Noise-Adaptive Conformal Classification with Marginal Coverage

Bortolotti, T.; Wang, Y. X. R.; Tong, X.; Menafoglio, A.; Vantini, S.; Sesia, M.

Noise-Adaptive Conformal Classification with Marginal Coverage

Teresa Bortolotti[1], Y. X. Rachel Wang[2], Xin Tong[3,4],
Alessandra Menafoglio[1], Simone Vantini[1], and Matteo Sesia[3,5]

Abstract

Conformal inference provides a rigorous statistical framework for uncertainty quantification in machine learning, enabling well-calibrated prediction sets with precise coverage guarantees for any classification model. However, its reliance on the idealized assumption of perfect data exchangeability limits its effectiveness in the presence of real-world complications, such as low-quality labels—a widespread issue in modern large-scale data sets. This work tackles this open problem by introducing an adaptive conformal inference method capable of efficiently handling deviations from exchangeability caused by random label noise, leading to informative prediction sets with tight marginal coverage guarantees even in those challenging scenarios. We validate our method through extensive numerical experiments demonstrating its effectiveness on synthetic and real data sets, including CIFAR-10H and BigEarthNet.

Keywords: conformal inference, classification, label noise, marginal validity.

1 Introduction

1.1 Background and Motivation

Conformal inference methods aim to provide rigorous uncertainty quantification for the predictions of any "black-box" machine learning model, avoiding parametric assumptions (Vovk et al., 2005). In classification, they seek a prediction set for the label of a new test point while guaranteeing a specified coverage level. This is typically achieved by studying the distribution of suitable residuals (or *non-conformity scores*) from the pre-trained model applied to an independent *calibration* data set, assuming exchangeability with the test data.

Perfect exchangeability, however, is not always realistic, due for example to *distribution shifts* between the available data and the future test points of interest, creating a need to relax the assumptions underlying conformal inference (Barber et al., 2023). This work tackles this challenge in scenarios where the observed labels may be randomly contaminated, causing the distribution of the calibration data to diverge from that of the test point, thereby invalidating classical conformal

---

[1]MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy.
[2]School of Mathematics and Statistics, University of Sydney, Sydney, Australia.
[3]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA.
[4]Business School, University of Hong Kong, Hong Kong, China.
[5]Department of Computer Science, University of Southern California, Los Angeles, CA, USA.

methods and, as we shall see, often leading to overly conservative predictions. Label noise is a pervasive problem, as modern datasets are often affected by privacy-preserving mechanisms, human annotation errors (e.g., in crowdsourcing platforms like Amazon Mechanical Turk), or measurement inaccuracies.

Our main contribution is to develop an adaptive method for constructing informative prediction sets that provably achieve tight *marginal coverage* in the presence of label noise. Marginal coverage is a widely pursued objective in conformal inference, ensuring that prediction sets contain the correct label for a predefined proportion of future test points.

While conformal inference methods can be designed to target different notions of coverage, in practice stronger guarantees may come at the expense of very large and, thus, uninformative prediction sets. For example, *label-conditional* coverage (Vovk et al., 2003), which requires prediction sets to contain the correct label for a sufficiently high proportion of test cases within each subgroup defined by the true labels, can lead to overly conservative prediction sets for the entire population in applications with many labels (Ding et al., 2024) or significant class imbalance, due to the statistical "bottleneck" created by rarer labels.

In contrast, marginal coverage often manages to strike a practical balance, providing both informative prediction sets and rigorous statistical guarantees even in challenging scenarios. However, without a suitably adaptive method, this balance tends to break down when the underlying data exchangeability assumption is violated, such as in the presence of random label noise. This challenge motivates this paper.

1.2 Main Contributions and Relation to Prior Work

Several recent studies have explored relaxing the exchangeability assumptions underlying conformal inference (Barber et al., 2023). Here, we focus on discussing this paper's connection to other recent works specifically addressing random label noise.

Einbinder et al. (2024) identified conditions where standard methods become too conservative in the presence of label noise, while Sesia et al. (2024) introduced an adaptive approach to obtain more informative prediction sets, also followed by Clarkson et al. (2024). Taking a complementary direction, Penso and Goldberger (2024) proposed computing more robust non-conformity scores that could potentially also be combined with our method.

Our work builds on Sesia et al. (2024) by addressing the more challenging problem of achieving *marginal* rather than label-conditional coverage. While marginal coverage is a weaker guarantee, achieving it efficiently under label noise requires a different approach, particularly to optimize the informativeness of the prediction sets. We overcome this challenge by developing new techniques to tightly bound the expected deviation between an empirical process estimating marginal coverage and its population counterpart—a practically useful theoretical innovation necessary to obtain informative prediction sets.

A direct extension of the label-conditional approach would seek to control the deviation for each label separately, combining the results via a union bound. While convenient—since the label-conditional process simplifies to a cumulative distribution function (CDF) of i.i.d. random variables, allowing the application of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality—this strategy is inefficient. It leads to prediction sets that are overly conservative by a factor of order $1/\sqrt{n_{\min}}$, where $n_{\min}$ is the sample size for the minority class. This inefficiency is particularly problematic

2

in settings with many classes or significant class imbalance—precisely where efficient marginal coverage is most critical.

We solve this problem by analyzing instead the deviation of an empirical process estimating *marginal* coverage under label noise from its population counterpart, without conditioning on specific labels. This analysis is more complex, as the process cannot be expressed as a CDF of i.i.d. random variables, necessitating different tools from empirical process theory, including Massart's lemma, Dudley's theorem, and Donsker's theorem. Although more involved, this approach is worthwhile as it yields a practically small finite-sample coverage inflation factor that scales as $1/\sqrt{n}$ with the calibration sample size n, independent of the number of classes. Thus, we obtain a method well-suited to produce informative prediction sets in applications with strong class imbalance or many labels.

1.3   Outline of the Paper

Section 2 outlines our method, covering the technical background, key theoretical results on the impact of label contamination on standard conformal prediction sets, and the adaptive calibration algorithm that we propose to correct for noise-induced coverage inflation. Sections 3 and 4 explain how to calibrate our algorithm in practice to obtain tight marginal coverage. Section 3 uses Massart's lemma and Dudley's theorem to derive practical finite-sample conservative estimates of the necessary calibration parameter, while Section 4 employs an even tighter asymptotic approach based on Donsker's theorem. Section 5 establishes a theoretical upper bound on the marginal coverage achieved with our method. Sections 6 and 7 evaluate our method empirically on synthetic and real data, respectively. Section 8 concludes with a discussion and some ideas for future work.

The Appendices in the Supplementary Material provide additional details. Appendix A reviews standard conformal classification with marginal coverage. Appendix B includes further implementation details and specific examples for some simple contamination models. Mathematical proofs are in Appendix C, and further numerical results are in Appendix D.

