Efficient Conformal Prediction under Data Heterogeneity

Vincent Plassier[1,2] **Nikita Kotelevskii**[3,6] **Aleksandr Rubashevskii**[3,6] **Fedor Noskov**[4]

Maksim Velikanov[2,5] **Alexander Fishkov**[3,6] **Samuel Horvath**[6] **Martin Takáč**[6]

Éric Moulines[2,6] **Maxim Panov**[6]

[1]Lagrange Mathematics and Computing Research Center [2]CMAP, Ecole Polytechnique, Paris
[3]Skolkovo Institute of Science and Technology [4]HSE University, Moscow
[5]Technology Innovation Institute [6]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi

Abstract

Conformal Prediction (CP) stands out as a robust framework for uncertainty quantification, which is crucial for ensuring the reliability of predictions. However, common CP methods heavily rely on the data exchangeability, a condition often violated in practice. Existing approaches for tackling non-exchangeability lead to methods that are not computable beyond the simplest examples. In this work, we introduce a new efficient approach to CP that produces provably valid confidence sets for fairly general non-exchangeable data distributions. We illustrate the general theory with applications to the challenging setting of federated learning under data heterogeneity between agents. Our method allows constructing provably valid personalized prediction sets for agents in a fully federated way. The effectiveness of the proposed method is demonstrated in a series of experiments on real-world datasets.

1 Introduction

Conformal Prediction (CP) has been shown to be a reliable method for quantifying uncertainty in machine learning models (Shafer and Vovk, 2008; Angelopoulos et al., 2023). However, the performance of prediction sets generated by CP can significantly decrease when the data is statistically heterogeneous. In particular, the presence of distributional shifts interferes with the assumption of exchangeability – an essential cornerstone of the methods of CP. The goal of this work is to develop efficient solutions that extend the CP framework to data with heterogeneous distributions.

Assume that we have N data samples $\mathcal{D}_N = \{X_i, Y_i\}_{i=1}^N$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$. Given the data $\mathcal{D}_N$, the goal of CP uncertainty quantification is to construct prediction sets $\mathcal{C}_\alpha(\mathbf{x}) \subseteq \mathcal{Y}$ for a new unseen object $\mathbf{x}$ and the desired miscoverage level $\alpha \in [0, 1]$. We say that $\mathcal{C}_\alpha$ is a valid (distribution-free) predictive interval if it holds that

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

Here the notation $\mathbb{P}$ denotes that the probability is computed with respect to $\{(X_i, Y_i)\}_{i=1}^N$ and (X_{N+1}, Y_{N+1}), which are assumed i.i.d. with common distribution P. The equation (1) essentially bounds the miscoverage rate on average over possible sets of calibration data and test points. However, if there is a high variability in the coverage probability as a function of the calibration data, the test coverage probability may be substantially below $1-\alpha$ for a particular calibration set. Therefore, for a given calibration set $\mathcal{D}_N$ the reliability of the confidence set can be assessed via the *empirical miscoverage rate*:

$$\alpha(\mathcal{D}_N) = \mathbb{P}\left(Y_{N+1} \notin \mathcal{C}_\alpha(X_{N+1}) \,|\, \mathcal{D}_N\right). \tag{2}$$

Understanding the distribution of $\alpha(\mathcal{D}_N)$ allows us to control the uncertainty. Its distribution was studied for split conformal prediction (SCP) in (Vovk, 2012,

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

Proposition 2a)), where it has been shown that:

$$\alpha(\mathcal{D}_N) \sim \mathrm{Beta}(\lceil (N+1)\alpha \rceil, \lceil (N+1)(1-\alpha) \rceil), \quad (3)$$

therefore ensuring the concentration of $\alpha(\mathcal{D}_N)$ around the target value: $\alpha(\mathcal{D}_N) \simeq \alpha + \mathrm{O}_{\mathbb{P}}(1/\sqrt{N})$, see also (Angelopoulos et al., 2023).

While classical CP methods (such as SCP) work well under data exchangeability, there are few established performance guarantees for non-exchangeable data. Among the notable exceptions, Tibshirani et al. (2019) in their seminal work developed methods that address the important case of *weighted exchangeable data.* However, their approach involves calculating certain weighted quantiles with weights that require difficult combinatorial calculations that are not feasible in practice beyond relatively simple examples. Moreover, their approach provides only the guarantees in expectation; the variability of coverage around its mean has not been addressed.

One of the most important examples of weighted-exchageable data is the case of distribution shift between calibration and test distributions (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021). However, computability issues prevent the inclusion of available data from the test distribution in these procedures, which should be advantageous in the case of domain adaptation. Another important example is federated learning in the presence of statistical heterogeneity between agents. In the first place, such heterogeneity cancels all standard guarantees of CP, since it conflicts with the principle of exchangeability, a fundamental assumption of CP. Again, the Tibshirani et al. (2019) approach can potentially be used. However, the problem of computational intractability remains.

Our work addresses these challenges through the following key contributions:

- We present a new method for constructing conformal prediction sets while accounting for heterogeneity in the data. The developed method introduces weights for the empirical distribution that can be computed efficiently, unlike those of Tibshirani et al. (2019). In addition, the validity of the prediction sets is demonstrated with high probability. For more details, see Section 2.
- We develop the application of the general method to the important practical situation of federated learning in the presence of statistical heterogeneity between agents. In particular, we propose a federated method for estimating importance weights and the resulting weighted quantile; see Section 3.
- Extensive empirical assessments are conducted using synthetic data and various benchmark computer vision datasets. The results of these experiments are discussed in Section 5.

2 Conformal Prediction under Data Heterogeneity

We consider the independent calibration data $\mathcal{D}_N = \{(X_k, Y_k)\}_{k=1}^N$, where

$$(X_k, Y_k) \sim P^k, \quad k = 1, \dots, N. \quad (4)$$

In this case, X_k represents the covariate, Y_k is the label and the distributions P^k can arbitrarily vary for different k in the general case. We also write by $Z_k = (X_k, Y_k)$ the pair of covariate and label and denote by $\mathcal{X}$ and $\mathcal{Y}$ the support of X and Y, respectively. The set $\mathcal{X}$ is often a subset of $\mathbb{R}^d$, while $\mathcal{Y}$ can be either finite in classification or a subset of $\mathbb{R}$. The goal is to construct a prediction set $\mathcal{C}_\alpha(\mathbf{x})$ for the input object $\mathbf{x}$ that conditionally on $\mathcal{D}_N$ yields a confidence level close to $1-\alpha$ with probability at least $1-\delta$, where $\alpha, \delta \in (0,1)$:

$$\mathbb{P}\left(Y_{N+1} \in \mathcal{C}_\alpha(X_{N+1}) \,|\, \mathcal{D}_N\right) = 1 - \alpha + \tau_{N,\delta}, \quad (5)$$

where the new observation (X_{N+1}, Y_{N+1}) comes from the distribution P^{N+1} and $\tau_{N,\delta}$ is a small number.

2.1 Basics of Conformal Prediction

Let us start from describing *Split Conformal Prediction (SCP)* which is de-facto standard approach in conformal prediction applications. This method assumes that the available data are split into two distinct subsets: a training dataset and a calibration dataset. A predictor $\hat{f}(\mathbf{x})$ is then trained on the training dataset, while the calibration dataset is used for generating the prediction sets. Using the predictor $\hat{f}(\mathbf{x})$, the so-called *score function* can be calculated as $V(\mathbf{x}, \mathbf{y}) = S(\hat{f}(\mathbf{x}), \mathbf{y})$. Here, the function S measures the discrepancy between the predicted label $\hat{f}(\mathbf{x})$ and the target $\mathbf{y}$. For example, in the case of regression, one can take $S(\hat{\mathbf{y}}, \mathbf{y}) = |\hat{\mathbf{y}} - \mathbf{y}|$, which leads to $V(\mathbf{x}, \mathbf{y}) = |\hat{f}(\mathbf{x}) - \mathbf{y}|$; see (Papadopoulos et al., 2011; Kato et al., 2023).

The core idea of SCP is to calculate the scores for the available calibration data and then construct the prediction set based on an appropriate quantile estimated from the empirical distribution of the scores. More precisely, the resulting confidence set is given by

$$\mathcal{C}_{\alpha,\mu}(\mathbf{x})\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \le Q_{1-\alpha}(\mu)\}, \quad (6)$$

where $\mu = \frac{1}{N+1}\delta_\infty + \frac{1}{N+1}\sum_{k=1}^N \delta_{V_k}$, $V_k = V(X_k, Y_k)$, and δ_v is the Dirac measure at $v \in \mathbb{R} \cup \{\infty\}$. Such a confidence set allows strong conformal guarantees under the assumption that all the calibration points $\{(X_k, Y_k)\}_{k=1}^N$ have the same distribution, i.e. $P^k \equiv P,\ k = 1, \ldots, N$:

$$0 \le \mathbb{P}\big(Y_{N+1} \in \mathcal{C}_{\alpha,\mu}(X_{N+1})\big) - 1 + \alpha < \tfrac{1}{N+1}, \qquad (7)$$

for the data point (X_{N+1}, Y_{N+1}) generated from the same distribution P.

Since the prediction sets are generated using a fixed calibration data set, some of them could lead to suboptimal decisions with test coverage much less than $1-\alpha$. This limitation can be particularly problematic in real-life scenarios where users work with a fixed calibration dataset. To enhance prediction reliability, it is essential to validate the coverage under most circumstances, ensuring its validity with a high probability, regardless of the specific calibration dataset (Vovk, 2012; Bian and Barber, 2023; Humbert et al., 2023).

CP applications beyond exchangeable cases. There exist multiple scenarios when standard SCP approach becomes not applicable as the data becomes non-exchageable. In particular, one can consider:

1. **Distribution shift between calibration and test data.** In this scenario, the calibration points are drawn as $(X_k, Y_k) \sim P^{\text{cal}}$, $k \in [N]$, whereas the new test point satisfies $(X_{N+1}, Y_{N+1}) \sim P^{\text{test}}$. If one knows the likelihood ratio $\mathrm{d}P^{\text{test}}/\mathrm{d}P^{\text{cal}}(\mathbf{x}, \mathbf{y})$, then the adaptation to distribution shift can be performed (Tibshirani et al., 2019).

2. **Domain adaptation.** In this scenario, in addition to the calibration points from a source distribution $(X_k, Y_k) \sim P^{\text{cal}}, k \in [N]$ one also has points from the test distribution $(X_{N+m}, Y_{N+m}) \sim P^{\text{test}}$, $m \in [M]$ with typically $M \ll N$. If the new test point satisfies $(X_{N+M+1}, Y_{N+M+1}) \sim P^{\text{test}}$, one can expect that the usage of both sets of calibration points should improve the resulting coverage compared to usage of only one of the sets.

3. **Federated learning with statistical heterogeneity between agents.** In a federated learning setting, 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. within client and that the statistical heterogeneity is due to the difference between client data distribution:

$$\forall k \in [N^i], \quad (X_k^i, Y_k^i) \sim P^i.$$

If one wants to perform personalized CP procedure towards the distribution of one of the agents $\star \in [n]$, the shifts between the distribution of this agent and the others should be taken into account.

These examples are the particular case of the independent but not identically distributed data considered in (4). In what follows we are going to present the general approach to CP that covers all these scenarios and allows for the efficient implementation.

Conformal prediction under covariate shift. In this work we focus on the particular case of covariate shift, while label shift or more general types of shifts can be considered in a similar way. The general approach for dealing with heterogeneous data distributions (due to covariate shift) was proposed by Tibshirani et al. (2019). The key idea is to account for the covariate shift by introducing *density ratios* $w_{\mathbf{x}}$ that quantify the shift for the particular input point $\mathbf{x}$. Finally, one can construct valid conformal confidence sets by considering the weighted quantile:

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

where $\bar{\mu}_{\mathbf{x}} = \bar{p}_{\mathbf{x},\mathbf{x}}\delta_\infty + \sum_{k=1}^N \bar{p}_{X_k,\mathbf{x}}\delta_{V_k}$. Here the weights $\bar{p}_{x,x'}$ depend on the density ratios w_x via a complex combinatorial formula that requires summation over $N!$ elements. This makes the method inapplicable in practice and motivates the need for alternative approaches.

2.2 Efficient Conformal Prediction under Data Shift

In this paper, we propose a general and efficiently computable approach for conformal prediction allowing for heterogeneity of the data. Consider the calibration measure $P^{\text{cal}} = \frac{1}{N}\sum_{k=1}^N P^k$ and the test measure $P^{\text{test}} \equiv P^{N+1}$. We propose to introduce density ratios of the following form:

$$\lambda(\mathbf{x}, \mathbf{y}) = \tfrac{\mathrm{d}P^{\text{test}}}{\mathrm{d}P^{\text{cal}}}(\mathbf{x}, \mathbf{y}) \qquad (8)$$

assuming that P^{test} is absolutely continuous with respect to P^{cal}, i.e., $P^{\text{test}} \ll P^{\text{cal}}$. The density ratios are given by $\lambda_k = \lambda(X_k, Y_k)$, where $(X_k, Y_k) \sim P^k$. In the following, we consider the prediction set defined by

$$\mathcal{C}_{\alpha,\mu}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x}, \mathbf{y}) \le Q_{1-\alpha}(\mu_{\mathbf{x},\mathbf{y}})\}, \qquad (9)$$

where the probability measure and the importance weights are given, with $V_k = V(X_k, Y_k)$, by

$$\mu_{\mathbf{x},\mathbf{y}} = p_{N+1}^{(\mathbf{x},\mathbf{y})}\delta_\infty + \sum_{k=1}^N p_k^{(\mathbf{x},\mathbf{y})}\delta_{V_k}, \qquad (10)$$

$$p_k^{(\mathbf{x},\mathbf{y})} = \tfrac{\lambda_k}{\lambda(\mathbf{x},\mathbf{y}) + \sum_{l=1}^N \lambda_l},\ p_{N+1}^{(\mathbf{x},\mathbf{y})} = \tfrac{\lambda(\mathbf{x},\mathbf{y})}{\lambda(\mathbf{x},\mathbf{y}) + \sum_{l=1}^N \lambda_l}. \qquad (11)$$

Interestingly, although these weights may appear similar to those proposed by Tibshirani et al. (2019), they are quite different. In fact, the importance weights proposed by Tibshirani et al. (2019, Lemma 3) rely on intractable sums of permutations when the distributions $\{P_k\}_{k=1}^N$ are distinct. At the same time the weights (11) can be computed in a straightforward way given the density ratios $\lambda(\mathbf{x}, \mathbf{y})$.

Theoretical guarantees. Since controlling the average miscoverage rate is not enough to ensure the validity of the method in most cases, in this part, we show that the miscoverage rate does not deviate much from the target value α for most of the calibration datasets $\mathcal{D}_N$. This is the result of the following theorem; see full proof and additional details in Appendix A, while sketch of the proof is given below.

Theorem 2.1. *Assume there are no ties between $\{V_k\}_{k\in[N+1]} \cup \{\infty\}$ almost surely. If $\{\lambda_k - \mathbb{E}\lambda_k\}_{k\in[N]}$ are sub-Gaussian random variables with parameters $\sigma_1, \ldots, \sigma_N$, then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds*

$$-\frac{\tau_{N,\delta} + 3N^{-1}\mathbb{E}\lambda_{N+1}}{1 + N^{-1}\mathbb{E}\lambda_{N+1}} < \alpha(\mathcal{D}_N) - \alpha < \tau_{N,\delta} + \sup_{v\in\mathbb{R}} \mathbb{P}\left(V_{N+1} = v\right),$$

where we denote

$$\tau_{N,\delta} = N^{-1}\sqrt{8\log(\tfrac{1}{6\delta})\textstyle\sum_{k=1}^N (4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}.$$

This theorem shows that the confidence level of $\mathcal{C}_{\alpha,\mu}(\mathbf{x})$ is close to $1-\alpha$, regardless of the calibration set $\mathcal{D}_N$. Specifically, it meets the condition in (5) with $\tau_{N,\delta} = \mathrm{O}(N^{-1/2})$. Under exchangeable data, this matches the standard deviation of the miscoverage rate distribution exhibited in (3). Therefore, we extend the guarantees of standard CP methods to our approach but significantly broadening the scope of applicability by allowing general distribution shifts. The following result concerns the bias of our method; see Appendix C for more details.

