bar{N}, \{\pi_i\}_{i\in[n]})$. Actually, we denote by $\bar{N}^i$ the multinomial count associated with agent i. We take $\bar{N}^i \wedge N^i$ calibration data from agent i and denote $V_k^i = V(X_k^i, Y_k^i)$. For any label $y \in \mathcal{Y}$, the weight $\bar{p}_{y,\mathbf{y}}^\star$ is given by:

$$\bar{p}_{y,\mathbf{y}}^\star = \frac{w_y^\star}{w_{\mathbf{y}}^\star + \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}^i} w_{Y_k^i}^\star}. \quad (5)$$

In addition, consider the following prediction set

$$\bar{\mu}_{\mathbf{y}}^\star = \bar{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}^i} \bar{p}_{Y_k^i, \mathbf{y}}^\star \delta_{V_k^i},$$
$$\mathcal{C}_{\alpha,\bar{\mu}^\star}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu}_{\mathbf{y}}^\star)\}. \quad (6)$$

Denote by $\|w^\star\|_\infty = \max_{y \in \mathcal{Y}}\{w_y^\star\}$. Using the new prediction set $\mathcal{C}_{\alpha,\bar{\mu}^\star}$, we obtain the following result.

Theorem 2.2. *Assume* **H1**. *Set* $\bar{N} = \lfloor N/2 \rfloor$ *and* $\pi_i =$

N^i/N, *for any* $i \in [n]$. *Then,*

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| \leq \frac{6}{N}$$

$$+ \frac{36 + 6\log N}{N}\|w^\star\|_\infty^2 + \frac{14\log N}{N}\sum_{i:\ \frac{N^i}{12} < \log N}\sqrt{N^i}.$$

The preceding theorem shows that $\mathcal{C}_{\alpha,\bar{\mu}^\star}(X^\star_{N^\star+1})$ contains the true label $Y^\star_{N^\star+1}$ with probability *close to* $1 - \alpha$. If $n = 1$ and $N \geq 46$, the set $\{i \in [n]: N^i < 12\log N\}$ is empty. In this case, the convergence rate reduces to $N^{-1}\log N$. More precisely, if each agent has the same number of calibration data, the convergence rate $N^{-1}\log N$ is ensured when $N \geq 12n\log N$. This is for example the case when $N^i = 200$ and $n \leq 86538$. On the other hand, if $n = N$, each agent has only one data point, and in this case the bound becomes $N^{-1}n(\log N)^{3/2}$.

Approximate Weights. Ideally, we would like to use the weights defined in equation (5) to compute valid prediction sets. However, these weights depend on the probability distribution of the labels for each agent, which in many scenarios must be estimated (and therefore known up to an error). Based on empirical estimation of these label probability distributions $\{\widehat{P}_Y^i\}_{i\in[n]}$, for each label y, define the likelihood ratio as follows:

$$\widehat{w}_y^\star = \frac{\widehat{P}_Y^\star(y)}{\sum_{i=1}^n \pi_i \widehat{P}_Y^i(y)}, \tag{7}$$

and denote by $\widehat{p}_{y,\mathbf{y}}^\star$ the weight defined in (5) with $w_y^\star$ replaced by $\widehat{w}_y^\star$. We also consider $\widehat{\mu}_{\mathbf{y}}$ defined as in (6) with $\bar{p}_{y,\mathbf{y}}^\star$ replaced by $\widehat{p}_{y,\mathbf{y}}^\star$. The prediction set becomes

$$\mathcal{C}_{\alpha,\widehat{\mu}}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y}: V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\}. \tag{8}$$

Since computing the exact weights $\bar{p}_{y,\mathbf{y}}^\star$ in (5) may not be feasible, we consider the approximation $\widehat{p}_{y,\mathbf{y}}^\star$ given in (7). We also construct a random variable $(\widehat{X}^\star_{N^\star+1}, \widehat{Y}^\star_{N^\star+1})$ as in (Lei and Candès, 2021) such that $\mathbb{P}(\widehat{Y}^\star_{N^\star+1} = y) = [\sum_{\tilde{y}\in\mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\text{cal}}(\tilde{y})]^{-1} \widehat{w}_y^\star P_Y^{\text{cal}}(y)$, where $\widehat{w}_y^\star$ is defined in (7); and $\widehat{X}^\star_{N^\star+1}|\widehat{Y}^\star_{N^\star+1}$ is drawn according to $P_{X|Y}$. The validity of the resulting prediction set is established in Lemma 2.3. Note that this approach makes the weights' computation feasible, at the cost of introducing one additional approximation.

Lemma 2.3. *For any* $\alpha \in (0,1)$, *we have*

$$\left|\mathbb{P}(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})) - \mathbb{P}(\widehat{Y}^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^\star_{N^\star+1}))\right|$$

$$\leq \frac{1}{2}\sum_{y\in\mathcal{Y}}\left|P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\text{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\text{cal}}(\tilde{y})}\right| := \text{R}, \tag{9}$$

where $\widehat{w}_y^\star$, $\mathcal{C}_{\alpha,\widehat{\mu}}$ *are defined in* (7) *and* (8), *respectively.*

When $\widehat{w}_y^\star$ is sufficiently close to $w_y^\star$, Lemma 2.3 shows that the approximate weights generate accurate prediction

sets (as discussed in Appendix C.1). The error disappears entirely when $\widehat{w}_y^\star = w_y^\star$ for all $y \in \mathcal{Y}$. Furthermore, using (Tibshirani et al., 2019, Corollary 1), we can establish that $\widehat{Y}^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^\star_{N^\star+1})$ with probability nearly $1 - \alpha$. Finally, similar ideas that developed for Theorem 2.2 on $Y^\star_{N^\star+1}$, in conjunction with Lemma 2.3, give a more accurate bound on the coverage validity.

Theorem 2.4. *Assume* **H**1. *For any* $i \in [n]$, *set* $\pi_i = N^i/N$ *and take* $\bar{N} = \lfloor N/2 \rfloor$. *Then,*

$$\left|\mathbb{P}\left(Y^\star_{N^\star+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^\star_{N^\star+1})\right) - 1 + \alpha\right| \leq \frac{36\|\widehat{w}^\star\|_\infty^2}{N(\mathbb{E}\widehat{w}_{Y^{\text{cal}}}^\star)^2}$$

$$+ \text{R} + \frac{6}{N} + \frac{2\log N}{N}\left(\frac{3\|\widehat{w}^\star\|_\infty^2}{(\mathbb{E}\widehat{w}_{Y^{\text{cal}}}^\star)^2} \vee 7\sum_{i:\ \frac{N^i}{12} < \log N}\sqrt{N^i}\right),$$

where $\widehat{w}_{Y_k^i}^\star$, R *are defined in* (7)-(9) *and* $Y^{\text{cal}} \sim P_Y^{\text{cal}}$.

This theorem provides a lower bound on the probability of coverage that is independent of the data distribution. A formal proof can be found in Appendix C.5. This result demonstrates that it is essential to include all agents with the most data. However, it also highlights a counterproductive effect when incorporating agents with few data.

Maximum Likelihood Estimation Weights. Denote by M_y^i the number of training data on agent i associated to label y. Consider the total number of local data $\mathsf{M}^\star = \sum_{y\in\mathcal{Y}}\mathsf{M}_y^i$, the number of training data with label y written by $\mathsf{M}_y = \sum_{i=1}^n \mathsf{M}_y^i$, and the total number of samples on all agents by $\mathsf{M} = \sum_{y\in\mathcal{Y}}\mathsf{M}_y$. When each agent independently learns its approximate label distribution based on counting the number of label in its training datasets, the empirical counterpart of (7) is given for any labels $(y,\mathbf{y}) \in \mathcal{Y}^2$ by

$$\widehat{w}_y^\star = \frac{\mathsf{M}\mathsf{M}_y^\star}{\mathsf{M}^\star\mathsf{M}_y}\mathbb{1}_{\mathsf{M}_y\geq 1}. \tag{10}$$

All the results in this article are given conditionally to the training dataset, meaning that they hold regardless of the specific training data. In order to determine the order of magnitude of the bound of Theorem 2.4, we analyze the average value of R. Given the number of training samples $\{\mathsf{M}^i\}_{i\in[n]}$, if we assume that each training point (X_k^i, Y_k^i) is distributed according to P^i, then taking the expectation over the training set yields:

$$\mathbb{E}[\text{R}] \leq \frac{6}{\sqrt{\mathsf{M}^\star}} + 12\sqrt{\frac{\log|\mathcal{Y}| + \log\mathsf{M}^\star}{\mathsf{M}\min_{y\in\mathcal{Y}}P_Y^{\text{cal}}(y)}}.$$

The proof is given in Appendix C.3. Interestingly, if the previous upper bound is plugged in Theorem 2.4 instead of Lemma 2.3, then the leading error of order $\text{O}(\mathsf{M}^{\star-1/2} \vee N^{-1}\log N)$ is due to the weights' estimates $\{\widehat{p}_{y,\mathbf{y}}^\star\}_{y\in\mathcal{Y}}$. This bound shows that we should not attempt to estimate the likelihood ratios for a single agent, especially when the square root number of local training data on agent $\star$ is small compared to the number of calibration data. Rather, we need

to do this for a group of agents that have approximately the same distribution, which will give us more stable estimators. The agent can benefit from learning simultaneous tasks by exploiting common structures (Caruana, 1998).

3. Privacy Preserving Federated CP

In the previous section, we constructed prediction sets that were valid in theory. However, their practical implementation in a federated environment posed challenges due to the reliance on estimations that are difficult to evaluate. In particular, estimating $Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})$ in order to derive the prediction set $\mathcal{C}_{\alpha,\widehat{\mu}}(\mathbf{x})$, defined in (8), is challenging because it requires knowledge of the global distribution $\widehat{\mu}_y$. This section is divided into two parts: (1) a new method is developed, called DP-FedCP, for estimating quantiles under the federated constraints; (2) then, a method for computing probabilities $\{\widehat{p}_{y,\mathbf{y}}^{\star}\}_{y,\mathbf{y}\in\mathcal{Y}}$ with differential privacy (DP) guarantees is presented.

Quantile Regression and Moreau-Yosida Regularization. Let $\alpha \in (0,1)$, we now propose to estimate the weighted $(1-\alpha)$-quantile of $\widehat{\mu}_{\mathbf{y}}$ defined in (13). To this end, we develop a federated optimization algorithm based on "pinball loss" minimization, a quantile regression techniques with asymmetric penalties (Koenker and Hallock, 2001). For $v \in \mathbb{R}$ and $q \in \mathbb{R}$ define the pinball loss as

$$S_{\alpha,v}(q) = (1-\alpha)(v-q)\mathbb{1}_{\{v \geq q\}} + \alpha(q-v)\mathbb{1}_{\{q>v\}}.$$

For any $\mathbf{y} \in \mathcal{Y}$, the $(1-\alpha)$-quantile of $\widehat{\mu}_{\mathbf{y}}$ is given by

$$Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}}) \in \arg\min_{q\in\mathbb{R}} \{\mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S_{\alpha,V}(q)]\}; \qquad (11)$$

e.g. see (Buhai, 2005). The pinball loss $S_{\alpha,v}$ is lower semicontinuous but not differentiable on $\mathbb{R}$. Hence, we consider the Moreau-Yosida inf-convolution (or *envelope*) $S_{\alpha,v}^{\gamma}$ instead of $S_{\alpha,v}$ – where γ is the regularization parameter; see e.g. (Moreau, 1963) and (Parikh et al., 2014, Chapter 3), whose expression is given by

$$S_{\alpha,v}^{\gamma}(q) = \min_{\tilde{q}\in\mathbb{R}}\left\{S_{\alpha,v}(\tilde{q}) + \frac{1}{2\gamma}(\tilde{q}-q)^2\right\}.$$

The function $S_{\alpha,v}^{\gamma}(\cdot)$ has an explicit expression given in (17). Note that the minima of $S_{\alpha,v}$ and $S_{\alpha,v}^{\gamma}$ coincide. We obtain the weighted quantile by considering $S_{\alpha,v}^{\gamma}$ instead of $S_{\alpha,v}$. An important property is that the inf-convolution of a proper lower semicontinuous convex function is a differentiable function whose derivative is Lipschitz; see (Rockafellar and Wets, 2009, Theorem 2.26). The original optimization problem given in (11) is replaced by a convex/smooth loss:

$$Q_{1-\alpha}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) \in \arg\min_{\mathbb{R}}\{S_{\alpha}^{\gamma}(q)\}, \qquad (12)$$

where $S_{\alpha}^{\gamma}\colon \mathbb{R} \to \mathbb{R}_+$ is the function given by

$$S_{\alpha}^{\gamma}\colon q \mapsto \mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S_{\alpha,V}^{\gamma}(q)].$$

Algorithm 1 DP-FedAvgQE

Input: initial quantile q_0, target significance level α, number of rounds T, learning rate η, Moreau regularization parameter γ, local gradients $\{\nabla S_{\alpha}^{i,\gamma}\}_{i\in[n]}$, local non-conformity scores $\{V_k^i\}_{k\in[N^i+1]}$, mixture weights $\{\lambda_{\mathbf{y}}^i\}_{i\in[n]}$, standard deviation of Gaussian mechanism noise σ_g, K number of local iteration.

for $t = 0$ **to** $T-1$ **do**

  $S_{t+1} \leftarrow$ random subset of $[n]$ **// Server side**

  for each agent $i \in S_{t+1}$ **do**   **// In parallel**

    Initialize quantile $q_{t,0}^i \leftarrow q_t$

    for $k = 0$ **to** $K-1$ **do**

      // Gradient with DP noise

      $g_{t,k}^i \leftarrow \nabla S_{\alpha}^{i,\gamma}(q_{t,k}^i) + z_{t,k}^i, \ z_{t,k}^i \sim \mathcal{N}(0,\sigma_g^2)$

      // Update local quantile

      $q_{t,k+1}^i \leftarrow q_{t,k}^i - \eta g_{t,k}^i$

    $(\Delta q_{t+1}^i, \Delta \bar{q}_{t+1}^i) \leftarrow (q_{t,K}^i - q_{t,0}^i, \sum_{k\in[K]}\frac{q_{t,k}^i}{K})$

  // On the central server

  $q_{t+1} \leftarrow q_t + \frac{n}{|S_{t+1}|}\sum_{i\in S_{t+1}}\Delta q_{t+1}^i$

  $\bar{q}_{t+1} \leftarrow \frac{t}{t+1}\bar{q}_t + \frac{n}{|S_{t+1}|}\sum_{i\in S_{t+1}}\frac{\lambda_{\mathbf{y}}^i \Delta\bar{q}_{t+1}^i}{t+1}$

Output: $\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) \leftarrow \bar{q}_T$.

For almost every value of $\alpha \in (0,1)$, there exists a unique minimizer of S_{α}^{γ}. This minimizer $Q_{1-\alpha}^{\gamma}(\widehat{\mu}_{\mathbf{y}})$ of the regularized loss function deviates from the true quantile. However, the error is controlled by the regularization parameter γ and is asymptotically exact when $\gamma \to 0$. More precisely (see Appendix A.2 for a proof) it holds that:

Theorem 3.1. *Let $\gamma > 0$ and $\alpha \in (0,1)$. Assume that for all $\{y_\ell\}_{\ell\in[N+1]} \in \mathcal{Y}^{[N+1]}$, $1-\alpha \notin \{W_k/W_{N+1}\}_{k\in[N+1]}$, where $W_k = \sum_{\ell=1}^{k}\widehat{w}_{y_\ell}^{\star}$. Then, we have $\left|Q_{1-\alpha}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\right| \leq \gamma$.*

The condition on α assumed in Theorem 3.1 ensures the uniqueness of the minimizer of S_{α}^{γ}.

Federated quantile computation. We now describe the Differentially Private Federated Average Quantile Estimation (DP-FedAvgQE) algorithm (see Algorithm 1), a novel method to compute quantile in a federated learning setting, with DP guarantees. We briefly described this method below. For each query $\mathbf{y} \in \mathcal{Y}$, we consider the distributions $\widehat{\mu}_{\mathbf{y}} = \sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\widehat{\mu}_{\mathbf{y}}^i$, where $\lambda_{\mathbf{y}}^i$ and $\widehat{\mu}_{\mathbf{y}}^i$ are given by

$$\begin{aligned}
\lambda_{\mathbf{y}}^i &= \frac{N^i}{N}\widehat{p}_{\mathbf{y},\mathbf{y}}^{\star} + \sum_{k=1}^{N^i \wedge \bar{N}^i}\widehat{p}_{Y_k^i,\mathbf{y}}^{\star}, \\
\widehat{\mu}_{\mathbf{y}}^i &= \frac{N^i\widehat{p}_{\mathbf{y},\mathbf{y}}^{\star}}{\lambda_{\mathbf{y}}^i N}\delta_1 + \sum_{k=1}^{N^i \wedge \bar{N}^i}\frac{\widehat{p}_{Y_k^i,\mathbf{y}}^{\star}}{\lambda_{\mathbf{y}}^i}\delta_{V_k^i}.
\end{aligned} \qquad (13)$$

To simplify the notation, for any client $i \in [n]$, we introduce the local loss function $S_{\alpha}^{i,\gamma}\colon q \in \mathbb{R} \mapsto \mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}^i}[S_{\alpha,V}^{\gamma}(q)]$; see (18) for explicit expression.

At each iteration $t \in [T]$, the server subsamples the participating agents $S_{t+1} \subseteq [n]$ independently from the past. Each selected agent $i \in S_{t+1}$ performs K local updates: (1) they independently compute their local gradient; (2) a Gaussian noise is added as in (14) to ensure the differential privacy. More precisely, for agent $i \in S_{t+1}$, at local iteration $k \in \{0, \ldots, K-1\}$, we define:

$$g_{t,k}^i = \nabla \mathrm{S}_\alpha^{i,\gamma}(q_{t,k}^i) + z_{t,k}^i, \qquad (14)$$

where $\{z_{t,k}^i : (t,k) \in \{0, \ldots, T-1\} \times [K]\}_{i \in [n]}$ are i.i.d. Gaussian random variables with zero mean and variance σ_g^2. For any agent $i \in S_{t+1}$, $g_{t,k}^i$ is an unbiased estimate of $\nabla \mathrm{S}_\alpha^{i,\gamma}(q_{t,k}^i)$. (3) The participating agents update their local quantiles $q_{t,k+1}^i \leftarrow q_{t,k}^i - \eta g_{t,k}^i$, where η is a positive stepsize; (4) then transmit $(\Delta q_{t+1}^i, \Delta \bar{q}_{t+1}^i) = (q_{t,K}^i - q_{t,0}^i, \sum_{k \in [K]} q_{t,k}^i / K)$ to the central server. The parameter Δq_{t+1}^i is used to update the common parameter q_t, while $\Delta \bar{q}_{t+1}^i$ is necessary to keep track of the average of the sampled parameters denoted $\bar{q}_t$; see (Nemirovski et al., 2009; Bubeck et al., 2015). (5) Finally, the server performs an online average to update $\bar{q}_t$ and computes the new parameter following

$$q_{t+1} = q_t + (n/|S_{t+1}|) \sum_{i \in S_{t+1}} \Delta q_{t+1}^i.$$

At the final stage, the central server output the quantile estimate is given by

$$\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}) = \sum_{t=1}^T (n/|S_t|) \sum_{i \in S_t} \lambda_{\mathbf{y}}^i \Delta \bar{q}_t^i / T. \quad (15)$$

Algorithm 1 is a Federated Averaging procedure (McMahan et al., 2017) applied to the Moreau envelope of the pinball loss. As we will see in Section 4, the addition of an independent Gaussian noise on the parameter at each update round provides differential privacy guarantees; see Theorem 4.4 for more details.

Remark 3.2. Privacy is also at risk when computing probabilities $\{\widehat{p}_{y,\mathbf{y}}^\star\}_{y,\mathbf{y} \in \mathcal{Y}}$. To compute the probabilities $\{\widehat{p}_{y,\mathbf{y}}^\star\}_{y,\mathbf{y} \in \mathcal{Y}}$ while preserving privacy, we need specific mechanisms to transmit the number of training labels $(\mathsf{M}_y^i)_{y \in \mathcal{Y}}$ from each agent i to the server. For this purpose, we use the method proposed in (Canonne et al., 2020). The idea is to add a discrete noise to the counts $\{\mathsf{M}_y^i : i \in [n]\}_{y \in \mathcal{Y}}$ and then transmit these noisy proxies. The resulting algorithm that combines the differentially-private count queries and federated quantile computation is given in Algorithm 2.

Remark 3.3. Algorithm 2 is designed to build a confidence set for the single agent $\star$. By vectorizing all computations, the algorithm can be scaled to compute a confidence set for each agent. This would result in an algorithm that remains linear in the number of clients but would be more efficient than computing several independent runs. From a practical perspective, complexity can be further improved by clustering clients into groups based on their label distributions and

---

Algorithm 2 DP-FedCP

Input: calibration dataset $\{(X_k^i, Y_k^i) : k \in [N^i]\}_{i \in [n]}$, covariate $\mathbf{x}$, communication round number T, subsampling number $\bar{N}$, Gaussian noise parameters $\sigma_g, \bar{\sigma} \geq 0$.
  for each agent $i \in [n] \cup \{\star\}$ **do**   // **In parallel**
    Set $\forall y \in \mathcal{Y}, \mathsf{M}_y^i \leftarrow$ number train data with label y
    Generate $\{\eta_y^i\}_{y \in \mathcal{Y}}$ i.i.d. according to $\mathcal{N}_{\mathbb{Z}}\left(0, \bar{\sigma}^2\right)$
    Send $\forall y \in \mathcal{Y}, \widehat{\mathsf{M}}_y^i \leftarrow \max(1, \mathsf{M}_y^i + \eta_y^i)$
    Compute & Send $\{V(X_k^i, Y_k^i) : k \in [N^i]\}_{i \in [n]}$
  // **On the central server**
  Aggregate $\widehat{\mathsf{M}}_y \leftarrow \sum_{i \in [n]} \widehat{\mathsf{M}}_y^i, \forall y \in \mathcal{Y}$
  Aggregate $\widehat{\mathsf{M}} \leftarrow \sum_{y \in \mathcal{Y}} \widehat{\mathsf{M}}_y$
  for each query $\mathbf{y} \in \mathcal{Y}$ **do**
    Sample $\{\bar{N}^i\}_{i \in [n]} \sim \mathcal{M}ulti(\bar{N}, \{N^i/N\}_{i \in [n]})$
    Compute $\widehat{p}_{y,\mathbf{y}}^\star$ as in (5) with $\widehat{w}_y^\star$ given in (10)
    Compute $\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}) \leftarrow$ DP-FedAvgQE
  Output: $\widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(\mathbf{x}) \leftarrow \{\mathbf{y} : V(\mathbf{x}, \mathbf{y}) \leq \widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})\}$.

---

performing conformal prediction on a group level.

Remark 3.4. The local loss functions $\mathrm{S}_\alpha^{i,\gamma}$ are expressed as the expectation of pinball loss functions. Since the sensitivity of these pinball loss functions is 1, there is no need to clip the gradient. It is sufficient adding Gaussian noise $\mathcal{N}(0, \sigma_g^2)$ to guarantee differential privacy. The value of σ_g is chosen to provide a suitable trade-off between privacy and utility, balancing the need for strong privacy protection with useful outputs. For an explicit setting of σ_g, refer to Theorem 4.4.

4. Theoretical Guarantees

Convergence guarantee. We provide a convergence guarantee for DP-FedAvgQE. Details of the proofs can be found in the supplementary paper. We show the convergence of $\{\widehat{Q}_{1-\alpha,t}^\gamma(\widehat{\mu}_{\mathbf{y}})\}_{t \in \mathbb{N}}$ to a minimizer which is unique under the assumptions discussed in Appendix A.2. We briefly sketch key steps from the theoretical derivations, since the local loss functions $\{\mathrm{S}_\alpha^{i,\gamma}\}_{i \in [n]}$ have different minimizers, this client drift/heterogeneity may slow down the convergence (Li et al., 2019). This dissimilarity is evaluated by the parameter $\zeta \geq 0$, which is given by

$$\zeta = \max_{i \in [n]} \|\nabla \mathrm{S}_\alpha^{i,\gamma} - \nabla \mathrm{S}_\alpha^\gamma\|_\infty^{1/2}.$$

The convergence analysis is performed for the estimate parameter $\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})$ given in (15). We provide below the statements without subsampling, i.e. $S_t = [n]$, given in Appendix B. Recall that $Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}})$ is provided in (12) and denote $\Delta = \mathbb{E}_{q_0}\|q_0 - Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}})\|^2$. The following results hold with fixed train/calibration datasets $(\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{cal}})$, and define their union by $\mathcal{D} = \mathcal{D}_{\mathrm{train}} \cup \mathcal{D}_{\mathrm{cal}}$.