2   Methods

2.1   Problem Statement and Notation

Consider $n + 1$ data points $(X_i, Y_i, \tilde{Y}_i)$, for $i \in [n+1] = \{1, \ldots, n+1\}$, where $X_i \in \mathbb{R}^d$ is a feature vector, $Y_i \in [K]$ is a latent categorical label, and $\tilde{Y}_i \in [K]$ is an observable label that we interpret as a contaminated version of Y_i. Assume the data are i.i.d. random samples from some unknown probability distribution. We aim to construct informative conformal prediction sets with marginal coverage for the true label Y_{n+1} of a test point with features X_{n+1}, leveraging the available observations $(X_i, \tilde{Y}_i)$ indexed by $[n]$, to which we collectively refer as $\mathcal{D}$. In order to establish precise marginal coverage guarantees, it is necessary to introduce some assumptions about the relation between the true and contaminated labels.

Assumption 1. $\tilde{Y}$ *is conditionally independent of* X *given* Y; *i.e.,* $\tilde{Y} \perp\!\!\!\perp X \mid Y$.

Although Assumption 1 may not always be exactly true for real data, it gives us a useful handle on the problem, without which it would be unclear how to adapt to label noise within a non-parametric conformal inference framework. While we rely on this simplification for methodological

development and theoretical analysis, later in this paper we will also evaluate our method on real data, where Assumption 1 may hold only approximately.

Additionally, we assume the contamination model is known, with T denoting the transition matrix describing the distribution of $\tilde{Y}$ given Y; that is, $T \in [0,1]^{K \times K}$ such that $T_{kl} = \mathbb{P}[\tilde{Y} = k \mid Y = l]$. Further, we assume T is invertible, a relatively mild assumption that typically holds as long as the label noise is not so strong as to destroy all information.

Assumption 2. *The matrix T is known and invertible, with inverse $W := T^{-1}$.*

Assumption 2 may hold exactly in some applications, including for example when data are intentionally contaminated by a known differential privacy algorithm (Dwork et al., 2006). More generally, Assumption 2, like Assumption 1, serves as a useful approximation that facilitates rigorous theoretical analysis without necessarily limiting the broader applicability of our methodology. In practice, the method developed in this paper can often be successfully applied using a suitable estimate of T in place of the true matrix—a key point that will be demonstrated empirically and discussed in more detail later in the paper.

Our goal is to build prediction sets $\hat{C}(X_{n+1})$ with guaranteed *marginal coverage*:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] \geq 1 - \alpha, \tag{1}$$

for a given value of $\alpha \in (0,1)$, while being as tight as possible; that is, the miscoverage probability should be close to α, especially if n is large. Note that the *marginal* probability in (1) is taken with respect to (X_{n+1}, Y_{n+1}) and the data in $\mathcal{D}$, both of which are random.

2.2 Relevant Technical Background

We adopt the standard approach in split-conformal inference and consider a *fixed* classification model $\hat{\pi}$, trained on a separate data set $\mathcal{D}^{\text{train}}$ independent of the test point and of the n calibration data points in $\mathcal{D}$. For any $k \in [K]$ and $x \in \mathbb{R}^d$, let $\hat{\pi}(x,k) \in [0,1]$ denote an estimate of the conditional probability of $\tilde{Y} = k \mid X = x$ computed by this model. Without loss of generality, assume that $\hat{\pi}$ is normalized; i.e., $\sum_{k=1}^{K} \hat{\pi}(x,k) = 1$ for any x.

To flexibly describe how the probability estimates output of the classification model may be converted into a prediction set for Y_{n+1} given X_{n+1}, we utilize the following notion of *prediction function*, using a notation similar to that of Sesia et al. (2024).

Definition 1 (Prediction function). *A prediction function $\mathcal{C}$ is a set-valued function, whose form may depend on the model $\hat{\pi}$, that takes as input $x \in \mathbb{R}^d$ and $\tau \in [0,1]$, and outputs a subset of $[K]$, while being monotone increasing in τ and satisfying $\mathcal{C}(x,1) = [K]$ for any x.*

Note that the dependence of a prediction function $\mathcal{C}$ on $\hat{\pi}$ will typically be kept implicit; i.e., $\mathcal{C}(x,\tau) = \mathcal{C}(x,\tau;\hat{\pi}) \subseteq [K]$. For any prediction function $\mathcal{C}$, we define the *non-conformity score* function $\hat{s} : \mathbb{R}^d \times [K] \mapsto [0,1]$, also implicitly depending on $\hat{\pi}$, as the function which outputs the smallest value of τ allowing the label k to be contained in the set $\mathcal{C}(x,\tau)$:

$$\hat{s}(x,k) = \inf\left\{\tau \in [0,1] : k \in \mathcal{C}(x,\tau)\right\}. \tag{2}$$

4

A simple example of $\mathcal{C}$ is the function that outputs the set of all labels $k \in [K]$ for which $\hat{\pi}(x, k)$ is sufficiently large; i.e., $\mathcal{C}(x, \tau; \hat{\pi}) := \{k \in [K] : \hat{\pi}(x, k) \geq 1 - \tau\}$, corresponding to scores $\hat{s}(x, k) = 1 - \hat{\pi}(x, k)$. However, our results also extend to other choices of prediction functions and non-conformity scores, including those proposed by Romano et al. (2020).

Having pre-trained the model $\hat{\pi}$ on the data in $\mathcal{D}^{\text{train}}$, a standard implementation of conformal inference utilizes the n calibration data points to compute non-conformity scores $\hat{s}(X_i, k)$ via (2), for all $k \in [K]$ and $i \in [n]$, using the chosen prediction function $\mathcal{C}$. These scores are utilized to calibrate a prediction set for the test point (X_{n+1}, Y_{n+1}) as follows. For a desired miscoverage probability $\alpha \in (0, 1)$, the threshold $\hat{\tau}$ is calculated as the $\lceil (1 + n) \cdot (1 - \alpha) \rceil$-th smallest value in $\{\hat{s}(X_i, \tilde{Y}_i)\}_{i=1}^n$. Then, the prediction set for Y_{n+1} is $\hat{C}(X_{n+1}) = \mathcal{C}(X_{n+1}, \hat{\tau})$. See Algorithm A1 in Appendix A for a summary of this method.

This procedure guarantees that $\hat{C}(X_{n+1})$ has *marginal coverage* in the sense of (1) in the absence of label noise—that is, assuming $Y = \tilde{Y}$ almost surely. See Proposition A1 for a formal statement of this result, which also provides an almost-matching coverage upper bound. In the next section, we study the behavior of the standard method when $\tilde{Y} \neq Y$.