Theorem 2.2. *Assume there are no ties between $\{V_k\}_{k\in[N+1]} \cup \{\infty\}$ almost surely. If for any $v \in \mathbb{R}$, $\lambda(Z)\mathbb{1}_{V(Z)<v} - \mathbb{E}[\lambda(Z)\mathbb{1}_{V(Z)<v}]$ is sub-Gaussian with parameter $\sigma \geq 0$, where $Z \sim P^{\mathrm{cal}}$. Then, it holds*

$$\left|\mathbb{E}[\alpha(\mathcal{D}_N)] - \alpha\right| \leq 19\sigma\sqrt{\tfrac{\log 4N}{N}} + \tfrac{18\mathbb{E}\lambda^2(Z_{N+1})}{N}.$$

Interestingly, the upper bound depends on λ, which does not directly capture the individual data distributions P^k, but rather the difference between the test distribution and the average of the local calibration distributions. The convergence rate of order $N^{-1/2}$ contrasts with the findings of Plassier et al. (2023), where their subsampling procedure achieved the standard centralized bias of order $\mathrm{O}(N^{-1})$. However, while subsampling techniques do improve bias, they also amplify the variance of the miscoverage rate, which can become a limiting factor. Also in the federated context, Lu et al. (2023) developed a method to generate prediction sets for $(X_{N+1}, Y_{N+1}) \sim \sum_{i=1}^n \frac{N^i+1}{N+n} P^i$. Their method achieves a bias that is inversely proportional to the proportion of data held by each agent, that scales as $\mathrm{O}(n/N)$. However, this result becomes less favorable when the number of agents exceeds $\sqrt{N}$, and additionally, their method does not allow for personalized prediction sets.

Sketch of Proof. We will briefly outline the main steps of the proof of Theorem 2.1. In particular, we will focus on upper-bounding the miscoverage error $\alpha(\mathcal{D}_N) - \alpha$. Detailed proofs can be found in the supplementary paper, where the lower bound is also analyzed. All the results are derived from the connection between coverage and the cumulative distribution function, which is defined for any data point $z = (x,y) \in \mathcal{X} \times \mathcal{Y}$ by

$$F_{N+1}(z) = \mathbb{E}\left[\mathbb{1}_{\{V_{N+1} \leq V(x,y)\}}\right].$$

We also introduce $\widehat{F}_{N+1}(z)$, an empirical approximation of $F_{N+1}(z)$:

$$\widehat{F}_{N+1}(z) = \sum_{k=1}^N p_k^{(x,y)} \mathbb{1}_{\{V_k \leq V(x,y)\}}. \tag{12}$$

For the first step of the proof, remark that

$$\mathbb{P}\left(\alpha(\mathcal{D}_N) < \alpha + \tau_{N,\delta} + \sup_{v\in\mathbb{R}} \mathbb{P}\{V_{N+1} = v\}\right) \geq \mathbb{P}\left(\sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{\widehat{F}_{N+1}(z) - F_{N+1}(z)\right\} < \tau_{N,\delta}\right).$$

As in (Bian and Barber, 2023), the problem consists in controlling difference between the empirical and the true cumulative distribution functions. Given that the weights $p_k^{(x,y)}$ depend on the entire calibration dataset, we introduce independent approximated weights $q_k = \lambda_k / \sum_{l=1}^N \mathbb{E}\lambda_l$. Based on those weights, consider

$$\widehat{G}_N(z) = \sum_{k=1}^N q_k \mathbb{1}_{\{V_k \leq V(x,y)\}}.$$

It can be shown that $\mathbb{E}[\widehat{G}_N(z)] = F_{N+1}(z)$. Therefore, the central part of the proof involves controlling the two right-hand side terms separately as follows:

$$\widehat{F}_{N+1}(z) - F_{N+1}(z) = \widehat{F}_{N+1}(z) - \widehat{G}_N(z) + \widehat{G}_N(z) - F_{N+1}(z).$$

Since the approximation weights q_k are chosen such that $q_k \approx p_k$, the term $\sup_{z\in\mathcal{X}\times\mathcal{Y}}\{\widehat{F}_{N+1}(z) - \widehat{G}_N(z)\}$ can be controlled using classical concentration inequalities; refer to Lemma A.4. However, controlling $\widehat{G}_N(z) - F_{N+1}(z)$ presents a more challenging task. As detailed in Appendix B, applying techniques similar to those used in the Dvoretzky–Kiefer–Wolfowitz (DKW) proof (Dvoretzky et al., 1956; Massart, 1990), it is possible to show that

$$\mathbb{P}\left(\sup_{z\in\mathcal{Z}}\left\{\widehat{G}_N(z) - F_{N+1}(z)\right\} \geq \tau_{N,\delta}\right) \leq 2\inf_{\theta>0}\left\{\mathrm{e}^{-\theta\tau_{N,\delta}}\prod_{k=1}^{N}\mathbb{E}\left[\cosh\left(\theta q_k\right)\right]\right\}. \quad (13)$$

Notably, using $\cosh(\theta q_k) \leq \mathrm{e}^{(\theta q_k)^2/2}$ with $q_k = 1/N$ leads to the standard DKW result. However, utilizing $\cosh(\theta q_k) \leq \exp(2^{-1}\theta^2\mathbb{E}[q_k^2])\cosh\{\theta(q_k - \mathbb{E}q_k)\}$ provides a sharper result:

$$\inf_{\theta>0}\left\{e^{-\theta\tau_{N,\delta}}\prod_{k=1}^{N}\mathbb{E}\left[\cosh\left(\theta q_k\right)\right]\right\} \leq \exp\left(-\frac{\tau_{N,\delta}^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{2\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}\right).$$

Combining the previous line with (13) enables controlling the deviation of $\widehat{G}_N(z) - F_{N+1}(z)$. These high-probability bounds for $\widehat{F}_{N+1}(z) - \widehat{G}_N(z)$ and $\widehat{G}_N(z) - F_{N+1}(z)$ are then unified to obtain the result of Theorem 2.1.

3 Federated Conformal Prediction under Covariate Shift

Let us illustrate the general approach with an application to the challenging problem of federated conformal prediction. We consider a system of n agents, and we further assume that the calibration data for these agents are heterogeneous due to the presence of a covariate shift:

$$(X_k^i, Y_k^i) \sim P^i = P_X^i \times P_{Y|X},\ k \in [N^i],\ i \in [n], \quad (14)$$

where $P_{Y|X}$ is the conditional distribution of the label given the covariate that is assumed identical among agents, P_X^i is the prior covariate density that may differ across agents, and N^i is the number of calibration point for the agent i.

Given calibration datasets $\mathcal{D}_i = \{(X_k^i, Y_k^i)\}_{k=1}^{N^i}, i \in [n]$, the goal is to construct a prediction set $\mathcal{C}_\alpha^\star(\mathbf{x})$ for an agent $\star \in [n]$ and input $\mathbf{x}$ with confidence level $1-\alpha \in (0,1)$. Ideally, for any new point $(X_{N+1}^\star, Y_{N+1}^\star)$ drawn from the distribution $P^\star$, the conditional coverage $\mathbb{P}\left(Y_{N+1}^\star \in \mathcal{C}_\alpha^\star(X_{N+1}^\star) \,|\, \mathcal{D}_N\right)$ should be near the confidence level $1-\alpha$ set by the user. Importantly, this confidence set should benefit not only the data available on the target agent, but also from the calibration data of other agents.

In our federated setup, the general density ratios (8) become

$$\lambda(\mathbf{x}) = \frac{P_X^\star(\mathbf{x})}{\sum_{i=1}^n \pi_i P_X^i(\mathbf{x})}, \quad (15)$$

where P_X^i represents the covariate density of agent $i \in [n]$, and $\pi_i = N^i/N$. Using these ratios (15), we can compute the general weighted confidence set (9) to obtain the confidence set $\mathcal{C}_\alpha^\star(\mathbf{x})$. Importantly, the resulting confidence set will fully satisfy the result of Theorem 2.1 and thus allows for the tight coverage guarantees.

Federated inference procedure. In practice, the confidence set $\mathcal{C}_\alpha^\star(\mathbf{x})$ from (9) can be computed in two steps:

(i) estimate the weights $p_k^{(\mathbf{x},\mathbf{y})} \equiv p_k^{(\mathbf{x})}$ that are involved in the definition of weighted empirical score distribution function in (9);

(ii) compute a quantile of the weighted empirical score distribution function $\mu_{\mathbf{x},\mathbf{y}} \equiv \mu_{\mathbf{x}}$ from (10).

Federated ratios estimation. The computation of importance weights $\{p_k^{(\mathbf{x},\mathbf{y})}\}_{k=1}^{N+1}$ in the empirical CDF relies on the determination of ratios $\lambda(\mathbf{x})$, which are initially unknown and should be estimated. For this, we utilize the Gaussian Mixture Model (GMM). In this process, each client independently computes its GMM parameters $\{\pi_y^i, \mu_y^i, \Sigma_y^i\}_{y\in\mathcal{Y}}$ using their local data $\mathcal{D}_i$:

$$\begin{aligned}\pi_y^i &= \frac{N_y^i}{N^i}, \qquad m_y^i = \frac{1}{N_y^i}\sum_k \phi(X_k^i),\\ \Sigma_y^i &= \frac{1}{N_y^i}\sum_k (\phi(X_k^i) - m_y^i)(\phi(X_k^i) - m_y^i)^\top.\end{aligned} \quad (16)$$

Here, ϕ denotes the mapping obtain while keeping the first layers of the trained neural network $\hat{f}$, and N_y^i is the number of data classified y on client i. These parameters are then transmitted to the other agents. With this information, the local agents compute the density ratio $\lambda_k^i = \lambda(X_k^i)$ on their local data, based on (15) with

$$P_X^\star(x) = \sum_{y\in\mathcal{Y}} \pi_y^\star \mathcal{N}(\phi(x); \mu_y^\star, \Sigma_y^\star),$$

$$P_X^{\mathrm{cal}}(x) = \sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} \pi_y^i \mathcal{N}(\phi(x); \mu_y^i, \Sigma_y^i),$$

where $\mathcal{N}(\phi(x); \mu_y^i, \Sigma_y^i)$ denotes the pdf of the Gaussian distribution with mean μ_y^i and covariance matrix Σ_y^i.

Federated quantile estimation. Once the density ratios are estimated, they can be used in the federated quantile estimation procedure described by Plassier et al. (2023). This algorithm relies on regularized pinball loss functions, which are smoothed using the Moreau-Yosida inf-convolution (Moreau, 1963). This smoothing enables the application of classical FL stochastic gradient methods. Theoretical guarantees on the coverage are provided, along with assurances of differential privacy. However, the developed approach demands calculating the quantile for any new query $\mathbf{x}$. To mitigate this issue, we approximate the importance weights $p_k^{(\mathbf{x},\mathbf{y})}$ by query-independent weights $\widehat{p}_k = (\sum_{l=1}^N \lambda_l)^{-1}\lambda_k$. As a result, the considered prediction set becomes:

$$\widehat{\mathcal{C}}_{\alpha,\mu}(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{Y} \colon V(\mathbf{x},\mathbf{y}) \leq \widehat{Q}_{1-\alpha}^{(\gamma)}\big(\textstyle\sum_{k=1}^N \widehat{p}_k \delta_{V_k}\big)\right\},$$

where $\widehat{Q}_{1-\alpha}^{(\gamma)}(\sum_{k=1}^N \widehat{p}_k \delta_{V_k})$ is the approximated quantiles obtained via the FL minimization of the γ-regularized pinball loss; refer to Algorithm 1 in the Supplementary Material for more details. Note that calculating $\widehat{\mathcal{C}}_{\alpha,\mu}(\mathbf{x})$ is straightforward, because the quantile $\widehat{Q}_{1-\alpha}^{(\gamma)}(\sum_{k=1}^N \widehat{p}_k \delta_{V_k})$ is independent of the query $\mathbf{x}$.

4 Related Work

The reliability of prediction sets has been extensively studied in centralized frameworks (Papadopoulos, 2008). However, their marginal capabilities may generate invalid confidence intervals for specific data subgroups. As a result, analyzing conditional conformal prediction methods is crucial, especially for high heteroscedasticity regions (Alaa et al., 2023). Exploring flexible coverage guarantees has been investigated by Foygel Barber et al. (2021). They studied potential relaxations of marginal coverage guarantees, however, achieving such guarantees is not always feasible (Vovk, 2012; Lei and Wasserman, 2014). In the case of exchangeable data, Bian and Barber (2023) examined theoretical guarantees in the *training-conditional* setting. They showed with high probability that the miscoverage rate is upper bounded by the value $2\alpha \pm \mathrm{O}_{\mathbb{P}}(1/\sqrt{N})$ for the K-fold CV+ method. In addition, they also demonstrated the unattainability of similar results for the full conformal method or the jackknife+ method without further assumptions.

Most conformal prediction methods are guaranteed to be valid under data exchangeability. However, this assumption can be restrictive. To alleviate this constraint, Tibshirani et al. (2019) proposed a method working with independent data, even when they follow different distributions. This approach is beneficial when dealing with data from diverse sources, which tend to produce data from different distributions. However, this method requires the determination of importance weights based on permutations of density ratios. If the distribution of only one datapoint differs from the calibration distribution, Lei and Candès (2021) demonstrated that approximations of these importance ratios is sufficient to achieve the desired coverage level. The error introduced by this ratio estimation can be controlled by the total variation distance between the true ratio and its approximation. Furthermore, Plassier et al. (2023) introduced a subsampling technique to simplify the combinatorial complexity of the importance weights computations. Based on a multinomial draw of the calibration datapoints, their method achieves a coverage error of order $\mathrm{O}(N^{-1}\log N)$, almost recovering the standard theoretical guarantee for exchangeable data (Vovk et al., 2005). However, we conjecture that the numerical advantage of the proposed method is primarily due to the small deviation of the miscoverage rather than the small bias.

This framework of data heterogeneity is particularly relevant in federated contexts where there may be shifts in data distribution among different agents. The concept of using calibration data from multiple agents to refine prediction sets has been the subject of several recent papers; see, among others, (Lu and Kalpathy-Cramer, 2021; Humbert et al., 2023; Lu et al., 2023; Plassier et al., 2023; Zhu et al., 2023). Plassier et al. (2023) introduced a federated conformal prediction method to account for label shifts. However, their methods require multiple rounds of communication between a central server and the agents. In contrast, Humbert et al. (2023) have developed a conformal approach with only one communication round. Their method relies on computing a quantile locally, which is then shared with the central server. Then, the central server computes the quantile of these quantiles to create prediction sets. Another approach to generating prediction sets when the test point follows the agent's mixture distribution is presented in (Lu et al., 2023).

5 Experiments

Domain Adaptation on Synthetic Data. We start with an initial experiment to demonstrate the potential of our proposed approach in addressing domain adaptation. For this, we consider synthetic 1-D data and a two layer neural network regression model. We use 100 calibration data points to build 90% prediction intervals for 20 test data. The data are sampled from either the distribution $P^1 = \mathcal{N}(3, 2)$ or drawn from $P^2 = \mathcal{N}(5, 2)$. The calibration set is sampled from

METHOD	P^{cal}	COVERAGE
This work	Mix	91.03 ± 4.86 %
Plassier et al. (2023)	Mix	92.28 ± 5.97 %
Tibshirani et al. (2019)	Mix	82.35 ± 3.94 %
Classic conformal	Mix	81.75 ± 5.55 %
Tibshirani et al. (2019)	P^1	92.96 ± 5.31 %
Classic conformal	P^1	79.16 ± 6.88 %
Classic conformal	P^2	90.26 ± 6.58 %

Table 1: Coverage on synthetic data. Mix corresponds to a 80/20 mixture of data from P^1 and P^2.

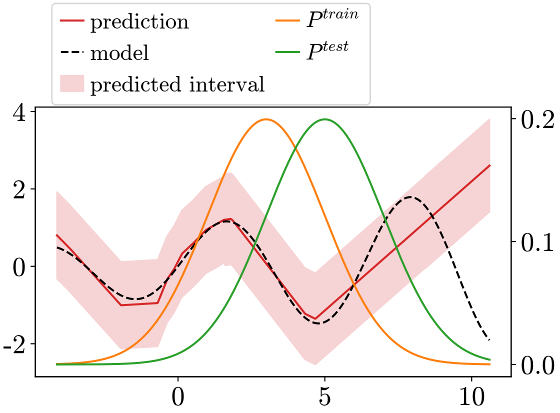


Figure 1: Example of synthetic data. We show the true dependence as a dotted line, neural network prediction and the predictive confidence interval are in red. Additionally, we present PDFs of the train and test distributions on a secondary vertical axis.

either P^1 (80 points), P^2 (20 points) or a mixture of both (80 + 20 points). The observations are modeled as $Y = (1 + 0.1|X|)\sin(X) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.5)$. We train a fully-connected two layers neural network with 10 neurons per layer, using the MSE loss function computed on 150 data points. To illustrate this, Figure 1 displays an example of the sampled data and the trained model. When $P^{\text{cal}} = P^1$, note that our method aligns with the approach proposed by Tibshirani et al. (2019, Corollary 1). This experiment is replicated 500 times, using the non-conformity score $V(\mathbf{x}, \mathbf{y}) = |\hat{f}(\mathbf{x}) - \mathbf{y}|$. The results are summarized in Table 1. Classical conformal method noticeably breaks the 90% coverage guarantee when the non-exchangeability is not satisfied. In contrast, our proposed method achieves the desired confidence level, while demonstrating a lower standard deviation than competitors.