Theorem 4.1. *Let* $\gamma \in (0,1]$, $S_t = [n]$ *and consider the stepsize* $\eta \in (0, \gamma/10]$. *Then, for* $t \in \{0, \dots, T-1\}$, $k \in \{0, \dots, K-1\}$, *we have*

$$\mathbb{E}\left[S_\alpha^\gamma(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})) \,|\, \mathcal{D}\right] - S_\alpha^\gamma(Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}}))$$
$$\leq (\eta K T)^{-1} \Delta + 14 \gamma^{-1} \eta^2 K (\sigma_g^2 + K \zeta^2).$$

The presence of heterogeneity among local datasets significantly influences convergence dynamics, particularly when the number of targets K, is significantly larger than 1. In such cases, the term $K^2 \zeta^2$ poses challenges by potentially hindering the effectiveness of numerous local steps. Consider the stepsize $\eta_\star$ defined by

$$\eta_\star = \min \left\{ \frac{\gamma}{10}, \left(\frac{\gamma \Delta}{13 K^2 T (\sigma_g^2 + \zeta^2 K)} \right)^{1/3} \right\}.$$

Setting $\eta = \eta_\star$, we obtain the following result.

Corollary 4.2. *Let* $\gamma \in (0,1]$, $S_t = [n]$ *and consider the stepsize* $\eta_\star$. *Then, for any* $t \in \{0, \dots, T-1\}$, $k \in \{0, \dots, K-1\}$, *we have*

$$\epsilon_{\mathrm{optim}}^{(\gamma)} = \mathbb{E}\left[S_\alpha^\gamma(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})) \,|\, \mathcal{D}\right] - S_\alpha^\gamma(Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}}))$$
$$\leq \frac{10\Delta}{\gamma K T} + \frac{5 \left(\sigma_g^2 + \zeta^2 K\right)^{1/3} \Delta^{2/3}}{(\gamma K T^2)^{1/3}}. \quad (16)$$

As shown in Corollary 4.2, $\epsilon_{\mathrm{optim}}^{(\gamma)}$ increases inversely proportional to γ. The smaller the regularization parameter γ, the smaller the stepsize η must be, and the more iterations are required to achieve the same accuracy. However, the error caused by the Moreau envelope vanishes for $\gamma \downarrow 0^+$, i.e. $Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}})$ approaches $Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})$. Thus, there is a tradeoff between the accuracy of the quantile approximation $\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})$ and the computational cost.

Conformal guarantees for `DP-FedCP`. We show that the confidence set $\widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star)$ provided by `DP-FedCP` constitutes valid coverage of $Y_{N^\star+1}^\star$. The theoretical derivations and complete statements are given in Appendix C. For all $i \in [n]$, denote by P_V^i the distribution of $V(X^i, Y^i)$ where $(X_i, Y_i) \sim P^i$, and consider $Y^{\mathrm{cal}} \sim P_Y^{\mathrm{cal}}$.

Theorem 4.3. *Assume there exist* $m, M > 0$ *such that for any* $i \in [n]$, P_V^i *admits a density* f_V^i *with respect to the Lebesgue measure that satisfies* $m \leq f_V^i \leq M$. *For any* $\alpha \in [0,1] \setminus \mathbb{Q}$, *it holds*

$$\left| \mathbb{P}(Y_{N^\star+1}^\star \in \widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star)) - \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star)) \right|$$
$$\leq 6M \sqrt{\frac{\log(N) \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y) \widehat{w}_y^\star}{m \min_{y \in \mathcal{Y}} \widehat{w}_y^\star} \left(\mathbb{E}[\epsilon_{\mathrm{optim}}^{(\gamma)} | \mathcal{D}_{\mathrm{train}}] + \gamma \right)}$$
$$+ \frac{2M \log N}{mN} + \frac{4 \operatorname{Var}(\widehat{w}_{Y^{\mathrm{cal}}}^\star)}{N(\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star)^2} + \frac{2\mathbb{E}\widehat{w}_{Y_{N^\star+1}^\star}^\star}{N\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star} + \frac{m}{2N \log N} + \frac{1}{N^2},$$

where $\epsilon_{\mathrm{optim}}^{(\gamma)}$ *is defined in* (16).

This result illustrates an interesting tradeoff introduced by the regularization parameter γ. As shown in Corollary 4.2, $\epsilon_{\mathrm{optim}}^{(\gamma)}$ increases inversely proportional to γ. Therefore, setting $\gamma \approx T^{-1/2}$ ensures a convergence rate of order $T^{-1/4}$ for the optimization procedure. In this case, the error term of order $\mathrm{O}(N^{-1} \log N)$ is guaranteed by choosing the number of iterations $T \approx N^4$. The condition $\alpha \in [0,1] \setminus \mathbb{Q}$ is a strong but unnecessary assumption. However, it provides a simple way to ensure that $\widehat{\mu}_{\mathbf{y}}$ has no jump at level $1 - \alpha$. Interestingly, the same condition on α is used in (Podkopaev and Ramdas, 2021, Corollary 1), where the authors explain why this condition cannot be avoided to ensure the consistency of the empirical quantile estimator.

Differential privacy guarantees. The (ϵ, δ)-differentially private nature of `DP-FedAvgQE` relies on two components: the additional Gaussian noise, combined with the bounded gradient which avoids extreme values/outliers. The parameter ϵ controls the level of privacy protection provided by a differentially private algorithm, by limiting the probability of inferring any information about an individual in a given dataset. However, there is a small chance that the algorithm may leak some information, even though this probability is kept under control by the parameter δ. Based on the Rényi differential privacy (Mironov, 2017), joined to agent subsampling mechanism (Balle et al., 2018), we establish the (ϵ, δ)-DP property following similar ideas to those of (Noble et al., 2022, Theorem 4.1). Detailed proof and definitions are provided in Appendix D.

Theorem 4.4. *If there is a constant number* $S \in [n]$ *of sampled agents, i.e.,* $S_t = S$, *for all* $t \in [T]$. *Then, for all* $\epsilon > 0$ *and* $\delta \in (0, 1 - (1+\sqrt{\epsilon})(1 - S/n)^T)$, *the Algorithm 1 is* (ϵ, δ)*-DP towards a third party when*

$$\sigma_g \geq 2 \sqrt{\frac{K \max_{i \in [n]} \lambda_{\mathbf{y}}^i}{\epsilon} \left(1 + \frac{24 S \sqrt{T} \log(1/\bar{\delta})}{\epsilon n} \right)},$$
$$\text{where} \quad \bar{\delta} = \frac{n}{S} \left[1 - \left(\frac{1-\delta}{1+\sqrt{\epsilon}} \right)^{1/T} \right].$$

5. Numerical experiments

We conducted the experimental study of `DP-FedCP` using both synthetic toy examples and real datasets. To perform a comprehensive evaluation, we compared our method with relevant baselines, namely `Unweighted Local` and `Unweighted Global` (see Appendix E for details). The `Unweighted Local` method computes the quantile based on the local validation data of the agent $\star$ and derives the local unweighted prediction set with $(1 - \alpha)$ confidence level, given by

$$\mathcal{C}_{\alpha, \bar{\mu}^{\mathrm{loc},\star}}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\bar{\mu}_{\mathbf{y}}^{\mathrm{loc},\star} \right) \right\},$$

where $\bar{\mu}_{\mathbf{y}}^{\mathrm{loc},\star} = \frac{1}{N^\star+1}\sum_{k=1}^{N^\star}\delta_{V_k^\star} + \frac{1}{N^\star+1}\delta_1$. This method is the adaptive classification technique with split-conformal calibration applied to agent $\star$, as introduced in (Romano et al., 2020) and also described in (Angelopoulos et al., 2021). On the other hand, the `Unweighted Global` method estimates the quantile based on aggregated non-conformity scores from all the agents, without taking into account the shift between calibration and target distributions. This method computes the $(1-\alpha)$-quantile in an analogous way to the "classical" conformal method recalled in (2).

For our experiments, we apply split-conformal calibration on the entire dataset, which requires all agents to report their non-conformity scores to a central server. We use the same non-conformity score $V(x,y)$ as considered in (Romano et al., 2020; Angelopoulos et al., 2021). Given the covariate x, the predictor $\hat{f}\colon \mathcal{X} \to \Delta_{|\mathcal{Y}|}$ estimates the probability of each class, and orders them from the most to the least likely label. The non-conformity score is then computed as the sum of all the probabilities greater than the true label y. Formally, the non-conformity scores are given by

$$\rho(X_k^i, Y_k^i) = \sum_{y\in\mathcal{Y}} \hat{f}(X_k^i)[y]\,\mathbb{1}_{\{\hat{f}(X_k^i)[y] > \hat{f}(X_k^i)[Y_k^i]\}},$$
$$V(X_k^i, Y_k^i) = \rho(X_k^i, Y_k^i) + U_k^i \times \hat{f}(X_k^i)[Y_k^i],$$

where $U_k^i \in [0,1]$ is a uniform random variable.

Simulated Data Experiment. In the first experiment, we demonstrate that it is necessary to consider label shifts between agents to obtain valid coverage of prediction sets. We consider a simple classification problem with 3 labels. The conditional distributions of the features given the class label are 3 two-dimensional Gaussian distributions with means $\boldsymbol{\theta}_1 = [-1,0], \boldsymbol{\theta}_2 = [1,0], \boldsymbol{\theta}_3 = [1,3]$ and with identity covariance matrices. We consider $n = 2$ agents with the distribution of labels $\{P_Y^1(y)\}_{y\in[3]} = \{0.8, 0.1, 0.1\}$ and $\{P_Y^2(y)\}_{y\in[3]} = \{0.1, 0.1, 0.8\}$. We use the Bayes classifier and consider calibration data with $(N^1, N^2) = (1000, 50)$. The inference is performed for agent 2.

We run independently 1000 experiments with different splits and record the obtained empirical coverage each time. Figure 1a shows the distribution of non-conformity scores for the different labels, and Figure 1b shows the empirical coverage of $(1-\alpha)$ prediction sets with $\alpha = 0.1$ using the `DP-FedCP` method (Algorithm 2) compared to `Unweighted Local` and `Unweighted Global`. We also included results obtained with `oracle-weights`, in which the conformal prediction sets are obtained using (85), i.e., assuming that the exact ratios $\{w_y^\star\}_{y\in\mathcal{Y}}$ are known.

The quantiles calculated via the `Unweighted Global` method are mostly due to the non-conformity scores from agent 1. This is due to the larger local dataset of agent 1, whose label distribution is very different from that of the target; see Figure 1a. The `Unweighted Local` method

(a) Non-conformity scores (b) Empirical coverage

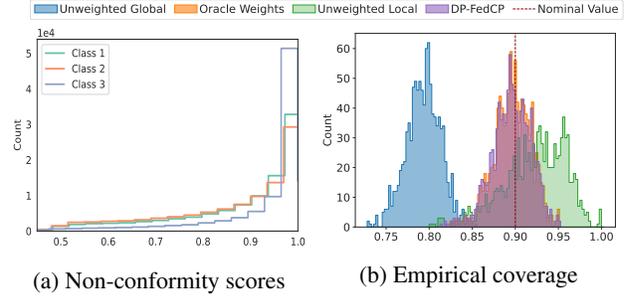Figure 1: Simulated data experiment with 2D data. Target confidence level $(1-\alpha) = 0.9$.
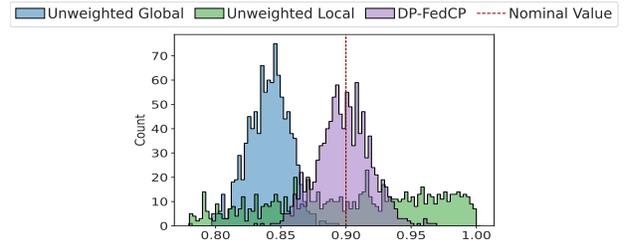

Figure 2: Empirical coverage on the CIFAR-10 data.

computes the quantiles based on the local data of agent 2, which has too little data to produce robust prediction sets. Therefore, `DP-FedCP` yields much better conformal prediction sets (see Figure 1b), which are little different from those obtained using the adaptive prediction set methods with oracle weights of (Podkopaev and Ramdas, 2022).

CIFAR-10 Experiments. We investigate the performance of `DP-FedCP` on the CIFAR-10 dataset. We use a ResNet-56 (He et al., 2016) pre-trained on the CIFAR-10 training dataset as the underlying classifier with temperature scaling $T = 1.6$. We also randomly split the CIFAR-10 test dataset into a calibration dataset and a test dataset, each containing 5000 points, and repeat the experiment 1000 times. The number of agents is $n = 10$, and the prediction set is learned for the agent $\star = 4$ that has the smallest number of data points. The distribution of labels for agent i is $P_Y^i(i) = 0.55$ and $P_Y^i(y) = 0.05$ for all $y \in [10]\backslash\{i\}$. We set the validation size for agent $\star$ to $N^\star = 50$, and for agent 2 the validation size is $N^2 = 2150$. The remaining agents have the same validation size of $N^i = 350$ for all $i \in [10]\backslash\{2,4\}$. The significance level α is set to 0.1. In this configuration, both `Unweighted Local` and `Unweighted Global` methods perform significantly worse than `DP-FedCP`; see Figure 2.

ImageNet Experiments. We use a pre-trained ResNet-152 (He et al., 2016) as a base model with temperature scaling $T = 10$. We perform 1000 runs with different splits of the 50K ImageNet test dataset into calibration and test
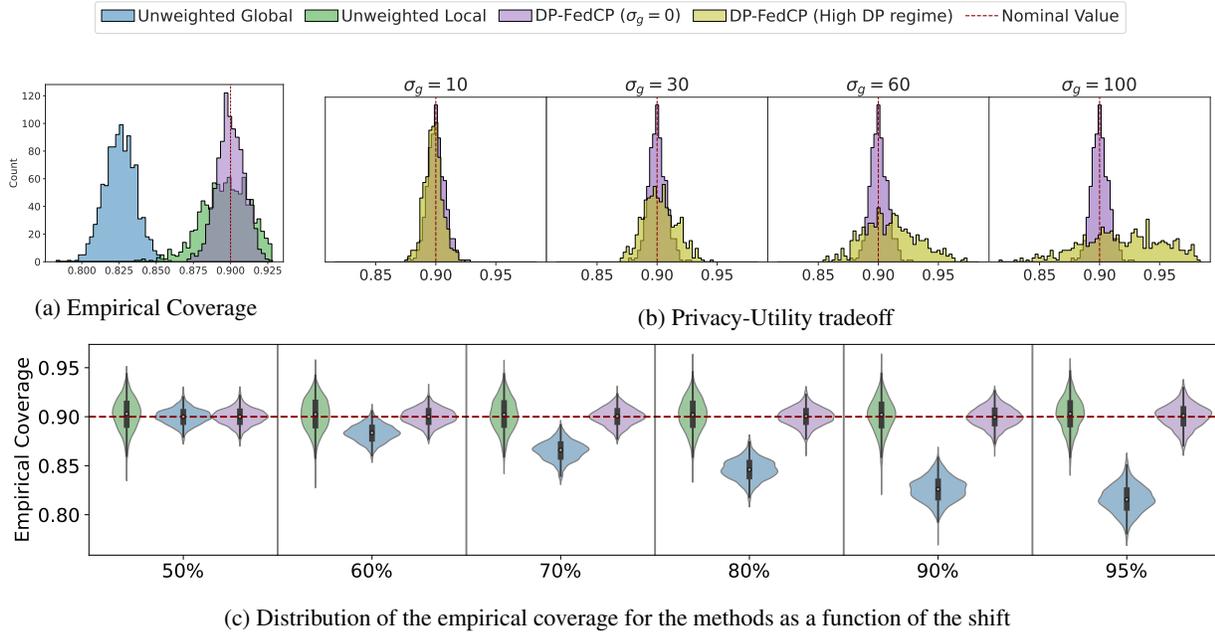
(a) Empirical Coverage

(b) Privacy-Utility tradeoff

(c) Distribution of the empirical coverage for the methods as a function of the shift

Figure 3: ImageNet experimental results: (a) Empirical coverage comparison of `DP-FedCP` with unweighted baselines (b) Empirical coverage comparison of `DP-FedCP` with non-DP version at different privacy parameter values (c) Effect of distribution shifts on empirical coverage for `DP-FedCP` and unweighted baselines.

datasets of size 40K and 10K samples, respectively. The calibration data is split into 11 agents. For agent $i \in [10]$, the size of the calibration dataset is $N^i = 3950$, while we $N^{11} = 500$. For ImageNet, the distribution of non-conformity scores $V(\hat{f}(X), Y)$ varies significantly as a function of the given label $Y = y$. In this experiment, we distribute the data between agents to ensure distinct non-conformity score distributions across agents, illustrated in Figures 6a and 6b. For this, we compute the mean of the non-conformity scores in function of the given label. We call G_1 the set of the 500 labels with the lowest means and G_2 the set of the remaining 500 labels. Agents $i \in [10]$ (`low-score` group) take 90% of their data from G_1 and the remaining 10% from G_2. Agent 11 takes 90% of its calibration data from G_2 and the remaining 10% from G_1.

We construct a prediction set with significance level $\alpha = 0.1$ for the distribution of the 11-th agent. Figure 3a shows the empirical coverage of the prediction sets. In contrast to unweighted alternatives, `DP-FedCP` achieves valid coverage. In Figure 3c, we evaluate the sensitivity of the different methods to the shift between G_1 and G_2. We repeat the previous experiment varying the shift parameter (90% in the first experiment) with 100 runs for each coefficient and show the Violin plot of the obtained empirical coverage. The experimental results show that `DP-FedCP` overcomes the challenge of obtaining valid conformal predictions in the presence of label shifts at a federated level compared to alternative methods.

Differential Privacy Experiments. We explore the trade-off between privacy and coverage quality. We conducted the ImageNet experiment with different values of σ_g in the set $\{10, 30, 60, 100\}$. The results of the experiment are shown in Figure 3b, which illustrates the tradeoff between the differential privacy parameter σ_g and the robustness of the method. In particular, we observe that as σ_g increases, the robustness of the method decreases.

6. Conclusion

We present a novel method called `DP-FedCP`, which is designed to construct personalized conformal prediction sets in a federated learning scenario. Unlike existing algorithms, the proposed method takes into account the label shifts between different agents, and computes prediction sets with a prescribed confidence level. The resulting sets are theoretically guaranteed to provide valid coverage, while ensuring differential privacy. Finally, we illustrate the strong performance of `DP-FedCP` in a series of benchmarks.

References

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. (2021). Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*.

Angelopoulos, A. N., Bates, S., Zrnic, T., and Jordan, M. I. (2022a). Private Prediction Sets. *Harvard Data Science Review*, 4(2). https://hdsr.mitpress.mit.edu/pub/deziirvg.

Angelopoulos, A. N., Krauth, K., Bates, S., Wang, Y., and Jordan, M. I. (2022b). Recommendation systems with distribution-free reliability guarantees. *arXiv preprint arXiv:2207.01609*.

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 31.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al. (2019). Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1:374–388.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Buhai, S. (2005). Quantile regression: overview and selected applications. *Ad Astra*, 4(4):1–17.

Canonne, C. L., Kamath, G., and Steinke, T. (2020). The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688.

Caruana, R. (1998). *Multitask learning*. Springer.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*.

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

Fannjiang, C., Bates, S., Angelopoulos, A., Listgarten, J., and Jordan, M. I. (2022). Conformal prediction for the design problem. *arXiv preprint arXiv:2202.03613*.

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

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Gillenwater, J., Joseph, M., and Kulesza, A. (2021). Differentially private quantiles. In *International Conference on Machine Learning*, pages 3713–3722. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hu, X. and Lei, J. (2020). A distribution-free test of covariate shift using conformal prediction. *arXiv preprint arXiv:2010.07147*.

Huang, W., Ye, M., and Du, B. (2022). Learn from others and be yourself in heterogeneous federated learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10133–10143.

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

Kairouz, P., Oh, S., and Viswanath, P. (2015). The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*.

Lipton, Z., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3122–3130. PMLR.

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

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

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE.

Moreau, J. J. (1963). Propriétés des applications «prox». *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 256:1069–1071.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Noble, M., Bellet, A., and Dieuleveut, A. (2022). Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 10110–10145. PMLR.

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

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239.

Pillutla, K., Laguel, Y., Malick, J., and Harchaoui, Z. (2022). Differentially private federated quantiles with the distributed discrete gaussian mechanism. In *International Workshop on Federated Learning: Recent Advances and New Challenges*.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR.

Podkopaev, A. and Ramdas, A. (2022). Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Triastcyn, A. and Faltings, B. (2019). Federated learning with bayesian differential privacy. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2587–2596. IEEE.

Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469.

Woodworth, B. E., Patel, K. K., and Srebro, N. (2020). Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.

Yoon, J., Park, G., Jeong, W., and Hwang, S. J. (2022). Bitwidth heterogeneous federated learning with progressive weight dequantization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25552–25565. PMLR.

Contents

A. Moreau Envelope for Quantile Computation

!TEX root = ../icml-2023.tex

A.1. Federated quantile using the Moreau envelope

Lemma A.1. *Let $\alpha \in [0, 1]$ and $(v, q) \in \mathbb{R}^2$, the Moreau envelope of the pinball loss with regularization parameter $\gamma > 0$ is given by*

$$S_{\alpha,v}^{\gamma}(q) = \begin{cases} (1-\alpha)(v-q) - \frac{\gamma(1-\alpha)^2}{2}; & \frac{v-q}{\gamma} > 1 - \alpha, \\ \frac{(q-v)^2}{2\gamma}; & 0 \leq \frac{q-v}{\gamma} + 1 - \alpha \leq 1, \\ \alpha(q-v) - \frac{\gamma\alpha^2}{2}; & \frac{q-v}{\gamma} > \alpha. \end{cases} \tag{17}$$

Moreover, its gradient is given by

$$\nabla S_{\alpha,v}^{\gamma}(q) = -(1-\alpha)\mathbb{1}_{\{q < v - \gamma(1-\alpha)\}} + \alpha\mathbb{1}_{\{q > v + \gamma\alpha\}} + \frac{1}{\gamma}(q-v)\mathbb{1}_{\{v-\gamma(1-\alpha) < q < v + \gamma\alpha\}}.$$

Proof. For all $\alpha \in [0, 1]$, $(v, q) \in \mathbb{R}^2$, recall that the pinball loss and its subgradient are given by

$$S_{\alpha,v}(q) = (1-\alpha)(v-q)\mathbb{1}_{\{v \geq q\}} + \alpha(q-v)\mathbb{1}_{\{q > v\}}, \qquad \partial S_{\alpha,v}(q) = \begin{cases} -(1-\alpha), & q < v \\ [-(1-\alpha), \alpha], & q = v \ . \\ \alpha, & q > v \end{cases}$$

Note that, by construction

$$S_{\alpha,v}^{\gamma}(q) = \min_{\tilde{q}\in\mathbb{R}}\left\{S_{\alpha,v}(\tilde{q}) + \frac{1}{2\gamma}(\tilde{q}-q)^2\right\} = \min_{\tilde{q}\in\mathbb{R}}\left\{(1-\alpha)(v-\tilde{q})\mathbb{1}_{\{v\geq\tilde{q}\}} + \alpha(\tilde{q}-v)\mathbb{1}_{\{v<\tilde{q}\}} + \frac{1}{2\gamma}(\tilde{q}-q)^2\right\}.$$

Denote $q_\star = \arg\min_{\tilde{q}\in\mathbb{R}}\{S_{\alpha,v}(\tilde{q}) + \frac{1}{2\gamma}(\tilde{q}-q)^2\}$ which exists and is unique (the function to be minimized is coercive and strongly convex). The stationary condition for the Moreau envelope is given by:

$$0 \in \partial S_{\alpha,v}(q_\star) + \frac{1}{\gamma}(q_\star - q), \quad \text{with} \quad \partial S_{\alpha,v}(q) = \begin{cases} -(1-\alpha), & q < v \\ [-(1-\alpha),\alpha], & q = v \\ \alpha, & q > v \end{cases}.$$

Considering the 3 different cases, we find that:

$$q_\star = \begin{cases} q + \gamma(1-\alpha), & q < v - \gamma(1-\alpha) \\ v, & q \in [v - \gamma(1-\alpha), v + \gamma\alpha] \\ q - \gamma\alpha, & q > v + \gamma\alpha \end{cases}.$$

We conclude the derivation by using the identity from Moreau envelope: $S_{\alpha,v}^{\gamma}(q) = S_{\alpha,v}(q_\star) + \frac{1}{2\gamma}(q_\star - q)^2$ and plugging in $q_\star$. □

To simplify the manuscript presentation, we now provide the definition of the local loss function $S_\alpha^{i,\gamma}: q \in \mathbb{R} \mapsto \mathbb{E}_{V\sim\widehat{\mu}_y^i}[S_{\alpha,V}^{\gamma}(q)] \in \mathbb{R}_+$. Recall the weights $\widehat{p}_{y,\mathbf{y}}^{\star}$ are given in (7), and also that

$$\lambda_{\mathbf{y}}^i = \frac{N^i}{N}\widehat{p}_{\mathbf{y},\mathbf{y}}^{\star} + \sum_{k=1}^{N^i}\widehat{p}_{Y_k^i,\mathbf{y}}^{\star}.$$

Therefore, for $q \in \mathbb{R}$, we have

$$S_\alpha^{i,\gamma}(q) = \frac{N^i\widehat{p}_{\mathbf{y},\mathbf{y}}^{\star}}{\lambda_{\mathbf{y}}^i N}S_{\alpha,1}^{\gamma}(q) + \sum_{k=1}^{N^i}\frac{\widehat{p}_{Y_k^i,\mathbf{y}}^{\star}}{\lambda_{\mathbf{y}}^i}S_{\alpha,V_k^i}^{\gamma}(q). \tag{18}$$

A.2. Moreau's approximation error

In this section, we consider fixed parameters $\alpha \in (0,1)$, $\gamma > 0$, $\{v_k\}_{k\in[N]}$ and $\{p_k\}_{k\in[N]} \in [0,1]^N$ satisfying $\sum_k p_k = 1$. We define $F := \sum_{k=1}^N p_k S_{\alpha,v_k}$ and $F_\gamma := \sum_{k=1}^N p_k S_{\alpha,v_k}^{\gamma}$, where S_{α,v_k} and S_{α,v_k}^{γ} are the pinball loss and its Moreau envelope defined for $v, q \in \mathbb{R}$ by (17). Without loss of generality, it is assumed that $\{v_k\}_{k\in[N]}$ is increasing since we can re-index $\{v_k\}_{k\in[N]}$ and if there exist $(j,j') \in [N]^2$ such that $j \neq j'$ and $v_j = v_{j'}$, we have $p_j S_{\alpha,v_j} + p_{j'}S_{\alpha,v_{j'}} = (p_j + p_j')S_{\alpha,v_j}$. Finally, for $k \in [N]$, denote

$$I_k = [v_k - \gamma(1-\alpha), v_k + \gamma\alpha]. \tag{19}$$

Lemma A.2. *If* $(1-\alpha) \notin \{\sum_{l=1}^k p_l\}_{k\in[n]}$, *then* F *admits a unique minimizer. Moreover, this minimizer belongs to* $\{v_k\}_{k\in[n]}$ *and we denote* $k_\star \in [n]$ *its index, i.e.,* $v_{k_\star} = \arg\min F$. *In addition,* F *is decreasing on* $(-\infty, v_{k_\star}]$ *and increasing on* $[v_{k_\star}, \infty)$. *The function* F_γ *also admits a unique minimizer denoted* $Q_{1-\alpha}^{\gamma} \in \mathbb{R}$, *and* F_γ *is decreasing on* $(-\infty, Q_{1-\alpha}^{\gamma}]$ *and increasing on* $[Q_{1-\alpha}^{\gamma}, \infty)$.