2.3   The Marginal Coverage Inflation Factor

For any $t \in [0, 1]$, define $F(t) := \mathbb{P}[\hat{s}(X, Y) \leq t]$ and $\tilde{F}(t) := \mathbb{P}[\hat{s}(X, \tilde{Y}) \leq t]$—the marginal CDFs of $\hat{s}(X, Y)$ and $\hat{s}(X, \tilde{Y})$, respectively, implicitly conditioning on the model $\hat{\pi}$ trained using $\mathcal{D}^{\text{train}}$. Then, for any $t \in [0, 1]$, we define the *marginal coverage inflation factor* as:

$$\Delta(t) := F(t) - \tilde{F}(t). \tag{3}$$

As made precise by the following theorem, the function Δ controls the effect of label noise on the marginal coverage guarantee (1) for standard conformal prediction sets.

Theorem 1. *Suppose $(X_i, Y_i, \tilde{Y}_i)$ are i.i.d. for all $i \in [n + 1]$. Fix any prediction function $\mathcal{C}$ satisfying Definition 1, and let $\hat{C}(X_{n+1})$ indicate the prediction set output by Algorithm A1 applied using the corrupted labels $\tilde{Y}_i$ instead of the clean labels Y_i, for all $i \in [n]$. Then,*

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] \geq 1 - \alpha + \mathbb{E}\left[\Delta(\hat{\tau})\right]. \tag{4}$$

Further, if the scores $\hat{s}(X_i, \tilde{Y}_i)$ used by Algorithm A1 are almost-surely distinct,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] \leq 1 - \alpha + \frac{1}{n+1} + \mathbb{E}\left[\Delta(\hat{\tau})\right]. \tag{5}$$

Theorem 1 applies to any contamination model and does not depend on Assumptions 1 or 2. However, under Assumption 1 and an additional condition on the distributions of the non-conformity scores—interpreted as the model being sufficiently accurate to typically identify the correct label as the most likely point prediction—we can also prove that the standard prediction sets generated by Algorithm A1 are conservative in the sense of (1) because $\Delta(\hat{\tau}) \geq 0$ almost-surely; see Corollary A1 in Appendix B.1. This robustness aligns with similar results by Barber et al. (2023), Einbinder et al. (2024), and Sesia et al. (2024).

Next, we build on Theorem 1 to develop an adaptive method that leverages an empirical estimate of the function Δ to achieve tight marginal coverage under label noise.

2.4 Adaptive Prediction Sets with Marginal Coverage

For any $k, l \in [K]$ and $t \in [0, 1]$, define

$$F_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \le t \mid Y = l\right], \qquad \tilde{F}_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \le t \mid \tilde{Y} = l\right]. \tag{6}$$

Above, $F_l^k(t)$ is the CDF of $\hat{s}(X, k)$, based on a fixed function $\hat{s}$ applied to a random X from the distribution of $X \mid Y = l$, while $\tilde{F}_l^k(t)$ is the corresponding CDF of $\hat{s}(X, k)$, with X conditioned on $\tilde{Y} = l$. For any $k \in [K]$, define also $\tilde{\rho}_k := \mathbb{P}[\tilde{Y} = k]$, denoting the marginal frequencies of the contaminated labels.

Under Assumptions 1 and 2, the coverage inflation factor (3) can be expressed in terms of quantities that are either known or estimable from the contaminated data; specifically,

$$\Delta(t) = \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \tilde{\rho}_l \tilde{F}_l^k(t) - \tilde{F}(t), \tag{7}$$

where W is the inverse of the label transition matrix T defined in Section 2.1. From (7), it follows directly that a natural empirical estimator of $\Delta(t)$ is

$$\hat{\Delta}(t) := \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \hat{\rho}_l \hat{F}_l^k(t) - \hat{F}(t), \tag{8}$$

where, for each $k, l \in [K]$, $\hat{F}_l^k$ is the empirical CDF of $\hat{s}(X_i, k)$ for $i \in \mathcal{D}_l = \{i \in [n] : \tilde{Y}_i = l\}$, $\hat{F}(t)$ is the empirical CDF of $\hat{s}(X_i, \tilde{Y}_i)$, while $\hat{\rho}_l = n_l/n$ is the proportion of observed labels equal to l. In summary,

$$\hat{F}(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[\hat{s}(X_i, \tilde{Y}_i) \le t\right], \qquad \hat{F}_l^k(t) := \frac{1}{n_l} \sum_{i \in \mathcal{D}_l} \mathbb{I}\left[\hat{s}(X_i, k) \le t\right], \qquad \hat{\rho}_l := \frac{n_l}{n}. \tag{9}$$

Further, to simplify the notation below, let $S_{(1)}, \ldots, S_{(n)}$ denote the (ascending) order statistics of the scores $\{\hat{s}(X_1, \tilde{Y}_1), \ldots, \hat{s}(X_n, \tilde{Y}_n)\}$.

Our adaptive calibration algorithm utilizes $\hat{\Delta}$ to compute a corrected threshold $\hat{\tau}$ as

$$\hat{\tau} = \begin{cases} S_{(\hat{i})} \text{ where } \hat{i} = \min\{i \in \hat{\mathcal{I}}\}, & \text{if } \hat{\mathcal{I}} \ne \emptyset, \\ 1, & \text{if } \hat{\mathcal{I}} = \emptyset, \end{cases} \tag{10}$$

where the set $\hat{\mathcal{I}}$ is defined as

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \ge 1 - \alpha - \hat{\Delta}(S_{(i)}) + \delta(n) \right\}, \tag{11}$$

and $\delta(n)$ is a finite-sample correction factor accounting for the errors involved in estimating Δ and $\hat{\tau}$ empirically using n contaminated data points. Finally, the adaptive prediction set output by our method is $\hat{C}(X_{n+1}) = \mathcal{C}(X, \hat{\tau})$. The overall procedure is outlined by Algorithm 1. We will show in the following that, when the correction factor $\delta(n)$ is properly set, Algorithm 1 achieves marginal coverage efficiently.

In Appendix B.2, we present an "optimistic" variation of Algorithm 1 that can generate even more informative prediction sets. Intuitively, the approach adaptively selects the smaller prediction

---

Algorithm 1: Contamination-adaptive classification with marginal coverage

---

> **Input:** Contaminated data set $\{(X_i, \tilde{Y}_i)\}_{i=1}^n$ with noisy labels $\tilde{Y}_i \in [K]$.
> The inverse W of the matrix T describing the contamination model.
> Pre-trained K-class classification model $\hat{\pi}$. Prediction function $\mathcal{C}$.
> Desired miscoverage probability $\alpha \in (0, 1)$.
> Finite-sample correction factor $\delta(n) \geq 0$.
> Unlabeled test point with features X_{n+1}.