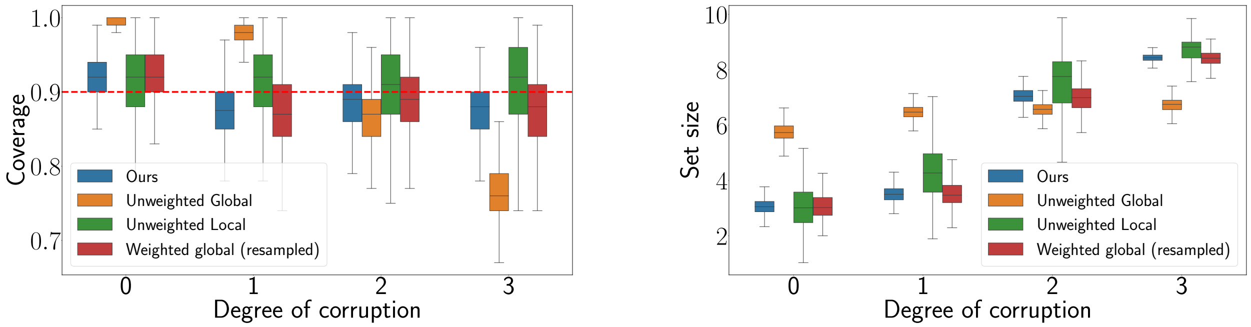
Federated Learning on CIFAR-10. In this experiment, we follow Section 3 and consider personalized federated learning problem with the CIFAR-10 dataset. To induce a covariate shift between agents, we apply augmentations using Gaussian blur at varying intensities. This strategy modifies the agent's local distributions P^i_X while keeping $P^i_{Y|X}$ intact. For this experiment, we introduce four different levels of data corruption. Level 0 corresponds to the original data, whereas levels 1-3 introduce increasing degrees of corruption; 1 being minimal and 3 being the most severe. We achieve this using square Gaussian kernels of sizes 3, 5, and 7. Additionally, the standard deviation of the elements in these filters is adjusted based on kernel size κ using the formula $\sigma = 0.3 \cdot \big((\kappa - 1) \cdot 0.5 - 1\big) + 0.8$. In this experiment, we consider 40 different distributed agents with 10 agents for each corruption level. Each agent owns a calibration set of 25 objects.

In our testing phase, we set $\alpha = 0.1$ and, for the nonconformity score, we adopt the adaptive prediction sets (APS) approach developed in (Romano et al., 2020; Angelopoulos et al., 2023):

$$V(\mathbf{x}, \mathbf{y}) = \textstyle\sum_{\{y:\ \hat{f}(\mathbf{x})_y \geq \hat{f}(\mathbf{x})_{\mathbf{y}}\}} \hat{f}(\mathbf{x})_y, \tag{17}$$

where $\hat{f}(\mathbf{x})_y$ is the predicted probability of label $y \in \mathcal{Y}$. We consider ResNet-18 model trained on the unaltered CIFAR-10 dataset. We found temperature scaling of the logits beneficial and set the temperature $T = 10$ in our experiments. Weights of the nonconformity scores are computed with the GMM approach described above. We compare our method against three baselines: "Unweighted Local", "Unweighted Global" and "Weighted global (resampled)" (Plassier et al., 2023). First two use standard conformal prediction with $\lambda = (\text{Num data} + 1)^{-1}$, while the latter uses specially constructed weights. "Unweighted Local" only uses the local agent's calibration data, while "Unweighted Global" combines calibration data from all agents. Our results are obtained from 100 independent runs, each using different random splits of an agent's data for calibration, testing, and for density model fitting. For each iteration, we evaluate the local test set of an agent, calculating both average coverage and average set size. The metrics are collected across 100 runs and visualized using boxplots.

Results for $\alpha = 0.1$ are shown in Figure 2a and Figure 2b. It is worth noting that while we had 40 agents, we aggregated results for clarity, presenting them by the level of data corruption. The figures show that except for "Unweighted Global", all methods achieve the target coverage. "Weighted global (resampled)" has a coverage of $92.28 \pm 5.97\%$ which is perfectly aligned with expectation to have increased variance compared to our method (see Table 1). In contrast, our method consistently shows less variation in set sizes than the others. This consistency arises from the federated pro-



(a)

(b)

Figure 2: CIFAR-10 experimental results: (a) The distribution of coverage percentage for each agent. It shows how closely the predicted values covers the true values across varying degrees of data corruption. (b) The distribution of set sizes for different agents. The plot illustrates the growth in set sizes as data corruption increases, emphasizing the relationship between data integrity and set size.

cess, when calibration scores from other agents, once appropriately weighted, contribute to the conformal procedure for a given agent.

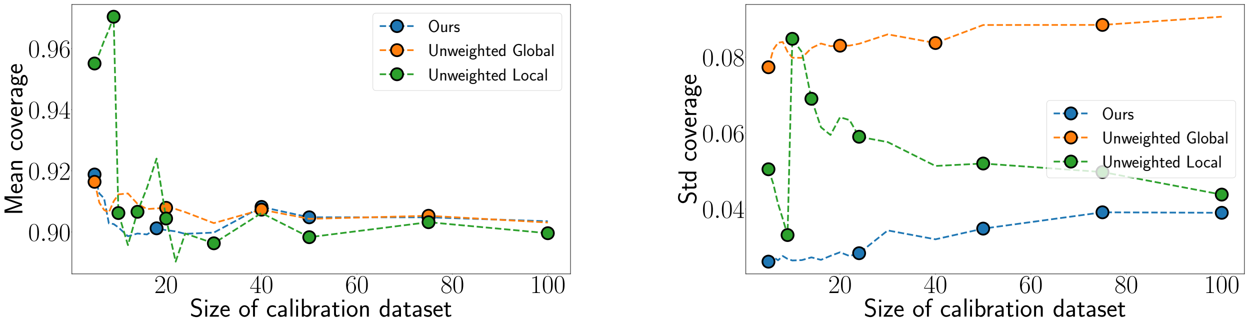
It is worth mentioning, that we do not present (Tibshirani et al., 2019) here among competitors. For this method, the only feasible approach is to sample a certain number of permutations and compute approximate weights based on them. Unfortunately, it leads to the very high variance of the weights which led us even to the bias in the mean coverage (82.35 on the "Mix" calibration data in our experiment on domain adaptation). When replicating the CIFAR10 experiment, sampling the permutations led us to even worse results.

Federated Learning on CIFAR-100. For CIFAR-100 dataset we performed a similar experiment but with slightly different hyperparameters and another model – ResNet50. However, here we study how the size of calibration dataset influences the results. Basically, we explore how mean and standard deviation of empirical coverage changes with the size of the calibration dataset. The results are presented in Figure 3a and Figure 3b. In each setup, we have 100 distributed agents which are split between 4 different levels of corruption. From these plots we see that our method has the smallest variance among others, and converges fast to the right mean with the increase of the calibration dataset. In contrast, other methods are much slower to converge with the "Unweighted Global" method having extremely high variance as it either significantly undercovers or overcovers depending on the level of data corruption.

Federated Learning on ImageNet. With the ImageNet dataset (Deng et al., 2009), we perform a similar set of experiments as with CIFAR-10 and CIFAR-100 datasets, but slightly modifying the experimental setup. We use ImageNet-C (Hendrycks and Dietterich, 2019) as the corrupted data with 5 different corruption levels from 1 to 5. Regarding the type of corruption, we select the "defocus blur" option. We additionally consider the 0 level as corresponding to the non-corrupted data, namely the original ImageNet data. To illustrate the severity of the corruption, the model's accuracy drops from 79% (non-corrupted data, level 0) to 9% (corrupted data, level 5). We use ResNet-50 pre-trained on the original ImageNet data. For the conformal procedure, we set the importance level $\alpha = 0.1$ and the temperature scaling parameter is chosen as $T = 10$. The APS non-conformity measure (17) is used.

We compare the performance of the introduced algorithm with local baselines at 5 different corruption levels in terms of coverage and average prediction set size. Data allocation scheme is as follows. ImageNet and ImageNet-C validation data are distributed among 20 agents without overlap, such that half of the agents contain clean, non-corrupted data, and the other half consists of corrupted images of the same level of corruption. Each agent contains 850 data samples, of which 50 are calibration samples, 300 test samples and 500 density model fitting samples. We conduct separate experiments with the above-mentioned setup for all levels of corruption. The results are presented as box plots with means and standard deviations obtained from 10 random splits of the data for each of these experiments with varying levels of corruption. The results of the algorithms at each corruption level are obtained by averaging the results from all agents within the same level of corruption.

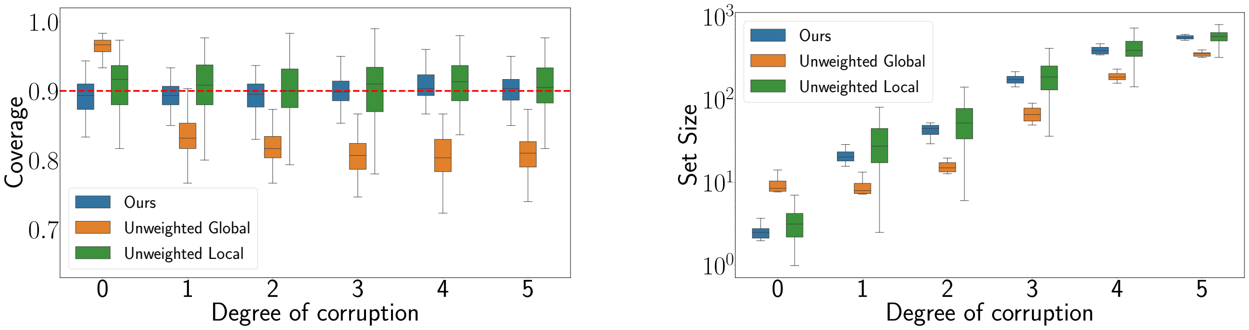
The Figure 4a shows the empirical coverage of the prediction sets at different levels of corruption. The resulting coverage of the introduced method is valid



(a)

(b)

Figure 3: CIFAR-100 experimental results: (a) Mean empirical coverage changes as function of the calibration dataset size. (b) Standard deviation of empirical coverage as function of the calibration dataset size.



(a)

(b)

Figure 4: ImageNet experimental results: (a) Empirical coverage of conformal prediction sets as a function of data corruption level. It shows how accurately do conformal sets capture the true classes of the data. (b) Average set size of conformal prediction sets as a function of data corruption level. The size of the sets increases with the level of corruption due to the increasing uncertainty of the model based on the corrupted data.

in comparison with “Unweighted Global” baseline, has lower variance and better accuracy compare to the “Unweighted Local” one. The Figure 4b demonstrates the average prediction set size of conformal sets as a function of the level of corruption. The presented method shows on average slightly smaller set sizes and much less variance of the results compared to an “Unweighted Local” baseline. By using information from other agents through importance weights, our method overcomes the problems of data heterogeneity compared to unweighted alternatives.

6 Conclusion

Our work introduces a new method for tailoring prediction sets on non-exchangeable data. By leveraging density ratio, the coverage validity is ensured for most calibration datasets. This approach opens up new possibilities for applications within the federated framework, where distributional shifts among agents are common. While our numerical experiments have primarily focused on covariate shift, it is important to emphasize the method’s potential applicability to a much broader range of scenarios.

Acknowledgements

The authors gratefully thank Pierre Humbert and Sylvain Arlot for their valuable feedback. The research on experimental evaluation of proposed methods (Section 5) was supported by the Russian Science Foundation grant 20-71-10135. Part of this work has been carried out under the auspice of the Lagrange Mathematics and Computing Research Center. The work of F. Noskov was prepared within the framework of HSE basic research program.

References

Alaa, A. M., Hussain, Z., and Sontag, D. (2023). Conformalized unconditional quantile regression. In *International Conference on Artificial Intelligence and Statistics*, pages 10690–10702. PMLR.

Angelopoulos, A. N., Bates, S., et al. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Bartl, D. and Mendelson, S. (2023). On a variance dependent dvoretzky-kiefer-wolfowitz inequality. *arXiv preprint arXiv:2308.04757*.

Bian, M. and Barber, R. F. (2023). Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066.

Boucheron, S., Lugosi, G., and Bousquet, O. (2003). Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482.

Ha, T., Dang, T. K., Dang, T. T., Truong, T. A., and Nguyen, M. T. (2019). Differential privacy in deep learning: an overview. In *2019 International Conference on Advanced Computing and Applications (ACOMP)*, pages 97–102. IEEE.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Humbert, P., Le Bars, B., Bellet, A., and Arlot, S. (2023). One-shot federated conformal prediction. In *International Conference on Machine Learning*, pages 14153–14177. PMLR.

Kato, Y., Tax, D. M., and Loog, M. (2023). A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383.

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

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)*.

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

Lu, C., Yu, Y., Karimireddy, S. P., Jordan, M., and Raskar, R. (2023). Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

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.

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

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In Fritzsche, P., editor, *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka.

Papadopoulos, H., Vovk, V., and Gammerman, A. (2011). Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840.

Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., and Panov, M. (2023). Conformal prediction for federated uncertainty quantification under label shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 27907–27947.

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

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.

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

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Zhu, M., Zecchin, M., Park, S., Guo, C., Feng, C., and Simeone, O. (2023). Federated inference with reliable uncertainty quantification over wireless channels via conformal prediction. *arXiv preprint arXiv:2308.04237*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [No]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents

A Proof of the conditional coverage

We assume that the calibration data $\{(X_k, Y_k)\}_{k\in[N+1]}$ are independent and

$$\forall k \in [N+1], \qquad (X_k, Y_k) \sim P^k.$$

In this context, X_k denotes the covariate, and Y_k represents the label. Additionally, we define $Z_k = (X_k, Y_k)$ to represent the pair consisting of a covariate and a label. The supports of X_k and Y_k are symbolized by $\mathcal{X}$ and $\mathcal{Y}$ respectively. Typically, the set $\mathcal{X}$ is a subset of $\mathbb{R}^d$, whereas $\mathcal{Y}$ may either be finite in the case of classification or represent a subset of $\mathbb{R}^d$. We denote by $P^{\mathrm{cal}} = (1/N)\sum_{k=1}^N P^k$ the calibration measure.

***H*1.** The data $\{Z_k\}_{k\in[N+1]}$ are pairwise mutually independent and $P^{N+1} \ll P^{\mathrm{cal}}$.

Moreover, we set

$$\begin{aligned} &\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad \lambda(x,y) = (\mathrm{d}P^{N+1}/\mathrm{d}P^{\mathrm{cal}})(x,y), \\ &\forall k \in [N+1], \quad \lambda_k = \lambda(Z_k). \end{aligned} \tag{18}$$

For $\alpha \in (0,1)$, we denote by $1-\alpha$ the confidence level and $\mathcal{D}_N = \{Z_k\}_{k\in[N]} \in (\mathcal{X}\times\mathcal{Y})^N$ the calibration dataset of size N. We denote by $V : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+$ a non-conformity score, and for $\beta \in [0,1]$, $Q_\beta(\mu)$ represents the β-quantile of the probability measure μ, and δ_v the Dirac measure at $v \in \mathbb{R}\cup\{\infty\}$. We define by $p_k^{(x,y)}$ the importance weights define by

$$\forall k \in [N], \quad p_k^{(x,y)} = \frac{\lambda_k}{\lambda(x,y) + \sum_{l=1}^N \lambda_l}, \qquad\qquad p_{N+1}^{(x,y)} = \frac{\lambda(x,y)}{\lambda(x,y) + \sum_{l=1}^N \lambda_l}.$$

In this paper, we consider the prediction set at $x \in \mathcal{X}$:

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

where

$$\mu^{(x,y)} = \sum_{k=1}^N p_k^{(x,y)} \delta_{V(Z_k)} + p_{N+1}^{(x,y)} \delta_\infty,$$

Given the calibration dataset $\mathcal{D}_N$, consider the conditional miscoverage rate

$$\alpha(\mathcal{D}_N) = \mathbb{P}\left(Y_{N+1} \notin \mathcal{C}_{\alpha,\mu}(X_{N+1}) \mid \mathcal{D}_N\right).$$

For $z \in \mathcal{X} \times \mathcal{Y}$, we denote

$$\widehat{F}_{N+1}(z) = \sum_{k=1}^{N} p_k^z \mathbb{1}_{\{V(Z_k) \leq V(z)\}}, \qquad F_{N+1}(z) = \mathbb{E}\left[\mathbb{1}_{\{V(Z_{N+1}) \leq V(z)\}}\right].$$

Define by G_{N+1} the cumulative density function of $V(Z_{N+1})$, and consider the associated quantile function:

$$\forall \gamma \in [0,1], \qquad G_{N+1}^{+}(\gamma) = \inf\left\{v \in \mathbb{R} \colon G_{N+1}(v) \geq \gamma\right\}.$$

In addition, define the supremum of the difference between the empirical and the true cumulative distribution functions:

$$\begin{aligned}\Delta(\mathcal{D}_N) &= \sup_{z \in \mathcal{X}\times\mathcal{Y}} \left\{\widehat{F}_{N+1}(z) - F_{N+1}(z)\right\}, \\ \tilde{\Delta}(\mathcal{D}_N) &= \sup_{z \in \mathcal{X}\times\mathcal{Y}} \left\{F_{N+1}(z) - \widehat{F}_{N+1}(z)\right\}.\end{aligned}$$