Proof. Note that F is differentiable on $\mathbb{R} \setminus \{v_k\}_{k\in[N]}$, and for all $q \in \mathbb{R} \setminus \{v_k\}_{k\in[N]}$, we have

$$F'(q) = \alpha\sum_{k:\,v_k<q}p_k - (1-\alpha)\sum_{k:\,v_k\geq q}p_k = \sum_{k:\,v_k<q}p_k - (1-\alpha).$$

Since $\alpha \in (0,1)$ with $(1-\alpha) \notin \{\sum_{l=1}^k p_l\}_{k\in[N]}$, we deduce that there exists a unique $v_{k_\star} \in \{v_k\}_{k\in[N]}$ such that, for $q \in \mathbb{R}$,

$$\sum_{k:\,v_k<q}p_k - (1-\alpha) \quad \text{is} \quad \begin{cases} <0 & \text{if } q < v_{k_\star}, \\ >0 & \text{if } q > v_{k_\star}. \end{cases}$$

Thus, from the continuity of F it follows that F is decreasing on $(-\infty, v_{k_\star}]$ and increasing on $[v_{k_\star}, \infty)$. Moreover, since $F'_\gamma = F'$ on $\mathbb{R} \setminus \{\cup_{k\in[N]}I_k\}$, its minimizer lies in $\cup_{k\in[N]}I_k$, where I_k is defined in (19). Finally, the strong convexity of F_γ on $\cup_{k\in[N]}I_k$ shows the uniqueness of $Q^\gamma_{1-\alpha} = \arg\min F_\gamma$ and also that F_γ is decreasing on $(-\infty, Q^\gamma_{1-\alpha}]$ and increasing on $[Q^\gamma_{1-\alpha}, \infty)$. $\qquad\square$

Denote by ∂F subgradient of F. If F is differentiable at $q \in \mathbb{R}$, then $\partial F(q) = \{F'(q)\}$. When $\partial F(q)$ is a singleton, by an abuse of notation, we use the same notation for the set and the unique element it contains.

Theorem A.3. *Assume* $(1 - \alpha) \notin \{\sum_{l=1}^k p_l\}_{k\in[N]}$, *then the unique minimizers* $(v_{k_\star}, Q^\gamma_{1-\alpha})$ *resp. of* (F, F_γ) *resp. satisfy* $|Q^\gamma_{1-\alpha} - v_{k_\star}| \leq \gamma$.

Proof. First, for any $k \in [N]$ such that $v_{k_\star} - \gamma(1 - \alpha) \in I_k$, we obtain

$$\frac{v_{k_\star} - v_k - \gamma(1-\alpha)}{\gamma} \leq \begin{cases} -(1-\alpha) & \text{if } v_k \geq v_{k_\star}, \\ \alpha & \text{else } v_k < v_{k_\star}. \end{cases} \tag{20}$$

Since Lemma A.2 shows that F is decreasing on $(-\infty, v_{k_\star}]$ and increasing on $[v_{k_\star}, \infty)$ where $v_{k_\star}$ is the unique minimizer of $F(v_{k_\star})$, the convexity of F implies that:

$$\partial F(q) \subset \begin{cases} (-\infty, 0) & \text{if } q < v_{k_\star}, \\ (0, \infty) & \text{if } q > v_{k_\star}. \end{cases} \tag{21}$$

Thus, (20) combined with (21) give that

$$F'_\gamma\big(v_{k_\star} - \gamma(1-\alpha)\big)$$
$$= \sum_{k:\, v_{k_\star} - \gamma(1-\alpha)\notin I_k} p_k S'_{\alpha, v_k}\big(v_{k_\star} - \gamma(1-\alpha)\big) + \sum_{k:\, v_{k_\star} - \gamma(1-\alpha)\in I_k} p_k \frac{\big(v_{k_\star} - v_k - \gamma(1-\alpha)\big)}{\gamma}$$
$$\leq \alpha \sum_{k:\, v_k < v_{k_\star}} p_k - (1-\alpha) \sum_{k:\, v_k \geq v_{k_\star}} p_k = \partial F(v_{k_\star} - \epsilon) < 0,$$

where $\epsilon = 2^{-1}\min_{k=1}^{N-1}\{v_{k+1} - v_k\}$. A similar reasoning shows that $F'_\gamma(v_{k_\star} + \gamma\alpha) \geq \partial F(v_{k_\star} + \epsilon) > 0$. Since F_γ is decreasing on $(-\infty, Q^\gamma_{1-\alpha}]$ and increasing on $[Q^\gamma_{1-\alpha}, \infty)$ by Lemma A.2, we have $F'_\gamma < 0$ on $(-\infty, v_{k_\star} - \gamma(1-\alpha)]$ and $F'_\gamma > 0$ on $[v_{k_\star} - \gamma\alpha, \infty)$. Therefore, we deduce that $Q^\gamma_{1-\alpha} \in I_{k_\star}$. Using that the interval $I_{k_\star}$ is of length γ, this implies that $|Q^\gamma_{1-\alpha} - v_{k_\star}| \leq \gamma$. $\qquad\square$

B. FL convergence guarantee: proof of Theorem 4.1

In this section, we suppose that $\{S_t\}_{t\in[T]}$ is a sequence of i.i.d. random variables, such that, for any $(i, i') \in [n]^2$, $i \in S_t$ and $i' \in S_t$ are independent if $i \neq i'$. For any $i \in [n]$, let $\{z^i_{t,k}: k \in \{0, \ldots, K\}\}_{t=0}^T$ be a sequence of i.i.d. standard Gaussian variables. Moreover, consider the local loss function $F^i : \mathbb{R} \to \mathbb{R}$ and denote, for $t \in \{0, \ldots, T\}, k \in \{0, \ldots, K\}$

$$F = \sum_{i=1}^n \lambda^i_{\mathbf{y}} F^i, \qquad\qquad g^i_{t,k} = \nabla F^i(q^i_{t,k}) + z^i_{t,k}.$$

In this section, we establish the convergence of the iterates given by Theorem B.3 under the following assumptions:

H2. The function $\sum_{i=1}^n \lambda^i_{\mathbf{y}} F^i$ admits at least a minimizer in $\mathbb{R}$, we denote $q_\star$ one of them, i.e., $q_\star \in \arg\min\{\sum_{i=1}^n \lambda^i_{\mathbf{y}} F^i\}$.

H3. For any $i \in [n]$, F^i is continuously differentiable and convex, i.e., for any $q, \tilde{q} \in \mathbb{R}$,

$$F^i(\tilde{q}) \leq F^i(q) + \langle \nabla F^i(q), \tilde{q} - q \rangle.$$

H4. For any $i \in [n], t \in \{0, \ldots, T\}, K \in \{0, \ldots K\}$, ∇F^i is continuously differentiable. In addition, there exist $H^i \geq 0$ such that the function ∇F^i is H^i-smooth, i.e., for any $q, \tilde{q} \in \mathbb{R}$,

$$\nabla F^i(\tilde{q}) \leq \nabla F^i(q) + \langle \nabla F^i(q), \tilde{q} - q \rangle + (H^i/2)\|\tilde{q} - q\|^2.$$

Moreover, denote $H = \max_{i \in [n]} \{H^i\}$.

We introduce the key assumptions appearing in the theoretical derivations below.

H5. For any $i \in [n]$, the variance of the gradients is uniformly bounded, for all $q \in \mathbb{R}$, we have

$$\mathbb{E}\left[\left\|\frac{\mathbb{1}_{i \in S_t}}{\mathbb{P}(i \in S_t)}\nabla F^i(q) - \nabla F^i(q)\right\|^2\right] \leq \xi^2, \qquad \mathbb{E}\left[\left\|\sum_{i \in S_t}\frac{\lambda_{\mathbf{y}}^i}{\mathbb{P}(i \in S_t)}\nabla F^i(q_\star) - \nabla F(q_\star)\right\|^2\right] \leq \xi_\star^2.$$

H6. The heterogeneity denoted ζ is bounded everywhere

$$\max_{i \in [n]}\left\{\|\nabla F^i - \nabla F\|_\infty\right\} \leq \zeta^2.$$

We prove Theorem B.3 using Lemma B.1 and Lemma B.2. Note that these results are close to Woodworth et al. (2020, Appendix C). However, we treat partial participation, i.e., $S_t \subseteq [n]$ and consider an objective function defined by importance weights $\{\lambda_{\mathbf{y}}^i\}_{i \in [n]}$. At time $t \in \{0, \ldots, T\}$, denote

$$\bar{q}_{t,k} = \sum_{i=1}^{n} \lambda_{\mathbf{y}}^i (\mathbb{1}_{S_{t+1}}(i)/\mathbb{P}(i \in S_{t+1}))q_{t,k}^i$$

the average of the local parameters defined in Algorithm 1. Finally, we introduce the following stepsize:

$$\eta_0 = \frac{1}{10}\min\left(\frac{1}{H}, \min_{i=1}^{n}\left\{\frac{\mathbb{P}(i \in S_t)}{\lambda_{\mathbf{y}}^i H^i}\right\}\right). \tag{22}$$

Lemma B.1. *Assume* **H2-H3-H4-H5** *and consider* $\eta \in (0, \eta_0]$. *Then, for any* $t \in \{0, \ldots, T-1\}$, $k \in \{0, \ldots, K-1\}$, *we have*

$$\mathbb{E}\left[F\left(\bar{q}_{t,k}\right) - F(q_\star)\right] \leq \frac{1}{\eta}\mathbb{E}\|\bar{q}_{t,k} - q_\star\|^2 - \frac{1}{\eta}\mathbb{E}\|\bar{q}_{t,k+1} - q_\star\|^2 + 2H\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\mathbb{E}\|\bar{q}_{t,k} - q_{t,k}^i\|^2$$

$$+ 3\eta\xi_\star^2 + \eta^2\sigma^2\sum_{i=1}^{n}\frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}.$$

Proof. Developing the squared norm, we find

$$\mathbb{E}\|\bar{q}_{t,k+1} - q_\star\|^2 = \mathbb{E}\left\|\bar{q}_{t,k} - \eta\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right) - q_\star\right\|^2 + \eta^2\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\left\{\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})}g_{t,k}^i - \nabla F^i\left(q_{t,k}^i\right)\right\}\right\|^2. \tag{23}$$

We start by upper bounding the first term, we have

$$\mathbb{E}\left\|\bar{q}_{t,k} - \eta\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right) - q_\star\right\|^2$$

$$= \mathbb{E}\|\bar{q}_{t,k} - q_\star\|^2 - 2\eta\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\mathbb{E}\left\langle\bar{q}_{t,k} - q_\star, \nabla F^i\left(q_{t,k}^i\right)\right\rangle + \eta^2\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right)\right\|^2. \tag{24}$$

Using **H4**, we know that F^i is H^i-smooth and thus $F = \sum_{i=1}^{n}\lambda_{\mathbf{y}}^i F^i$ is $\bar{H}$-smooth, where $\bar{H} = \sum_{i=1}^{n}\lambda_{\mathbf{y}}^i H^i$. Following (Nesterov, 2003), the smoothness and convexity of F imply that

$$\|\nabla F\left(\bar{q}_{t,k}\right) - \nabla F(q_\star)\|^2 \leq 2\bar{H}\left(F\left(\bar{q}_{t,k}\right) - F(q_\star)\right).$$

For any $a, b \in \mathbb{R}$, using that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the last term of (24) can be upper bounded as follows:

$$\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\nabla F^i\left(q_{t,k}^i\right)\right\|^2 \leq 2\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\left\{\nabla F^i\left(q_{t,k}^i\right) - \nabla F^i\left(\bar{q}_{t,k}\right)\right\}\right\|^2 + 2\mathbb{E}\left\|\sum_{i=1}^{n}\lambda_{\mathbf{y}}^i\left\{\nabla F^i\left(\bar{q}_{t,k}\right) - \nabla F^i\left(q_\star\right)\right\}\right\|^2$$

$$\leq 2 \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\| \nabla F^{i} \left(q_{t,k}^{i} \right) - \nabla F^{i} \left(\bar{q}_{t,k} \right) \right\|^{2} + 2 \mathbb{E} \left\| \nabla F \left(\bar{q}_{t,k} \right) - \nabla F \left(q_{\star} \right) \right\|^{2}$$

$$\leq 2 \sum_{i=1}^{n} (H^{i})^{2} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\| q_{t,k}^{i} - \bar{q}_{t,k} \right\|^{2} + 4 \bar{H} \mathbb{E} \left[F \left(\bar{q}_{t,k} \right) - F \left(q_{\star} \right) \right]. \tag{25}$$

Regarding the inner product in (24), **H**3 and **H**4 show

$$- \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\langle \bar{q}_{t,k} - q_{\star}, \nabla F^{i} \left(q_{t,k}^{i} \right) \right\rangle = - \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\langle q_{t,k}^{i} - q_{\star}, \nabla F^{i} \left(q_{t,k}^{i} \right) \right\rangle + \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\langle q_{t,k}^{i} - \bar{q}_{t,k}, \nabla F^{i} \left(q_{t,k}^{i} \right) \right\rangle$$

$$\leq - \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left[F^{i} \left(q_{t,k}^{i} \right) - F^{i} \left(q_{\star} \right) \right] + \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left[F^{i} \left(q_{t,k}^{i} \right) - F^{i} \left(\bar{q}_{t,k} \right) + \frac{H^{i}}{2} \mathbb{E} \left\| q_{t,k}^{i} - \bar{q}_{t,k} \right\|^{2} \right]$$

$$\leq - \mathbb{E} \left[F \left(\bar{q}_{t,k} \right) - F \left(q_{\star} \right) \right] + \frac{1}{2} \sum_{i=1}^{n} H^{i} \lambda_{\mathbf{y}}^{i} \left\| q_{t,k}^{i} - \bar{q}_{t,k} \right\|^{2}. \tag{26}$$

Moreover, recall that the estimator $(\mathbb{1}_{S_{t+1}}(i) / \mathbb{P}(i \in S_{t+1})) \nabla F^{i}$ is unbiased, i.e.,

$$\mathbb{E} \left[\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i}(q_{t,k}^{i}) - \nabla F^{i} \left(q_{t,k}^{i} \right) \right] = 0, \tag{27}$$

and we also have

$$\frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(q_{t,k}^{i} \right) - \nabla F^{i} \left(q_{t,k}^{i} \right) = \left\{ \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(q_{\star} \right) - \nabla F^{i} \left(q_{\star} \right) \right\}$$

$$+ \left\{ \frac{\mathbb{1}_{i \in S_{t+1}}}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(\bar{q}_{t,k} \right) - \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(q_{\star} \right) - \nabla F^{i} \left(\bar{q}_{t,k} \right) + \nabla F^{i} \left(q_{\star} \right) \right\}$$

$$+ \left\{ \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(q_{t,k}^{i} \right) - \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(\bar{q}_{t,k} \right) - \nabla F^{i} \left(q_{t,k}^{i} \right) + \nabla F^{i} \left(\bar{q}_{t,k} \right) \right\}. \tag{28}$$

We now upper bound the second term of (23). Since for all random variable X, $E[\|X - EX\|^2] \leq \mathbb{E}\|X\|^2$, combining (27) and (28) gives

$$\mathbb{E} \left\| \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \left\{ \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} g_{t,k}^{i} - \nabla F^{i} \left(q_{t,k}^{i} \right) \right\} \right\|^{2} = \sum_{i=1}^{n} (\lambda_{\mathbf{y}}^{i})^{2} \mathbb{E} \left\| \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(q_{t,k}^{i} \right) - \nabla F^{i} \left(q_{t,k}^{i} \right) \right\|^{2}$$

$$+ \sum_{i=1}^{n} (\lambda_{\mathbf{y}}^{i})^{2} \mathbb{E} \left\| \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} z_{t,k}^{i} \right\|^{2}$$

$$\leq 3 \sum_{i=1}^{n} \frac{(\lambda_{\mathbf{y}}^{i})^{2} (1 - \mathbb{P}(i \in S_{t+1}))}{\mathbb{P}(i \in S_{t+1})} \left[\mathbb{E} \left\| \nabla F^{i} \left(q_{t,k}^{i} \right) - \nabla F^{i} \left(\bar{q}_{t,k} \right) \right\|^{2} + \mathbb{E} \left\| \nabla F^{i} \left(\bar{q}_{t,k} \right) - \nabla F^{i} \left(q_{\star} \right) \right\|^{2} \right]$$

$$+ 3 \mathbb{E} \left\| \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \frac{\mathbb{1}_{S_{t+1}}(i)}{\mathbb{P}(i \in S_{t+1})} \nabla F^{i} \left(q_{\star} \right) - \nabla F \left(q_{\star} \right) \right\|^{2} + \sum_{i=1}^{n} \frac{(\lambda_{\mathbf{y}}^{i})^{2}}{\mathbb{P}(i \in S_{t+1})} \mathbb{E} \left\| z_{t,k}^{i} \right\|^{2}$$

$$\leq 3 \sum_{i=1}^{n} \frac{(\lambda_{\mathbf{y}}^{i})^{2}}{\mathbb{P}(i \in S_{t+1})} \left\{ (H^{i})^{2} \mathbb{E} \left\| q_{t,k}^{i} - \bar{q}_{t,k} \right\|^{2} + 2 H^{i} \mathbb{E} \left[F^{i} \left(\bar{q}_{t,k} \right) - F^{i} \left(q_{\star} \right) \right] \right\}$$

$$+ 3 \xi_{\star}^{2} + \sum_{i=1}^{n} \frac{(\lambda_{\mathbf{y}}^{i})^{2}}{\mathbb{P}(i \in S_{t+1})} \sigma^{2}$$

$$\leq 3 \sum_{i=1}^{n} \frac{(\lambda_{\mathbf{y}}^{i})^{2}}{\mathbb{P}(i \in S_{t+1})} (H^{i})^{2} \mathbb{E} \left\| q_{t,k}^{i} - \bar{q}_{t,k} \right\|^{2} + 6 \sum_{i=1}^{n} \frac{(\lambda_{\mathbf{y}}^{i})^{2}}{\mathbb{P}(i \in S_{t+1})} H^{i} \mathbb{E} \left[F^{i} \left(\bar{q}_{t,k} \right) - F^{i} \left(q_{\star} \right) \right]$$

$$+ 3\xi_\star^2 + \sigma^2 \sum_{i=1}^n \frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}. \tag{29}$$

Plugging (25)-(26)-(29) back into (23) with $\eta \leq \eta_0$, it holds

$$\mathbb{E} \left\| \bar{q}_{t,k+1} - q_\star \right\|^2 \leq \mathbb{E} \left\| \bar{q}_{t,k} - q_\star \right\|^2 + \eta \sum_{i=1}^n \left(1 + 2\eta H^i + 3\eta \frac{\lambda_{\mathbf{y}}^i H^i}{\mathbb{P}(i \in S_{t+1})} \right) \lambda_{\mathbf{y}}^i H^i \mathbb{E} \left\| q_{t,k}^i - \bar{q}_{t,k} \right\|^2$$

$$+ \eta \left(4\eta \bar{H} + 6\eta \max_{i=1}^n \left\{ \frac{(\lambda_{\mathbf{y}}^i)^2 H^i}{\mathbb{P}(i \in S_{t+1})} \right\} - 2 \right) \mathbb{E} \left[F(\bar{q}_{t,k}) - F(q_\star) \right] + 3\eta^2 \xi_\star^2 + \eta^2 \sigma^2 \sum_{i=1}^n \frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}$$

$$\leq \mathbb{E} \left\| \bar{q}_{t,k} - q_\star \right\|^2 + 2H\eta \sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E} \left\| q_{t,k}^i - \bar{q}_{t,k} \right\|^2 - \eta \mathbb{E} \left[F(\bar{q}_{t,k}) - F(q_\star) \right] + 3\eta^2 \xi_\star^2 + \eta^2 \sigma^2 \sum_{i=1}^n \frac{(\lambda_{\mathbf{y}}^i)^2}{\mathbb{P}(i \in S_{t+1})}.$$

$\square$

Lemma B.2. *Assume H2-H3-H4-H5-H6, and for all $t \in [T]$, suppose that $S_t = [n]$. We consider $\eta \in (0, 2/ \sum_{i=1}^n \lambda_{\mathbf{y}}^i H^i]$. Then, for any $t \in \{0, \ldots, T-1\}$, $k \in \{0, \ldots, K-1\}$, we have*

$$\sum_{i=1}^n \lambda_{\mathbf{y}}^i \mathbb{E} \left\| q_{t,k}^i - \bar{q}_{t,k} \right\|^2 \leq 6K\eta^2 \left(\sigma^2 + \xi^2 + K\zeta^2 \right).$$