1 Compute the scores $\hat{s}(X_i, k)$ for all $i \in [n]$ and $k \in [K]$, based on $\mathcal{C}$ and $\hat{\pi}$ using (2).

2 Compute the empirical contaminated label frequency $\hat{\rho}_k = n_k/n$, for all $k \in [K]$.

3 Compute the empirical CDFs $\hat{F}$ and $\hat{F}_l^k$, for all $k, l \in [K]$, using (9).

4 Compute $\hat{\Delta}(S_{(i)})$ for all $i \in [n]$, as in (8).

5 Sort $\{\hat{s}(X_i, \tilde{Y}_i) : i \in [n]\}$ into $(S_{(1)}, S_{(2)}, \ldots, S_{(n)})$, in ascending order.

6 Construct the set $\hat{\mathcal{I}} \subseteq [K]$ as in (11).

7 Evaluate $\hat{\tau}$ based on $\hat{\mathcal{I}}$ as in (10).

8 Evaluate $\hat{C}(X_{n+1}) = \mathcal{C}(X_{n+1}, \hat{\tau}; \hat{\pi})$.

> **Output:** Conformal prediction set $\hat{C}(X_{n+1})$ for Y_{n+1}.

---

set between those produced by Algorithm 1 and the standard conformal approach (Algorithm A1). Despite its seemingly greedy nature, this method preserves marginal coverage as long as the standard conformal inference approach is conservative in the presence of label noise—a relatively mild assumption as demonstrated in Appendix B.1.

2.5 Lower Bound on Coverage of Adaptive Prediction Sets

Define the following zero-mean empirical process,

$$\hat{\psi}(t) := \sum_{k=1}^K \sum_{l=1}^K W_{kl} \left(\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right), \tag{12}$$

and the corresponding expected supremum

$$\delta^*(n) = \mathbb{E} \left[\sup_{t \in [0,1]} \hat{\psi}(t) \right]. \tag{13}$$

The following result states that, when applied with a correction factor $\delta(n) \geq \delta^*(n)$, Algorithm 1 outputs prediction sets that are marginally valid in finite samples.

Theorem 2. *Suppose $\{(X_i, Y_i, \tilde{Y}_i)\}_{i=1}^{n+1}$ are i.i.d. and that Assumptions 1 and 2 hold. For any prediction function $\mathcal{C}$, let $\hat{C}(X_{n+1})$ be the prediction set output by Algorithm 1 applied with a correction factor $\delta(n)$. Then, $\mathbb{P}[Y_{n+1} \in \hat{C}(X_{n+1})] \geq 1 - \alpha$ if $\delta(n) \geq \delta^*(n)$.*

Further, we prove in Section 5 that the closer $\delta(n)$ is to $\delta^*(n)$, the tighter the prediction sets output by Algorithm 1 are. Therefore, $\delta^*(n)$ represents the ideal correction factor for our method. Unfortunately, this quantity is not readily available because it depends on the distribution of $\hat{\psi}(t)$, which is unknown and potentially very complicated.

The following sections hence focus on addressing this challenge, presenting two alternative solutions for calculating a practical correction factor $\delta(n)$ that satisfies $\delta(n) \geq \delta^*(n)$ without being too conservative. In Section 3, we obtain a relatively tight finite-sample upper bound for $\delta^*(n)$. In Section 4, we use asymptotic theory to obtain an even closer approximation of $\delta^*(n)$. While theoretically valid only in the large-n limit, the latter approach often performs better in practice than the finite-sample approach.

3 Implementation using Finite-Sample Techniques

In this section, we explain how to replace the ideal correction factor $\delta^*(n)$ required by Algorithm 1 with an upper bound that is guaranteed to be conservative in finite samples.

3.1 Correction Factor for the Randomized Response Model

To simplify the exposition, we begin by considering the special case of the classical randomized response model (Warner, 1965). In this case, the transition matrix T is characterized by a scalar noise parameter $\epsilon \in [0, 1)$ and takes the form $T = (1 - \epsilon) \cdot I + \epsilon/K \cdot J$, where I denotes the $(K \times K)$ dimensional identity matrix and J is the $(K \times K)$ matrix of all ones, and thus $W = T^{-1} = 1/(1 - \epsilon) \cdot I - \epsilon/[K(1 - \epsilon)] \cdot J$.

In this case, it is not difficult to derive a finite-sample upper bound for the ideal correction factor $\delta^*(n)$, ensuring $1/\sqrt{n}$ scaling. Let $U_1, \ldots, U_n$ be i.i.d. uniform random variables on $[0, 1]$, and denote their ascending order statistics as $U_{(1)}, \ldots, U_{(n)}$. Then, define

$$c(n) := \mathbb{E}\left[\sup_{i \in [n]} \left\{ \frac{i}{n} - U_{(i)} \right\}\right],$$

a constant that satisfies $c(n) \leq \sqrt{\pi/(2n)}$ (see Lemma A3) and can be precisely computed for any n using a simple Monte Carlo simulation. The following result establishes that $c(n)$ is a conservative estimate of $\delta^*(n)$ under the randomized response model.

Theorem 3. *Under the assumptions of Theorem 2, if the transition matrix T is in the form $T = (1 - \epsilon) \cdot I + \epsilon/K \cdot J$ for some $\epsilon \in [0, 1)$, then $c(n) \geq \delta^*(n)$.*

Therefore, applying Algorithm 1 with $\delta(n) = c(n)$ achieves finite-sample coverage efficiently, regardless of the number of possible classes K, provided the label contamination follows the randomized response model. While this serves as a useful starting point, extending the solution to more general contamination models is more challenging.

3.2 Correction Factor for General Contamination Models

The proof of Theorem 3 exploits the structure of the inverse transition matrix W under the randomized response model to derive a bound for $\delta^*(n)$ that is independent of K and scales well with n. This motivates seeking a similar result for the general case by decomposing the empirical process (12) into a well-behaved component, analogous to the randomized response case, and a more challenging remainder requiring a distinct analytical approach.

8

Consider a matrix $\bar{W} \in \mathbb{R}^{K \times K}$, parameterized by $K+1$ coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_K)$, with entries $\bar{W}_{kl} = \beta_0 \mathbb{I}[k = l] + \beta_k / K$, and let Ω denote the matrix difference between W and $\bar{W}$, such that $\Omega_{kl} = W_{kl} - \bar{W}_{kl}$. This parametrization is motivated by the fact that, if $W = \bar{W}$, the expected supremum of the empirical process $\hat{\psi}(t)$ defined in (12) can be analyzed using the same approach as in Theorem 3. Notably, the matrix W under the randomized response model is a special case of $\bar{W}$, corresponding to a specific choice of β.