Lemma A.1. *Assume that there are no ties between the $\{V(Z_k) \colon k \in [N+1]\}$ and $V(Z_{N+1}) < \infty$ almost surely. For any $\delta > 0$, we have*

$$\begin{aligned}&\mathbb{P}\left(\alpha(\mathcal{D}_N) \geq \alpha + \delta + \sup_{v \in \mathbb{R}} \mathbb{P}\left\{V(Z_{N+1}) = v\right\}\right) \leq \mathbb{P}\left(\Delta(\mathcal{D}_N) \geq \delta\right), \\ &\mathbb{P}\big(\alpha(\mathcal{D}_N) \leq \alpha - \delta\big) \leq \mathbb{P}\big(\tilde{\Delta}(\mathcal{D}_N) \geq \delta\big).\end{aligned}$$

Proof. For any $x \in \mathcal{X}$, we have

$$\begin{aligned}\mathcal{C}_{\alpha,\mu}(x) &= \left\{y \in \mathcal{Y} \colon V(x,y) \leq Q_{1-\alpha}(\mu^{(x,y)})\right\} \\ &= \left\{y \in \mathcal{Y} \colon V(x,y) < Q_{1-\alpha}(\mu^{(x,y)})\right\} \cup \left\{y \in \mathcal{Y} \colon V(x,y) = Q_{1-\alpha}(\mu^{(x,y)})\right\}.\end{aligned}$$

Moreover, note that

$$\begin{aligned}\left\{y \in \mathcal{Y} \colon V(x,y) < Q_{1-\alpha}(\mu^{(x,y)})\right\} &= \left\{y \in \mathcal{Y} \colon \sum_{k=1}^{N} p_k^{(x,y)} \mathbb{1}_{V(Z_k) \leq V(x,y)} < 1-\alpha\right\} \\ &= \left\{y \in \mathcal{Y} \colon \widehat{F}_{N+1}(x,y) < 1-\alpha\right\},\end{aligned}$$

and remark that

$$\mathbb{E}\left[\mathbb{P}\left(V(Z_{N+1}) = Q_{1-\alpha}(\mu^{(Z_{N+1})}) \mid \mathcal{D}_N\right)\right] \leq \mathbb{P}\left(V(Z_{N+1}) \in \{V(Z_k)\}_{k \in [N]} \cup \{\infty\}\right).$$

Hence, the assumptions on $\{V(Z_k) \colon k \in [N+1]\}$ imply that almost surely:

$$\mathbb{P}\left(V(Z_{N+1}) = Q_{1-\alpha}(\mu^{(Z_{N+1})}) \mid \mathcal{D}_N\right) = 0.$$

Therefore, we deduce that

$$\begin{aligned}\alpha(\mathcal{D}_N) &= \mathbb{P}\left(Y_{N+1} \notin \mathcal{C}_{\alpha,\mu}(X_{N+1}) \mid \mathcal{D}_N\right) \\ &= 1 - \mathbb{P}\left(V(Z_{N+1}) < Q_{1-\alpha}(\mu^{(Z_{N+1})}) \mid \mathcal{D}_N\right) - \mathbb{P}\left(V(Z_{N+1}) = Q_{1-\alpha}(\mu^{(Z_{N+1})}) \mid \mathcal{D}_N\right) \\ &= \mathbb{P}\left(\widehat{F}_{N+1}(Z_{N+1}) \geq 1-\alpha \mid \mathcal{D}_N\right) \\ &= \mathbb{P}\left(F_{N+1}(Z_{N+1}) - \left(F_{N+1}(Z_{N+1}) - \widehat{F}_{N+1}(Z_{N+1})\right) \geq 1-\alpha \mid \mathcal{D}_N\right). \qquad (19)\end{aligned}$$

For any $\beta \in (0,1)$ and $v \in \mathbb{R}$, remark that $G_{N+1}(v) \geq \beta$ if and only if $v \geq G_{N+1}^{+}(\beta)$. Therefore, using the shorthand notation $V_{N+1} = V(Z_{N+1})$, it follows that

$$\mathbb{P}\left(G_{N+1}(V_{N+1}) < \beta\right) = \mathbb{P}\left(V_{N+1} < G_{N+1}^{+}(\beta)\right)$$

$$= G_{N+1}\left(G^{+}_{N+1}(\beta)\right) - \mathbb{P}\left(V_{N+1} = G^{+}_{N+1}(\beta)\right).$$

Therefore, we deduce that

$$\begin{aligned}
\mathbb{P}\left(F_{N+1}(Z_{N+1}) \geq \beta\right) &= 1 - \mathbb{P}\left(G_{N+1}(V_{N+1}) < \beta\right) \\
&= 1 - \beta + \{\beta - \mathbb{P}\left(G_{N+1}(V_{N+1}) < \beta\right)\} \\
&= 1 - \beta + \left\{\beta - G_{N+1}\left(G^{+}_{N+1}(\beta)\right) + \mathbb{P}\left(V_{N+1} = G^{+}_{N+1}(\beta)\right)\right\} \\
&\in [1 - \beta, 1 - \beta + \mathbb{P}\left(V_{N+1} = G^{+}_{N+1}(\beta)\right)]. \qquad (20)
\end{aligned}$$

The previous upper bound also holds for $\beta \leq 0$ or $\beta = 1$, whereas the previous lower bound holds for $\beta \geq 0$. From (19), we obtain

$$\alpha(\mathcal{D}_N) \leq \mathbb{P}\left(F_{N+1}(Z_{N+1}) \geq 1 - \alpha - \Delta(\mathcal{D}_N) \mid \mathcal{D}_N\right).$$

Therefore, applying (20) with $\beta = 1 - \alpha - \Delta(\mathcal{D}_N)$ concludes the first part of the proof since $\beta \in (-\infty, 1]$. The second part of the proof follows from (19) and (20), with $\beta = 1 - \alpha + \tilde{\Delta}(\mathcal{D}_N)$, since $\beta \geq 0$. □

Theorem A.2. *Assume that **H**1 holds, and suppose there are no ties between the $\{V(Z_k) \colon k \in [N+1]\}$ and $V(Z_{N+1}) < \infty$ almost surely. If we suppose that $\{\lambda_k - \mathbb{E}\lambda_k\}_{k\in[N]}$ are sub-Gaussian with parameters $\sigma_1, \ldots, \sigma_N \geq 0$, then, for any $\delta > 0$ it follows*

$$\mathbb{P}\left(\alpha(\mathcal{D}_N) < \alpha + \frac{\sqrt{8\log(\frac{1}{3\delta})\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}}{N} + \sup_{v\in\mathbb{R}} \mathbb{P}\{V(Z_{N+1}) = v\}\right) \geq 1 - \delta,$$

$$\mathbb{P}\left(\alpha(\mathcal{D}_N) > \alpha - \frac{3\mathbb{E}\lambda_{N+1} + \sqrt{8\log(\frac{1}{3\delta})\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}}{N + \mathbb{E}\lambda_{N+1}}\right) \geq 1 - \delta.$$

Proof. This proof is split in two part. In part one, we demonstrate the first inequality. In part two, we prove the second inequality.

Proof of part 1. For any $k \in [N+1]$, introduce the weights $q_k = \lambda_k / \sum_{l=1}^{N} \mathbb{E}\lambda_l$. The random variables $\{q_k\}_{k\in[N+1]}$ are mutually pairwise independent approximations of the importance weights $\{p^z_k\}_{k\in[N+1]}$. Note that

$$\begin{aligned}
\Delta(\mathcal{D}_N) \leq \sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{\sum_{k=1}^{N} (p^z_k - q_k)\, \mathbb{1}_{\{V(Z_k)\leq V(z)\}}\right\} + \sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{\sum_{k=1}^{N} \left(q_k \mathbb{1}_{\{V(Z_k)\leq V(z)\}} - \mathbb{E}\left[q_k \mathbb{1}_{\{V(Z_k)\leq V(z)\}}\right]\right)\right\} \\
+ \sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{\sum_{k=1}^{N} \mathbb{E}\left[q_k \mathbb{1}_{\{V(Z_k)\leq V(z)\}}\right] - F_{N+1}(z)\right\}. \qquad (21)
\end{aligned}$$

Using Lemma A.3 we obtain

$$\sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{\sum_{k=1}^{N} \mathbb{E}\left[q_k \mathbb{1}_{\{V(Z_k)\leq V(z)\}}\right] - F_{N+1}(z)\right\} = 0. \qquad (22)$$

Let δ denote a positive real. Applying Theorem B.1 shows that

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{\sum_{k=1}^{N} \left(q_k \mathbb{1}_{\{V(Z_k)\leq V(z)\}} - \mathbb{E}\left[q_k \mathbb{1}_{\{V(Z_k)\leq V(z)\}}\right]\right)\right\} \geq \delta\right) \leq 2\inf_{\theta>0}\left\{e^{-\theta\delta}\prod_{k=1}^{N}\mathbb{E}\left[\cosh\left(\theta q_k\right)\right]\right\}. \qquad (23)$$

Let $\theta > 0$, and denote by $\{\epsilon_k\}_{k\in[N]}$ a sequence of i.i.d. Rademacher random variables. The independence of $\{\lambda_k\}_{k\in[N]}$ implies that

$$\prod_{k=1}^{N}\mathbb{E}\left[\cosh\left(\theta q_k\right)\right] = \prod_{k=1}^{N}\left(2^{-1}\mathbb{E}\left[\exp\left(\theta q_k\right)\right] + 2^{-1}\mathbb{E}\left[\exp\left(-\theta q_k\right)\right]\right) = \prod_{k=1}^{N}\mathbb{E}\left[\exp\left(\theta\epsilon_k q_k\right)\right].$$

For all $x \in \mathbb{R}$, note that $\cosh(x) \leq \exp(x^2/2)$. Thus, we deduce that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\theta\epsilon_k q_k\right)\right] &= \mathbb{E}\left[\exp\left(\theta\epsilon_k \mathbb{E}q_k\right)\right]\mathbb{E}\left[\exp\left(\theta\epsilon_k\left[q_k - \mathbb{E}q_k\right]\right)\right] \\ &\leq \exp\left(2^{-1}\theta^2\mathbb{E}[q_k]^2\right)\mathbb{E}\left[\exp\left(\theta\epsilon_k\left[q_k - \mathbb{E}q_k\right]\right)\right].\end{aligned}$$

Using the σ_k-sub-Gaussianity of $\lambda_k - \mathbb{E}\lambda_k$, it yields that

$$\begin{aligned}\prod_{k=1}^{N}\mathbb{E}\left[\cosh\left(\theta q_k\right)\right] &\leq \exp\left(\frac{\theta^2}{2}\sum_{k=1}^{N}\mathbb{E}[q_k]^2\right)\mathbb{E}\left[\mathbb{E}\left[\exp\left(\theta\sum_{k=1}^{N}\epsilon_k\left[q_k - \mathbb{E}q_k\right]\right) \;\middle|\; \{\epsilon_k\}_{k\in[N]}\right]\right] \\ &\leq \exp\left(\frac{\theta^2}{2(\sum_{l=1}^{N}\mathbb{E}\lambda_l)^2}\sum_{k=1}^{N}\mathbb{E}[\lambda_k]^2 + \frac{2\theta^2}{(\sum_{l=1}^{N}\mathbb{E}\lambda_l)^2}\sum_{k=1}^{N}\sigma_k^2\right). \qquad (24)\end{aligned}$$

Now, consider the specific choice of θ_N given by

$$\theta_N = \frac{\delta(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{\sum_{k=1}^{N}(4\sigma_N^2 + \mathbb{E}[\lambda_k]^2)}.$$

Combining (24) with the expression of θ_N, it follows that

$$\begin{aligned}\inf_{\theta>0}\left\{e^{-\theta\delta}\prod_{k=1}^{N}\mathbb{E}\left[\cosh\left(\theta q_k\right)\right]\right\} &\leq \exp\left(-\delta\theta_N + \frac{\theta_N^2\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}{2(\sum_{l=1}^{N}\mathbb{E}\lambda_l)^2}\right) \\ &= \exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{2\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}\right).\end{aligned}$$

Furthermore, since we assumed the σ_k-sub-Gaussianity of $\lambda_k - \mathbb{E}\lambda_k$, applying Lemma A.4 with $\tilde{\sigma}_N^2 = \sum_{k=1}^{N}\sigma_k^2$ implies that

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}\times\mathcal{Y}}\left\{\sum_{k=1}^{N}(p_k^z - q_k)\,\mathbb{1}_{\{V(Z_k)\leq V(z)\}}\right\} \geq \delta\right) \leq \exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{2\sum_{k=1}^{N}\sigma_k^2}\right). \qquad (25)$$

Plugging (22)-(23)-(25) into (21) gives

$$\mathbb{P}\left(\Delta(\mathcal{D}_N) \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{8\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}\right) + \exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{8\sum_{k=1}^{N}\sigma_k^2}\right). \qquad (26)$$

Let $\gamma > 0$, and denote by $\delta_0(\gamma), \delta_1(\gamma)$ the positive real numbers defined by

$$\delta_0(\gamma) = \frac{\sqrt{8\log(\frac{1}{\gamma})\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}}{\sum_{k=1}^{N}\mathbb{E}\lambda_k}, \qquad \delta_1(\gamma) = \frac{\sqrt{8\log(\frac{1}{\gamma})\sum_{k=1}^{N}\sigma_k^2}}{\sum_{k=1}^{N}\mathbb{E}\lambda_k}.$$

With this notation, remark that

$$\begin{aligned}\forall\delta \geq \delta_0(\gamma), &\qquad \exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{8\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}\right) \leq \gamma, \\ \forall\delta \geq \delta_1(\gamma), &\qquad \exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{8\sum_{k=1}^{N}\sigma_k^2}\right) \leq \gamma.\end{aligned}$$

Therefore, for any $\delta \geq \delta_0(\gamma) \vee \delta_1(\gamma)$, Equation (26) shows

$$\mathbb{P}\left(\Delta(\mathcal{D}_N) \geq \delta\right) \leq 3\gamma.$$

Finally, combining $\sum_{k=1}^{N}\mathbb{E}\lambda_k = N$ with Lemma A.1 concludes the first part of the proof.

Proof of part 2. For the second part of the proof, consider the weights $\tilde{q}_k = \lambda_k / \sum_{l=1}^{N+1} \mathbb{E}\lambda_l$. Furthermore, remark that

$$\tilde{\Delta}(\mathcal{D}_N) \le \sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(\tilde{q}_k - p_k^z\right) \mathbb{1}_{\{V(Z_k)\le V(z)\}} \right\} + \sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(\mathbb{E}\left[\tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}}\right] - \tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}} \right) \right\} + \sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{ F_{N+1}(z) - \sum_{k=1}^{N} \mathbb{E}\left[\tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}}\right] \right\}. \tag{27}$$

First, using Lemma A.4, $\forall \delta \ge \mathbb{E}\lambda_{N+1} / \sum_{k=1}^{N+1} \mathbb{E}\lambda_k$ it follows

$$\sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(\tilde{q}_k - p_k^z\right) \mathbb{1}_{\{V(Z_k)\le V(z)\}} \right\} \le \exp\left(- \frac{\left(\delta \sum_{k=1}^{N} \mathbb{E}\lambda_k + (1-\delta)\mathbb{E}\lambda_{N+1}\right)^2}{2\sum_{k=1}^{N} \sigma_k^2} \right). \tag{28}$$

Moreover, Lemma A.3 shows that

$$\begin{aligned} F_{N+1}(z) - \sum_{k=1}^{N} \mathbb{E}\left[\tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}}\right] &= \sum_{k=1}^{N} \mathbb{E}\left[\left(\frac{1}{\sum_{l=1}^{N} \mathbb{E}\lambda_l} - \frac{1}{\sum_{l=1}^{N+1} \mathbb{E}\lambda_l} \right) \lambda_k \mathbb{1}_{\{V(Z_k)\le V(z)\}} \right] \\ &= \frac{\mathbb{E}\lambda_{N+1}}{\sum_{l=1}^{N+1} \mathbb{E}\lambda_l} \sum_{k=1}^{N} \frac{\mathbb{E}\left[\lambda_k \mathbb{1}_{\{V(Z_k)\le V(z)\}}\right]}{\sum_{l=1}^{N} \mathbb{E}\lambda_l}. \end{aligned}$$

Therefore, we deduce that

$$\sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{ F_{N+1}(z) - \sum_{k=1}^{N} \mathbb{E}\left[\tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}}\right] \right\} \le \frac{\mathbb{E}\lambda_{N+1}}{\sum_{l=1}^{N+1} \mathbb{E}\lambda_l}. \tag{29}$$