Proof. Let $\epsilon > 0$, for any $i, i' \in [n]$ and any $k \in [K]$,

$$\mathbb{E} \left\| q_{t,k}^i - q_{t,k}^{i'} \right\|^2 - 2\eta^2(\sigma^2 + \xi^2) = \mathbb{E} \left\| q_{t,k-1}^i - q_{t,k-1}^{i'} - \eta \left(g_{t,k-1}^i - g_{t,k-1}^{i'} \right) \right\|^2 - 2\eta^2(\sigma^2 + \xi^2)$$

$$= \mathbb{E} \left\| q_{t,k-1}^i - q_{t,k-1}^{i'} - \eta \left(\nabla F^i(q_{t,k-1}^i) - \nabla F^{i'}(q_{t,k-1}^{i'}) \right) \right\|^2$$

$$+ \eta^2 \mathbb{E} \left\| \left(\nabla F^i(q_{t,k-1}^i) - \nabla F^{i'}(q_{t,k-1}^{i'}) \right) - \left(g_{t,k-1}^i - g_{t,k-1}^{i'} \right) \right\|^2 - 2\eta^2(\sigma^2 + \xi^2)$$

$$\leq \mathbb{E} \left\| q_{t-1}^i - q_{t-1}^{i'} - \eta \left(\nabla F(q_{t-1}^i) - \nabla F(q_{t-1}^{i'}) \right) + \eta \left(\nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i) - \nabla F(q_{t-1}^{i'}) + \nabla F^{i'}(q_{t-1}^{i'}) \right) \right\|^2$$

$$\leq \left(1 + \frac{1}{\epsilon} \right) \mathbb{E} \left\| q_{t-1}^i - q_{t-1}^{i'} - \eta \left(\nabla F(q_{t-1}^i) - \nabla F(q_{t-1}^{i'}) \right) \right\|^2$$

$$+ (1+\epsilon)\eta^2 \mathbb{E} \left\| \nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i) - \nabla F(q_{t-1}^{i'}) + \nabla F^{i'}(q_{t-1}^{i'}) \right\|^2$$

$$\leq \left(1 + \frac{1}{\epsilon} \right) \mathbb{E} \left\| q_{t-1}^i - q_{t-1}^{i'} \right\|^2 + (1+\epsilon)\eta^2 \mathbb{E} \left\| \nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i) \right\|^2$$

$$+ (1+\epsilon)\eta^2 \mathbb{E} \left\| \nabla F(q_{t-1}^{i'}) - \nabla F^{i'}(q_{t-1}^{i'}) \right\|^2$$

$$- 2(1+\epsilon)\eta^2 \mathbb{E} \left\langle \nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i), \nabla F(q_{t-1}^{i'}) - \nabla F^{i'}(q_{t-1}^{i'}) \right\rangle. \tag{30}$$

The third inequality is implied by the co-coercivity: $\eta \in (0, 2/ \sum_{i=1}^n \lambda_{\mathbf{y}}^i H^i]$, $\forall (q, \tilde{q}) \in \mathbb{R}^2$,

$$\left\| q - \tilde{q} - \eta \left(\nabla F(q) - \nabla F(\tilde{q}) \right) \right\|^2$$

$$= \left\| \tilde{q} - q \right\|^2 - \eta \left[2 \left\langle q - \tilde{q}, \nabla F(q) - \nabla F(\tilde{q}) \right\rangle + \eta \left\| \nabla F(q) - \nabla F(\tilde{q}) \right\|^2 \right] \leq \left\| \tilde{q} - q \right\|^2,$$

and we also have

$$\sum_{i=1}^n \sum_{i'=1}^n \lambda_{\mathbf{y}}^i \lambda_{\mathbf{y}}^{i'} \mathbb{E} \left\langle \nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i), \nabla F(q_{t-1}^{i'}) - \nabla F^{i'}(q_{t-1}^{i'}) \right\rangle$$

$$= \mathbb{E} \left\| \sum_{i=1}^n \lambda_{\mathbf{y}}^i \left(\nabla F(q_{t-1}^i) - \nabla F^i(q_{t-1}^i) \right) \right\|^2 \geq 0.$$

Therefore, summing (30) gives that

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} \lambda_{\mathbf{y}}^{i} \lambda_{\mathbf{y}}^{i'} \mathbb{E} \big\| q_{t,k}^{i} - q_{t,k}^{i'} \big\|^2 \leq \left(1 + \frac{1}{\epsilon}\right) \sum_{i=1}^{n} \sum_{i'=1}^{n} \lambda_{\mathbf{y}}^{i} \lambda_{\mathbf{y}}^{i'} \mathbb{E} \big\| q_{t-1}^{i} - q_{t-1}^{n} \big\|^2 + 2\eta^2 \left(\sigma^2 + \xi^2 + (1 + \epsilon)\zeta^2\right).$$

Set $\epsilon = K - 1$, since for any $i, i' \in [n]$, $x_{t,0}^{i} = x_{t,0}^{i'}$, we get

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} \lambda_{\mathbf{y}}^{i} \lambda_{\mathbf{y}}^{i'} \mathbb{E} \big\| q_{t,k}^{i} - q_{t,k}^{i'} \big\|^2 \leq 2\eta^2 \left(\sigma^2 + \xi^2 + (1 + \epsilon)\zeta^2\right) \sum_{k'=0}^{K-1} \left(1 + \frac{1}{\epsilon}\right)^{k'}$$

$$\leq 6K\eta^2 \left(\sigma^2 + \xi^2 + (1 + \epsilon)\zeta^2\right).$$

Since $\sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} = 1$, the Jensen's inequality yields that $\sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \big\| q_{t,k}^{i} - \bar{q}_{t,k} \big\|^2 \leq \sum_{i=1}^{n} \sum_{i'=1}^{n} \lambda_{\mathbf{y}}^{i} \lambda_{\mathbf{y}}^{i'} \mathbb{E} \big\| q_{t,k}^{i} - q_{t,k}^{i'} \big\|^2$, which concludes the proof. $\qquad \square$

In addition, with the previous notations consider the stepsize

$$\eta_{\star} = \min \left\{ \eta_0, \left(\frac{\mathbb{E} \left\| \bar{q}_0 - q_{\star} \right\|^2}{14HK^2T \left[\sigma^2 + K\zeta^2\right]}\right)^{1/3} \right\}, \tag{31}$$

and define the average parameter

$$\hat{q}_T = \frac{1}{T} \sum_{t=0}^{T-1} \left\{ \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \left[\frac{1}{K} \sum_{k=0}^{K-1} q_{t,k}^{i} \right] \right\}. \tag{32}$$

Theorem B.3. *Assume* **H2**-**H3**-**H4**-**H5**-**H6**. *We consider* $\eta \in (0, \eta_0]$ *with* $S_t = [n]$, *for all* $t \in [T]$. *Then, for any* $t \in \{0, \ldots, T-1\}$, $k \in \{0, \ldots, K-1\}$, *we have*

$$\mathbb{E} F(\hat{q}_T) - F(q_{\star}) \leq \frac{\mathbb{E} \left\| \bar{q}_0 - q_{\star} \right\|^2}{\eta KT} + 2\eta^2 \left[6H(K\zeta)^2 + \sigma^2 \left(6HK + \max_{i \in [n]} \lambda_{\mathbf{y}}^{i}\right)\right],$$

where $\eta_0, \hat{q}_T$ *are given in* (22) *and* (32). *Moreover, for* $\eta = \eta_{\star}$ *defined in* (31) *and* $H \geq K^{-1} \max_{i \in [n]} \lambda_{\mathbf{y}}^{i}$, *it follows*

$$\mathbb{E} F(\hat{q}_T) - F(q_{\star}) \leq \frac{\mathbb{E} \left\| \bar{q}_0 - q_{\star} \right\|^2}{\eta_0 KT} + \frac{5(\mathbb{E} \left\| \bar{q}_0 - q_{\star} \right\|^2)^{2/3} \left[H\left(\sigma^2 + K\zeta^2\right)\right]^{1/3}}{(KT^2)^{1/3}}.$$

Proof. For any $\eta \leq \eta_0$, using Lemma B.1 we have

$$\mathbb{E} \left[F(\bar{q}_{t,k})\right] - F(q_{\star}) \leq \frac{1}{\eta} \mathbb{E} \left\| \bar{q}_{t,k} - q_{\star} \right\|^2 - \frac{1}{\eta} \mathbb{E} \left\| \bar{q}_{t,k+1} - q_{\star} \right\|^2 + 2H \sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\| \bar{q}_{t,k} - q_{t,k}^{i} \right\|^2 + 2\eta^2 \sigma^2 \max_{i \in [n]} \{\lambda_{\mathbf{y}}^{i}\}. \tag{33}$$

Moreover, by Lemma B.2 it follows that

$$\sum_{i=1}^{n} \lambda_{\mathbf{y}}^{i} \mathbb{E} \left\| q_{t,k}^{i} - \bar{q}_{t,k} \right\|^2 \leq 6K\eta^2 (\sigma^2 + K\zeta^2). \tag{34}$$

Combining (33) and (34), we obtain

$$\mathbb{E}[F(\bar{q}_{t,k}) - F(q_{\star})] \leq \frac{1}{\eta} \mathbb{E} \left\| \bar{q}_{t,k} - q_{\star} \right\|^2 - \frac{1}{\eta} \mathbb{E} \left\| \bar{q}_{t,k+1} - q_{\star} \right\|^2 + 12H(K\eta\zeta)^2 + 2(\eta\sigma)^2 \left(6HK + \max_{i \in [n]} \{\lambda_{\mathbf{y}}^{i}\}\right).$$

Moreover, telescoping proves that

$$\sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \left[\mathbb{E} \left\| \bar{q}_{t,k} - q_{\star} \right\|^2 - \mathbb{E} \left\| \bar{q}_{t,k+1} - q_{\star} \right\|^2\right] \leq \mathbb{E} \left\| \bar{q}_0 - q_{\star} \right\|^2.$$

Therefore, the convexity **H3** gives that

$$\mathbb{E} \left[F\left(\frac{1}{KT} \sum_{t=1}^{KT} \bar{q}_{t,k}\right) - F(q_{\star})\right] \leq \frac{1}{KT} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \mathbb{E} \left[F(\bar{q}_{t,k}) - F(q_{\star})\right]$$

$$\leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta KT} + 2\eta^2 \left[6H(K\zeta)^2 + \sigma^2 \left(6HK + \max_{i\in[n]}\{\lambda_{\mathbf{y}}^i\}\right)\right].$$

Finally, the choice of η provided in (31) ensures that

$$\mathbb{E}\left[F\left(\hat{q}_T\right)\right] - F(q_\star) \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta_0 KT} + \frac{5(\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2)^{2/3}\left[H\left(\sigma^2 + K\zeta^2\right)\right]^{1/3}}{(KT^2)^{1/3}}.$$

$\square$

Now, we denote $\alpha \in (0,1)$ the confidence level, and consider the functions defined for $t \in \{0,\dots,T\}, k \in \{0,\dots,K\}$ by

$$F = S_{\alpha,\widehat{\mu}_{\mathbf{y}}}^\gamma, \qquad\qquad\qquad F^i = S_{\alpha,\widehat{\mu}_{\mathbf{y}}^i}^\gamma.$$

Denote by $q_\star$ the minimizer of $S_{\alpha,\widehat{\mu}_{\mathbf{y}}}^\gamma = \sum_{i=1}^n \lambda_{\mathbf{y}}^i S_{\alpha,\widehat{\mu}_{\mathbf{y}}^i}^\gamma$, which always exists. with the following variables:

$$H^i = \frac{1}{\gamma}, \qquad\qquad\qquad \zeta = \max_{i\in[n]}\|\nabla\, \mathrm{S}_\alpha^{i,\gamma} - \nabla\, \mathrm{S}_\alpha^\gamma\|_\infty^{1/2}. \tag{35}$$

Corollary B.4. *Let* $\gamma \in (0, (\max_{i\in[n]}\lambda_{\mathbf{y}}^i)^{-1}K]$ *and consider the stepsize* $\eta_0 = \gamma/10$. *Then, for any* $t \in \{0,\dots,T-1\}$, $k \in \{0,\dots,K-1\}$, *we have*

$$\mathbb{E}S_{\alpha,\widehat{\mu}_{\mathbf{y}}}^\gamma\left(\hat{q}_T\right) - S_{\alpha,\widehat{\mu}_{\mathbf{y}}}^\gamma(q_\star) \leq \frac{\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2}{\eta_0 KT} + \frac{5(\sigma^2 + K\zeta^2)^{1/3}(\mathbb{E}\left\|\bar{q}_0 - q_\star\right\|^2)^{2/3}}{(\gamma KT^2)^{1/3}},$$

where $\hat{q}_T$ *is provided in* (32).

Proof. Since **H2**-**H3**-**H4**-**H5**-**H6** are satisfied with $\{H^i\}_{i\in[n]}, \zeta$ provided in (35), applying Theorem B.3 concludes the proof. $\square$

C. Theoretical Coverage Guarantee

C.1. General coverage guarantee

Consider an increasing sequence $\{v_k\}_{k\in[N+1]} \in (\mathbb{R}\cup\{+\infty\})^{N+1}$ and $\{p_k\}_{k\in[N+1]} \in \Delta_{N+1}$. For any $\alpha \in [0,1]$, recall that

$$Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N+1} p_k\delta_{v_k}\right) = \inf\left\{t \in [-\infty,\infty]\colon \mathbb{P}\left(V \leq t\right) \geq 1-\alpha, \text{ where } V \sim \textstyle\sum_{k=1}^{N+1} p_k\delta_{v_k}\right\}.$$

Lemma C.1. *Let* $\{v_\ell\}_{\ell\in[N+1]}$ *be an increasing sequence and* $\{p_\ell\}_{\ell\in[N+1]} \in \Delta_{N+1}$. *If* $V \sim \sum_{l=1}^{N+1} p_l\delta_{v_l}$, *then, for all* $\alpha \in [0,1)$, *we have*

$$1-\alpha \leq \mathbb{P}\left(V \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N+1} p_k\delta_{v_k}\right)\right) < 1-\alpha + \max_{k=1}^{N+1}\{p_k\}.$$

Proof. Fix $\alpha \in [0,1)$, and by convention set $\sum_{k=1}^0 p_k = 0$. There exists $k \in [N+1]$, such that $1-\alpha \in (\sum_{l=1}^{k-1} p_l, \sum_{l=1}^k p_l]$, hence $Q_{1-\alpha}(\sum_{l=1}^{N+1} p_l\delta_{v_l}) = v_k$. This last identity implies that

$$1-\alpha \leq \mathbb{P}\left(V \leq Q_{1-\alpha}\left(\textstyle\sum_{l=1}^{N+1} p_l\delta_{v_l}\right)\right) = \textstyle\sum_{l=1}^k p_l < 1-\alpha + \max_{k=1}^{N+1}\{p_k\}.$$

$\square$

Denote $\{(X_k,Y_k)\}_{k\in[N+1]}$ a set of pairwise independent random variables and for $k \in [N+1]$, denote $Z_k = (X_k,Y_k)$, $V_k = V(X_k,Y_k)$. Let $\mathfrak{S}_{N+1}$ be the set of all permutations of $[N+1]$ and consider $\mathfrak{S}_{N+1}^k = \{\sigma \in \mathfrak{S}_{N+1}\colon \sigma(N+1) = k\}$.

Moreover, write f the joint density of $\{Z_k\}_{k \in [N+1]}$, and for all $k \in [N+1]$ define

$$p_k^{z_{1:N+1}} = \begin{cases} \frac{1}{N+1} & \text{if } \sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) = 0 \\ \frac{\sum_{\sigma \in \mathfrak{S}_{N+1}^k} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})} & \text{otherwise} \end{cases}. \tag{36}$$

Lemma C.2. *For any $\alpha \in [0,1)$, we have*

$$\int \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}$$

$$= \int \left[\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} \right] \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}.$$

Proof. First, let's show the invariance of $\sigma \in \mathfrak{S}_{N+1} \mapsto Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_{\sigma(k)}^{z_{\sigma(1):\sigma(N+1)}} \delta_{v_{\sigma(k)}}\right) \in \mathbb{R}$. For that, fix $\tilde{\sigma} \in \mathfrak{S}_{N+1}$. The invariance is immediate when $\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) = 0$. Therefore, assume that $\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \neq 0$. We get

$$\sum_{k=1}^{N+1} p_{\tilde{\sigma}(k)}^{z_{\tilde{\sigma}(1):\tilde{\sigma}(N+1)}} \delta_{v_{\tilde{\sigma}(k)}} = \sum_{k=1}^{N+1} \frac{\sum_{\sigma \in \mathfrak{S}_{N+1}^{\tilde{\sigma}(k)}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})} \delta_{v_{\tilde{\sigma}(k)}}$$

$$= \sum_{k=1}^{N+1} \frac{\sum_{\sigma \in \mathfrak{S}_{N+1}^{k}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})} \delta_{v_k} = \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}.$$

Moreover, we can write

$$\int \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}$$

$$= \sum_{\sigma \in \mathfrak{S}_{N+1}} \int \mathbb{1}_{v_{\sigma(N+1)} \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\sigma(\bar{k})}^{z_{\sigma(1):\sigma(N+1)}} \delta_{v_{\sigma(\bar{k})}}\right)} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \frac{\mathrm{d}z_{\sigma(1)} \cdots \mathrm{d}z_{\sigma(N+1)}}{(N+1)!}$$

$$= \sum_{k=1}^{N+1} \sum_{\sigma \in \mathfrak{S}_{N+1}^k} \int \mathbb{1}_{v_{\sigma(N+1)} \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\sigma(\bar{k})}^{z_{\sigma(1):\sigma(N+1)}} \delta_{v_{\sigma(\bar{k})}}\right)} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \frac{\mathrm{d}z_{\sigma(1)} \cdots \mathrm{d}z_{\sigma(N+1)}}{(N+1)!}$$

$$= \sum_{k=1}^{N+1} \int \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} \left[\sum_{\sigma \in \mathfrak{S}_{N+1}^k} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}$$

$$= \sum_{k=1}^{N+1} \int \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} p_k^{z_{1:N+1}} \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}.$$

$\square$

Given $\mathrm{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ define

$$D_N^{\mathrm{z}} = (z_1, \ldots, z_N, \mathrm{z}), \qquad \qquad \mu_{\mathbf{y}}^N = p_{N+1}^{D_N^{\mathrm{z}}} \delta_1 + \sum_{k=1}^{N} p_k^{D_N^{\mathrm{z}}} \delta_{V_k},$$

and consider the prediction set given by

$$\mathcal{C}_{\alpha, \mu^N}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\mu_{\mathbf{y}}^N) \right\}.$$

Theorem C.3. *Assume there are no ties between $\{V_k\}_{k \in [N+1]}$ almost surely. Then, for any $\alpha \in [0,1)$, we have*

$$1 - \alpha \leq \mathbb{P}\left(Y_{N+1} \in \mathcal{C}_{\alpha, \mu^N}(X_{N+1})\right) \leq 1 - \alpha + \mathbb{E}\left[\max_{k=1}^{N+1} \{p_k^{Z_{1:N+1}}\} \right],$$

where $p_k^{Z_{1:N+1}}$ is defined in (36).

Proof. Let α be in $[0,1)$ and for any $(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}$, denote $z_k = (x_k, y_k)$, $v_k = V(x_k, y_k)$. First, we can write

$$\mathbb{P}\left(Y_{N+1} \in \mathcal{C}_{\alpha,\mu^N}(X_{N+1})\right) = \mathbb{P}\left(Y_{N+1} \in \{\mathbf{y} \in \mathcal{Y} \colon V(X_{N+1}, \mathbf{y}) \leq Q_{1-\alpha}(\mu_{\mathbf{y}}^N)\}\right)$$

$$= \mathbb{P}\left(V(X_{N+1}, Y_{N+1}) \leq Q_{1-\alpha}(\mu_{Y_{N+1}}^N)\right)$$

$$\overset{(\star)}{=} \mathbb{P}\left(V(X_{N+1}, Y_{N+1}) \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{Z_{1:N+1}} \delta_{V_k}\right)\right)$$

$$= \int_{(\mathcal{X} \times \mathcal{Y})^{N+1}} \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{V_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}.$$

Where $(\star)$ holds since

$$\left(V_{N+1} \leq Q_{1-\alpha}(\mu_{Y_{N+1}}^N)\right)$$

$$\Longleftrightarrow \left(p_{N+1}^{D_N^{(X_{N+1}, Y_{N+1})}} \delta_{1 \leq V_{N+1}} + \sum_{k=1}^{N} p_k^{D_N^{(X_{N+1}, Y_{N+1})}} \delta_{V_k \leq V_{N+1}} \leq \alpha\right)$$

$$\Longleftrightarrow \left(p_{N+1}^{D_N^{(X_{N+1}, Y_{N+1})}} \delta_{V_{N+1} \leq V_{N+1}} + \sum_{k=1}^{N} p_k^{D_N^{(X_{N+1}, Y_{N+1})}} \delta_{V_k \leq V_{N+1}} \leq \alpha\right)$$

$$\Longleftrightarrow \left(V(X_{N+1}, Y_{N+1}) \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{Z_{1:N+1}} \delta_{V_k}\right)\right).$$

Define the set $E \subset (\mathcal{X} \times \mathcal{Y})^{N+1}$ of points such that the non-conformity scores are pairwise distinct:

$$E = \{(z_1, \ldots, z_{N+1}) \in (\mathcal{X} \times \mathcal{Y})^{N+1} \colon \textstyle\prod_{k < \ell}(v(x_k, y_k) - v(x_\ell, y_\ell)) \neq 0\},$$

$$E^c = (\mathcal{X} \times \mathcal{Y})^{N+1} \setminus E,$$

$$F = \{(z_1, \ldots, z_{N+1}) \in (\mathcal{X} \times \mathcal{Y})^{N+1} \colon \textstyle\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)}) \neq 0\}$$

In addition, combining Lemma C.2 with the no-tie assumption on $\{V_k\}_{k \in [N+1]}$ gives that

$$\int \mathbb{1}_{v_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}\right)} f(z_1, \ldots, z_{N+1}) \mathrm{d}z_1 \cdots \mathrm{d}z_{N+1} \tag{37}$$

$$= \int_{E \cap F} \left[\sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)}\right] \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}.$$

Consider the random variable $V \sim \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \delta_{v_k}$, we have

$$\mathbb{P}\left(V \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)\right) = \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)}.$$

Therefore, applying Lemma C.1 on $(z_1, \ldots, z_{N+1}) \in E \cap F$ implies that

$$1 - \alpha \leq \sum_{k=1}^{N+1} p_k^{z_{1:N+1}} \mathbb{1}_{v_k \leq Q_{1-\alpha}\left(\sum_{\bar{k}=1}^{N+1} p_{\bar{k}}^{z_{1:N+1}} \delta_{v_{\bar{k}}}\right)} \leq 1 - \alpha + \max_{k=1}^{N+1}\{p_k^{z_{1:N+1}}\}. \tag{38}$$

Lastly, using that

$$\int_{E \cap F} \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!}$$

$$= \int_E \left[\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\sigma(1)}, \ldots, z_{\sigma(N+1)})\right] \frac{\mathrm{d}z_1 \cdots \mathrm{d}z_{N+1}}{(N+1)!} = 1, \tag{39}$$

and combining the bounds (38)-(39) with (37) yields the result. $\square$

C.2. Proof of Theorem 2.1

First, recall that

$$\mathcal{I} = \{(\star, N^\star + 1)\} \cup \left\{(i,k) \colon i \in [n], k \in [N^i]\right\}.$$

For any $\{(\mathrm{x}_k^i, \mathrm{y}_k^i)\}_{(i,k)\in\mathcal{I}} \in (\mathcal{X} \times \mathcal{Y})^{N+1}$, we define the set

$$D_N^{(\mathrm{x}_{N^\star+1}^\star, \mathrm{y}_{N^\star+1}^\star)} = \{(\mathrm{x}_k^i, \mathrm{y}_k^i) \colon i \in [n], k \in [N^i]\} \cup \{(\mathrm{x}_{N^\star+1}^\star, \mathrm{y}_{N^\star+1}^\star)\}.$$

We consider a bijection (ϕ, φ) between the set $[N+1]$ and $\mathcal{I}$. This bijection is defined for any $k \in [N]$ as follows:

$$(\phi(k), \varphi(k)) = \begin{cases} (j, \ell) & \text{if } 1 \leq \ell := k - \sum_{i=1}^{j-1} N^i \leq \sum_{i=1}^{j} N^i \\ (\star, N^\star + 1) & \text{otherwise} \end{cases}.$$

Recall that $\forall i \in [n]$ and $y \in \mathcal{Y}$, the likelihood ratio is given by

$$w_y^i = \frac{P_Y^i(y)}{P_Y^{\mathrm{cal}}(y)},$$

and for all $(i,k) \in \mathcal{I}$ we write

$$\mathfrak{P}_k^i = \left\{\sigma \in \mathfrak{S}_{N+1} \colon \phi(\sigma(N+1)) = i, \varphi(\sigma(N+1)) = k\right\},$$

$$W_k^i(D_N^{(\mathrm{x}_{N^\star+1}^\star, \mathrm{y}_{N^\star+1}^\star)}) = w_{\mathrm{y}_k^i}^\star \sum_{\sigma \in \mathfrak{P}_k^i} \prod_{\ell=1}^{N} w_{\mathrm{y}_{\varphi(\ell)}^{\phi(\ell)}}^{\phi(\ell)}. \tag{40}$$

Given the set of points $D_N^{(\mathbf{x},\mathbf{y})}$, for all $(i,k) \in \mathcal{I}$ define

$$p_{i,k}^\star = \frac{W_k^i(D_N^{(\mathbf{x},\mathbf{y})})}{\sum_{\ell=1}^{N+1} W_{\varphi(\ell)}^{\phi(\ell)}(D_N^{(\mathbf{x},\mathbf{y})})}. \tag{41}$$

Finally, define the probability measure and the prediction set given by

$$\mu_{\mathbf{y}}^\star = p_{\star, N^\star+1}^\star \delta_1 + \sum_{i=1}^{n} \sum_{k=1}^{N^i} p_{i,k}^\star \delta_{V_k^i},$$

$$\mathcal{C}_{\alpha,\mu^\star}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\mu_{\mathbf{y}}^\star)\right\}.$$

Theorem C.4. *If H1 holds, then for any $\alpha \in [0,1)$, we have*

$$1 - \alpha \leq \mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\mu^\star}(X_{N^\star+1}^\star)\right) \leq 1 - \alpha + \mathbb{E}\left[\max_{(i,k)\in\mathcal{I}} \{p_{k,i}^\star\}\right],$$

where $p_{i,k}^\star$ is defined in (41).