With this setup, we can introduce our general finite-sample upper bound for $\delta^*(n)$, derived using chaining arguments (Massart, 2000; Wainwright, 2019), which takes the form:

$$\delta^{\text{FS}}(n) := \inf_{\beta \in \mathbb{R}^{K+1}} \left\{ c(n) \left(\beta_0 + \frac{\sum_{k=1}^{K} \beta_k}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\}, \tag{14}$$

where $B(K, n, \beta)$ is given by the following expression, which depends on β through Ω:

$$B(K, n, \beta) := 2 \min \left\{ \max_{l \in [K]} \sum_{k=1}^{K} |\Omega_{kl}| \sqrt{\log(Kn + 1)}, 24 \max_{k,l \in [K]} |\Omega_{kl}| \frac{2 \log K + 1}{2 \log K - 1} \sqrt{2K \log K} \right\}. \tag{15}$$

The definition of $\delta^{\text{FS}}(n)$ in (14) involves an optimization over β, which can be efficiently solved as a convex optimization problem, as detailed in Appendix B.3. This makes our upper bound practically computable for any contamination model.

To prove that $\delta^{\text{FS}}(n) \geq \delta^*(n)$, we impose an additional mild condition on the smoothness of the non-conformity score distributions, alongside the assumptions of Theorem 2. For any $k \in [K]$, let $\tilde{F}^k(t) := \mathbb{P}[\hat{s}(X, k) \leq t]$ represent the marginal score CDF for class k.

Assumption 3. *For any* $k \in [K]$, $\tilde{F}^k$ *is differentiable on the interval* $(0, 1)$, *and the corresponding density* $\tilde{f}^k$ *is uniformly bounded with* $\|\tilde{f}^k\|_\infty \leq \tilde{f}^k_{\max}$, *for some* $\tilde{f}^k_{\max} > 0$. *Further,* $\tilde{f}^k_{\min} := \inf_{t \in (0,1)} \tilde{f}^k(t) > 0$.

Note that this assumption merely requires that $\tilde{F}^1, \ldots, \tilde{F}^k$ are continuous and with bounded densities. This can always be ensured in practice by adding a small amount of random noise to the scores computed by any classifier.

Theorem 4. *Under the assumptions of Theorem 2, suppose also that Assumption 3 holds. Then,* $\delta^{\text{FS}}(n) \geq \delta^*(n)$, *where* $\delta^{\text{FS}}(n)$ *is defined as in* (14).

Having established the practicality and validity of $\delta^{\text{FS}}(n)$ as a general upper bound for $\delta^*(n)$, we turn to discussing its nice scaling properties with respect to n and K. We begin by noting that the behavior of $c(n)$, described by Lemma A3, implies

$$\delta^{\text{FS}}(n) \leq \sqrt{\frac{\pi}{2n}} \cdot \inf_{\beta \in \mathbb{R}^{K+1}} \left(\beta_0 + \frac{\sum_{k=1}^{K} \beta_k}{K} + \sqrt{\frac{2}{\pi}} B(K, n, \beta) \right).$$

Therefore, if K is fixed, it is clear that $\delta^{\text{FS}}(n)$ decays as $1/\sqrt{n}$, as desired, since $B(K, n, \beta)$ does not grow with n. By contrast, the scaling of $\delta^{\text{FS}}(n)$ with respect to K requires a little more attention because it depends on the specific structure of the contamination model.

In the special case of the randomized response model, we can simply choose β so that $B(K, n, \beta) = 0$, recovering $\delta^{\text{FS}}(n) = c(n) \leq \sqrt{\pi/(2n)}$ as in Section 3.2. In general, however, the term $B(K, n, \beta)$

may (slowly) increase with K, potentially reducing the efficiency of our conformal prediction sets. It is easy to see that, in the worst-case, $\delta^{\text{FS}}(n) = \mathcal{O}(\min\{\sqrt{(K \log K)/n}, \sqrt{\log(Kn)/n}\})$. Concretely, however, this worst-case behavior is quite pessimistic. In many interesting special cases beyond the randomized response model, including for example the *two-level* randomized response model studied in Appendix B.4, our finite-sample correction factor does not increase with K asymptotically.

While these results show our correction factor scales well theoretically—contrast this with the $\mathcal{O}(\sqrt{K/n})$ behavior resulting from a DKW-based analysis in the style of Sesia et al. (2024)—they are not sufficient on their own to demonstrate the practical utility of our method. This will be shown in Sections 6 and 7, where we numerically demonstrate that Algorithm 1, with the correction factor $\delta^{\text{FS}}(n)$, produces informative prediction sets and significantly outperforms standard conformal predictions under label noise.

That being said, the practical performance of Algorithm 1 can be further enhanced by replacing $\delta^{\text{FS}}(n)$ with an asymptotic approximation of $\delta^*(n)$, as discussed next.

4   Implementation using Asymptotic Techniques

4.1   Convergence in Distribution and High-Level Strategy

In this section, we replace the ideal correction factor $\delta^*(n)$ in Algorithm 1 with an asymptotic approximation, derived by analyzing the large-sample behavior of the empirical process $\hat{\psi}$. As we will see in Section 6, this approach performs well in practice, even though the approximation is not theoretically guaranteed to be conservative in finite samples.

Theorem 5. *Under the assumptions of Theorem 4, consider the empirical process $\hat{\psi}$ defined in (12). Let $\text{GBB}(t)$ denote a zero-mean Gaussian process, indexed by $t \in [0, 1]$, with covariance function $G : [0, 1]^2 \mapsto \mathbb{R}$ such that $G(t_1, t_2) = \mathbb{E}[f_{t_1} f_{t_2}] - \mathbb{E}[f_{t_1}] \mathbb{E}[f_{t_2}]$, for any $t_1, t_2 \in [0, 1]$, where $f_t := \sum_{k,l=1}^{K} W_{kl} \mathbb{I}[\hat{s}(X_i, k) \leq t, \tilde{Y}_i = l]$ for any $t \in [0, 1]$. Then, $\sqrt{n} \hat{\psi}(t) \xrightarrow{d} \text{GBB}(t)$ as $n \to \infty$, with $\xrightarrow{d}$ representing convergence in distribution.*

This result suggests approximating $\delta^*(n)$ with

$$\delta^{\text{asy}}(n) := \frac{1}{\sqrt{n}} \mathbb{E}\left[\sup_{t \in [0,1]} \text{GBB}(t)\right]. \tag{16}$$

The Generalized Brownian Bridge (GBB) defined in Theorem 5 generalizes the F-Brownian bridge (see Van der Vaart (2000), Chapter 19.1). While the distribution of the supremum for the F-Brownian bridge matches that of the classical Brownian bridge and its expected value can be computed in closed form, computing (16) for the GBB is more challenging. This is because its covariance function depends on the distribution of the non-conformity scores, which must be estimated empirically, and the distribution of its supremum does not appear to be analytically tractable.