Furthermore, taking back (24) with $\tilde{q}_k$ instead of q_k demonstrates that

$$\begin{aligned} \prod_{k=1}^{N} \mathbb{E}\left[\cosh\left(\theta \tilde{q}_k\right)\right] &\le \exp\left(\frac{\theta^2}{2} \sum_{k=1}^{N} \mathbb{E}[\tilde{q}_k]^2 \right) \mathbb{E}\left[\mathbb{E}\left[\exp\left(\theta \sum_{k=1}^{N} \epsilon_k \left[\tilde{q}_k - \mathbb{E}\tilde{q}_k\right] \right) \;\middle|\; \{\epsilon_k\}_{k\in[N]} \right] \right] \\ &\le \exp\left(\frac{\theta^2}{2(\sum_{l=1}^{N+1} \mathbb{E}\lambda_l)^2} \sum_{k=1}^{N} \mathbb{E}[\lambda_k]^2 + \frac{2\theta^2}{(\sum_{l=1}^{N+1} \mathbb{E}\lambda_l)^2} \sum_{k=1}^{N} \sigma_k^2 \right). \end{aligned} \tag{30}$$

Consider the specific choice of θ_{N+1} given by

$$\theta_{N+1} = \frac{\delta (\sum_{k=1}^{N+1} \mathbb{E}\lambda_k)^2}{\sum_{k=1}^{N} \left(4\sigma_N^2 + \mathbb{E}[\lambda_k]^2\right)}.$$

Setting θ_{N+1} into (30), it yields that

$$\begin{aligned} \inf_{\theta>0} \left\{ e^{-\theta\delta} \prod_{k=1}^{N} \mathbb{E}\left[\cosh\left(\theta \tilde{q}_k\right)\right] \right\} &\le \exp\left(-\delta\theta_{N+1} + \frac{\theta_{N+1}^2 \sum_{k=1}^{N} (4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}{2(\sum_{l=1}^{N+1} \mathbb{E}\lambda_l)^2} \right) \\ &= \exp\left(- \frac{\delta^2 (\sum_{k=1}^{N+1} \mathbb{E}\lambda_k)^2}{2\sum_{k=1}^{N} (4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)} \right). \end{aligned} \tag{31}$$

Therefore, applying Theorem B.1 implies that

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}\times\mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(\tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}} - \mathbb{E}\left[\tilde{q}_k \mathbb{1}_{\{V(Z_k)\le V(z)\}}\right] \right) \right\} \ge \delta \right) \le 2\exp\left(- \frac{\delta^2 (\sum_{k=1}^{N+1} \mathbb{E}\lambda_k)^2}{2\sum_{k=1}^{N} (4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)} \right). \tag{32}$$

Eventually, plugging (28)-(29) and (31) into (27) gives

$$\begin{aligned}\mathbb{P}\left(\tilde{\Delta}(\mathcal{D}_N) \geq \delta\right) &\leq 2\exp\left(-\frac{(\delta - \mathbb{E}\lambda_{N+1}/\sum_{l=1}^{N+1}\mathbb{E}\lambda_l)^2(\sum_{k=1}^{N+1}\mathbb{E}\lambda_k)^2}{8\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}\right)\\ &+ \exp\left(-\frac{\{(\delta - \mathbb{E}\lambda_{N+1}/\sum_{l=1}^{N+1}\mathbb{E}\lambda_l)\sum_{k=1}^{N}\mathbb{E}\lambda_k + (\delta - \mathbb{E}\lambda_{N+1}/\sum_{l=1}^{N+1}\mathbb{E}\lambda_l - 2)\mathbb{E}\lambda_{N+1}\}^2}{8\sum_{k=1}^{N}\sigma_k^2}\right) + \mathbb{1}_{\frac{\mathbb{E}\lambda_{N+1}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda_l}>\delta}.\end{aligned} \tag{33}$$

Let $\gamma > 0$, and denote by $\tilde{\delta}_0(\gamma), \tilde{\delta}_1(\gamma)$ the positive real numbers defined by

$$\tilde{\delta}_0(\gamma) = \frac{\mathbb{E}\lambda_{N+1} + \sqrt{8\log(\frac{1}{\gamma})\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}}{\sum_{k=1}^{N+1}\mathbb{E}\lambda_k}, \qquad \tilde{\delta}_1(\gamma) = \frac{3\mathbb{E}\lambda_{N+1} + \sqrt{8\log(\frac{1}{\gamma})\sum_{k=1}^{N}\sigma_k^2}}{\sum_{k=1}^{N+1}\mathbb{E}\lambda_k}.$$

With this notation, remark that

$$\begin{aligned}&\forall \delta \geq \tilde{\delta}_0(\gamma), \qquad \exp\left(-\frac{(\delta - \mathbb{E}\lambda_{N+1}/\sum_{l=1}^{N+1}\mathbb{E}\lambda_l)^2(\sum_{k=1}^{N+1}\mathbb{E}\lambda_k)^2}{8\sum_{k=1}^{N}(4\sigma_k^2 + \mathbb{E}[\lambda_k]^2)}\right) \leq \gamma,\\ &\forall \delta \geq \tilde{\delta}_1(\gamma), \qquad \exp\left(-\frac{\left\{\left(\delta - \frac{\mathbb{E}\lambda_{N+1}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda_l}\right)\sum_{k=1}^{N}\mathbb{E}\lambda_k + \left(\delta - \frac{\mathbb{E}\lambda_{N+1}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda_l} - 2\right)\mathbb{E}\lambda_{N+1}\right\}^2}{8\sum_{k=1}^{N}\sigma_k^2}\right) \leq \gamma,\\ &\forall \delta \geq \tilde{\delta}_0(\gamma) \wedge \tilde{\delta}_1(\gamma), \qquad \mathbb{1}_{\frac{\mathbb{E}\lambda_{N+1}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda_l}>\delta} = 0.\end{aligned}$$

Therefore, for any $\delta \geq \tilde{\delta}_0(\gamma) \vee \tilde{\delta}_1(\gamma)$, (33) shows

$$\mathbb{P}\left(\tilde{\Delta}(\mathcal{D}_N) \geq \delta\right) \leq 3\gamma.$$

Eventually, combining $\sum_{k=1}^{N}\mathbb{E}\lambda_k = N$ with Lemma A.1 concludes the second part of the proof. □

Lemma A.3. *For any $k \in [N]$, set $q_k = \lambda_k / \sum_{l=1}^{N}\mathbb{E}\lambda_l$. If **H**1 holds, then, we have*

$$F_{N+1}(z) = \sum_{k=1}^{N}\mathbb{E}\left[q_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}}\right].$$

Proof. Let $g \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a bounded measurable function and set $z \in \mathcal{X} \times \mathcal{Y}$. We get

$$\begin{aligned}\mathbb{E}\left[g(Z_{N+1})\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{d}P^k}{\mathrm{d}P^{\mathrm{cal}}}(Z_{N+1})g(Z_{N+1})\right]\\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\frac{\mathrm{d}P^k}{\mathrm{d}P^{\mathrm{cal}}}(Z_{N+1})g(Z_{N+1})\right]\\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\frac{\mathrm{d}P^{N+1}}{\mathrm{d}P^{\mathrm{cal}}}(Z_k)g(Z_k)\right]\\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\lambda_k g(Z_k)\right].\end{aligned}$$

Setting $g\colon \tilde{z} \mapsto 1$ gives $\sum_{k=1}^{N}\mathbb{E}\lambda_k = N$ and setting $g\colon \tilde{z} \mapsto \mathbb{1}_{\{V(\tilde{z}) \leq V(z)\}}$ shows that $F_{N+1}(z) = \sum_{k=1}^{N}\mathbb{E}\left[q_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}}\right]$. □

Lemma A.4. *Assume that $\sum_{k=1}^{N}(\lambda_k - \mathbb{E}\lambda_k)$ is $\tilde{\sigma}_N$-sub-Gaussian with parameter $\tilde{\sigma}_N \geq 0$ and denote $q_k = \lambda_k / \sum_{l=1}^{N}\mathbb{E}\lambda_l$. Then, for any $\delta > 0$, it follows*

$$\mathbb{P}\left(\sup_{z \in \mathcal{X} \times \mathcal{Y}}\left\{\sum_{k=1}^{N}(p_k^z - q_k)\mathbb{1}_{\{V(Z_k) \leq V(z)\}}\right\} \geq \delta\right) \leq \exp\left(-\frac{\delta^2(\sum_{k=1}^{N}\mathbb{E}\lambda_k)^2}{2\tilde{\sigma}_N^2}\right).$$

Moreover, if we consider $\tilde{q}_k = \lambda_k / \sum_{l=1}^{N+1} \mathbb{E}\lambda_l$*. Then, for* $\delta \geq \mathbb{E}\lambda_{N+1} / \sum_{l=1}^{N+1} \mathbb{E}\lambda_l$*, it holds*

$$\mathbb{P}\left(\sup_{z \in \mathcal{X} \times \mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(\tilde{q}_k - p_k^z \right) \mathbb{1}_{\{V(Z_k) \leq V(z)\}} \right\} \geq \delta \right) \leq \exp\left(-\frac{(\delta \sum_{k=1}^{N} \mathbb{E}\lambda_k + (\delta - 1)\mathbb{E}\lambda_{N+1})^2}{2\tilde{\sigma}_N^2} \right).$$

Proof. Let $z \in \mathcal{X} \times \mathcal{Y}$, and to simplify notation consider $\Lambda_N(z) := \sum_{l=1}^{N} \lambda_l + \lambda(z)$. By calculation, we obtain

$$\begin{aligned} \sum_{k=1}^{N} \left(p_k^z - q_k \right) \mathbb{1}_{\{V(Z_k) \leq V(z)\}} &= \sum_{k=1}^{N} \frac{\sum_{l=1}^{N} \mathbb{E}\lambda_l - \Lambda_N(z)}{(\sum_{l=1}^{N} \mathbb{E}\lambda_l)\Lambda_N(z)} \lambda_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}} \\ &= \left(\frac{\sum_{l=1}^{N} (\mathbb{E}\lambda_l - \lambda_l)}{\sum_{l=1}^{N} \mathbb{E}\lambda_l} - \frac{\lambda(z)}{\sum_{l=1}^{N} \mathbb{E}\lambda_l} \right) \times \frac{\sum_{k=1}^{N} \lambda_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}}}{\Lambda_N(z)} \\ &= (B - A)\, C, \end{aligned} \tag{34}$$

where we denote

$$A = \frac{\lambda(z)}{\sum_{k=1}^{N} \mathbb{E}\lambda_k}, \qquad B = \frac{\sum_{k=1}^{N} (\mathbb{E}\lambda_k - \lambda_k)}{\sum_{k=1}^{N} \mathbb{E}\lambda_k}, \qquad C = \frac{\sum_{k=1}^{N} \lambda_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}}}{\Lambda_N(z)}.$$

Since $A \geq 0$ and $C \in [0, 1]$, it follows that

$$\begin{aligned} (B - A)C &\leq \max\{0, B - A\} \\ &\leq \max\{0, B\} = \max\left\{ 0, 1 - \frac{\sum_{k=1}^{N} \lambda_k}{\sum_{k=1}^{N} \mathbb{E}\lambda_k} \right\}. \end{aligned}$$

Therefore, using that $\sum_{k=1}^{N} (\lambda_k - \mathbb{E}\lambda_k)$ is $\tilde{\sigma}_N$-sub-Gaussian, it yields

$$\begin{aligned} \mathbb{P}\left(\sup_{z \in \mathcal{X} \times \mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(p_k^z - q_k \right) \mathbb{1}_{\{V(Z_k) \leq V(z)\}} \right\} \geq \delta \right) &\leq \mathbb{P}\left(\max\left\{ 0, 1 - \frac{\sum_{k=1}^{N} \lambda_k}{\sum_{k=1}^{N} \mathbb{E}\lambda_k} \right\} \geq \delta \right) \\ &\leq \mathbb{P}\left(\sum_{k=1}^{N} \lambda_k \leq (1 - \delta) \sum_{k=1}^{N} \mathbb{E}\lambda_k \right) \\ &\leq \exp\left(-\frac{\delta^2 (\sum_{k=1}^{N} \mathbb{E}\lambda_k)^2}{2\tilde{\sigma}_N^2} \right). \end{aligned}$$

This proves the first part of the lemma. Now, we will demonstrate the second part of the lemma. Consider $S_N = \sum_{k=1}^{N} \lambda_k$, $E_{N+1} = \sum_{k=1}^{N+1} \mathbb{E}\lambda_k$ and remark that

$$\begin{aligned} \sum_{k=1}^{N} \left(\tilde{q}_k^z - p_k \right) \mathbb{1}_{\{V(Z_k) \leq V(z)\}} &= \sum_{k=1}^{N} \frac{E_{N+1} - S_N - \lambda(z)}{E_{N+1}(S_N + \lambda(z))} \lambda_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}} \\ &= \frac{E_{N+1} - S_N - \lambda(z)}{E_{N+1}} \times \frac{\sum_{k=1}^{N} \lambda_k \mathbb{1}_{\{V(Z_k) \leq V(z)\}}}{S_N + \lambda(z)} \\ &\leq \max\left\{ 0, \frac{E_{N+1} - S_N - \lambda(z)}{E_{N+1}} \times \frac{S_N}{S_N + \lambda(z)} \right\}. \end{aligned}$$

Since the function $f : \lambda \in \mathbb{R}_+ \mapsto \frac{E_{N+1} - S_N - \lambda}{E_{N+1}} \times \frac{S_N}{S_N + \lambda} \in \mathbb{R}$ is non-increasing, we deduce that

$$\frac{E_{N+1} - S_N - \lambda}{E_{N+1}} \times \frac{S_N}{S_N + \lambda} \leq 1 - \frac{S_N}{E_{N+1}}$$

Thus, we obtain that

$$\mathbb{P}\left(\sup_{z \in \mathcal{X} \times \mathcal{Y}} \left\{ \sum_{k=1}^{N} \left(\tilde{q}_k - p_k^z \right) \mathbb{1}_{\{V(Z_k) \leq V(z)\}} \right\} \geq \delta \right) \leq \mathbb{P}\left(\max\left\{ 0, \frac{E_{N+1} - S_N - \lambda(z)}{E_{N+1}} \times \frac{S_N}{S_N + \lambda(z)} \right\} \geq \delta \right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\frac{E_{N+1} - S_N - \lambda(z)}{E_{N+1}} \times \frac{S_N}{S_N + \lambda(z)} \geq \delta\right) \\
&\leq \mathbb{P}\left(1 - \frac{S_N}{E_{N+1}} \geq \delta\right) \\
&= \mathbb{P}\left(\sum_{k=1}^{N} (\mathbb{E}\lambda_k - \lambda_k) \geq \delta \sum_{k=1}^{N} \mathbb{E}\lambda_k + (\delta - 1)\mathbb{E}\lambda_{N+1}\right).
\end{aligned}
$$

For $\delta \geq \mathbb{E}\lambda_{N+1} / \sum_{k=1}^{N+1} \mathbb{E}\lambda_k$, using the Hoeffding's inequality combined with the previous inequality concludes the proof:

$$
\mathbb{P}\left(\sum_{k=1}^{N} (\mathbb{E}\lambda_k - \lambda_k) \geq \delta \sum_{k=1}^{N} \mathbb{E}\lambda_k + (\delta - 1)\mathbb{E}\lambda_{N+1}\right) \leq \exp\left(-\frac{(\delta \sum_{k=1}^{N} \mathbb{E}\lambda_k + (\delta - 1)\mathbb{E}\lambda_{N+1})^2}{2\tilde{\sigma}_N^2}\right).
$$

□

B Extension of the DKW theorem

Consider $\{g_k \colon z \mapsto g_k(z)\}_{k \in [N]}$ a family of real valued functions such that the random variables $\{g_k(Z_k)\}_{k \in [N]}$ are mutually pairwise independent, and define the empirical cumulative function given for $z \in \mathcal{Z}$, by

$$
\hat{G}(z) = \sum_{k=1}^{N} g_k(Z_k) \mathbb{1}_{V(Z_k) \leq V(z)}, \qquad G(z) = \mathbb{E}\big[\hat{G}(z)\big].
$$

In this paragraph we control $\mathbb{P}(\sup_{z \in \mathcal{Z}} \{\hat{G}_N(z) - G(z)\} \geq \epsilon)$. The tools utilized in this proof are closed to the ones developed in (Dvoretzky et al., 1956; Massart, 1990; Bartl and Mendelson, 2023).