Proof. By independence, the joint density f of $\{(X_\ell^j, Y_\ell^j) \colon (j,\ell) \in \mathcal{I}\}$ with respect to $(P_{X|Y} \times P_Y^{\mathrm{cal}})^{\otimes(N+1)}$ is given for $\{(x_k^i, y_k^i) \colon (i,k) \in \mathcal{I}\} \in (\mathcal{X} \times \mathcal{Y})^{N+1}$ by

$$f\left((x_1^1, y_1^1), \ldots, (x_{N^1}^1, y_{N^1}^1), \ldots, (x_1^n, y_1^n), \ldots, (x_{N^n}^n, y_{N^n}^n), (x_{N^\star+1}^\star, y_{N^\star+1}^\star)\right) = w_{y_{N^\star+1}^\star}^\star \prod_{j=1}^{n} \prod_{\ell=1}^{N^j} w_{y_\ell^j}^j.$$

Using the definition of $p_{i,k}^\star$ (41), for all $(i,k) \in \mathcal{I}$ we have

$$\begin{aligned} p_{i,k}^\star &= \frac{W_k^i(D_{N+1})}{\sum_{\ell=1}^{N+1} W_{\varphi(\ell)}^{\phi(\ell)}(D_{N+1})} \\ &= \frac{\sum_{\sigma \in \mathfrak{P}_k^i} f(z_{\varphi\circ\sigma(1)}^{\phi\circ\sigma(1)}, \ldots, z_{\varphi\circ\sigma(N+1)}^{\phi\circ\sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\varphi\circ\sigma(1)}^{\phi\circ\sigma(1)}, \ldots, z_{\varphi\circ\sigma(N+1)}^{\phi\circ\sigma(N+1)})} \end{aligned}$$

22

$$= \frac{\sum_{\sigma \in \mathfrak{S}_{N+1}: \, \sigma(N+1) = (\phi, \varphi)^{-1}(i,k)} f(z_{\varphi \circ \sigma(1)}^{\phi \circ \sigma(1)}, \ldots, z_{\varphi \circ \sigma(N+1)}^{\phi \circ \sigma(N+1)})}{\sum_{\sigma \in \mathfrak{S}_{N+1}} f(z_{\varphi \circ \sigma(1)}^{\phi \circ \sigma(1)}, \ldots, z_{\varphi \circ \sigma(N+1)}^{\phi \circ \sigma(N+1)})}.$$

Therefore, applying Theorem C.3 concludes the proof. $\qquad \square$

C.3. Proof of Lemma 2.3 and Equation (2)

In this section, we consider a probability measure P_Y^{cal} dominating $P_Y^\star$ and suppose that the likelihood ratios are given by $w_y^\star = P_Y^\star(y)/P_Y^{\mathrm{cal}}(y)$. Let $\{\widehat{w}_y^\star\}_{y \in \mathcal{Y}}$ be fixed, and $\forall y \in \mathcal{Y}$, define $\nu_y = [\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})]^{-1} \widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)$. Denote $\widehat{Y}_{N^\star+1}$ a multinomial random variable of parameter ν independent of the calibration dataset $\{(X_k^i, Y_k^i) \colon k \in [N^i]\}_{i \in [n]}$ and write $\mathcal{P}(\mathcal{Y})$ the partition of the set $\mathcal{Y}$. Moreover, denote $\widehat{P}_Y^\star$ the probability distribution of $\widehat{Y}_{N^\star+1}$, and consider $\widehat{X}_{N^\star+1}$ such that $(\widehat{X}_{N^\star+1}, \widehat{Y}_{N^\star+1}) \sim P_{X|Y} \times \widehat{P}_Y^\star$.

Lemma C.5. *For any prediction set $\mathcal{C} \colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ independent of $(X_{N^\star+1}^\star, Y_{N^\star+1}^\star)$ and $(\widehat{X}_{N^\star+1}^\star, \widehat{Y}_{N^\star+1}^\star)$, we have*

$$\left| \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star)) - \mathbb{P}(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star)) \right| \leq \frac{1}{2} \sum_{y \in \mathcal{Y}} \left| P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right|.$$

Proof. Developing the left-hand side as follows, we get

$$\left| \mathbb{P} \left(Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star) \right) - \mathbb{P} \left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star) \right) \right|$$

$$= \left| \mathbb{E} \left[\mathbb{1}_{Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star)} \right] - \mathbb{E} \left[\mathbb{1}_{\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star)} \right] \right|$$

$$= \left| \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y) \left(w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right) \int \mathbb{E} \mathbb{1}_{y \in \mathcal{C}(x)} \mathrm{d} P_{X|Y=y}(x) \right|$$

$$\leq \frac{1}{2} \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y) \left| w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right|.$$

Finally, using that $P_Y^{\mathrm{cal}}(y) w_y^\star = P_Y^\star(y)$ concludes the proof. $\qquad \square$

Remark C.6. If for some probability atoms $\{m_i\}_{i \in [n]} \in \Delta_n$, we know good approximations $\widehat{w}_y^\star$ of the likelihood ratios $[(P_Y^{\mathrm{cal}})^{-1}(\sum_{i=1}^n m_i P_Y^i)](y)$. Then, Lemma C.5 implies the following result:

$$\left| \mathbb{P} \left(Y_{N^\star+1}^\star \in \mathcal{C}(X_{N^\star+1}^\star) \right) - \mathbb{P} \left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}(\widehat{X}_{N^\star+1}^\star) \right) \right|$$

$$\leq \mathrm{d}_{\mathrm{TV}} \left(P_Y^\star, \sum_{i=1}^n m_i P_Y^i \right) + \frac{1}{2} \sum_{y \in \mathcal{Y}} \left| \left(\sum_{i=1}^n m_i P_Y^i \right)(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right|.$$

Lemma C.7. *If $|\mathcal{Y}| \geq 2$ and $\mathsf{M} \in \mathbb{N}^\star$, then we have*

$$|\mathcal{Y}| \exp \left(-\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\} \right) \wedge 1 \leq \sqrt{\frac{2 \log |\mathcal{Y}|}{(\log 2)\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}}}.$$

Proof. Introduce the set $E = \{2 \log |\mathcal{Y}| \leq \mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}\}$, we obtain

$$|\mathcal{Y}| \exp \left(-\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\} \right) \wedge 1 \leq \mathbb{1}_E \exp \left(-\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}/2 \right) + \mathbb{1}_{E^c}. \tag{42}$$

We also have that for all $x \geq 0$, that $\mathrm{e}^{-x} \leq 1/\sqrt{x}$. Using this inequality on the first right-side term implies that

$$\exp \left(-\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}/2 \right) \leq \sqrt{\frac{2}{\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}}} \leq \sqrt{\frac{\log |\mathcal{Y}|}{\log 2}} \sqrt{\frac{2}{\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}}}.$$

Moreover, remark that

$$\mathbb{1}_{E^c} \leq \mathbb{1}_{E^c} \sqrt{\frac{2 \log |\mathcal{Y}|}{\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}}}.$$

Finally, plugging the two previous inequalities in (42) concludes the proof. $\qquad \square$

Recall that M_y^i denotes the number of training data on agent i associated to label $y \in \mathcal{Y}$. Consider the total number of local data $\mathsf{M}^\star = \sum_{y \in \mathcal{Y}} \mathsf{M}_y^i$, the number of training data with label y is given by $\mathsf{M}_y = \sum_{i=1}^n \mathsf{M}_y^i$, and the total number of samples on all agents is written by $\mathsf{M} = \sum_{y \in \mathcal{Y}} \mathsf{M}_y$. Recall that the likelihood ratios and the weights are given for any labels $(y, \mathbf{y}) \in \mathcal{Y}^2$ by

$$\widehat{w}_y^\star = \begin{cases} \frac{\mathsf{M} \mathsf{M}_y^\star}{\mathsf{M}^\star \mathsf{M}_y} & \text{if } \mathsf{M}_y \geq 1 \\ 0 & \text{otherwise} \end{cases}, \qquad \widehat{p}_{y,\mathbf{y}}^\star = \frac{(\mathsf{M}_y^\star/\mathsf{M}_y) \cdot \mathbb{1}_{\mathsf{M}_y \geq 1}}{\mathsf{M}^\star + (\mathsf{M}_\mathbf{y}^\star/\mathsf{M}_\mathbf{y}) \cdot \mathbb{1}_{\mathsf{M}_\mathbf{y} \geq 1}}.$$

For any $y \in \mathcal{Y}$, we also consider $\nu \in \Delta_{|\mathcal{Y}|}^{\mathbb{Q}} = \{p' \in \mathbb{Q}_+^{|\mathcal{Y}|} : \sum_{y \in \mathcal{Y}} p_y' = 1\}$ defined by

$$\nu_y = \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})}.$$

For any parameter $p \in \Delta_{|\mathcal{Y}|}^{\mathbb{Q}}$, denote M_p a multinomial random variable independent of the training/calibration datasets, and define $\widehat{Y}_{N^\star+1}^\star = M_\nu$. For any set A in the partition of $\mathcal{Y}$, we have $M_\nu^{-1}(A) = \cup_{p \in \Delta_{|\mathcal{Y}|}^{\mathbb{Q}}} \{\nu^{-1}(\{p\}) \cap M_p^{-1}(A)\}$. Therefore, $\widehat{Y}_{N^\star+1}^\star$ is a valid random variable. Given the target coverage level $1 - \alpha$, recall that the prediction set is defined for any $\mathbf{x} \in \mathcal{X}$ by

$$\widehat{\mu}_\mathbf{y}^{\mathrm{MLE}} = \widehat{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}^i} \widehat{p}_{Y_k^i,\mathbf{y}}^\star \delta_{V_k^i},$$

$$\mathcal{C}_{\alpha,\widehat{\mu}^{\mathrm{MLE}}}(\mathbf{x}) = \left\{ \mathbf{y} : V(\mathbf{x},\mathbf{y}) \leq Q_{1-\alpha}\left(\widehat{\mu}_\mathbf{y}^{\mathrm{MLE}}\right) \right\}.$$

Since the considered likelihood ratios are now depending on the training dataset, it is no longer possible to apply Lemma C.5. However, conditioning by the training dataset, a similar reasoning shows that

$$\left| \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\mathrm{MLE}}}(X_{N^\star+1}^\star)) - \mathbb{P}(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\mathrm{MLE}}}(\widehat{X}_{N^\star+1}^\star)) \right| \leq \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{E} \left| P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right|. \qquad (43)$$

By utilizing the following lemma combined with (43), we can control the difference between the probabilities of the events $Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\mathrm{MLE}}}(X_{N^\star+1}^\star)$ and $\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\mathrm{MLE}}}(\widehat{X}_{N^\star+1}^\star)$.

Theorem C.8. *For any $\alpha \in (0,1)$, we have*

$$\sum_{y \in \mathcal{Y}} \mathbb{E} \left| P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right| \leq \frac{6}{\sqrt{\mathsf{M}^\star}} + 12 \sqrt{\frac{\log |\mathcal{Y}| + \log \mathsf{M}^\star}{\mathsf{M} \min_{y \in \mathcal{Y}} \{P_Y^{\mathrm{cal}}(y)\}}}.$$

Proof. For any $y \in \mathcal{Y}$, introduce the following quantities: $\hat{f}_y^\star = \mathsf{M}_y^\star/\mathsf{M}^\star$, $f^\star = \{P_Y^\star(y)\}_{y \in \mathcal{Y}}$, $\hat{f}^\star = \{\hat{f}_y^\star\}_{y \in \mathcal{Y}}$, and $\widehat{r} = \{(P_Y^{\mathrm{cal}}(y)\mathsf{M}/\mathsf{M}_y)\mathbb{1}_{\mathsf{M}_y > 0}\}_{y \in \mathcal{Y}}$. We denote $\odot$ the Hadamard product, i.e., for any vectors $a, b \in \mathbb{R}^{|\mathcal{Y}|}$, $a \odot b$ is the vector of the component-wise product between a and b. We now bound the quantity in the right-hand side of the previous inequality, we obtain

$$\sum_{y \in \mathcal{Y}} \mathbb{E} \left| P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\mathrm{cal}}(y)}{\sum_{\tilde{y} \in \mathcal{Y}} \widehat{w}_{\tilde{y}}^\star P_Y^{\mathrm{cal}}(\tilde{y})} \right| = \sum_{y \in \mathcal{Y}} \mathbb{E} \left| f_y^\star - \hat{f}_y^\star + \hat{f}_y^\star - \frac{\hat{f}_y^\star \widehat{r}_y}{\sum_{\tilde{y} \in \mathcal{Y}} \hat{f}_{\tilde{y}}^\star \widehat{r}_{\tilde{y}}} \right|$$

$$\leq \mathbb{E} \left\| f^\star - \hat{f}^\star + \hat{f}^\star - \frac{\hat{f}^\star \odot \widehat{r}}{1 + \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle} \right\|_1$$

$$\leq \mathbb{E} \| f^\star - \hat{f}^\star \|_1 + \mathbb{E} \left\| \frac{\hat{f}^\star \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle - \hat{f}^\star \odot (\widehat{r} - \mathbb{1})}{1 + \langle \hat{f}^\star, \widehat{r} - \mathbb{1} \rangle} \right\|_1.$$

First, we establish the following equality that will be injected in the computation of $\mathbb{E}\|f^\star - \hat{f}^\star\|_1$.

$$\|f^\star - \hat{f}^\star\|_1 = \max_{u \in [-1/2, 1/2]^{|\mathcal{Y}|}} \langle f^\star - \hat{f}^\star, u + \mathbb{1}/2 \rangle = \max_{u \in [0,1]^{|\mathcal{Y}|}} \langle f^\star - \hat{f}^\star, u \rangle. \tag{44}$$

Then, using the result provided by (Agrawal and Jia, 2017, Lemma C.2), for any $\delta \in (0, 1)$, we get

$$\mathbb{P}\left(\max_{u \in [0,1]^{|\mathcal{Y}|}} \langle f^\star - \hat{f}^\star, u \rangle \geq \sqrt{\frac{-2 \log \delta}{\mathsf{M}^\star}} \right) \leq \delta. \tag{45}$$

Looking back to $\mathbb{E}\|f^\star - \hat{f}^\star\|_1$ and using the two previous identities, the following lines hold

$$\begin{aligned}
\mathbb{E}\|f^\star - \hat{f}^\star\|_1 &= \int_{t \geq 0} \mathbb{P}\left(\|f^\star - \hat{f}^\star\|_1 \geq t \right) \mathrm{d}t \\
&\leq \delta + \int_{t \geq \delta} \mathbb{P}\left(\|f^\star - \hat{f}^\star\|_1 \geq t \right) \mathrm{d}t \\
&\leq \delta + \int_{t \geq \delta} \exp\left(-\mathsf{M}^\star t^2 / 2 \right) \mathrm{d}t \qquad \text{using (44)-(45)} \\
&\leq \delta + \frac{1}{\sqrt{\mathsf{M}^\star}} \int_{t \geq \delta\sqrt{\mathsf{M}^\star}} \exp\left(-t^2/2 \right) \mathrm{d}t.
\end{aligned}$$

After optimizing for $\delta > 0$, we can retrieve the following upper bound

$$\mathbb{E}\|f^\star - \hat{f}^\star\|_1 \leq \frac{1.4}{\sqrt{\mathsf{M}^\star}}.$$

Let $\epsilon \in (0, 1/2]$, we have that

$$\left\| \frac{\hat{f}^\star \langle \hat{f}^\star, \hat{r} - \mathbb{1} \rangle - \hat{f}^\star \odot (\hat{r} - \mathbb{1})}{1 + \langle \hat{f}^\star, \hat{r} - \mathbb{1} \rangle} \right\|_1 \leq 2 \mathbb{1}_{\|\hat{r} - \mathbb{1}\|_\infty > \epsilon} + \frac{2\epsilon}{1 - \epsilon}.$$

Taking the expectation for both sides, it shows

$$\mathbb{E}\left\| \frac{\hat{f}^\star \langle \hat{f}^\star, \hat{r} - \mathbb{1} \rangle - \hat{f}^\star \odot (\hat{r} - \mathbb{1})}{1 + \langle \hat{f}^\star, \hat{r} - \mathbb{1} \rangle} \right\|_1 \leq 2 \mathbb{P}\left(\|\hat{r} - \mathbb{1}\|_\infty > \epsilon \right) + \frac{2\epsilon}{1 - \epsilon}.$$

We now upper bound the first term in the right-hand side of the inequality, we obtain

$$\begin{aligned}
\mathbb{P}\left(\|\hat{r} - \mathbb{1}\|_\infty > \epsilon \right) &= \mathbb{P}\left(\max_{y \in \mathcal{Y}} \left| \frac{P_Y^{\mathrm{cal}}(y) \mathbb{1}_{\mathsf{M}_y > 0}}{\mathsf{M}_y / \mathsf{M}} - 1 \right| > \epsilon \right) \\
&\leq \sum_{y \in \mathcal{Y}} \mathbb{P}\left(\left| \frac{P_Y^{\mathrm{cal}}(y) \mathbb{1}_{\mathsf{M}_y > 0}}{\mathsf{M}_y / \mathsf{M}} - 1 \right| > \epsilon \right) \\
&\leq \sum_{y \in \mathcal{Y}} \left\{ \mathbb{P}\left(\frac{\mathsf{M}_y}{\mathsf{M}} < \frac{P_Y^{\mathrm{cal}}(y) \mathbb{1}_{\mathsf{M}_y > 0}}{1 + \epsilon} \right) + \mathbb{P}\left(\frac{\mathsf{M}_y}{\mathsf{M}} > \frac{P_Y^{\mathrm{cal}}(y)}{1 - \epsilon} \right) \right\}. \tag{46}
\end{aligned}$$

Since the random variable M_y is the sum of independent Bernoulli random variables, using the Chernoff bound it follows that

$$\begin{aligned}
\mathbb{P}\left(\mathsf{M}_y / \mathsf{M} < P_Y^{\mathrm{cal}}(y) / (1 + \epsilon) \right) &\leq \exp\left(-2\epsilon^2 N P_Y^{\mathrm{cal}}(y) / 9 \right), \\
\mathbb{P}\left(\mathsf{M}_y / \mathsf{M} > P_Y^{\mathrm{cal}}(y) / (1 - \epsilon) \right) &\leq \exp\left(-4\epsilon^2 N P_Y^{\mathrm{cal}}(y) / 3 \right).
\end{aligned} \tag{47}$$

Therefore, combining (46) with (47) gives

$$\mathbb{P}\left(\|\hat{r} - \mathbb{1}\|_\infty > \epsilon \right) \leq \sum_{y \in \mathcal{Y}} \left[\mathbb{P}\left(\mathsf{M}_y = 0 \right) + 2 \exp\left(-\epsilon^2 N P_Y^{\mathrm{cal}}(y) / 5 \right) \right].$$

Putting all the previous results together, we obtain

$$\sum_{y\in\mathcal{Y}} P_Y^{\text{cal}}(y)\mathbb{E}\left|w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\text{cal}}(\tilde{y})}\right| \le \frac{1.4}{\sqrt{\mathsf{M}^\star}} + 4\epsilon + 4\sum_{y\in\mathcal{Y}}\exp\left(-2\epsilon^2 NP_Y^{\text{cal}}(y)/9\right) + \sum_{y\in\mathcal{Y}}(1-P_Y^{\text{cal}}(y))^{\mathsf{M}}. \quad (48)$$

Consider the following quantity

$$\epsilon = \frac{3}{2}\sqrt{\frac{2\log|\mathcal{Y}| + \log\mathsf{M}^\star}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\text{cal}}(y)\}}}.$$

If $\epsilon \le 1/2$, it yields that

$$\sum_{y\in\mathcal{Y}}\exp\left(-2\epsilon^2 NP_Y^{\text{cal}}(y)/9\right) \le \sum_{y\in\mathcal{Y}}\exp\left(-\log|\mathcal{Y}| - (1/2)\log\mathsf{M}^\star\right) = \frac{1}{\sqrt{\mathsf{M}^\star}}.$$

Therefore, combining this last inequality with (43)-(48) implies that

$$\left|\mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(X_{N^\star+1}^\star)) - \mathbb{P}(\widehat{Y} \in \mathcal{C}_{\alpha,\widehat{\mu}^{\text{MLE}}}(X_{N^\star+1}^\star))\right| \le \frac{3}{\sqrt{\mathsf{M}^\star}} + 3\sqrt{\frac{2\log|\mathcal{Y}| + \log\mathsf{M}^\star}{\mathsf{M}\min_{y\in\mathcal{Y}}\{P_Y^{\text{cal}}(y)\}}}$$

$$+ |\mathcal{Y}|\left(1-\min_{y\in\mathcal{Y}}\{P_Y^{\text{cal}}(y)\}\right)^{\mathsf{M}} \wedge 1.$$

Otherwise, if $\epsilon > 1/2$, then, the last inequality immediately holds since the right-hand term is greater than 1. Lastly, applying Lemma C.7 concludes the proof $\qquad\square$

C.4. Proof of Theorem 4.3

First, for all $i \in [n]$, denote by $F_V^i : u \in [0,1] \mapsto \mathbb{P}(V(X,Y) \le u) \in [0,1]$ the cumulative distribution function of $V(X^i, Y^i)$, where $(X^i, Y^i) \sim P^i$. Recall that $N = \sum_{i=1}^n N^i$, $\mathcal{I} = (i,k): i \in [n], k \in [N^i] \cup \{(\star, N^\star+1)\}$, and also that there is almost surely no ties between the $\{V(X_k^i, Y_k^i)\}_{(i,k)\in\mathcal{I}}$. To simplify the notation, we re-index $\{(X_k^i, Y_k^i, V_k^i)\}_{(i,k)\in\mathcal{I}}$ into $\{X_k, Y_k, V_k\}_{k\in[N+1]}$, sorted such that $\{V_k\}_{k\in[N+1]}$ is non-decreasing.

Now, we consider $\alpha \in [0,1] \setminus \mathbb{Q}$. Using Theorem A.3, this condition ensures the existence and uniqueness of $k_{\text{opt}} \in [N+1]$ such that $V_{k_{\text{opt}}} = \arg\min_{q\in\mathbb{R}}\{\mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S_{\alpha,V}(q)]\}$, and this condition also proves the existence and uniqueness of $Q_{1-\alpha}^\gamma$ minimizing $\{\mathbb{E}_{V\sim\widehat{\mu}_{\mathbf{y}}}[S_{\alpha,V}^\gamma(q)]: q \in \mathbb{R}\}$. Moreover, for any $k \in [N+1] \setminus \{k_{\text{opt}}\}$, define

$$\widehat{\rho}_{k_c} = \begin{cases} \sum_{\ell:\, V_\ell\in(V_{k_{\text{opt}}}, V_k]}\widehat{p}_{Y_\ell, Y_{N^\star+1}^\star}^\star, & \text{if } k > k_{\text{opt}} \\ \sum_{\ell:\, V_\ell\in[V_k, V_{k_{\text{opt}}})}\widehat{p}_{Y_\ell, Y_{N^\star+1}^\star}^\star, & \text{if } k < k_{\text{opt}} \end{cases}. \quad (49)$$

Lemma C.9. *Let $\alpha \in [0,1] \setminus \mathbb{Q}$, for any $V_k \in [\min(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}), V_{k_{\text{opt}}}), \max(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}), V_{k_{\text{opt}}})]$, we have*

$$|\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}) - V_k| \le \widehat{\rho}_{k_c}^{-1}\left(\mathrm{S}_\alpha^\gamma(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}_\alpha^\gamma(Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}}))\right) + \widehat{\rho}_{k_c}^{-1}\gamma,$$

where $\widehat{\rho}_{k_c}$ is given in (49).