To address these challenges, we propose a discretization strategy for evaluating (16). First, the interval $[0, 1]$ is divided into a finite grid of N points, replacing the continuous process $\text{GBB}(t)$ with a Gaussian random vector of length N. Second, the covariance matrix of this random vector is approximated using an empirical estimate of the covariance function in Theorem 5 based on

the observed contaminated data. Third, a Monte Carlo estimate of the expected supremum of
$\mathrm{GBB}(t)$ is computed by generating multiple independent realizations of the Gaussian vector and
averaging their maximum values. Finally, a numerical extrapolation strategy is applied to reduce
the discretization bias introduced by the finite grid size N. We will now describe the key steps in
more detail.

4.2 Empirical Covariance Estimate

According to Theorem 5, for any $t_1, t_2 \in [0, 1]$, the covariance between $\mathrm{GBB}(t_1)$ and $\mathrm{GBB}(t_2)$ can
be written as $G(t_1, t_2) = \mathbb{E}\left[f_{t_1} f_{t_2}\right] - \mathbb{E}\left[f_{t_1}\right] \mathbb{E}\left[f_{t_2}\right]$, where

$$\mathbb{E}\left[f_t\right] = \sum_{l,k=1}^{K} W_{kl} \mathbb{P}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l\right] = \sum_{l,k=1}^{K} W_{kl} \tilde{\rho}_l \tilde{F}_l^k(t),$$

$$\mathbb{E}\left[f_{t_1} f_{t_2}\right] = \sum_{l,k,k'=1}^{K} W_{kl} W_{k'l} \tilde{\rho}_l \mathbb{P}\left[\hat{s}(X_i, k) \le t_1, \hat{s}(X_i, k') \le t_2 \mid \tilde{Y}_i = l\right].$$

An intuitive plug-in estimate of $G(t_1, t_2)$ is $\hat{G}(t_1, t_2) = \hat{\mathbb{E}}[f_{t_1} f_{t_2}] - \hat{\mathbb{E}}[f_{t_1}]\hat{\mathbb{E}}[f_{t_2}]$, where

$$\hat{\mathbb{E}}[f_t] = \sum_{l,k=1}^{K} W_{kl} \hat{\rho}_l \frac{1}{n_l} \sum_{i \in \mathcal{D}_l} \mathbb{I}\left[\hat{s}(X_i, k) \le t\right],$$

$$\hat{\mathbb{E}}[f_{t_1} f_{t_2}] = \sum_{l,k,k'=1}^{K} W_{kl} W_{k'l} \hat{\rho}_l \frac{1}{n_l} \sum_{i \in \mathcal{D}_l} \mathbb{I}\left[\hat{s}(X_i, k) \le t_1, \hat{s}(X_i, k) \le t_2\right],$$

with $\mathcal{D}_l$ denoting the subset of indices $i \in [n]$ such that $\tilde{Y}_i = l$, and $n_l = |\mathcal{D}_l|$.

4.3 Monte Carlo Simulation of the Generalized Brownian Bridge

For a given grid resolution parameter $h \in (0, 1)$, define a sequence $0 = t_1 < t_2 = h < t_3 = 2h < \cdots < t_N = 1$ discretizing the interval $[0, 1]$ at $N = 1/h$ equal-spaced points. Define also the
covariance matrix $\hat{\Sigma} \in \mathbb{R}^{N \times N}$ such that $\hat{\Sigma}_{ij} = \hat{G}(t_i, t_j)$ for all $i, j \in [N]$, and let $\xi^{(1)}, \ldots, \xi^{(M)} \in \mathbb{R}^N$ denote M i.i.d. realizations of an N-dimensional Gaussian random vector $\xi \sim \mathcal{N}(0, \hat{\Sigma})$, with
$\xi^{(m)} = (\xi_1^{(m)}, \ldots, \xi_N^{(m)})$ for each $m \in [M]$. Then, as a preliminary Monte Carlo estimate of the
factor $\delta^{\mathrm{asy}}(n)$ defined in (16), we can compute

$$\hat{\delta}^{\mathrm{asy}}(n, h) := \frac{1}{\sqrt{n}} \cdot \frac{1}{M} \sum_{m=1}^{M} \max_{i \in [N]} \xi_i^{(m)}, \tag{17}$$

highlighting explicitly the dependence of $\hat{\delta}^{\mathrm{asy}}(n, h)$ on the grid resolution parameter h.

Since $\hat{\delta}^{\mathrm{asy}}(n, h)$ provides a downward-biased estimate of (16) for any $h > 0$, it is desirable
to choose h as small as possible. However, the computational cost of evaluating (17) increases
rapidly with $N = 1/h$. To balance this trade-off, we employ Richardson extrapolation (Richardson
and Glazebrook, 1911), a numerical technique that recursively combines estimates of $\hat{\delta}^{\mathrm{asy}}(n, h)$
obtained at coarser resolutions to produce a single estimate $\hat{\delta}^{\mathrm{asy}}(n)$ with accuracy comparable to

that of a finer grid but at lower computational cost. Further implementation details can be found in Appendix B.5.

Appendix B.5 also provides a rule of thumb for selecting the hyperparameters h and M in this Monte Carlo procedure, as well as the appropriate order of Richardson extrapolation. This leads to a principled solution for computing an estimate $\hat{\delta}^{\mathrm{asy}}(n)$ of $\delta^{\mathrm{asy}}(n)$, to be used in Algorithm 1. Since $\delta^{\mathrm{asy}}(n)$, and therefore $\hat{\delta}^{\mathrm{asy}}(n)$, scales as $\mathcal{O}(1/\sqrt{n})$ regardless of K, this implementation is particularly well-suited for classification problems with many classes.

5 Theoretical Upper Bound on Coverage

In order to prove that the prediction sets output by Algorithm 1 are not overly conservative, we need to make three additional technical assumptions.