Theorem B.1. *For any $\epsilon > 0$, the following inequality holds*

$$
\mathbb{P}\left(\sup_{z \in \mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\} \geq \epsilon\right) \leq 2 \inf_{\theta > 0} \left\{ \mathrm{e}^{-\theta\epsilon} \prod_{k=1}^{N} \mathbb{E}\left[\cosh\left(\theta g_k(Z_k)\right)\right] \right\}.
$$

Proof. First, for any $\theta > 0$, applying Markov's inequality gives

$$
\mathbb{P}\left(\sup_{z \in \mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\} \geq \epsilon\right) \leq \mathrm{e}^{-\theta\epsilon} \mathbb{E}\left[\exp\left(\theta \sup_{z \in \mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\}\right)\right]. \tag{35}
$$

Moreover, Lemma B.3 shows that

$$
\mathbb{E}\left[\exp\left(\theta \sup_{z \in \mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\}\right)\right] \leq 2 \exp\left(2\theta^2 \sum_{k=1}^{N} (g_k(Z_k))^2\right).
$$

Plugging the previous inequality into (35), and using the independence between $\{g_k(Z_k)\}_{k=1}^N$ implies

$$
\mathbb{P}\left(\sup_{z \in \mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\}\right) \leq 2 \inf_{\theta > 0} \left\{ \mathrm{e}^{-\theta\epsilon} \prod_{k=1}^{N} \mathbb{E}\left[\cosh\left(\theta g_k(Z_k)\right)\right] \right\}.
$$

□

Lemma B.2. *Let $\{\epsilon_i\}_{i \in [n]}$ be i.i.d Rademacher random variables taking values in $\{-1, 1\}$, then for any $\theta > 0$ and $\{p_j\}_{j \in [N]} \in \mathbb{R}^N$, we have*

$$
\mathbb{E}\left[\exp\left(\theta \sup_{0 \leq i \leq N} \sum_{j=1}^{i} p_j \epsilon_j\right)\right] \leq 2 \prod_{k=1}^{N} \cosh\left(\theta p_k\right).
$$

By convention, we consider $\sum_{j=1}^{0} p_j \epsilon_j = 0$.

Proof. First we will show that

$$\forall t \in \mathbb{R}_+, \qquad \mathbb{P}\left(\max_{i=0}^{N}\sum_{j=1}^{i} p_j\epsilon_j \geq t\right) \leq 2\mathbb{P}\left(\sum_{j=1}^{N} p_j\epsilon_j \geq t\right). \tag{36}$$

To prove the previous inequality, for any $i \in \{1, \dots, N\}$ consider the following set

$$E_i = \left\{\sum_{j=1}^{i} p_j\epsilon_j \geq t\right\} \bigcap_{l=1}^{i-1} \left\{\sum_{j=1}^{l} p_j\epsilon_j < t\right\},$$

and write

$$E_0 = \begin{cases} \emptyset & \text{if } t > 0 \\ \Omega & \text{if } t \leq 0 \end{cases}.$$

Observe that $\{E_i\}_{i=0}^{N}$ are pairwise disjoint and also that

$$\left\{\max_{i=0}^{N}\sum_{j=1}^{i} p_j\epsilon_j \geq t\right\} = \bigcup_{i=0}^{N} E_i. \tag{37}$$

Secondly, note that

$$\bigcup_{i=0}^{N}\left(E_i \cap \left\{\sum_{j=i+1}^{N} p_j\epsilon_j \geq 0\right\}\right) \subset \left\{\sum_{j=1}^{N} p_j\epsilon_j \geq t\right\}. \tag{38}$$

Since $\sum_{j=i+1}^{N} p_j\epsilon_j$ is symmetric around 0, we have

$$\mathbb{P}\left(\sum_{j=i+1}^{N} p_j\epsilon_j \geq 0\right) \geq \frac{1}{2}. \tag{39}$$

Moreover, by independence, we get

$$\mathbb{P}\left(E_i \cap \left\{\sum_{j=i+1}^{N} p_j\epsilon_j \geq 0\right\}\right) = \mathbb{P}\left(E_i\right)\mathbb{P}\left(\sum_{j=i+1}^{N} p_j\epsilon_j \geq 0\right) \geq \frac{\mathbb{P}\left(E_i\right)}{2}. \tag{40}$$

Thus, combining (37) and (38) with (40) implies that

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{N} p_j\epsilon_j \geq t\right) &\geq \sum_{i=0}^{N}\mathbb{P}\left(E_i \cap \left\{\sum_{j=i+1}^{N} p_j\epsilon_j \geq 0\right\}\right) \\
&\geq \sum_{i=0}^{N}\frac{\mathbb{P}\left(E_i\right)}{2} = \frac{1}{2}\mathbb{P}\left(\cup_{i=0}^{N} E_i\right) \\
&= \frac{1}{2}\mathbb{P}\left\{\sup_{i=0}^{N}\sum_{j=1}^{i} p_j\epsilon_j \geq t\right\}.
\end{aligned}$$

Therefore, the last line concludes the proof of (36). Note that for any differentiable function f such that $f' \geq 0$, the Fubini-Tonelli's theorem shows that

$$\begin{aligned}
\mathbb{E}\left[f(U)\mathbb{1}_{\{U\geq 0\}}\right] &= \mathbb{E}\left[\left(f(0) + \int_0^U f'(t)\mathrm{d}t\right)\mathbb{1}_{\{U\geq 0\}}\right] \\
&= f(0)\mathbb{E}\left[\mathbb{1}_{\{U\geq 0\}}\right] + \mathbb{E}\left[\int_0^\infty f'(t)\mathbb{1}_{\{U\geq t\}}\mathrm{d}t\mathbb{1}_{\{U\geq 0\}}\right] \\
&= f(0)\mathbb{P}\left(U \geq 0\right) + \int_0^\infty f'(t)\mathbb{P}\left(U \geq t\right)\mathrm{d}t.
\end{aligned} \tag{41}$$

Now, we apply this previous formula to the function $f(t) = \mathrm{e}^{\theta t}$ and the random variable $U = \max_{i=0}^{N} \sum_{j=1}^{i} p_j \epsilon_j$. We get

$$\begin{aligned}\mathbb{E}\left[f\left(U\right)\right] &= \mathbb{E}\left[f(U)\mathbb{1}_{U<0}\right] + \mathbb{E}\left[f(U)\mathbb{1}_{U\geq 0}\right] \\ &= \theta \int_0^{\infty} \mathrm{e}^{\theta t} \mathbb{P}\left(U \geq t\right) \mathrm{d}t \\ &\overset{(36)}{\leq} 1 + 2\theta \int_0^{\infty} \mathrm{e}^{\theta t} \mathbb{P}\left(\sum_{j=1}^{N} p_j \epsilon_j \geq t\right) \mathrm{d}t.\end{aligned} \tag{42}$$

Once again, applying (41) with $V = \sum_{j=1}^{N} p_j \epsilon_j$, it yields

$$\begin{aligned}\theta \int_0^{\infty} \mathrm{e}^{\theta t} \mathbb{P}\left(V \geq t\right) \mathrm{d}t &= \mathbb{E}\left[f(V)\mathbb{1}_{\{V\geq 0\}}\right] - \mathbb{P}\left(V \geq 0\right) \\ &\overset{(39)}{\leq} \mathbb{E}\left[f(V)\right] - \frac{1}{2}.\end{aligned} \tag{43}$$

Hence, plugging (43) inside (42) gives

$$\mathbb{E}\left[\exp\left(\theta \max_{i=0}^{N} \sum_{j=1}^{i} p_j \epsilon_j\right)\right] \leq 2\mathbb{E}\left[\exp\left(\theta \sum_{j=1}^{N} p_j \epsilon_j\right)\right] = 2 \prod_{j=1}^{N} \mathbb{E}\left[\exp\left(\theta p_j \epsilon_j\right)\right].$$

The proof is finished using that $\mathbb{E}[\exp(\theta p_j \epsilon_j)] = \cosh(\theta p_j)$. □

Lemma B.3. *Let $\theta > 0$, we have*

$$\mathbb{E}\left[\exp\left(\theta \sup_{z\in\mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\}\right)\right] \leq 2 \prod_{k=1}^{N} \mathbb{E}\left[\cosh\left(\theta g_k(Z_k)\right)\right].$$

Proof. Let $\theta > 0$ be fixed, since $t \mapsto \mathrm{e}^{\theta t}$ is increasing, the supremum can be inverted with the exponential:

$$\mathbb{E}\left[\exp\left(\theta \sup_{z\in\mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\}\right)\right] = \mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(\theta \left\{\hat{G}_N(z) - G(z)\right\}\right)\right].$$

For any $k \in [N]$, consider $\tilde{Z}_k$ an independent copy of the random variable Z_k. The linearity of the expectation gives

$$\begin{aligned}&\sum_{k=1}^{N} \left(g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} - \mathbb{E}\left[g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)}\right]\right) \\ &\qquad\qquad = \mathbb{E}\left[\sum_{k=1}^{N} \left(g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} - g_k(\tilde{Z}_k)\mathbb{1}_{V(\tilde{Z}_k)\leq V(z)}\right) \;\middle|\; \{Z_k\}_{k=1}^{N}\right].\end{aligned}$$

Therefore, the Jensen's inequality implies

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\theta \sup_{z\in\mathcal{Z}} \left\{\hat{G}_N(z) - G(z)\right\}\right)\right] \\ &= \mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(\theta \mathbb{E}\left[\sum_{k=1}^{N} \left\{g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} - g_k(\tilde{Z}_k)\mathbb{1}_{V(\tilde{Z}_k)\leq V(z)}\right\} \;\middle|\; \{Z_k\}_{k=1}^{N}\right]\right)\right] \\ &\leq \mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(\theta \sum_{k=1}^{N} \left\{g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} - g_k(\tilde{Z}_k)\mathbb{1}_{V(\tilde{Z}_k)\leq V(z)}\right\}\right)\right].\end{aligned}$$

Let $\{\epsilon_k\}_{k\in[N]}$ be i.i.d. random Rademacher variables independent of $\{(Z_k, \tilde{Z}_k)\}_{k=1}$, we have

$$\mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(\theta \sum_{k=1}^{N}\left\{g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} - g_k(\tilde{Z}_k)\mathbb{1}_{V(\tilde{Z}_k)\leq V(z)}\right\}\right)\right]$$
$$= \mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(\theta \sum_{k=1}^{N}\epsilon_k\left\{g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} - g_k(\tilde{Z}_k)\mathbb{1}_{V(\tilde{Z}_k)\leq V(z)}\right\}\right)\right].$$

Using the Cauchy-Schwarz's inequality, we deduce that

$$\mathbb{E}\left[\exp\left(\theta \sup_{z\in\mathcal{Z}}\left\{\hat{G}_N(z) - G(z)\right\}\right)\right] \leq \mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(2\theta\sum_{k=1}^{N}\epsilon_k g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)}\right)\right].$$

Given the random variables $\{V(Z_k)\}_{k=1}^N$, denote by σ the permutation of $[N]$ such that $V(Z_{\sigma(1)}) \leq \cdots \leq V(Z_{\sigma(N)})$. In particular, it holds

$$\sum_{k=1}^{N}\epsilon_k g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)} = \begin{cases} 0 & \text{if } V(z) < V(Z_{\sigma(1)}) \\ \sum_{j=1}^{i}\epsilon_{\sigma(j)}p_{\sigma(j)}^{Z_{\sigma(j)}} & \text{if } V(Z_{\sigma(i)}) \leq V(z) < V(Z_{\sigma(i+1)}) \\ \sum_{j=1}^{N}\epsilon_{\sigma(j)}p_{\sigma(j)}^{Z_{\sigma(j)}} & \text{if } V(z) \geq V(Z_{\sigma(N)}) \end{cases}.$$

Thus, can rewrite the supremum as

$$\sup_{z\in\mathcal{Z}} \exp\left(2\theta\sum_{k=1}^{N}\epsilon_k g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)}\right) \leq \sup_{0\leq i\leq n} \exp\left(2\theta\sum_{j=1}^{i}\epsilon_{\sigma(j)}p_{\sigma(j)}^{Z_{\sigma(j)}}\right).$$

Applying Lemma B.2, we finally obtain that

$$\mathbb{E}\left[\sup_{z\in\mathcal{Z}} \exp\left(2\theta\sum_{k=1}^{N}\epsilon_k g_k(Z_k)\mathbb{1}_{V(Z_k)\leq V(z)}\right) \,\middle|\, \{Z_k\}_{k=1}^N\right]$$
$$\leq \mathbb{E}\left[\sup_{0\leq i\leq N} \exp\left(2\theta\sum_{j=1}^{i}\epsilon_{\sigma(j)}p_{\sigma(j)}^{Z_{\sigma(j)}}\right) \,\middle|\, \{Z_k\}_{k=1}^N\right] \leq 2\prod_{k=1}^{N}\cosh\left(\theta g_k(Z_k)\right).$$

□

C Proof of Theorem 2.2

The objective of this section is to investigate the bias introduced by the conformal procedure developed in Section 2. Due to distribution shift, studying this bias is different from the case of exchangeable data. To address this, we introduce a sequence $\{\tilde{Z}_k\}_{k\in[N]}$ of independent and identically distributed random variables following the distribution P^{cal}. To simplify notation, we consider $\tilde{Z}_{N+1} = Z_{N+1}$. Note that, unlike the other data points, $\tilde{Z}_{N+1}$ is drawn from P^{N+1}. For all $k \in [N+1]$, we define:

$$p_k = \frac{\lambda(Z_k)}{\sum_{l=1}^{N+1}\lambda(Z_l)}, \qquad q_k = \frac{\lambda(\tilde{Z}_k)}{\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)}. \tag{44}$$

Furthermore, let's introduce the following two random variables:

$$\Gamma = \sum_{k=1}^{N} q_k\mathbb{1}_{\tilde{V}_k<V_{N+1}}, \quad \delta = \sum_{k=1}^{N} p_k\mathbb{1}_{V_k<V_{N+1}} - \sum_{k=1}^{N} q_k\mathbb{1}_{\tilde{V}_k<V_{N+1}}, \tag{45}$$

where $\tilde{V}_k = V(\tilde{Z}_k)$, $V_k = V(Z_k)$ and $Z_k = (X_k, Y_k) \sim P^k$.

Lemma C.1. *Assume there are no ties between there are no ties between $\{V_k\}_{k\in[N+1]} \cup \{\infty\}$ almost surely. For any $\epsilon \geq 0$, it holds*

$$-\epsilon - \mathbb{P}(\delta \leq -\epsilon) \leq \mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N} p_k\delta_{V_k} + p_{N+1}\delta_\infty\right)\right) - 1 + \alpha \leq \epsilon + \mathbb{P}(\delta \geq \epsilon) + \mathbb{E}\left[\max_{k=1}^{N+1}\{q_k\}\right].$$

Proof. First, using the definition of the quantile combined with (45), we have

$$\left(V_{N+1} < Q_{1-\alpha}\left(\textstyle\sum_{k=1}^N p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right) \iff (\Gamma + \delta > \alpha)\,.$$

Therefore, using the no ties assumption on $\{V_k \colon k \in [N+1]\} \cup \{\infty\}$, it holds that

$$\begin{aligned}\mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^N p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right) &= \mathbb{P}\left(V_{N+1} < Q_{1-\alpha}\left(\textstyle\sum_{k=1}^N p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right)\\ &= \mathbb{E}\left[\mathbb{1}_{\Gamma>\alpha}\right] + \mathbb{E}\left[\mathbb{1}_{\Gamma+\delta>\alpha} - \mathbb{1}_{\Gamma>\alpha}\right]. \end{aligned} \tag{46}$$

Moreover, for any $\epsilon \geq 0$, remark that

$$\mathbb{1}_{\Gamma>\alpha+\epsilon} - \mathbb{1}_{\delta\leq-\epsilon} \leq \mathbb{1}_{\Gamma+\delta>\alpha} \leq \mathbb{1}_{\Gamma>\alpha-\epsilon} + \mathbb{1}_{\delta\geq\epsilon}. \tag{47}$$

Thus, combining (46) with (47) gives

$$\mathbb{P}\left(\Gamma > \alpha+\epsilon\right) - \mathbb{P}\left(\delta \leq -\epsilon\right) \leq \mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^N p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right) \leq \mathbb{P}\left(\Gamma > \alpha-\epsilon\right) + \mathbb{P}\left(\delta \geq \epsilon\right). \tag{48}$$

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

$$(X > \tilde{\alpha}) \iff \left(V_{N+1} < Q_{1-\tilde{\alpha}}\left(\textstyle\sum_{k=1}^N q_k \delta_{V_k} + q_{N+1}\delta_\infty\right)\right).$$

Once again, the no ties assumption implies that

$$\begin{aligned}\mathbb{P}(X > \tilde{\alpha}) &= \mathbb{P}(V_{N+1} < Q_{1-\tilde{\alpha}}(\textstyle\sum_{k=1}^N q_k \delta_{V_k} + q_{N+1}\delta_{V_\infty}))\\ &= \mathbb{P}(V_{N+1} \leq Q_{1-\tilde{\alpha}}(\textstyle\sum_{k=1}^N q_k \delta_{V_k} + q_{N+1}\delta_{V_\infty}))\\ &= \mathbb{P}(V_{N+1} \leq Q_{1-\tilde{\alpha}}(\textstyle\sum_{k=1}^{N+1} q_k \delta_{V_k})).\end{aligned}$$

Applying the result from (Tibshirani et al., 2019, Lemma 3), it gives that

$$0 \leq \mathbb{P}\left(V_{N+1} \leq Q_{1-\tilde{\alpha}}\left(\textstyle\sum_{k=1}^{N+1} q_k \delta_{V_k}\right)\right) - 1 + \tilde{\alpha} \leq \mathbb{E}\left[\max_{k=1}^{N+1}\{q_k\}\right].$$