Proof. First, suppose that $V_{k_{\text{opt}}} \le V_k < \widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})$. Since $\partial S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(V_k) = \widehat{\rho}_{k_c} + \partial S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(V_{k_{\text{opt}}})$, the convexity of $S_{\alpha,\widehat{\mu}_{\mathbf{y}}}$ implies that

$$\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}) - V_k \le \widehat{\rho}_{k_c}^{-1}\left(S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})) - S_{\alpha,\widehat{\mu}_{\mathbf{y}}}(V_k)\right)$$

$$\le \widehat{\rho}_{k_c}^{-1}\left(\mathrm{S}_\alpha^\gamma(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}_\alpha^\gamma(Q_{1-\alpha}^\gamma(\widehat{\mu}_{\mathbf{y}}))\right) + \widehat{\rho}_{k_c}^{-1}\gamma. \quad (50)$$

The last inequality holds since $\|\mathrm{S}_\alpha^\gamma - S_{\alpha,\widehat{\mu}_{\mathbf{y}}}\|_\infty \le \gamma/2$. Moreover, (50) is immediately satisfied when $V_k = \widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})$. Therefore, (50) holds for all $V_k \in [V_{k_{\text{opt}}}, \widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})]$. Finally, the same lines show that (50) is also satisfied when $\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}}) \le V_k \le V_{k_{\text{opt}}}$. $\qquad\square$

For any $\gamma > 0$ and $T \in \mathbb{N}^\star$, recall that $Q_{1-\alpha}^\gamma(\widehat{\mu}_\mathbf{y})$ and $\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_\mathbf{y})$ are defined in (3), (15). Consider

$$C_T^\gamma = \frac{2N \sum_{y \in \mathcal{Y}} P_Y^{\text{cal}}(y) \widehat{w}_y^\star}{\min_{y \in \mathcal{Y}} \widehat{w}_y^\star} \left[\mathbb{E} \, \text{S}_\alpha^\gamma(\widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_\mathbf{y})) - \text{S}_\alpha^\gamma(Q_{1-\alpha}^\gamma(\widehat{\mu}_\mathbf{y})) + \gamma \right] \tag{51}$$

and define

$$k_{\text{c}} = \begin{cases} k_{\text{opt}} + \text{Ent}\left(\sqrt{\frac{mNC_T^\gamma}{2 \log N}} \right) & \text{if } k_{\text{opt}} + \text{Ent}\left(\sqrt{\frac{mNC_T^\gamma}{2 \log N}} \right) \leq N+1 \\ k_{\text{opt}} - \text{Ent}\left(\sqrt{\frac{mNC_T^\gamma}{2 \log N}} \right) & \text{otherwise} \end{cases} . \tag{52}$$

Lemma C.10. *Assume there exists $m > 0$ such that for any $i \in [n]$, P_V^i admits a density f_V^i with respect to the Lebesgue measure that satisfies $m \leq f_V^i$. Fix $\alpha \in [0,1] \setminus \mathbb{Q}$, and suppose that $C_T^\gamma < 2m^{-1}N \log N$. We have $k_{\text{c}} \in [N+1]$, and with probability at least $m(2N \log N)^{-1} + N^{-2}$ the next inequality holds*

$$|V_{k_{\text{c}}} - V_{k_{\text{opt}}}| \leq \sqrt{\frac{2C_T^\gamma \log N}{mN}} + \frac{2 \log N}{mN}.$$

Proof. Since we suppose $\alpha \in [0,1] \setminus \mathbb{Q}$, we can apply Theorem A.3 to prove the existence and uniqueness of $k_{\text{opt}} \in [N+1]$ such that $Q_{1-\alpha}(\widehat{\mu}_\mathbf{y}) = V_{k_{\text{opt}}}$. Since by assumption we have

$$\frac{mNC_T^\gamma}{2 \log N} < (N+1)^2.$$

Therefore, it holds that

$$\text{Ent}\left(\sqrt{\frac{mNC_T^\gamma}{2 \log N}} \right) \leq N.$$

Hence, we deduce that $k_{\text{c}} \in [N+1]$. Next, consider

$$L_N = \frac{2 \log N}{mN}$$

and denote by P_N the partitioned obtained by splitting the interval $[0,1]$ into intervals of length L_N. Finally, define

$$A_N = \left\{ \forall S \in P_N, \exists (i,k) \in \mathcal{I}, V_k^i \in S \right\}.$$

Note that $|P_N| = \lceil 1/L_N \rceil \leq mN/(2 \log N) + 1$ and $\log(1 - mL_N) \leq -mL_N$, thus

$$\begin{aligned} \mathbb{P}(A_N) &\leq \sum_{S \in P_N} \prod_{(i,k) \in \mathcal{I}} \mathbb{P}\left(V_k^i \notin S \right) \\ &\leq \sum_{S \in P_N} \prod_{i=1}^n \mathbb{P}\left(V_1^i \notin S \right)^{N^i + \mathbb{1}_\star(i)} \\ &\leq |P_N|(1 - mL_N)^{N+1} \\ &\leq (mN/(2 \log N) + 1) \exp\left(-m(N+1)L_N \right) \\ &\leq \frac{m}{2N \log N} + \frac{1}{N^2}. \end{aligned}$$

Without loss of generality, we can assume that $k_{\text{opt}} < k_{\text{c}}$. Denote $K = \{k_{\text{opt}}, \ldots, k_{\text{c}}\}$ the indices between k_{c} and k_{opt}. Consider $\mathcal{S} = \{I \in P_N : \exists k \in K, V_k \in I\}$, on the event A_N we get

$$\begin{aligned} |V_{k_{\text{c}}} - V_{k_{\text{opt}}}| &\leq \sum_{I \in \mathcal{S}} |I| \\ &\leq L_N(|k_{\text{c}} - k_{\text{opt}}| + 1) \\ &\leq L_N \sqrt{\frac{mNC_T^\gamma}{2 \log N}} + L_N = \sqrt{C_T^\gamma L_N} + L_N. \end{aligned}$$

$\square$

Lemma C.11. *For any $(i,k) \in \mathcal{I}$, assume that (X_k^i, Y_k^i) is distributed according to $P_{X|Y} \times P_Y^i$ and suppose the random variables are pairwise independent. We have*

$$\mathbb{P}\left(\min_{y \in \mathcal{Y}} \widehat{p}_{y, Y_{N^\star+1}^\star}^\star < \frac{\min_{y \in \mathcal{Y}} \widehat{w}_y^\star}{2N \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y) \widehat{w}_y^\star}\right) \leq \frac{4 \operatorname{Var}(\widehat{w}_{Y^{\mathrm{cal}}}^\star)}{N(\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star)^2} + \frac{2\mathbb{E}\widehat{w}_{Y_{N^\star+1}^\star}^\star}{N\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star}.$$

Proof. First, recall that $I = \{(i,k) \colon i \in [n], k \in [N^i]\} \cup \{(\star, N^\star + 1)\}$. We have

$$\mathbb{E}\left[\sum_{(i,k)\in\mathcal{I}} \widehat{w}_{Y_k^i}^\star\right] = \sum_{i=1}^{n} \sum_{y\in\mathcal{Y}} \left(N^i + \mathbb{1}_\star(i)\right) P_Y^i(y)\widehat{w}_y^\star$$

$$= \sum_{y\in\mathcal{Y}} P_Y^\star(y)\widehat{w}_y^\star + N \sum_{y\in\mathcal{Y}} \left(\sum_{i=1}^{n} \pi_i P_Y^i(y)\right) \widehat{w}_y^\star$$

$$= \sum_{y\in\mathcal{Y}} \left[P_Y^\star(y) + N P_Y^{\mathrm{cal}}(y)\right] \widehat{w}_y^\star.$$

Therefore, using the Bienaymé-Tchebytchev inequality implies that

$$\mathbb{P}\left(\min_{(i,k)\in\mathcal{I}} \{\widehat{p}_{Y_k^i, Y_{N^\star+1}^\star}^\star\} < \frac{\min_{y\in\mathcal{Y}} \widehat{w}_y^\star}{2N \sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^\star}\right)$$

$$= \mathbb{P}\left(\min_{(i,k)\in\mathcal{I}} \{\widehat{w}_{Y_k^i}^\star\} < \frac{\min_{y\in\mathcal{Y}} \widehat{w}_y^\star}{2N \sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^\star} \sum_{(i,k)\in\mathcal{I}} \widehat{w}_{Y_k^i}^\star\right)$$

$$\leq \mathbb{P}\left(\sum_{(i,k)\in\mathcal{I}} \widehat{w}_{Y_k^i}^\star \geq 2N \sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^\star\right)$$

$$\leq \mathbb{P}\left(\sum_{i=1}^{n}\sum_{k=1}^{N^i} \left(\widehat{w}_{Y_k^i}^\star - \mathbb{E}\widehat{w}_{Y_k^i}^\star\right) \geq \frac{N}{2} \sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^\star\right) + \mathbb{P}\left(\widehat{w}_{Y_{N^\star+1}^\star}^\star \geq \frac{N}{2} \sum_{y\in\mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^\star\right)$$

$$\leq \frac{4 \operatorname{Var}(\widehat{w}_{Y^{\mathrm{cal}}}^\star)}{N(\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star)^2} + \frac{2\mathbb{E}\widehat{w}_{Y_{N^\star+1}^\star}^\star}{N\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star}.$$

$\square$

Theorem C.12. *Assume there exist $m, M > 0$ such that for any $i \in [n]$, P_V^i admits a density f_V^i with respect to the Lebesgue measure that satisfies $m \leq f_V^i \leq M$. Let $\alpha \in [0,1] \setminus \mathbb{Q}$, and suppose that $C_T^\gamma < 2m^{-1}N \log N$. It holds*

$$\left|\mathbb{P}(Y_{N^\star+1}^\star \in \widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star)) - \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star))\right| \leq 3M\sqrt{\frac{2C_T^\gamma \log N}{mN}}$$

$$+ \frac{2M\log N}{mN} + \frac{4\operatorname{Var}(\widehat{w}_{Y^{\mathrm{cal}}}^\star)}{N(\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star)^2} + \frac{2\mathbb{E}\widehat{w}_{Y_{N^\star+1}^\star}^\star}{N\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^\star} + \frac{m}{2N\log N} + \frac{1}{N^2}, \quad (53)$$

where C_T^γ is defined in (51).

Proof. Using the definitions of $\widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star), \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star)$ provided in Algorithm 2 and (8), we can write

$$\mathbb{P}(Y_{N^\star+1}^\star \in \widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^\gamma(X_{N^\star+1}^\star)) - \mathbb{P}(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star))$$

$$= \mathbb{P}\left(V(X_{N^\star+1}^\star, Y_{N^\star+1}^\star) \leq \widehat{Q}_{1-\alpha,T}^\gamma(\widehat{\mu}_{\mathbf{y}})\right) - \mathbb{P}\left(V(X_{N^\star+1}^\star, Y_{N^\star+1}^\star) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\right). \quad (54)$$

Recall that $k_{\mathrm{opt}} = \arg\min_{k \in [N+1]} \mathbb{E}_{V \sim \widehat{\mu}_{\mathbf{y}}}[S_{\alpha,V}^{\gamma}(V_k)]$ is a random variable and consider the event

$$B_N = \left\{ |V_{k_{\mathrm{c}}} - V_{k_{\mathrm{opt}}}| \leq \sqrt{C_T^{\gamma} L_N} + L_N \right\} \cap \left\{ \min_{y \in \mathcal{Y}} \widehat{p}_{y,Y_{N^{\star}+1}^{\star}}^{\star} \geq \frac{\min_{y \in \mathcal{Y}} \widehat{w}_y^{\star}}{2N \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^{\star}} \right\}$$

and denote by B_N^c its complement. Since $V(X_{N^{\star}+1}^{\star}, Y_{N^{\star}+1}^{\star})$ and $\{\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}), Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\}$ are independent, we have

$$\left| \mathbb{P}(V(X_{N^{\star}+1}^{\star}, Y_{N^{\star}+1}^{\star}) \leq \widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) - \mathbb{P}(V(X_{N^{\star}+1}^{\star}, Y_{N^{\star}+1}^{\star}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) \right|$$

$$= \left| \mathbb{E}\left[\mathbb{1}_{V(X_{N^{\star}+1}^{\star}, Y_{N^{\star}+1}^{\star}) \leq \widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}})} - \mathbb{1}_{V(X_{N^{\star}+1}^{\star}, Y_{N^{\star}+1}^{\star}) \leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})} \right] \right|$$

$$\leq \mathbb{P}\left(B_N^c\right) + \mathbb{E}\left[\mathbb{1}_{B_N} \left| F_{V^{\star}}(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) - F_{V^{\star}}(Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) \right| \right]. \quad (55)$$

The following inequality holds

$$\left| F_{V^{\star}}(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) - F_{V^{\star}}(Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) \right| \leq \| F_{V^{\star}}(\cdot + \widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) - F_{V^{\star}} \|_{\infty}.$$

Thus, using that $F_{V^{\star}}$ is M-Lipschitz, we get

$$\mathbb{E}\left[\mathbb{1}_{B_N} \| F_{V^{\star}}(\cdot + \widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})) - F_{V^{\star}} \|_{\infty} \right] \leq M \mathbb{E}\left[\mathbb{1}_{B_N} |\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})| \right]. \quad (56)$$

Furthermore, we have

$$\mathbb{E}\left[\mathbb{1}_{B_N} |\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})| \right]$$

$$\leq \mathbb{E}\left[\mathbb{1}_{B_N} |V_{k_{\mathrm{c}}} - V_{k_{\mathrm{opt}}}| \right] + \mathbb{E}\left[\mathbb{1}_{V_{k_{\mathrm{c}}} \in [\min(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}), V_{k_{\mathrm{opt}}}), \max(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}), V_{k_{\mathrm{opt}}})]} |\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - V_{k_{\mathrm{c}}}| \right].$$

Applying Lemma C.9, this implies that

$$\mathbb{1}_{V_{k_{\mathrm{c}}} \in [\min(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}), V_{k_{\mathrm{opt}}}), \max(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}), V_{k_{\mathrm{opt}}})]} |\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - V_{k_{\mathrm{c}}}|$$

$$\leq \begin{cases} \widehat{\rho}_{k_{\mathrm{c}}}^{-1} \left(\mathrm{S}_{\alpha}^{\gamma}(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}_{\alpha}^{\gamma}(Q_{1-\alpha}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) \right) + \widehat{\rho}_{k_{\mathrm{c}}}^{-1} \gamma & \text{if } k_{\mathrm{c}} \neq k_{\mathrm{opt}} \\ 0 & \text{otherwise} \end{cases},$$

where recall that $\widehat{\rho}_{k_{\mathrm{c}}}$ is defined in (49) and also that $\mathcal{I} = \{(i,k) \colon i \in [n], k \in [N^i]\} \cup \{(\star, N^{\star}+1)\}$. Moreover, on the event B_N, we immediately have that

$$\widehat{\rho}_{k_{\mathrm{c}}} \geq |k_{\mathrm{c}} - k_{\mathrm{opt}}| \min_{(i,k) \in \mathcal{I}} \{ \widehat{p}_{Y_k^i, Y_{N^{\star}+1}^{\star}}^{\star} \} \geq \frac{|k_{\mathrm{c}} - k_{\mathrm{opt}}| \min_{y \in \mathcal{Y}} \widehat{w}_y^{\star}}{2N \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^{\star}}.$$

Finally, recall that k_{c} is given in (52) and suppose that $C_T^{\gamma} < 2m^{-1}N \log N$. Therefore, using the bound provided in Lemma C.10 implies that

$$\mathbb{E}\left[\mathbb{1}_{B_N} |\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}}) - Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})| \right]$$

$$\leq \mathbb{E}\left[\mathbb{1}_{B_N} |V_{k_{\mathrm{c}}} - V_{k_{\mathrm{opt}}}| \right] + \mathbb{E}\left[\frac{\left(\mathbb{E}\,\mathrm{S}_{\alpha}^{\gamma}(\widehat{Q}_{1-\alpha,T}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) - \mathrm{S}_{\alpha}^{\gamma}(Q_{1-\alpha}^{\gamma}(\widehat{\mu}_{\mathbf{y}})) + \gamma \right) \mathbb{1}_{k_{\mathrm{c}} \neq k_{\mathrm{opt}}}}{(2N \sum_{y \in \mathcal{Y}} P_Y^{\mathrm{cal}}(y)\widehat{w}_y^{\star})^{-1} |k_{\mathrm{c}} - k_{\mathrm{opt}}| \min_{y \in \mathcal{Y}} \widehat{w}_y^{\star}} \right]$$

$$\leq \sqrt{\frac{2C_T^{\gamma} \log N}{mN}} + \frac{2 \log N}{mN} + \frac{C_T^{\gamma} \mathbb{1}_{mNC_T^{\gamma} \geq 2 \log N}}{\mathrm{Ent}\left(\sqrt{\frac{mNC_T^{\gamma}}{2 \log N}} \right)}. \quad (57)$$

Combining (54)-(55)-(56)-(57) shows that

$$\left| \mathbb{P}(Y_{N^{\star}+1}^{\star} \in \widehat{\mathcal{C}}_{\alpha,\widehat{\mu}}^{\gamma}(X_{N^{\star}+1}^{\star})) - \mathbb{P}(Y_{N^{\star}+1}^{\star} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^{\star}+1}^{\star})) \right| \leq \mathbb{P}\left(B_N^c\right) + M\left(3\sqrt{\frac{2C_T^{\gamma} \log N}{mN}} + \frac{2 \log N}{mN} \right). \quad (58)$$

Using Lemma C.11 gives that

$$\mathbb{P}\left(B_N^c\right) \le \frac{4\operatorname{Var}(\widehat{w}_{Y^{\mathrm{cal}}}^{\star})}{N(\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^{\star})^2} + \frac{2\mathbb{E}\widehat{w}_{Y_{N^{\star}+1}^{\star}}^{\star}}{N\mathbb{E}\widehat{w}_{Y^{\mathrm{cal}}}^{\star}} + \frac{m}{2N\log N} + \frac{1}{N^2}. \tag{59}$$

Lastly, plugging (59) into (58) concludes the proof when $C_T^\gamma < 2m^{-1}N\log N$. However, if $C_T^\gamma \ge 2m^{-1}N\log N$ then (53) immediately holds. Thus, (53) always holds. □

C.5. Proofs of Theorem 2.2 and Theorem 2.4

Recall that $\{\pi_i\}_{i\in[n]} \in \Delta_n$ and $P^{\mathrm{cal}} = \sum_{i=1}^n \pi_i P^i$. Moreover, draw $(\widehat{X}_{N^{\star}+1}^{\star}, \widehat{Y}_{N^{\star}+1}^{\star})$ according to $P_{X|Y} \times \widehat{P}_Y^{\star}$, where $\widehat{P}_Y^{\star}$ is defined in Appendix C.3 and denote $\widehat{V}_{N+1} = V(\widehat{X}_{N^{\star}+1}^{\star}, \widehat{Y}_{N^{\star}+1}^{\star})$. In the following paragraph, we explain how to construct a sequence $\{V_k\}_{k\in[N]}$ of i.i.d. random variables distributed according to $P^{\mathrm{cal}}(V)$ – see Lemma C.13, where $P^{\mathrm{cal}}(V)$ denotes the distribution of $V(X,Y)$ with $(X,Y) \sim P^{\mathrm{cal}}$. We also explain the construction of a bijection $\psi : \{(i,k)\colon i\in[n], k\in[N^i]\} \to [N]$. For all $k\in[N]$, draw M_k according to a categorical random variable with parameter $\{\pi_i\}_{i\in[n]}$ and define $\bar{N}_k^i = \sum_{l=1}^k \mathbb{1}_{\{i\}}(M_l)$. If $\bar{N}_k^{M_k} \le N^{M_k}$ then define $(X_k, Y_k) \leftarrow (X_{\bar{N}_k^{M_k}}^{M_k}, Y_{\bar{N}_k^{M_k}}^{M_k})$ and $\psi(M_k, \bar{N}_k^{M_k}) = k$. Else $\bar{N}_k^{M_k} > N^{M_k}$, then draw (X_k, Y_k) according to P^{M_k} – where we recall that P^{M_k} is the distribution of calibration of agent $M_k \in [n]$. If there exists $k\in[N]$ such that $\bar{N}_k^{M_k} > N^{M_k}$, consider

$$J_0 = \left\{k\in[N]\colon \bar{N}_k^{M_k} > N^{M_k}\right\}, \qquad J_1 = \left\{(i,k)\in[n]\times\mathbb{N}\colon \bar{N}_N^i < k \le N^i\right\}.$$

Since the following inequalities hold:

$$\operatorname{Card}(J_0) + \sum_{i=1}^n \min(N^i, \bar{N}_N^i) = N, \qquad \operatorname{Card}(J_1) + \sum_{i=1}^n \min(N^i, \bar{N}_N^i) = N,$$

we deduce that $\operatorname{Card}(J_0) = \operatorname{Card}(J_1)$. Moreover, using the existence of $k\in[N]$ such that $\bar{N}_k^{M_k} > N^{M_k}$, we deduce that $J_0 \ne \emptyset$. Therefore, there exists a bijection $\varphi : J_0 \to J_1$. We have previously defined ψ on $\{(i,k)\colon i\in[n], k\in[N^i]\} \setminus J_1$. For any $k\in J_0$, define $\psi(\varphi(k)) = k$. Remark, ψ is now correctly defined on $\{(i,k)\colon i\in[n], k\in[N^i]\} \to [N]$.

Lemma C.13. *Denote $P^{\mathrm{cal}}(V)$ the distribution of $V(X,Y)$ with $(X,Y) \sim P^{\mathrm{cal}}$. The sequence $\{V_k\}_{k\in[N]}$ is a sequence of i.i.d. random variables distributed according to $P^{\mathrm{cal}}(V)$.*

Proof. Let $h : \mathbb{R}^N \to \mathbb{R}$ be a continuous and bounded function, we have

$$\mathbb{E}\left[h(V_1, \ldots, V_N)\right] = \sum_{i_1,\ldots,i_N\in[n]} \left(\prod_{k'=1}^N \pi_{i_{k'}}\right) \int h(v_1, \ldots, v_N) \prod_{k=1}^N \mathrm{d}P^{i_k}(v_k)$$

$$= \sum_{i_1,\ldots,i_{N-1}\in[n]} \left(\prod_{k'=1}^{N-1} \pi_{i_{k'}}\right) \int h(v_1, \ldots, v_N) \prod_{k=1}^{N-1} \mathrm{d}P^{i_k}(v_k) \left(\sum_{i=1}^n \pi_i \mathrm{d}P^i\right)(v_N)$$

$$= \sum_{i_1,\ldots,i_{N-1}\in[n]} \left(\prod_{k'=1}^{N-1} \pi_{i_{k'}}\right) \int h(v_1, \ldots, v_N) \prod_{k=1}^{N-1} \mathrm{d}P^{i_k}(v_k)\mathrm{d}P^{\mathrm{cal}}(v_N)$$

$$= \cdots = \int h(v_1, \ldots, v_N) \prod_{k=1}^N \mathrm{d}P^{\mathrm{cal}}(v_k).$$

This last line concludes the proof. □

In the following, we consider general likelihood ratios $\{\widehat{w}_y^{\star}\}_{y\in\mathcal{Y}}$ and recall that $w_y^{\star} = P_Y^{\star}(y)/P_Y^{\mathrm{cal}}(y)$. In addition, let

$\bar{N} \in [N]$, denote for any $i \in [n], k \in [N^i \wedge \bar{N}^i_{\bar{N}}], k' \in [\bar{N}], V_{k'} = V(X_{k'}, Y_{k'})$ and define

$$\bar{D}_{\widehat{Y}^\star_{N^\star+1}} = \{\widehat{Y}^\star_{N^\star+1}\} \cup \{Y_k : k \in [\bar{N}]\}, \quad \widehat{D}_{\widehat{Y}^\star_{N^\star+1}} = \{\widehat{Y}^\star_{N^\star+1}\} \cup \{Y_k : i \in [n], k \in [N^i \wedge \bar{N}^i_{\bar{N}}]\}$$

$$\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_{k'}} = \frac{\widehat{w}^\star_{Y_{k'}}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l}}, \qquad \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i} = \frac{\widehat{w}^\star_{Y_k^i}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{j=1}^{n} \sum_{l=1}^{N^j \wedge \bar{N}^j_{\bar{N}}} \widehat{w}^\star_{Y_l^j}}. \tag{60}$$

Moreover, consider $\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} = \widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} (\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l})^{-1}$ and $\widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} = \widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} (\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} +$

$\sum_{j=1}^{n} \sum_{l=1}^{N^j \wedge \bar{N}^j_{\bar{N}}} \widehat{w}^\star_{Y_l^j})^{-1}$. Lastly, define

$$X = \sum_{k=1}^{\bar{N}} \bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k} \mathbb{1}_{V_k < \widehat{V}_{N+1}},$$

$$\delta = \sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i} \mathbb{1}_{V_k^i < \widehat{V}_{N+1}} - \sum_{k=1}^{\bar{N}} \bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k} \mathbb{1}_{V_k < \widehat{V}_{N+1}}. \tag{61}$$

Lemma C.14. *For any $\{\pi_i\}_{i \in [n]} \in \Delta_n$, it holds*

$$-\epsilon - \mathbb{P}(\delta \le -\epsilon) \le \mathbb{P}\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i} \delta_{V_k^i} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1\right)\right) - 1 + \alpha$$

$$\le \mathbb{E}\left[\max_{k=1}^{\bar{N}+1} \{\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k}\}\right] + \epsilon + \mathbb{P}(\delta \ge \epsilon),$$

where $\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k}$ and $\widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i}$ are defined in (60).