Assumption 4. *The marginal CDFs F and $\tilde{F}$ are differentiable on $(0,1)$, and the corresponding densities f and $\tilde{f}$ are uniformly bounded with $\|f\|_\infty \le f_{\max}$ and $\|\tilde{f}\|_\infty \le \tilde{f}_{\max}$, for some $f_{\max}, \tilde{f}_{\max} > 0$. Further, $f_{\min} := \inf_{t \in (0,1)} f(t) > 0$ and $\tilde{f}_{\min} := \inf_{t \in (0,1)} \tilde{f}(t) > 0$.*

Assumption 5. *The CDF $\tilde{F}_k^k$ satisfies $\max_{l \ne k} \tilde{F}_l^k(t) \le \tilde{F}_k^k(t)$ for all $t \in [0,1]$ and $k \in [K]$.*

Assumption 6. *The factor $\Delta(t)$ defined in (3) is bounded from below by*

$$\inf_{t \in (0,1)} \Delta(t) \ge -\alpha + \delta(n) + \sqrt{\frac{\log(2n)}{2n}} + d(n),$$

where $d(n) := n^{-1/4} \cdot \inf_\beta \{ \sqrt{\pi/2}[|\beta_0| + (\sum_{k=1}^K |\beta_k|)/K] + B(K, n, \beta) \}$.

Assumption 4 requires that the marginal distributions of the non-conformity scores with respect to the true and the noisy labels are continuous with strictly positive and bounded densities. Assumption 5 states that, despite the contamination, the classifier tends to assign stochastically smaller scores $\hat{s}(X, k)$ when $\tilde{Y} = k$ compared to when $\tilde{Y} \ne k$, suggesting that it is reasonably accurate. Assumption 6 may be interpreted as a requirement that, without suitable corrections, standard conformal prediction sets would be sufficiently conservative due to label noise. Note that, as long as $\delta(n)$ scales sufficiently well with n, the threshold goes to zero when $n \to \infty$, implying that this assumption becomes easier to satisfy in the large sample limit.

Theorem 6. *Under the setup of Theorem 2, let $\hat{C}(X_{n+1})$ be the prediction set output by Algorithm 1 based on the inverse W of the model matrix T. Suppose that Assumptions 4- 6 hold, and $\tilde{\rho}_k > 0$ for all $k \in [K]$. Then, $\mathbb{P}[Y_{n+1} \in \hat{C}(X_{n+1})] \le 1 - \alpha + \delta(n) + \varphi(n)$, where*

$$\varphi(n) = 3\delta^{**}(n) + \frac{2}{n} + \frac{1}{\sqrt[4]{n}} + \frac{1}{n+1} \cdot \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^K |V_{kl}| \right) - 1 \right], \tag{18}$$

with $\delta^{**}(n) := \mathbb{E}[\sup_{t \in [0,1]} |\hat{\psi}(t)|] = \mathcal{O}(1/\sqrt{n})$ and $V \in \mathbb{R}^{K \times K}$ indicating the inverse of $M \in \mathbb{R}^{K \times K}$, defined such that $M_{kl} = T_{kl} \rho_l / \tilde{\rho}_k$.

Combined with the lower bound in Theorem 2, this result establishes that the prediction sets generated by Algorithm 1 are asymptotically tight, with their coverage converging to the desired $1 - \alpha$ level as $n \to \infty$. Additionally, it underscores the importance of the efforts in Sections 3 and 4 to minimize the correction factor $\delta(n)$ subject to $\delta(n) \ge \delta^*(n)$.

6 Numerical Experiments with Synthetic Data

In this section, we apply Algorithm 1 to synthetic data sets. These experiments focus on evaluating our method in controlled settings where the label contamination model is known and satisfies Assumption 1. This assumption is realistic in certain applications, such as when a clean data set is intentionally contaminated through a controlled process, for instance, by differential privacy algorithms. More generally, however, Assumption 1 may not always hold exactly, or the model matrix T may need to be approximated. The performance of our method in such more challenging settings will be examined in Section 7.

6.1 Setup and Methods under Comparison

We simulate classification data with $K = 4$ labels and $d = 20$ features from a Gaussian mixture distribution using the standard `make_classification` function from the Scikit-Learn Python package (Pedregosa et al., 2011). This creates $2K$ clusters of normally distributed points, with unit variance, centered about the vertices of a 25-dimensional hypercube with sides of length 2, and then randomly assigns an equal number of clusters to each of the K classes. Note that this leads to uniform label frequencies; i.e., $\rho_k := \mathbb{P}[Y = k] = 1/K$ for all $k \in [K]$. Conditional on the simulated data, the contaminated labels $\tilde{Y}$ are generated following a pre-specified contamination process. In the following experiments, this process is the two-level randomized response model defined in Appendix B.4.2, for which we consider different values of the parameters ϵ and ν.

A random forest classifier implemented by Scikit-Learn is trained on $10,000$ independent observations with contaminated labels generated as described above. The classifier is then applied to an i.i.d. calibration data set of size n, whose labels are also similarly contaminated, evaluating generalized inverse quantile non-conformity scores (Romano et al., 2020) as detailed in Appendix A.2. These scores are transformed into prediction sets for 2,000 independent unlabeled test points using the following four alternative approaches.

(i) The *Standard* conformal inference approach that ignores the label contamination. (ii) *Adaptive*: Algorithm 1, applied with the finite-sample correction factor $\delta^{\mathrm{FS}}(n)$ described in Section 3. (iii) *Adaptive (simplified)*: a less efficient implementation of the latter, with $\delta^{\mathrm{FS}}(n)$ evaluated based on the parameter β which optimizes the correction for the randomized response model, regardless of whether that may be optimal for the true contamination model at hand. This helps highlight the distinct importance of solving the optimization problem described in Section 3.2. (iv) *Adaptive (asymptotic)*: Algorithm 1 applied with the estimated asymptotic correction factor $\hat{\delta}^{\mathrm{asy}}(n)$ described in Section 4. All adaptive approaches assume perfect knowledge of the transition matrix T and target 90% marginal coverage, with $\alpha = 0.1$.

In Section 6.5, we will also compare with the adaptive method proposed by Sesia et al. (2024), which focuses on label-conditional coverage, to explicitly highlight the advantages of targeting marginal coverage in applications with relatively small or imbalanced data sets.