Therefore, setting $\tilde{\alpha}$ as follows:

$$\begin{cases}\tilde{\alpha} = \alpha + \epsilon, & \forall \epsilon \in [0, 1-\alpha)\\ \tilde{\alpha} = \alpha - \epsilon, & \forall \epsilon \in [0, \alpha)\end{cases};$$

using (48) we obtain

$$\begin{cases}1 - (\alpha+\epsilon) - \mathbb{P}\left(\delta \leq -\epsilon\right) \leq \mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}(\textstyle\sum_{k=1}^N p_k \delta_{V_k} + p_{N+1}\delta_\infty)\right), & \forall \epsilon \in [0, 1-\alpha)\\ \mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}(\textstyle\sum_{k=1}^N p_k \delta_{V_k} + p_{N+1}\delta_\infty)\right) \leq 1 - \alpha + \epsilon + \mathbb{P}\left(\delta \geq \epsilon\right), & \forall \epsilon \in [0, \alpha)\end{cases},$$

The proof is concluded since the previous inequalities hold $\forall \epsilon \geq 0$. □

Let $\sigma_{\mathrm{cal},\mathbb{1}}, \sigma_{\mathrm{cal}} > 0$ and $\forall k \in [N+1]$ consider $(\sigma_{k,\mathbb{1}}, \sigma_k) \in (\mathbb{R}_+)^2$. Moreover, define

$$\epsilon_N = 8\sqrt{\frac{2\log 4N}{N}}\left(\sigma_{\mathrm{cal},\mathbb{1}}^2 \vee \sigma_{\mathbb{1}} \vee \sqrt{\sigma_{\mathrm{cal}}^2 + \sigma^2}\right). \tag{49}$$

Lemma C.2. *Assume that* $\frac{1}{N}\sum_{k=1}^N\{\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<v} - \mathbb{E}[\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<v}]\}$, $\frac{1}{N}\sum_{k=1}^N\{\lambda(Z_k)\mathbb{1}_{V_k<v} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}]\}$, $\frac{1}{N}\sum_{k=1}^N\{\lambda(\tilde{Z}_k) - \mathbb{E}\lambda(\tilde{Z}_k)\}$ *and* $\frac{1}{N}\sum_{k=1}^N\{\lambda(Z_k) - \mathbb{E}\lambda(Z_k)\}$ *are sub-Gaussian with parameters* $\{\sigma_{\mathbb{1}}, \sigma_{\mathrm{cal},\mathbb{1}}, \sigma, \sigma_{\mathrm{cal}}\}$ *respectively,* $\forall v \in \mathbb{R}$*. It holds*

$$\mathbb{P}\left(\delta > \epsilon_N\right) \vee \mathbb{P}\left(\delta < -\epsilon_N\right) \leq \frac{1 + 4\,\mathrm{Var}\,\lambda(\tilde{Z}_1)}{N} + \frac{4\sum_{l=1}^N \mathrm{Var}\,\lambda(Z_l) + 8\,\mathrm{Var}\,\lambda(Z_{N+1})}{N^2},$$

where δ *is defined in* (45).

Proof. First, recall that $\lambda = \mathrm{d}P^{N+1}/\mathrm{d}P^k$ and $P^{\mathrm{cal}} = (1/N)\sum_{k=1}^N P^k$. Since $Z_k \sim P^k$ and $\tilde{Z}_k \sim P^{\mathrm{cal}}$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^N \lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \,\middle|\, V_{N+1}\right] &= \mathbb{E}\left[\sum_{k=1}^N \int \lambda(z)\mathbb{1}_{V(z)<V_{N+1}}\mathrm{d}P^k(z) \,\middle|\, V_{N+1}\right] \\
&= \mathbb{E}\left[N\int \lambda(z)\mathbb{1}_{V(z)<V_{N+1}}\mathrm{d}P^{\mathrm{cal}}(z) \,\middle|\, V_{N+1}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^N \lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}} \,\middle|\, V_{N+1}\right]. \qquad (50)
\end{aligned}$$

Using definition of the importance weights p_k given in (44), note that

$$\sum_{k=1}^N p_k \mathbb{1}_{V_k<V_{N+1}} = \frac{\sum_{k=1}^N \lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}}}{\sum_{l=1}^{N+1}\lambda(Z_l)}, \quad \sum_{k=1}^N q_k \mathbb{1}_{V_k<V_{N+1}} = \frac{\sum_{k=1}^N \lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}}}{\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)}.$$

Therefore, (50) implies that

$$\begin{aligned}
\delta &= \frac{\left(\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)\right)\sum_{k=1}^N \lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} - \left(\sum_{l=1}^{N+1}\lambda(Z_l)\right)\sum_{k=1}^N \lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}}}{\left(\sum_{l=1}^{N+1}\lambda(Z_l)\right)\left(\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)\right)} \\
&= \frac{\sum_{k=1}^N\{\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \mid V_{N+1}]\}}{\sum_{l=1}^{N+1}\lambda(Z_l)} \\
&+ \frac{\sum_{l=1}^{N+1}\{\lambda(Z_l)-\lambda(\tilde{Z}_l)\}}{\left(\sum_{l=1}^{N+1}\lambda(Z_l)\right)\left(\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)\right)}\mathbb{E}\left[\sum_{k=1}^N \lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \,\middle|\, V_{N+1}\right] \\
&\quad - \frac{\sum_{k=1}^N\{\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}} - \mathbb{E}[\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}} \mid V_{N+1}]\}}{\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)}. \qquad (51)
\end{aligned}$$

Moreover, define the following set:

$$A_N = \left\{\sum_{l=1}^{N+1}\lambda(Z_l) \le \frac{1}{2}\mathbb{E}\left[\sum_{l=1}^{N+1}\lambda(Z_l)\right]\right\} \bigcup \left\{\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l) \le \frac{1}{2}\mathbb{E}\left[\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)\right]\right\}.$$

Hence, the Bienaymé-Tchebytchev's inequality shows that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{l=1}^{N+1}\lambda(Z_l) \le \frac{1}{2}\mathbb{E}\left[\sum_{l=1}^{N+1}\lambda(Z_l)\right]\right) \le \mathbb{P}\left(\sum_{l=1}^{N+1}\{\mathbb{E}\lambda(Z_l)-\lambda(Z_l)\} \ge \frac{1}{2}\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l)\right) \le \frac{4\sum_{l=1}^{N+1}\operatorname{Var}\lambda(Z_l)}{(\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l))^2}, \\
&\mathbb{P}\left(\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l) \le \frac{1}{2}\mathbb{E}\left[\sum_{l=1}^{N+1}\lambda(\tilde{Z}_l)\right]\right) \le \frac{4\sum_{l=1}^{N+1}\operatorname{Var}\lambda(\tilde{Z}_l)}{(\sum_{l=1}^{N+1}\mathbb{E}\lambda(\tilde{Z}_l))^2}.
\end{aligned}$$

Thus, summing these two inequalities gives that

$$\mathbb{P}(A_N) \le \frac{4\sum_{l=1}^{N+1}\{\operatorname{Var}\lambda(Z_l)+\operatorname{Var}\lambda(\tilde{Z}_l)\}}{(\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l))^2}. \qquad (52)$$

Let $\epsilon > 0$ be fixed. Since the $\{\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \mid V_{N+1}]\}$ are assumed $\sigma_{k,\mathbb{1}}$-sub-Gaussian, it holds:

$$\mathbb{P}\left(\frac{\sum_{k=1}^N\{\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \mid V_{N+1}]\}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l)} \ge \epsilon\right) \le \exp\left(-\frac{\epsilon^2(\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l))^2}{2N\sigma^2_{\mathrm{cal},\mathbb{1}}}\right). \qquad (53)$$

Similarly, the Hoeffding's inequality implies that

$$\mathbb{P}\left(\frac{\sum_{k=1}^N\{\mathbb{E}[\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}} \mid V_{N+1}] - \lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}}\}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda(\tilde{Z}_l)} \ge \epsilon\right) \le \exp\left(-\frac{\epsilon^2(\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l))^2}{2N\sigma^2_{\mathbb{1}}}\right). \qquad (54)$$

Moreover, define $\tau \in [0,1]$ by:

$$\tau = \frac{\mathbb{E}[\sum_{k=1}^{N} \lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \,|\, V_{N+1}]}{\sum_{l=1}^{N+1} \mathbb{E}\lambda(Z_l)}.$$

Using that $\sum_{l=1}^{N} \mathbb{E}\lambda(Z_l) = \sum_{l=1}^{N} \mathbb{E}\lambda(\tilde{Z}_l)$, we obtain

$$\mathbb{P}\left(\frac{\sum_{l=1}^{N}\{\lambda(Z_l)-\lambda(\tilde{Z}_l)\}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l)} \geq \frac{\epsilon}{\tau}\right) \leq \exp\left(-\frac{\epsilon^2(\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l))^2}{2N\sigma_{\text{cal}}^2 + 2N\sigma^2}\right). \tag{55}$$

Therefore, combining (51)-(52)-(53)-(54)-(55) with $\epsilon = \epsilon_N$, we obtain

$$\begin{aligned}\mathbb{P}\left(\delta \geq \frac{\epsilon_N}{4} + \frac{\epsilon_N}{4} + \frac{\epsilon_N}{2}\right) &\leq \mathbb{P}\left(A_N\right) + \mathbb{P}\left(\frac{\sum_{k=1}^{N}\{\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<V_{N+1}} \,|\, V_{N+1}]\}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l)} \geq \frac{\epsilon_N}{8}\right) \\ &+ \mathbb{P}\left(\frac{\sum_{k=1}^{N}\{\mathbb{E}[\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}} \,|\, V_{N+1}] - \lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<V_{N+1}}\}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda(\tilde{Z}_l)} \geq \frac{\epsilon_N}{8}\right) + \mathbb{P}\left(\frac{\sum_{l=1}^{N}\{\lambda(Z_l)-\lambda(\tilde{Z}_l)\}}{\sum_{l=1}^{N+1}\mathbb{E}\lambda(Z_l)} \geq \frac{\epsilon_N}{8}\right).\end{aligned}$$

Finally, a similar reasoning for $\mathbb{P}(\delta \leq -\epsilon_N)$ and using $\sum_{l=1}^{N}\mathbb{E}\lambda(Z_l) = N$ concludes the proof. □

Lemma C.3. *Assume for $v \in \mathbb{R}$, that $\lambda(\tilde{Z}_1)\mathbb{1}_{\tilde{V}_1<v} - \mathbb{E}[\lambda(\tilde{Z}_1)\mathbb{1}_{\tilde{V}_1<v}]$ is σ-sub-Gaussian with parameter $\sigma \geq 0$. Then, $\frac{1}{N}\sum_{k=1}^{N}\{\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<v} - \mathbb{E}[\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<v}]\}$, $\frac{1}{N}\sum_{k=1}^{N}\{\lambda(Z_k)\mathbb{1}_{V_k<v} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}]\}$, $\frac{1}{N}\sum_{k=1}^{N}\{\lambda(\tilde{Z}_k) - \mathbb{E}\lambda(\tilde{Z}_k)\}$ and $\frac{1}{N}\sum_{k=1}^{N}\{\lambda(Z_k) - \mathbb{E}\lambda(Z_k)\}$ are σ-sub-Gaussian.*

Proof. Let $v \in \mathbb{R}$ and $\gamma \in \mathbb{R}$. By concavity of the logarithm, the Jensen's inequality implies

$$\frac{1}{N}\sum_{k=1}^{N}\log\mathbb{E}\left[\exp\left\{\gamma\lambda(Z_k)\mathbb{1}_{V_k<v}\right\}\right] \leq \log\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\exp\left\{\gamma\lambda(Z_k)\mathbb{1}_{V_k<v}\right\}\right].$$

Therefore, by the increasing property of the exponential, we obtain

$$\prod_{k=1}^{N}\mathbb{E}\left[\exp\left\{\gamma\lambda(Z_k)\mathbb{1}_{V_k<v}\right\}\right] \leq \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\exp\left\{\gamma\lambda(Z_k)\mathbb{1}_{V_k<v}\right\}\right]^N = \prod_{k=1}^{N}\mathbb{E}\left[\exp\left\{\gamma\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<v}\right\}\right].$$

Since $\sum_{k=1}^{N}\mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}] = \sum_{k=1}^{N}\mathbb{E}[\lambda(\tilde{Z}_k)\mathbb{1}_{\tilde{V}_k<v}]$, multiplying by $\exp(-\gamma\sum_{k=1}^{N}\mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}])$ the previous inequality shows that $\frac{1}{N}\sum_{k=1}^{N}\{\lambda(Z_k)\mathbb{1}_{V_k<v} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}]\}$ is σ-sub-Gaussian. Moreover, applying Fatou's lemma gives that

$$\begin{aligned}\mathbb{E}\left[\liminf_{v} \mathrm{e}^{\frac{\gamma}{N}\sum_{k=1}^{N}\{\lambda(Z_k)\mathbb{1}_{V_k<v} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}]\}}\right] &\leq \liminf_{v}\mathbb{E}\left[\mathrm{e}^{\frac{\gamma}{N}\sum_{k=1}^{N}\{\lambda(Z_k)\mathbb{1}_{V_k<v} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}]\}}\right] \\ &\leq \exp\left(\frac{\gamma^2\sigma^2}{2}\right).\end{aligned}$$

For $k \in [N]$, since $V_k < \infty$ almost surely, the dominated convergence theorem combined with the continuity of the exponential function imply that

$$\liminf_{v}\exp\left(\frac{\gamma}{N}\sum_{k=1}^{N}\{\lambda(Z_k)\mathbb{1}_{V_k<v} - \mathbb{E}[\lambda(Z_k)\mathbb{1}_{V_k<v}]\}\right) = \exp\left(\frac{\gamma}{N}\sum_{k=1}^{N}\{\lambda(Z_k) - \mathbb{E}[\lambda(Z_k)]\}\right).$$

Thus, we deduce that $\frac{1}{N}\sum_{k=1}^{N}\{\lambda(Z_k) - \mathbb{E}[\lambda(Z_k)]\}$ is σ-sub-Gaussian. □

In order to simplify the calculation, we also provide the statement when assuming that λ is bounded by $\|\lambda\|_\infty$.

Theorem C.4. *Assume there are no ties between $\{V_k\}_{k\in[N+1]} \cup \{\infty\}$ almost surely. If $\exists\sigma \geq 0$, such that $\forall v \in \mathbb{R}$, $\lambda(\tilde{Z}_1)\mathbb{1}_{\tilde{V}_1<v} - \mathbb{E}[\lambda(\tilde{Z}_1)\mathbb{1}_{\tilde{V}_1<v}]$ is sub-Gaussian with parameter σ. Then, it holds*

$$\left|\mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\sum_{k=1}^{N} p_k\delta_{V_k} + p_{N+1}\delta_\infty\right)\right) - 1 + \alpha\right| \leq \begin{cases} 35\,\|\lambda\|_\infty^2\sqrt{\frac{\log 4N}{N}} & \text{if } \lambda \text{ is bounded} \\ \frac{18\mathbb{E}\lambda^2(Z_{N+1})}{N} + 19\sigma\sqrt{\frac{\log 4N}{N}} & \text{otherwise}\end{cases}.$$

Proof. Let's start by using Lemma C.1 with $\epsilon = \epsilon_N$ defined in (49), it gives

$$\begin{aligned} -\epsilon_N - \mathbb{P}\left(\delta \leq -\epsilon_N\right) \leq \mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N} p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right) - 1 + \alpha \\ \leq \epsilon_N + \mathbb{P}\left(\delta \geq \epsilon_N\right) + \mathbb{E}\left[\max_{k=1}^{N+1}\{q_k\}\right]. \end{aligned} \tag{56}$$

Therefore, the previous inequalities combined with Lemma C.2 and Lemma C.3 imply that

$$\begin{aligned} &\left|\mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N} p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right) - 1 + \alpha\right| \\ &\quad\leq \frac{1 + 4\operatorname{Var}\lambda(\tilde{Z}_1)}{N} + \frac{4\sum_{l=1}^{N}\operatorname{Var}\lambda(Z_l) + 8\operatorname{Var}\lambda(Z_{N+1})}{N^2} + 16\sigma\sqrt{\frac{\log 4N}{N}} + \mathbb{E}\left[\max_{k=1}^{N+1}\{q_k\}\right]. \end{aligned} \tag{57}$$

Furthermore, applying the result from (Plassier et al., 2023, Lemma C.17) implies that

$$\mathbb{E}\left[\max_{k=1}^{N+1} q_k\right] \leq \begin{cases} \frac{2\|\lambda\|_\infty}{N} + \frac{4}{N}\left(\operatorname{Var}\lambda(\tilde{Z}_1) + \frac{\operatorname{Var}\lambda(Z_{N+1})}{N}\right) \leq \frac{10\|\lambda\|_\infty^2}{N} & \text{if } \lambda \text{ is bounded} \\ \frac{\sigma\sqrt{8\log(N+1)}}{N} + \frac{2}{N}\left(\mathbb{E}\lambda(Z_{N+1}) + 2\operatorname{Var}\lambda(\tilde{Z}_1) + \frac{2\operatorname{Var}\lambda(Z_{N+1})}{N}\right) & \text{otherwise} \end{cases}. \tag{58}$$