Proof. By the definition of the quantile combined with (61), we get

$$\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i} \delta_{V_k^i} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1\right)\right) \iff (X + \delta > \alpha).$$

Therefore, it holds that

$$\mathbb{P}\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i} \delta_{V_k^i} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1\right)\right) = \mathbb{E}\left[\mathbb{1}_{X>\alpha}\right] + \mathbb{E}\left[\mathbb{1}_{X+\delta>\alpha} - \mathbb{1}_{X>\alpha}\right]. \tag{62}$$

Remark that

$$\mathbb{1}_{X>\alpha+\epsilon} - \mathbb{1}_{X>\alpha} - \mathbb{1}_{\delta \le -\epsilon} \le \mathbb{1}_{X+\delta>\alpha} - \mathbb{1}_{X>\alpha} \le \mathbb{1}_{X>\alpha-\epsilon} - \mathbb{1}_{X>\alpha} + \mathbb{1}_{\delta \ge \epsilon}. \tag{63}$$

Thus, combining (63) with (62) gives

$$\mathbb{P}(X > \alpha + \epsilon) - \mathbb{P}(\delta \le -\epsilon)$$

$$\le \mathbb{P}\left(\widehat{V}_{N+1} \le Q_{1-\alpha}\left(\sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i_{\bar{N}}} \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k^i} \delta_{V_k^i} + \widehat{p}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1\right)\right)$$

$$\le \mathbb{P}(X > \alpha - \epsilon) + \mathbb{P}(\delta \ge \epsilon). \tag{64}$$

Consider $\tilde{\alpha} \in (0, 1)$,

$$(X > \tilde{\alpha}) \iff \left(\widehat{V}_{N+1} \le Q_{1-\tilde{\alpha}}\left(\bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{\widehat{Y}^\star_{N^\star+1}} \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}_{Y_k} \delta_{V_k}\right)\right).$$

Hence $\mathbb{P}(X > \tilde{\alpha}) = \mathbb{P}(\widehat{V}_{N+1} \leq Q_{1-\tilde{\alpha}}(\bar{p}_{\widehat{Y}^\star_{N^\star+1}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_{V_k}))$. Applying Theorem C.3 gives that

$$0 \leq \mathbb{P}\left(\widehat{V}_{N+1} \leq Q_{1-\tilde{\alpha}}\left(\bar{p}_{\widehat{Y}^\star_{N^\star+1}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \delta_{V_k}\right)\right) - 1 + \tilde{\alpha} \leq \mathbb{E}\left[\max_{k=1}^{\bar{N}+1} \{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\}\right].$$

Therefore, plugging the previous inequality into (64) concludes the proof. $\qquad\square$

Recall that $(\widehat{w}^\star_{Y_1} - \mathbb{E}\widehat{w}^\star_{Y_1})$ is σ-sub Gaussian if for any $s \in \mathbb{R}$, the following inequality holds

$$\int_{y \in \mathcal{Y}} \exp\left(s\left[\widehat{w}^\star_y - \int_{y' \in \mathcal{Y}} \widehat{w}^\star_{y'} \mathrm{d}P_Y^{\mathrm{cal}}(y')\right]\right) \mathrm{d}P_Y^{\mathrm{cal}}(y) \leq \exp\left(\frac{\sigma^2 s^2}{2}\right).$$

Lemma C.15. *Assume the random variable $(\widehat{w}^\star_{Y_1} - \mathbb{E}\widehat{w}^\star_{Y_1})$ is σ-sub Gaussian with parameter $\sigma \geq 0$. For any $\epsilon \geq 8\sigma^2 \log N/(\bar{N}\mathbb{E}[\widehat{w}^\star_{Y_1}]^2)$ and $\{\pi_i\}_{i \in [n]} \in \Delta_n$, we have*

$$\mathbb{P}\left(|\delta| > \epsilon\right) \leq \mathbb{P}\left(\sum_{i=1}^n \left(\bar{N}_{\bar{N}}^i - N^i\right)_+ > \frac{N\epsilon}{4}\right) + \frac{4\operatorname{Var}\left(\widehat{w}^\star_{Y_1}\right)}{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]^2} + \frac{1}{N},$$

where δ is defined in (61).

Proof. First, denote $\mathcal{J}_0 = J_0 \cap [\bar{N}]$ and remark that $\operatorname{Card}(\mathcal{J}_0) = \sum_{i=1}^n (\bar{N}_{\bar{N}}^i - N^i)_+$. Moreover, recall that δ is defined by

$$\delta = \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}_{\bar{N}}^i} \left(\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}} - \bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\right) \mathbb{1}_{V_k^i < V_{N+1}} - \sum_{k \in \mathcal{J}_0} \bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \mathbb{1}_{V_k < \widehat{V}_{N+1}}.$$

Using definition of the weighs $\bar{p}^{\bar{D}}$ given in (60) and the definition of ψ provide at the beginning of this section, note that

$$\sum_{k \in \mathcal{J}_0} \bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}} \mathbb{1}_{V_k < \widehat{V}_{N+1}} \leq \frac{\sum_{k \in \mathcal{J}_0} \widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l}}. \tag{65}$$

From the definition of $\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}}$ and $\bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}$ provided in (60), we obtain

$$\sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}_{\bar{N}}^i} \left(\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}} - \bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\right) \mathbb{1}_{V_k^i < V_{N+1}} = \sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}_{\bar{N}}^i} \left(\frac{\widehat{w}^\star_{Y_k^i} \mathbb{1}_{V_k^i < V_{N+1}}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{j=1}^n \sum_{l \in J_i} \widehat{w}^\star_{Y_l^j}} - \frac{\widehat{w}^\star_{Y_k^i} \mathbb{1}_{V_k^i < V_{N+1}}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l}}\right),$$

$$= \left(\frac{1}{\sum_{l \in [\bar{N}+1] \setminus \mathcal{J}_0} \widehat{w}^\star_{Y_l}} - \frac{1}{\sum_{l \in [\bar{N}+1] \setminus \mathcal{J}_0} \widehat{w}^\star_{Y_l} + \sum_{l \in \mathcal{J}_0} \widehat{w}^\star_{Y_l}}\right) \sum_{k \in [\bar{N}] \setminus \mathcal{J}_0} \widehat{w}^\star_{Y_k} \mathbb{1}_{V_k^i < V_{N+1}}.$$

Therefore, we know that

$$\left|\sum_{i=1}^n \sum_{k=1}^{N^i \wedge \bar{N}_{\bar{N}}^i} \left(\widehat{p}_{Y_k^i}^{\widehat{D}_{\widehat{Y}^\star_{N^\star+1}}} - \bar{p}_{Y_{\psi(i,k)}}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\right) \mathbb{1}_{V_k^i < V_{N+1}}\right| \leq \frac{\sum_{k \in \mathcal{J}_0} \widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l}}.$$

Hence, plugging the previous line into (61) combined with (65) gives

$$|\delta| \leq \frac{2 \sum_{k \in \mathcal{J}_0} \widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}^\star_{N^\star+1}} + \sum_{l=1}^{\bar{N}} \widehat{w}^\star_{Y_l}}. \tag{66}$$

Moreover, define the following event:

$$E_N = \left\{\sum_{i=1}^n \left(\bar{N}_{\bar{N}}^i - N^i\right)_+ \leq \frac{N\epsilon}{4}\right\}.$$

Next, using the event E_N, we decompose the following probability

$$\mathbb{P}\left(\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}\geq\frac{\epsilon}{2};E_N\right)\leq\mathbb{P}\left(\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}\geq\frac{\epsilon\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{4};E_N\right)$$
$$+\mathbb{P}\left(\widehat{w}^\star_{\widehat{Y}_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}<\frac{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{2}\right).\quad(67)$$

Since the $\{\widehat{w}^\star_{Y_k}\}_{k\in[N]}$ are σ-sub Gaussian, the first term of the previous right-hand side inequality is upper bounded thanks to Hoeffding's inequality

$$\mathbb{P}\left(\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}\geq\frac{\epsilon\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{4};E_N\right)=\mathbb{E}\left[\mathbb{P}\left(\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}\geq\frac{\epsilon\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{4}\,\bigg|\,\mathrm{Card}(\mathcal{J}_0)\right)\mathbb{1}_{E_N}\right]$$
$$\leq\exp\left(-\frac{(\epsilon\bar{N}/4)^2\mathbb{E}[\widehat{w}^\star_{Y_1}]^2}{2\mathrm{Card}(\mathcal{J}_0)\sigma^2}\right)\leq\frac{1}{N},\quad(68)$$

where the last inequality holds by setting $\epsilon\geq 8\sigma^2\log N/(\bar{N}\mathbb{E}[\widehat{w}^\star_{Y_1}]^2)$. Moreover, since $\widehat{w}^\star_{Y_1}\geq 0$ almost surely, from the Chebyshev inequality we deduce that

$$\mathbb{P}\left(\widehat{w}^\star_{\widehat{Y}_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}<\frac{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]}{2}\right)\leq\frac{4\,\mathrm{Var}\left(\widehat{w}^\star_{Y_1}\right)}{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]^2}.\quad(69)$$

Therefore, combining (67), (68) and (69) shows

$$\mathbb{P}\left(\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}\geq\frac{\epsilon}{4};E_N\right)\leq\frac{1}{N}+\frac{4\,\mathrm{Var}\left(\widehat{w}^\star_{Y_1}\right)}{\bar{N}\mathbb{E}\left[\widehat{w}^\star_{Y_1}\right]^2}.\quad(70)$$

Lastly, using (66) we derive the next inequality:

$$\mathbb{P}\left(|\delta|>\epsilon\right)\leq 1-\mathbb{P}\left(E_N\right)+\mathbb{P}\left(\frac{\sum_{k\in\mathcal{J}_0}\widehat{w}^\star_{Y_k}}{\widehat{w}^\star_{\widehat{Y}_{N^\star+1}}+\sum_{l=1}^{\bar{N}}\widehat{w}^\star_{Y_l}}\geq\frac{\epsilon}{2};E_N\right).$$

Plugging (70) into the previous line completes the proof. $\qquad\square$

Lemma C.16. *For any $i\in[n]$, consider $\pi_i=N^i/N$ and $2\bar{N}\leq N$. We have*

$$\mathbb{P}\left(\sum_{i=1}^n\left(\bar{N}^i_{\bar{N}}-N^i\right)_+>\frac{7}{4}\log(nN)\sum_{j:\,\frac{N^j}{6}<\log(nN)}\sqrt{N^j}\right)\leq\frac{1}{N}.\quad(71)$$

Proof. First, define $\epsilon=7[\log(nN)/N]\sum_{j\in A}\sqrt{N^j}$ where we consider the following set

$$A=\left\{i\in[n]\colon N^i<6\log(nN)\right\}.$$

If $A\neq\emptyset$, for all $i\in A$ take

$$\alpha_i=\frac{\sqrt{\pi_i}}{\sum_{j\in A}\sqrt{\pi_j}}.$$

Using the union bound, we get

$$\mathbb{P}\left(\sum_{i=1}^n\left(\bar{N}^i_{\bar{N}}-N^i\right)_+>\frac{N\epsilon}{4}\right)\leq\sum_{i\in[n]\setminus A}\mathbb{P}\left(\bar{N}^i_{\bar{N}}\geq N^i\right)+\sum_{i\in A}\mathbb{P}\left(\bar{N}^i_{\bar{N}}\geq N^i+\frac{\alpha_i N\epsilon}{4}\right).\quad(72)$$

If $i \in [n] \setminus A$, then applying the Chernoff bound gives

$$\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i\right) \leq \exp\left(-\frac{\pi_i(N - \bar{N})}{1 + 2\bar{N}/(N - \bar{N})}\right) \leq \exp\left(-\frac{\pi_i N}{6}\right) \leq \frac{1}{nN}. \tag{73}$$

If $i \in A$, then applying the Bernstein inequality gives

$$\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i + \frac{\alpha_i N\epsilon}{4}\right) \leq \exp\left(-\frac{(\alpha_i N\epsilon/4)^2}{2\bar{N}\pi_i(1 - \pi_i) + \alpha_i N\epsilon/6}\right). \tag{74}$$

Moreover, we can suppose that $n \geq 2$ otherwise (71) immediately holds. Thus, $nN \geq 4$ and using the fact that $2\bar{N} \leq N$ we deduce that

$$4\left(\frac{2}{3} + \sqrt{\frac{2\bar{N}}{N\log(nN)}}\right) \leq 7.$$

Therefore, it follows

$$N\epsilon \geq 4\log(nN)\left(\frac{2}{3} + \sqrt{\frac{2\bar{N}}{N\log(nN)}}\right)\sum_{j \in A}\sqrt{N^j}.$$

The last inequality is enough to ensure that

$$\frac{(N\epsilon)^2/8}{4(\bar{N}/N)(\sum_{j \in A}\sqrt{N^j})^2 + N\epsilon\sum_{j \in A}\sqrt{N^j}/3} \geq \log(nN). \tag{75}$$

Moreover, using the definition of α_i implies

$$\frac{(\alpha_i N\epsilon/4)^2}{2\bar{N}\pi_i(1 - \pi_i) + \alpha_i N\epsilon/6} \geq \frac{(N\epsilon)^2/8}{4(\bar{N}/N)(\sum_{j \in A}\sqrt{N^j})^2 + N\epsilon\sum_{j \in A}\sqrt{N^j}/3}.$$

Hence, combining (74) with (75) shows that

$$\mathbb{P}\left(\bar{N}^i_{\bar{N}} \geq N^i + \frac{\alpha_i N\epsilon}{4}\right) \leq \frac{1}{nN}. \tag{76}$$

Finally, plugging (73) and (76) into (72) yields the result. $\qquad\square$

Lemma C.17. *Recall that* $\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}^\star_{N^\star+1}}}\}_{k \in [\bar{N}]}$ *is defined in* (60) *and for the sake of simplicity define* $Y_{\bar{N}+1} = \widehat{Y}^\star_{N^\star+1}$. *If* $\|\widehat{w}\|_\infty < \infty$, *then*

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\right] \leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\left(\mathrm{Var}(\widehat{w}_{Y_1}) + \frac{\mathrm{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}}\right).$$

Moreover, if the random variables $(\widehat{w}_{Y_k} - \mathbb{E}\widehat{w}_{Y_k})_{k \in [\bar{N}+1]}$ *are* σ-*subGaussian with parameter* $\sigma \geq 0$, *then*

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\right] \leq \frac{\sigma\sqrt{8\log(\bar{N}+1)}}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{2}{\bar{N}}\left(1 \vee \frac{\mathbb{E}\widehat{w}_{Y_{\bar{N}+1}}}{\mathbb{E}\widehat{w}_{Y_1}} + \frac{2\,\mathrm{Var}(\widehat{w}_{Y_1})}{(\mathbb{E}\widehat{w}_{Y_1})^2} + \frac{2\,\mathrm{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\right).$$

Proof. By definition of the probabilities $\{\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\}_{k \in [\bar{N}+1]}$, we have

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\{\bar{p}_{Y_k}^{\bar{D}_{Y_{\bar{N}+1}}}\}\right] = \mathbb{E}\left[\frac{\max_{k=1}^{\bar{N}+1}\{\widehat{w}_{Y_k}\}}{\sum_{l=1}^{\bar{N}+1}\widehat{w}_{Y_l}}\right].$$

The expectation is split by introducing $\mathcal{A}_{N+1} = \{\sum_{l=1}^{\bar{N}+1}\widehat{w}_{Y_l} \leq (\bar{N}\mathbb{E}\widehat{w}_{Y_1} + \mathbb{E}\widehat{w}_{Y_{\bar{N}+1}})/2\}$, we obtain

$$\mathbb{E}\left[\frac{\max_{k=1}^{\bar{N}+1}\{\widehat{w}_{Y_k}\}}{\sum_{l=1}^{\bar{N}+1}\widehat{w}_{Y_l}}\right] \leq \frac{2}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}^c_{N+1}} \cdot \max_{k=1}^{\bar{N}+1}\{\widehat{w}_{Y_k}\}\right] + \mathbb{P}(\mathcal{A}_{N+1}). \tag{77}$$

Using the Chebyshev's inequality it follows that

$$\mathbb{P}\left(\mathcal{A}_{N+1}\right) = \mathbb{P}\left(2\sum_{k=1}^{\bar{N}+1}\widehat{w}_{Y_k} < \bar{N}\mathbb{E}\widehat{w}_{Y_1} + \mathbb{E}\widehat{w}_{Y_{\bar{N}+1}}\right)$$

$$\leq \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1} + \mathbb{E}\widehat{w}_{Y_{\bar{N}+1}}/\bar{N})^2}\left(\text{Var}(\widehat{w}_{Y_1}) + \frac{\text{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}}\right). \quad (78)$$

When $\widehat{w}$ is bounded, combining (77) with (78) gives

$$\mathbb{E}\left[\max_{k=1}^{\bar{N}+1}\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\}\right] \leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \mathbb{P}\left(\mathcal{A}_{N+1}\right)$$

$$\leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\left(\text{Var}(\widehat{w}_{Y_1}) + \frac{\text{Var}(\widehat{w}_{Y_{\bar{N}+1}})}{\bar{N}}\right).$$

Otherwise, if we suppose that the $(\widehat{w}_{Y_k} - E\widehat{w}_{Y_k})_{k\in[\bar{N}+1]}$ are σ-sub-Gaussian, then applying the result given in (Boucheron et al., 2013, Section 2.5) combined with (77)-(78) concludes the proof since it follows that $\mathbb{E}[\max_{k=1}^{\bar{N}+1}\{\widehat{w}_{Y_k}\}] \leq \max_{k=1}^{\bar{N}+1}\{\mathbb{E}\widehat{w}_{Y_k}\} + \sigma\sqrt{2\log(\bar{N}+1)}$. $\qquad\square$

Finally, for any point $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, define the following measure and prediction set

$$\widehat{\mu}_{\mathbf{y}} = \widehat{p}_{\mathbf{y}}^{\widehat{D}_{\mathbf{y}}}\delta_1 + \sum_{i=1}^n\sum_{k=1}^{N^i\wedge\bar{N}_{\bar{N}}^i}\widehat{p}_{Y_k^i}^{\widehat{D}_{\mathbf{y}}}\delta_{V_k^i}$$

$$\mathcal{C}_{\alpha,\widehat{\mu}}(\mathbf{x}) = \{\mathbf{y}\in\mathcal{Y}\colon V(\mathbf{x},\mathbf{y})\leq Q_{1-\alpha}(\widehat{\mu}_{\mathbf{y}})\}.$$

Theorem C.18. *For any $i\in[n]$, let $\pi_i = N^i/N$ and consider $\bar{N} = \lfloor N/2 \rfloor$. If $\|\widehat{w}^\star\|_\infty < \infty$, then it holds*

$$\left|\mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star)\right) - 1 + \alpha\right| \leq \frac{1}{2}\sum_{y\in\mathcal{Y}}\left|P_Y^\star(y) - \frac{\widehat{w}_y^\star P_Y^{\text{cal}}(y)}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\text{cal}}(\tilde{y})}\right|$$

$$+ \frac{6}{N} + \frac{36\|\widehat{w}\|_\infty^2}{N(\mathbb{E}\widehat{w}_{Y_1})^2} + \frac{2\log N}{N}\left(\frac{3\|\widehat{w}\|_\infty^2}{(\mathbb{E}\widehat{w}_{Y_1}^\star)^2}\vee 7\sum_{j\colon\frac{N^j}{12}<\log N}\sqrt{N^j}\right).$$

Proof. Using Lemma C.5 implies that

$$\left|\mathbb{P}\left(Y_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(X_{N^\star+1}^\star)\right) - \mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star)\right)\right| \leq \frac{1}{2}\sum_{y\in\mathcal{Y}}P_Y^{\text{cal}}(y)\left|w_y^\star - \frac{\widehat{w}_y^\star}{\sum_{\tilde{y}\in\mathcal{Y}}\widehat{w}_{\tilde{y}}^\star P_Y^{\text{cal}}(\tilde{y})}\right|. \quad (79)$$

Since $\{\widehat{w}_y^\star\}_{y\in\mathcal{Y}}$ are bounded, we deduce that $\widehat{w}_{Y_1}^\star$ is σ-subGaussian with $\sigma = 2^{-1}(\max_{y\in\mathcal{Y}}\{\widehat{w}_y^\star\} - \min_{y\in\mathcal{Y}}\{\widehat{w}_y^\star\})$. Therefore, the inequality derived in Lemma C.16 combined with Lemma C.15 provide an upper bound on $\mathbb{P}(|\delta| > \epsilon)$ with

$$\epsilon = \frac{8\sigma^2\log N}{\bar{N}\mathbb{E}[\widehat{w}_{Y_1}^\star]^2}\vee\frac{7\log(nN)\sum_{j\colon\frac{N^j}{6}<\log(nN)}\sqrt{N^j}}{N}.$$

Plugging this result into the bound derived in Lemma C.14 shows that

$$\left|\mathbb{P}\left(\widehat{Y}_{N^\star+1}^\star \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}_{N^\star+1}^\star)\right) - 1 + \alpha\right| \leq \mathbb{E}\left[\bar{p}_{\widehat{Y}_{N^\star+1}^\star}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\vee\max_{k=1}^{\bar{N}}\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\}\right] + \frac{2}{N} + \frac{4\,\text{Var}\left(\widehat{w}_{Y_1}^\star\right)}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1}^\star)^2}$$

$$+ \frac{8\sigma^2\log N}{\bar{N}\mathbb{E}[\widehat{w}_{Y_1}^\star]^2}\vee\frac{7\log(nN)\sum_{j\colon\frac{N^j}{6}<\log(nN)}\sqrt{N^j}}{N}. \quad (80)$$

Moreover, applying Lemma C.17, we deduce that

$$\mathbb{E}\left[\bar{p}_{\widehat{Y}_{N^\star+1}^\star}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\vee\max_{k=1}^{\bar{N}}\{\bar{p}_{Y_k}^{\bar{D}_{\widehat{Y}_{N^\star+1}^\star}}\}\right] \leq \frac{2\|\widehat{w}\|_\infty}{\bar{N}\mathbb{E}\widehat{w}_{Y_1}} + \frac{4}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\left(\text{Var}(\widehat{w}_{Y_1}) + \frac{\text{Var}(\widehat{w}_{\widehat{Y}_{N^\star+1}^\star})}{\bar{N}}\right). \quad (81)$$

Therefore, combining (80) with (81) implies that

$$\left| \mathbb{P}\left(\widehat{Y}^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^{\star}_{N^{\star}+1})\right) - 1 + \alpha \right| \leq \frac{2}{\bar{N}}\left(\frac{\bar{N}}{N} + \frac{\|\widehat{w}\|_{\infty}}{\mathbb{E}\widehat{w}_{Y_1}} + \frac{4\operatorname{Var}\left(\widehat{w}^{\star}_{Y_1}\right)}{(\mathbb{E}\widehat{w}^{\star}_{Y_1})^2} + \frac{2\operatorname{Var}(\widehat{w}_{\widehat{Y}^{\star}_{N^{\star}+1}})}{\bar{N}(\mathbb{E}\widehat{w}_{Y_1})^2}\right)$$

$$+ \frac{8\sigma^2 \log N}{\bar{N}(\mathbb{E}\widehat{w}^{\star}_{Y_1})^2} \vee \frac{14 \log N \sum_{j\,:\,\frac{N^j}{6} < \log(nN)} \sqrt{N^j}}{N}.$$

If $N \geq 6$, then remark that $N/\bar{N} \leq 3$. Thus, we deduce that

$$\left| \mathbb{P}\left(\widehat{Y}^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^{\star}_{N^{\star}+1})\right) - 1 + \alpha \right| \leq \frac{6}{N}\left(1 + \frac{\|\widehat{w}\|_{\infty}}{\mathbb{E}\widehat{w}_{Y_1}} + \frac{4\operatorname{Var}\left(\widehat{w}^{\star}_{Y_1}\right)}{(\mathbb{E}\widehat{w}^{\star}_{Y_1})^2} + \frac{\operatorname{Var}(\widehat{w}_{\widehat{Y}^{\star}_{N^{\star}+1}})}{(\mathbb{E}\widehat{w}_{Y_1})^2}\right)$$

$$+ \frac{24\sigma^2 \log N}{N(\mathbb{E}\widehat{w}^{\star}_{Y_1})^2} \vee \frac{14 \log N \sum_{j\,:\,\frac{N^j}{12} < \log N} \sqrt{N^j}}{N}. \quad (82)$$

Finally, using the next inequality combined with (79) and (82) concludes the proof

$$\mathbb{P}(Y^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^{\star}_{N^{\star}+1})) = \mathbb{P}(Y^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^{\star}_{N^{\star}+1})) - \mathbb{P}(\widehat{Y}^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^{\star}_{N^{\star}+1})) + \mathbb{P}(\widehat{Y}^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(\widehat{X}^{\star}_{N^{\star}+1})).$$

$\square$

The proof of the following result is similar to that of Theorem C.18.

Theorem C.19. *For any $i \in [n]$, let $\pi_i = N^i/N$ and set $\bar{N} = \lfloor N/2 \rfloor$. Moreover, for $y \in \mathcal{Y}$ consider $\widehat{w}^{\star}_y = w^{\star}_y$. If $\|w^{\star}\|_{\infty} < \infty$, then it holds*

$$\left| \mathbb{P}\left(Y^{\star}_{N^{\star}+1} \in \mathcal{C}_{\alpha,\widehat{\mu}}(X^{\star}_{N^{\star}+1})\right) - 1 + \alpha \right| \leq \frac{6}{N} + \frac{36 + 6\log N}{N}\|w^{\star}\|^2_{\infty} + \frac{14 \log N}{N} \sum_{j\,:\,\frac{N^j}{12} < \log N} \sqrt{N^j}.$$

Proof. First, for $y \in \mathcal{Y}$ recall that $w^{\star}_y = (P^{\star}_Y/P^{\text{cal}}_Y)(y)$. Thus, we remark that $\mathbb{E}w^{\star}_{Y_1} = 1$. Therefore, applying Theorem C.18 concludes the proof. $\square$

D. Differential privacy guarantee: proof of Theorem 4.4

In this section, we recall the definition of being (ϵ, δ)-DP (Dwork et al., 2014). The idea behind differential privacy is to ensure that no attacker can determine with high confidence whether a particular individual's data is included in the dataset or not. Often, a controlled amount of random noise is added to the data, so that any individual data point becomes indistinguishable from the noise. This ensures that the probability distribution of an algorithm's output does not change significantly when a single individual's data is added or removed.

Definition D.1. For any $\epsilon > 0$ and $\delta \in [0, 1)$, a randomized mechanism $\mathcal{A}$ is said to be (ϵ, δ)-DP, if for all neighboring datasets D, D', and for any event E:

$$\mathbb{P}\left(\mathcal{A}(D) \in E\right) \leq \exp(\epsilon)\mathbb{P}\left(\mathcal{A}(D') \in E\right) + \delta.$$

The following result gives the noise level sufficient to ensure the (ϵ, δ)-DP regarding the third-party attacker. This type of attacker is an external entity who does not have access to the private data but can observe the algorithm outputs. This attacker tries to infer sensitive information about individuals by analyzing the output.

Theorem D.2. *Assume there is a constant number $S \in [n]$ of sampled agents, i.e., $S_t = S$, for all $t \in [T]$. Then, for all $\epsilon > 0$ and $\delta \in (0, 1 - (1 + \sqrt{\epsilon})(1 - S/n)^T)$, the Algorithm 1 is (ϵ, δ)-DP towards a third party if*

$$\sigma_g \geq 2\sqrt{\frac{K \max_{i\in[n]} \lambda^i_{\mathbf{y}}}{\epsilon}\left(1 + \frac{24S\sqrt{T}\log(1/\bar{\delta})}{\epsilon n}\right)}, \qquad \bar{\delta} = \frac{n}{S}\left[1 - \left(\frac{1 - \delta}{1 + \sqrt{\epsilon}}\right)^{1/T}\right].$$

Proof. The loss function $\nabla S^{i;\gamma}_{\alpha}$ has a sensitivity of 1. Therefore, for any $\tilde{\alpha} > 1$, we know that $\nabla S^{i;\gamma}_{\alpha} + \mathcal{N}(0, \sigma^2_g)$ is

$(\tilde{\alpha}, \tilde{\alpha}/2\sigma_g^2)$-RDP (Mironov, 2017, Corollary 3). By assumption, note that $\bar{\delta} \in (0, 1)$ and for any $t \in [T]$, consider

$$\epsilon_t = \frac{nK\tilde{\alpha} \max_{i \in [n]} \lambda_{\mathbf{y}}^i}{2(t+1)S\sigma_g^2} - \frac{\log \bar{\delta}}{\tilde{\alpha} - 1},$$

$$\tilde{\alpha} = 1 + \sigma_g \sqrt{\frac{2\sqrt{T}S \log(1/\bar{\delta})}{nK \max_{i \in [n]} \lambda_{\mathbf{y}}^i}}.$$

After K local iterations, using the RDP composition result, the mechanism becomes $(\tilde{\alpha}, K\tilde{\alpha}/2\sigma_g^2)$-RDP. Using the aggregation step on the server, the mechanism is now $(\tilde{\alpha}, nK\tilde{\alpha} \max_{i \in [n]} \lambda_{\mathbf{y}}^i/2(t+1)S\sigma_g^2)$-RDP. Based on the RDP to DP conversion, we know that the mechanism is $(\epsilon_t, \bar{\delta})$-DP. Define $f \colon x \in \mathbb{R} \mapsto \log\{1 + (S/n)(\mathrm{e}^x - 1)\} \in \mathbb{R}$. For any $x \in \mathbb{R}$, we have

$$f'(x) = \frac{S}{n + 2x(n - S)}.$$

In addition, since the agents are subsampled, it yields the $(\tilde{\epsilon}_t, \tilde{\delta}_t)$-DP (Balle et al., 2018, Theorem 9), where we denote

$$\tilde{\epsilon}_t = f(\epsilon_t), \qquad\qquad\qquad \tilde{\delta}_t = \frac{S\bar{\delta}}{n}.$$

Since $f'(x) \leq S/n$ on $\mathbb{R}_+$, we deduce that

$$\tilde{\epsilon}_t \leq \frac{K\tilde{\alpha} \max_{i \in [n]} \lambda_{\mathbf{y}}^i}{2(t+1)\sigma_g^2} - \frac{S \log \bar{\delta}}{n(\tilde{\alpha} - 1)}.$$

For all $a, b \in \mathbb{R}$, using that $(a + b)^2 \leq 2a^2 + 2b^2$ gives that

$$\sum_{t=1}^{T} \tilde{\epsilon}_t^2 \leq \frac{2\left(K\tilde{\alpha} \max_{i \in [n]} \lambda_{\mathbf{y}}^i\right)^2}{4\sigma_g^4} \sum_{t=2}^{T+1} \frac{1}{t^2} + \frac{2TS^2 \log(1/\bar{\delta})^2}{n^2(\tilde{\alpha} - 1)^2}. \tag{83}$$

Plugging the definition of $\tilde{\alpha}$ into (83) combined with $\sum_{t=2}^{T+1} t^{-2} \leq 1$ show that

$$\sum_{t=1}^{T} \tilde{\epsilon}_t^2 \leq \frac{\left(K \max_{i \in [n]} \lambda_{\mathbf{y}}^i\right)^2}{\sigma_g^4} \left(1 + \frac{3\sigma_g^2 \sqrt{T}S \log(1/\bar{\delta})}{nK \max_{i \in [n]} \lambda_{\mathbf{y}}^i}\right)$$

$$= \frac{\left(K \max_{i \in [n]} \lambda_{\mathbf{y}}^i\right)^2}{\sigma_g^4} + \frac{3\sqrt{T}KS \log(1/\bar{\delta}) \max_{i \in [n]} \lambda_{\mathbf{y}}^i}{n\sigma_g^2}. \tag{84}$$

By assumption, recall that

$$\sigma_g \geq 2\sqrt{\frac{K \max_{i \in [n]} \lambda_{\mathbf{y}}^i}{\epsilon}} \sqrt{1 + \frac{24\sqrt{T}S \log(1/\bar{\delta})}{\epsilon n}}.$$

Therefore, we obtain

$$\frac{\sigma_g^4 \epsilon^2}{16} \geq \left(K \max_{i \in [n]} \lambda_{\mathbf{y}}^i\right)^2 + \frac{3\sigma_g^2 \sqrt{T}SK \log(1/\bar{\delta}) \max_{i \in [n]} \lambda_{\mathbf{y}}^i}{n}.$$

The previous inequality combined with (84) implies that

$$\sum_{t=1}^{T} \tilde{\epsilon}_t^2 \leq \epsilon^2/16.$$

Define the following quantity

$$\tilde{\delta} = \frac{1 - \delta}{\prod_{t=1}^{T}(1 - \tilde{\delta}_t)} - 1.$$

Since $1 - (1 - \delta)^{1/T} \leq \tilde{\delta}_t \leq 1 - [(1 - \delta)/2]^{1/T}$, we can verify that $\tilde{\delta} \in [0, 1]$ and also $\tilde{\delta} = \sqrt{\epsilon}$. Using the assumption

$\delta \leq 1 - (1 + \sqrt{\epsilon})(1 - S/n)^T$, it yields

$$\mathrm{e} + \tilde{\delta}^{-1} \sqrt{\sum_{t=1}^{T} \epsilon_t^2} \leq \mathrm{e} + 1 \leq \mathrm{e}^2.$$

Thus, we have

$$2 \sum_{t=1}^{T} \tilde{\epsilon}_t^2 \log \left(\mathrm{e} + \tilde{\delta}^{-1} \sqrt{\sum_{t=1}^{T} \epsilon_t^2} \right) \leq \frac{\epsilon^2}{4}.$$

Since $(\exp(\tilde{\epsilon}_t) - 1)/(\exp(\tilde{\epsilon}_t) + 1) \leq \tilde{\epsilon}_t$, we deduce that

$$\sum_{t=1}^{T} \frac{\exp(\tilde{\epsilon}_t) - 1}{\exp(\tilde{\epsilon}_t) + 1} \leq \sum_{t=1}^{T} \tilde{\epsilon}_t^2 \leq \frac{\epsilon^2}{16}.$$

Finally, applying (Kairouz et al., 2015, Theorem 3.5) concludes the proof. □

Note that the local loss function $S_\alpha^{i,\gamma} \colon q \in \mathbb{R} \mapsto \mathbb{E}_{V \sim \widehat{\mu}_y^i}[S_{\alpha,V}^{\gamma}(q)] \in \mathbb{R}_+$ is expressed as the expectation of pinball loss functions. Since the sensitivity of these pinball loss functions is 1, we do not need to clip the gradient. It is sufficient to add additional Gaussian noise $\mathcal{N}(0, \sigma_g)$ to guarantee differential privacy. The value of σ_g is chosen to provide a suitable trade-off between privacy and utility, balancing the need for strong privacy protection with the requirement for useful output.

E. Additional numerical results

E.1. Algorithm design

The objective is to generate valid federated prediction sets for the testset by leveraging the calibration datasets of the agents. In this section, we present four different algorithms that were compared in our experiments. The first two are unweighted algorithms (`Unweighted Local` and `Unweighted Global`), while the other two are weighted benchmarks: `Oracle Weights` which uses true weights, and `Estimated Weights` which employs estimated weights. Additionally, we propose the `DP-FedCP` method, available in both basic and differentially private versions. The main differences between these approaches lie in the way the resulting quantile is computed, such as the importance given to the set of non-conformity scores and their corresponding weights. Moreover, the approaches' ability to maintain coverage and privacy guarantees varies.

Unweighted Local. The Unweighted method assigns equal weight to all non-conformity scores, regardless of the agent or label. As a result, the resulting quantile and prediction sets are only influenced by the local scores of the querying agent. Specifically, the `Unweighted Local` approach only employs the local calibration dataset of Agent $\star$ to compute the corresponding prediction set. This approach is easily computable because it does not involve any exchange of information between agents. Formally, the confidence set can be expressed as follows:

$$\bar{\mu}^{\mathrm{loc},\star} = \frac{1}{N^\star + 1} \sum_{k=1}^{N^\star} \delta_{V_k^\star} + \frac{1}{N^\star + 1} \delta_1,$$

$$\mathcal{C}_{\alpha, \bar{\mu}^{\mathrm{loc},\star}}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha} \left(\bar{\mu}^{\mathrm{loc},\star} \right) \right\}.$$

Unweighted Global. The `Unweighted Global` approach calculates the quantile by using all non-conformity scores gathered on the central server from all agents. This approach violates FL constraints since all non-conformity scores are shared with the central server. Furthermore, each collected score is given equal weight, without taking into account any

label shift. The corresponding prediction set is represented as follows:

$$\bar{\mu} = \frac{1}{N+1} \sum_{i=1}^{n} \sum_{k=1}^{N^i} \delta_{V_k^i} + \frac{1}{N+1} \delta_1,$$

$$\mathcal{C}_{\alpha,\bar{\mu}}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu})\}.$$

Oracle Weights. This approach utilize importance weights to compute the prediction set. The `Oracle Weights` method has access to the true distribution of all the agents, enabling it to calculate the exact likelihood ratios. For this method, a number of $\bar{N} = \lfloor N/2 \rfloor$ data points of the calibration dataset is randomly selected — denoted as $(X_k, Y_k)_{k \in [\bar{N}]}$. This sumbsampling is based on a multinomial random variable with parameter $(\bar{N}, \{N^i/N\}_{i \in [n]})$; more details are provided in Section 2 (see for example Theorem 2.2). For any label $y \in \mathcal{Y}$, the prediction set is determined by:

$$\bar{\mu}_{\mathbf{y}}^\star = \bar{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \sum_{k=1}^{\bar{N}} \bar{p}_{Y_k,\mathbf{y}}^\star \delta_{V_k}, \tag{85}$$

$$\mathcal{C}_{\alpha,\bar{\mu}^\star}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}(\bar{\mu}_{\mathbf{y}}^\star)\right\}.$$

Estimated Weights. The empirical equivalent of `Oracle Weights` based on client label counts is `Estimated Weights`. However, two sources of error are introduced: (1) the calibration subsampling and (2) the likelihood ratio estimations (refer to (10)). Similar to `Oracle Weights`, we draw a multinomial distribution $\{\bar{N}^i\}_{i \in [n]} \sim \mathcal{M}(\bar{N}, \{N^i/N\}_{i \in [n]})$ and subsample $N^i \wedge \bar{N}^i$ calibration data from each client i. The resulting prediction set is represented by:

$$\widehat{p}_{y,\mathbf{y}}^\star = \frac{(\mathsf{M}_y^\star/\mathsf{M}_y) \cdot \mathbb{1}_{\mathsf{M}_y \geq 1}}{(\mathsf{M}_{\mathbf{y}}^\star/\mathsf{M}_{\mathbf{y}}) \cdot \mathbb{1}_{\mathsf{M}_{\mathbf{y}} \geq 1} + \sum_{\tilde{y} \in \mathcal{Y}} \mathrm{Card}(\{(i,k) \in [n] \times [N^i] \colon k \leq \bar{N}^i, Y_k^i = y\}) \times (\mathsf{M}_{\tilde{y}}^\star/\mathsf{M}_{\tilde{y}}) \cdot \mathbb{1}_{\mathsf{M}_{\tilde{y}} \geq 1}},$$

$$\widehat{\mu}_{\mathbf{y}}^{\text{\tiny MLE}} = \widehat{p}_{\mathbf{y},\mathbf{y}}^\star \delta_1 + \sum_{i=1}^{n} \sum_{k=1}^{N^i \wedge \bar{N}^i} \widehat{p}_{Y_k^i,\mathbf{y}}^\star \delta_{V_k^i},$$

$$\mathcal{C}_{\alpha,\widehat{\mu}^{\text{\tiny MLE}}}(\mathbf{x}) = \left\{\mathbf{y} \colon V(\mathbf{x}, \mathbf{y}) \leq Q_{1-\alpha}\left(\widehat{\mu}_{\mathbf{y}}^{\text{\tiny MLE}}\right)\right\}.$$

E.2. Additional Details on Numerical Experiments

To provide a comprehensive overview of our experimental methodology, we present additional details about the federated optimization parameters, along with supplementary figures to complement those presented in Section 5.

Optimization parameters. For all experiments, we split the initial dataset $\mathcal{D}$ across the clients $(\mathcal{D}_i)_{i \in [n]}$ and the test dataset $\mathcal{D}_{\text{test}}$ as detailed in Section 5. Label shift is then simulated by resampling using the clients' local distributions $\{P^y\}_{y \in \mathcal{Y}}$. For ImageNet experiments, we also assume that we have labels sampled from the target client's distribution for weights' approximation. The optimization parameters taken for `DP-FedCP` experiments are $T = 200$ iterations, $\gamma = 1e^{-6}$ regularization parameter, $\eta = 1e^{-3}$ stepsize, and $K = 20$ local iteration rounds. We sample all clients during each communication round with the server.

Simulated Data Experiments. The generated data consists of 3 Gaussians, two of which significantly overlap each other, see Figure 4. This data design is chosen such that we obtain different distributions of non-conformity scores for different classes, which is directly related to the different degrees of model confidence for different data samples.

CIFAR-10 Experiments. Using the CIFAR-10 data, we demonstrate a comparison of the empirical coverage for all considered methods: `Unweighted Local`, `Unweighted Global`, `Oracle Weights`, `Estimated Weights` and the proposed `DP-FedCP` method version with ($\sigma_g = 0$), see Figure 5a. `DP-FedCP` along with weighted baselines, shows valid coverage results, unlike unweighted baselines. At the same time, both weighted algorithms are extremely similar in performance to `DP-FedCP`. Unlike weighted baselines, `DP-FedCP` is federated and privacy preserving.

ImageNet Experiments. The ImageNet experiment is designed to have very different score distributions across agents. The
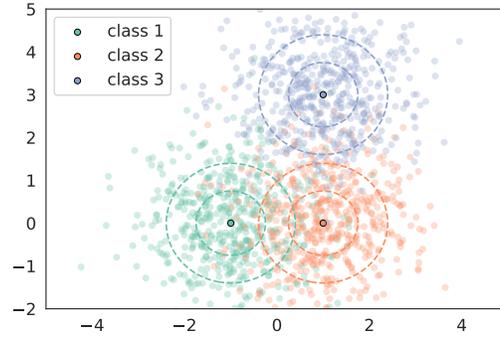
Figure 4: Simulated data distribution: 3 two-dimensional Gaussians with means $\boldsymbol{\theta}_1 = [-1, 0], \boldsymbol{\theta}_2 = [1, 0], \boldsymbol{\theta}_3 = [1, 3]$ and identity covariance matrices.

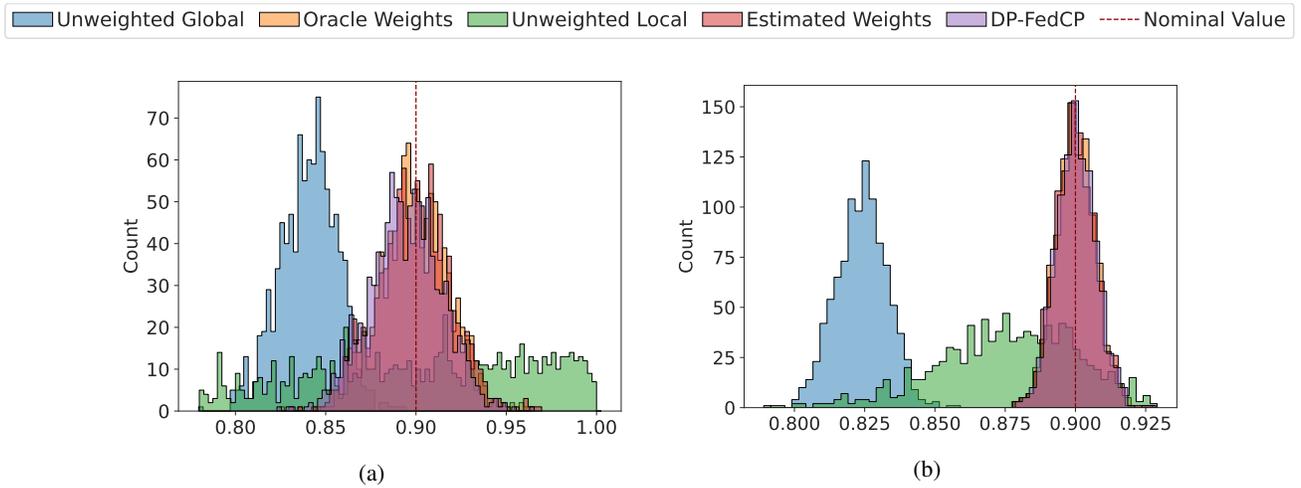

Figure 5: Experimental results with all benchmarks. (a) CIFAR-10 empirical coverage. (b) ImageNet empirical coverage.

(a) Distribution of scores for labels in G1

(b) Distribution of scores for labels in G2
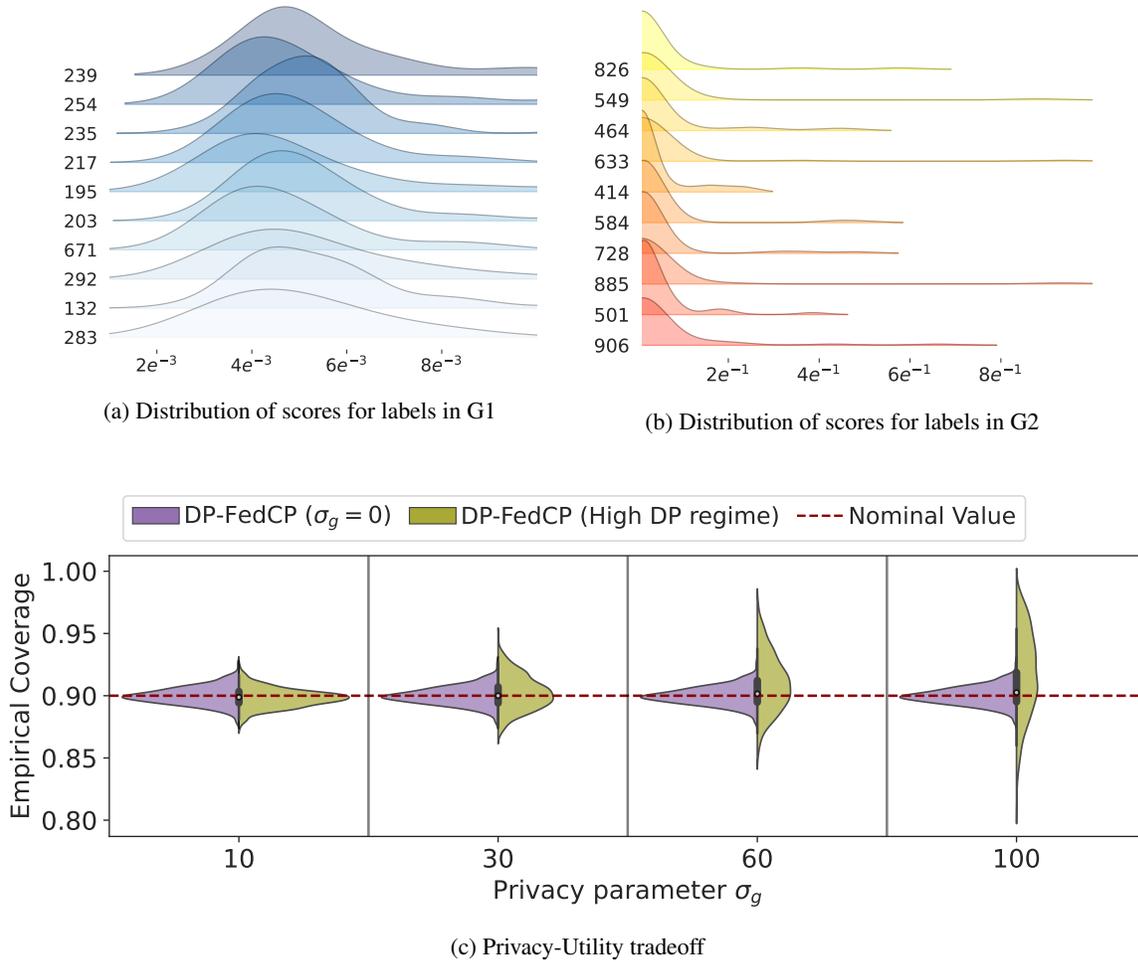

(c) Privacy-Utility tradeoff

Figure 6: ImageNet complementary results. (a) Score distributions of the classes with the lowest non-conformity scores. (b) Score distributions of the classes with the highest non-conformity scores. (c) Violin plot of the empirical coverage distribution for `DP-FedCP` with different DP regimes.

grouping scheme of clients into a `low-score` group (Figure 6a) and a `high-score` group that consists of the querying agent (Figure 6b) creates adversial heterogeneity, possible in real-life scenarios, under which unweighted methods are more prone to perform very poorly. Comparing the `DP-FedCP` method with weighted and unweighted baselines on ImageNet data, we note the same behavior as on CIFAR-10 dataset, see Figure 5b. Only algorithms that account for shifts between agents' data achieve the desired empirical coverage. There is an additional violin plot for the ImageNet differential privacy study that demonstrates the effect of the DP parameter σ_g on the resulting empirical coverage; see Figure 6c.