Therefore, if λ is bounded by $\|\lambda\|_\infty$, then the sub-Gaussian parameters are controled as follows:

$$\sigma \leq \|\lambda\|_\infty / 2;$$

see (Boucheron et al., 2003, Lemma 2.2) for more details. Hence, plugging (58) into (57) concludes the first part of the proof. Now, let's proof the second part of the theorem. For that, remark

$$\mathbb{E}\lambda(\tilde{Z}_1) = 1, \qquad\qquad \operatorname{Var}\lambda(\tilde{Z}_1) = \mathbb{E}[\lambda^2(\tilde{Z}_1)] - 1, \tag{59}$$

and note that

$$\mathbb{E}\lambda(Z_{N+1}) = \int \left(\frac{\mathrm{d}P^{N+1}}{\mathrm{d}P^{\mathrm{cal}}}\right)^2 (z)\,\mathrm{d}P^{\mathrm{cal}}(z) = \mathbb{E}[\lambda^2(\tilde{Z}_1)]. \tag{60}$$

Therefore, we obtain that

$$\operatorname{Var}\lambda(Z_{N+1}) = \mathbb{E}[\lambda^2(Z_{N+1})] - \mathbb{E}\left[\lambda(Z_{N+1})\right]^2 = \mathbb{E}[\lambda^3(\tilde{Z}_1)] - \mathbb{E}[\lambda^2(\tilde{Z}_1)]^2. \tag{61}$$

Combining with (59)-(60)-(61) with (58) shows that

$$\mathbb{E}\left[\max_{k=1}^{N+1} q_k\right] \leq \frac{\sigma\sqrt{8\log(N+1)}}{N} + \frac{2}{N}\left(\left\{3 - \frac{2}{N}\right\}\mathbb{E}\lambda^2(\tilde{Z}_1) + \frac{2\mathbb{E}\lambda^3(\tilde{Z}_1)}{N} - 2\right). \tag{62}$$

Using the Cauchy-Schwarz's inequality, it holds

$$\sum_{k=1}^{N}\operatorname{Var}\lambda(Z_k) = \sum_{k=1}^{N}\int\left(\frac{\mathrm{d}P^{N+1}}{\mathrm{d}P^{\mathrm{cal}}}\right)^2 (z)\,\mathrm{d}P^k(z) - \sum_{k=1}^{N}\mathbb{E}\left[\lambda(Z_k)\right]^2 \leq N\mathbb{E}[\lambda^2(\tilde{Z}_1)] - N.$$

The previous inequality combined with (57) and (62) implies that

$$\begin{aligned} &\left|\mathbb{P}\left(V_{N+1} \leq Q_{1-\alpha}\left(\textstyle\sum_{k=1}^{N} p_k \delta_{V_k} + p_{N+1}\delta_\infty\right)\right) - 1 + \alpha\right| \\ &\quad\leq \frac{8\mathbb{E}\lambda^3(\tilde{Z}_1) + 8\mathbb{E}\lambda^2(\tilde{Z}_1) - 8(\mathbb{E}\lambda^2(\tilde{Z}_1))^2 - 7}{N} + 16\sigma\sqrt{\frac{\log 4N}{N}} + \mathbb{E}\left[\max_{k=1}^{N+1}\{q_k\}\right] \\ &\quad\leq \frac{18\mathbb{E}\lambda^3(\tilde{Z}_1)}{N} + 19\sigma\sqrt{\frac{\log 4N}{N}}. \end{aligned} \tag{63}$$

□

Algorithm 1 Federated Quantile Estimation

```
Input: significance level $\alpha$, number of rounds $T$, learning rate $\eta$, Moreau regularization parameter $\gamma$, parameter of the DP noise $\sigma$, number of local iteration $K$.
// In parallel on the local agents
for each agent $i = 0$ to $n$ do
    Estimate and transmit the GMM parameters $\{\pi_y^i, m_y^i, \Sigma_y^i\}_{y\in\mathcal{Y}}$ as in (16)
    Compute $\forall(X_k^i, Y_k^i)$, $V_k^i = V(X_k^i, Y_k^i)$ and $\lambda_k^i = (\mathrm{d}P_X^\star/\mathrm{d}P_X^{\mathrm{cal}})(X_k^i)$ using (15)
    Transfer $\Lambda^i = \sum_{k=1}^{N^i} \lambda_k^i$ to the central server
for $t = 0$ to $T-1$ do
    // On the central server
    $S_{t+1} \leftarrow$ random subset of $[n]$
    // In parallel on the local agents
    for each agent $i \in S_{t+1}$ do
        Initialize quantile $q_{t,0}^i \leftarrow q_t$
        for $k = 0$ to $K-1$ do
            // Gradient with DP noise
            $g_{t,k}^i \leftarrow \nabla\mathrm{loss}_i^{(\gamma)}(q_{t,k}^i) + z_{t,k}^i$, $z_{t,k}^i \sim \mathcal{N}(0,\sigma^2)$
            // 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
    $q_{t+1} \leftarrow q_t + \frac{n}{|S_{t+1}|}\sum_{i\in S_{t+1}}(\frac{\Lambda^i}{\sum_{j=1}^n \Lambda^j})\Delta q_{t+1}^i$
    $\bar{q}_{t+1} \leftarrow \frac{t}{t+1}\bar{q}_t + \frac{n}{|S_{t+1}|}\sum_{i\in S_{t+1}}(\frac{\Lambda^i}{\sum_{j=1}^n \Lambda^j})\frac{\Delta\bar{q}_{t+1}^i}{t+1}$
Output: $\widehat{Q}_{1-\alpha}^{(\gamma)} \leftarrow \bar{q}_T$.
```

D Details on the federated quantile computation

In this section, we detail the algorithm implemented for Section 5. This Algorithm 1 is divided into two parts. In the first part, the density ratios $\lambda(X_k^i)$ are estimated from the local data of each agent. Then, the quantile $\widehat{Q}_{1-\alpha}^{(\gamma)}$ is determined using the procedure developed by Plassier et al. (2023). This method is based on the FedAvg algorithm with Gaussian noise to ensure the differential privacy (McMahan et al., 2017; Dwork, 2006; Ha et al., 2019). Given a regularization parameter $\gamma > 0$, each agent uses its local data to calculate its loss function $\mathrm{loss}_i^{(\gamma)}$ whose gradient is:

$$\nabla\mathrm{loss}_i^{(\gamma)}(q) = \frac{1}{\Lambda^i}\sum_{k=1}^{N^i}\lambda(X_k^i)\nabla S_{\alpha,V_k^i}^{(\gamma)}(q), \qquad \Lambda^i = \sum_{k=1}^{N^i}\lambda(X_k^i)$$

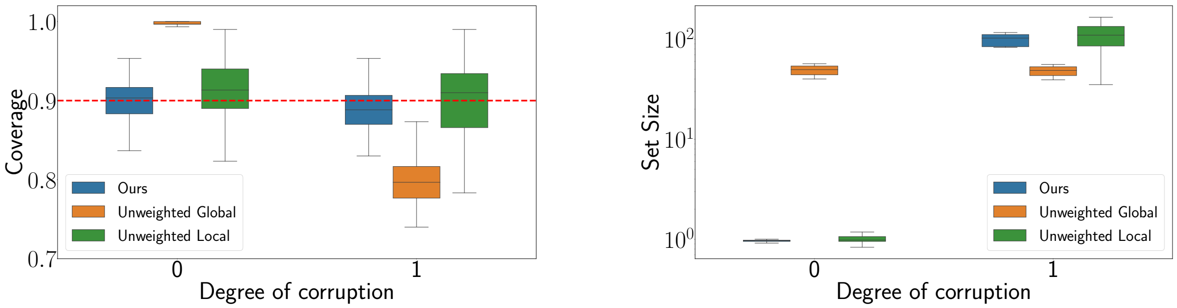
$$\nabla S_{\alpha,v}^{\gamma}(q) = -(1-\alpha)\mathbb{1}_{\{q<v-\gamma(1-\alpha)\}} + \alpha\mathbb{1}_{\{q>v+\gamma\alpha\}} + \frac{1}{\gamma}(q-v)\mathbb{1}_{\{v-\gamma(1-\alpha)<q<v+\gamma\alpha\}}.$$

The quantile $Q_{1-\alpha}^{(\gamma)}$ is obtained solving

$$Q_{1-\alpha}^{(\gamma)} \in \arg\min\left\{\sum_{i=1}^n \Lambda^i\,\nabla\mathrm{loss}_i^{(\gamma)}\right\}.$$

Based on an approximation $\widehat{Q}_{1-\alpha}^{(\gamma)}$, we generate the following prediction set:

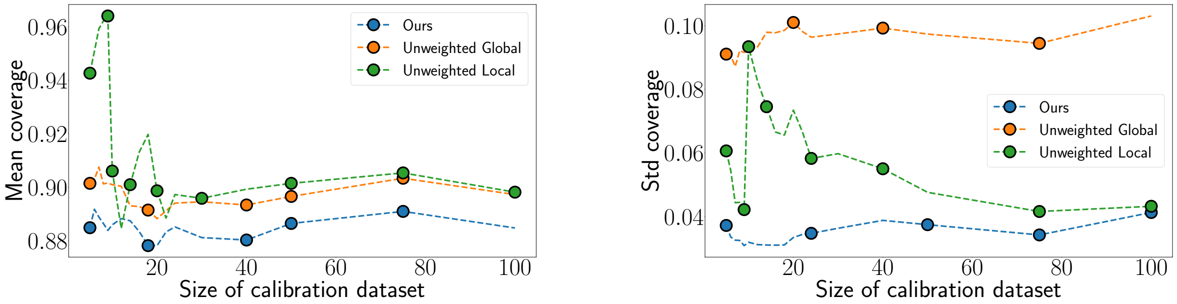
$$\widehat{\mathcal{C}}_{\alpha,\mu}(\mathbf{x}) = \left\{\mathbf{y}\in\mathcal{Y}\colon V(\mathbf{x},\mathbf{y}) \le \widehat{Q}_{1-\alpha}^{(\gamma)}\right\}.$$



(a)

(b)

Figure 5: ImageNet & ImageNet-R experimental results: (a) Empirical coverage of conformal prediction sets for non-corrupted and corrupted data. It shows how accurately do conformal sets capture the true classes of the data. (b) Average set size of conformal prediction sets for non-corrupted and corrupted data. The size of the sets increases with the level of corruption due to the increasing uncertainty of the model based on the corrupted data.



(a)

(b)

Figure 6: CIFAR-10 experimental results. (a) Mean empirical coverage changes as a function of the calibration dataset size. (b) Standard deviation of empirical coverage as function of the calibration dataset size.

E Additional Experiments

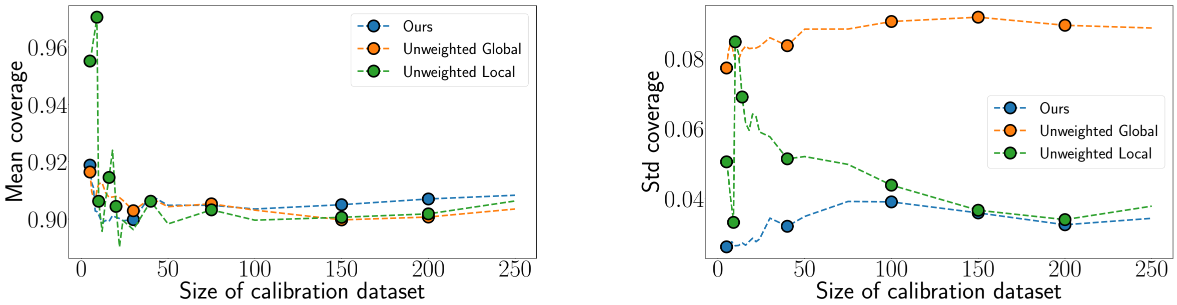
E.1 Additional experiment on ImageNet

We also experiment with ImageNet-R (Hendrycks et al., 2021) as a corrupted data source. Using the same data partitioning scheme as with ImageNet-C, we evaluate the performance of the presented algorithm under real distribution shifts. Specifically, we have 20 clients, half of them contain original ImageNet data, and the other half consist of ImageNet-R data samples. Each client has 850 data samples, consisting of 500 model fitting samples, 50 calibration samples and 300 test samples.

The results obtained with ImageNet-R are consistent with those obtained using the ImageNet-C dataset. In particular, in Figure 5a the coverage of our method best approaches the nominal value of $1-\alpha$ for both non-shifted data and shifted data. Also, in the Figure 5b, the average prediction set sizes of the introduced method have less variance than the local baseline.

E.2 Additional experiment on CIFAR-10

Following the same setup as for the CIFAR-100 in Figure 3a and Figure 3b, we conduct the same experiment for the CIFAR-10. To recap, we have 100 clients, 100 data points were allocated for test, while the size of calibration dataset varied. Results for this experiment are presented in Figure 6a and Figure 6b. As before for CIFAR-100, we see that our approach benefits from the federated collaborative procedure, which results in almost perfect coverage with smaller variance.



(a)

(b)

Figure 7: CIFAR-100 experimental results with extended axis on the size of calibration dataset. (a) Mean empirical coverage changes as a function of the calibration dataset size. (b) Standard deviation of empirical coverage as function of the calibration dataset size.

E.3 Additional experiment on CIFAR-100

In Figure 3b from the main text, it may seem that variance of our approach constantly increases. We decided to continue the experiment just to check if the variances will intersect at some point. We present the additional plot in Figure 7a and Figure 7b. We see, that this effect was rather random, and our approach still optimal from the variance point of view.

E.4 Comparison with Tibshirani's weights

Tibshirani et al. (2019)'s algorithm relies on importance weights derived from the density of $(Z_{\sigma(1)}, \dots, Z_{\sigma(N+1)})$ calculated over all possible permutations σ, that in general case requires the evaluation of sums of $N!$ terms, where N represents the total number of calibration points. Tibshirani's expression simplifies in very special cases, such as when the distribution of the test point is different from the distribution of the calibration data, which are i.i.d. (an example discussed in Tibshirani et al. (2019)). Our proposed conformal prediction set **aligns with Tibshirani's when the calibration data are i.i.d.**. However, our main contribution is that our prediction set **remains valid even for the non-i.i.d. calibration data**, i.e. in the presence of label and covariate shifts **inside** the calibration data. To show this, we provide conditional coverage guarantees and avoid the intractability of combinatorial Tibshirani's CP set.

Plassier et al. (2023) method consists in subsampling the calibration data to generate i.i.d. samples from the mixture distribution, and then using the "simplified" Tibshirani's weights. However, the subsampling cost leads to an increase in variance (and also the theory becomes more convoluted compared to our paper).

We conducted additional experiments comparing the work of Plassier et al. (2023) and Tibshirani et al. (2019). For Tibshirani's method, the only feasible approach is to sample a certain number of permutations and compute approximate weights based on them. Unfortunately, it leads to the very high variance of the weights which led us even to the bias in the mean coverage (82.48 on the "Mix" calibration data in our experiment on domain adaptation). For Plassier's method we observed a coverage of $92.28 \pm 5.97\%$ which is perfectly aligned with expectation to have increased variance compared to our method; see Figure 1. When replicating the CIFAR10 experiment, sampling the permutations led us to even worse results and we only present the comparison with Plassier's method which again gives an expected increase in variance (see figures below).

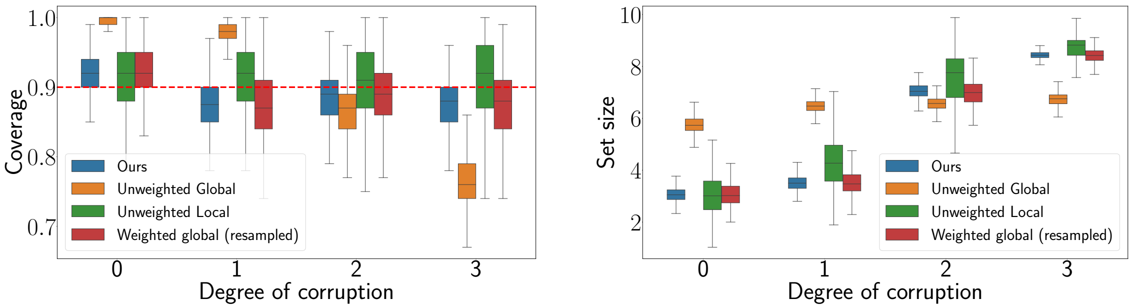


Figure 8: Coverage comparison on CIFAR-10. (Left) Empirical coverage in function of the data corruption level. (Right) Average set size of conformal prediction sets as a function of data corruption level.