6.2 The Impact of the Label Contamination Strength

Figure 1 compares the four methods in terms of coverage and prediction set size, averaged over 20 independent repetitions of each experiment. The results are shown as a function of n and stratified by the strength ϵ of the label contamination process. In these experiments, the parameter ν of the

contamination model is fixed equal to 0.2, which corresponds to a moderate deviation from the simple randomized response model.
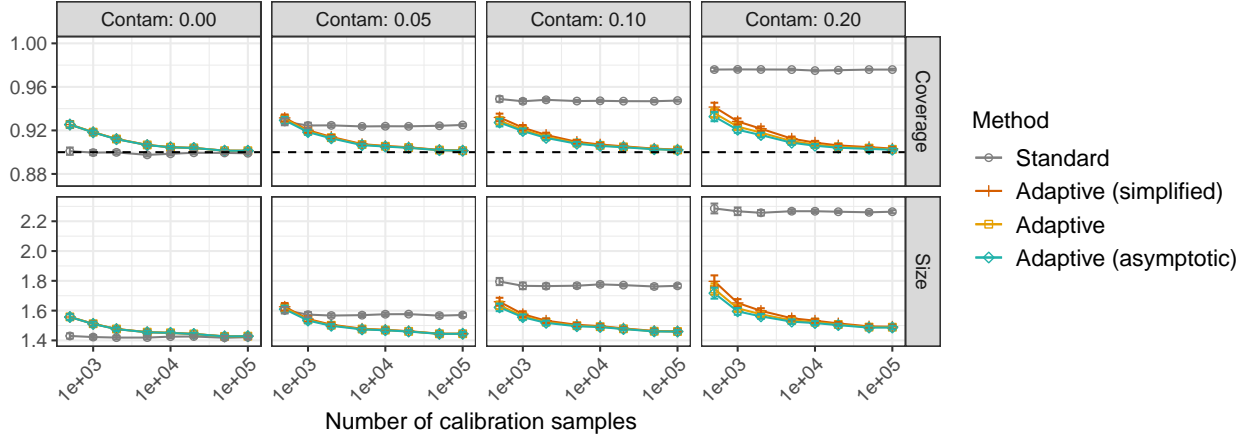

Figure 1: Performances of different conformal methods on simulated data with labels contaminated by a two-level randomized response model, as function of the number of calibration samples. The empirical coverage and average size of the prediction sets are stratified based on the strength ϵ of the label contamination process. The number of classes is $K = 4$.

As expected, the *Standard* approach is too conservative when $\epsilon > 0$. Notably, its over-conservativeness persists regardless of increases in the calibration sample size. In contrast, all adaptive methods successfully produce increasingly more informative prediction sets as the calibration sample size grows. Note that Figure 1 shows no significant performance differences among the three alternative implementations of our adaptive method. This is likely due to the label contamination process being not too far from the randomized response model. Below, we will explore how deviations from the randomized response model impact the performances of different implementations. Figures A9 and A10 in Appendix D provide additional results with similar conclusions, from experiments using the randomized response model and the block-randomized response model as contamination models.

6.3 The Impact of the Label Contamination Model

Figure 2 reports on experiments similar to those of Figure 1 but conducted fixing $\epsilon = 0.1$ while varying the parameter $\nu \in [0, 1]$, which controls the discrepancy between the two-level randomized response model and the simple randomized response model ($\nu = 0$).

Unsurprisingly, *Adaptive (simplified)* becomes noticeably more conservative than the fully optimized implementation of the *Adaptive* method as ν increases, particularly for small n. This occurs because *Adaptive (simplified)* relies on a finite-sample correction factor optimized for the randomized response model. The *Asymptotic* method increasingly outperforms the other adaptive implementations as ν grows, primarily because the latter depend on a finite-sample correction factor that, despite our efforts to optimize the constants in the proof of Theorem 4, still relies on theoretical chaining techniques that become less tight when the contamination model deviates significantly from the randomized response model. Exploring alternative mathematical tools to derive even tighter finite-sample corrections remains an open question. Figure A11 in Appendix D gives an alternative visualization of these results, by displaying coverage as function of ν for different
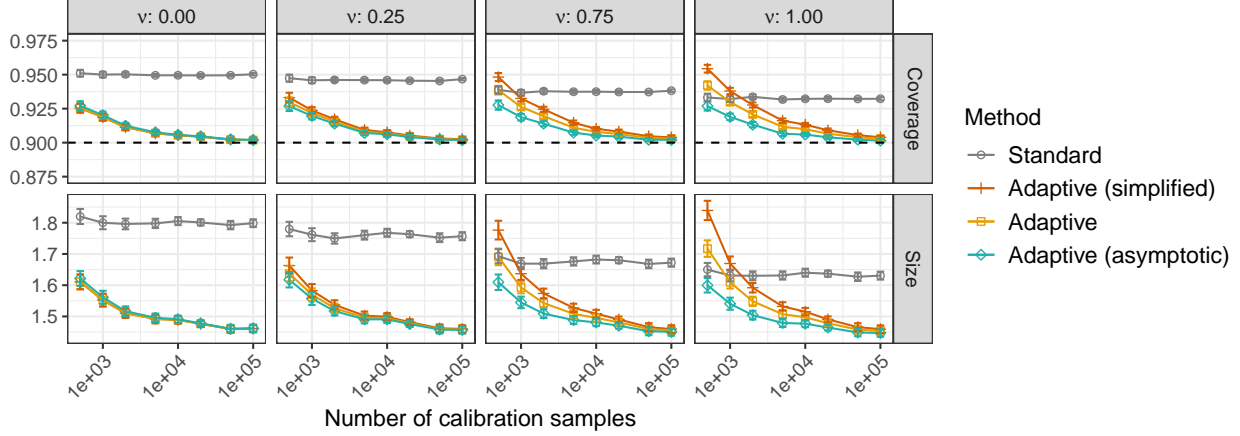
Figure 2: Performances of different conformal methods on simulated data with labels contaminated by a two-level randomized response model, as function of the number of calibration samples. The reported empirical coverage and average size of the prediction sets are stratified based on the contamination model parameter ν. Other details are as in Figure 1.

calibration sample sizes.

Appendix D.3 focuses on how an increase in the number of classes impacts the performances of our adaptive methods. The key conclusion is that while the performance of the *Adaptive (simplified)* and *Adaptive* implementations generally deteriorates as K grows, particularly when n is small and the label contamination strength is high, the *Asymptotic* implementation outperforms the *Standard* method and the other corrections, especially in scenarios with significant deviations from the simple randomized response model.

6.4   The Advantage of Optimistic Calibration

Figure 3 compares the performance of the *Adaptive* method from previous experiments with its optimistic counterpart, *Adaptive+*, introduced in Section 2.4 and detailed in Appendix B.2. The results indicate that, despite the additional assumptions underlying the optimistic method being challenging to rigorously verify, *Adaptive+* performs well in practice. It consistently produces more informative prediction sets compared to the *Standard* and *Adaptive* methods. This advantage is particularly evident in scenarios where the *Adaptive* method is outperformed by the *Standard* method, such as when n is small or K is large, particularly in settings where the contamination model deviates from the randomized response model and the label contamination strength is high.

6.5   The Advantage of Targeting Marginal Coverage

In this section, we include in the comparison the "optimistic" implementation of the adaptive method proposed by Sesia et al. (2024), called *Adaptive+ (label-cond)*, which constructs conformal prediction sets with label-conditional rather than marginal coverage while accounting for label noise. To facilitate a direct comparison, all methods are still evaluated using the same two metrics: empirical marginal coverage and average prediction set size.

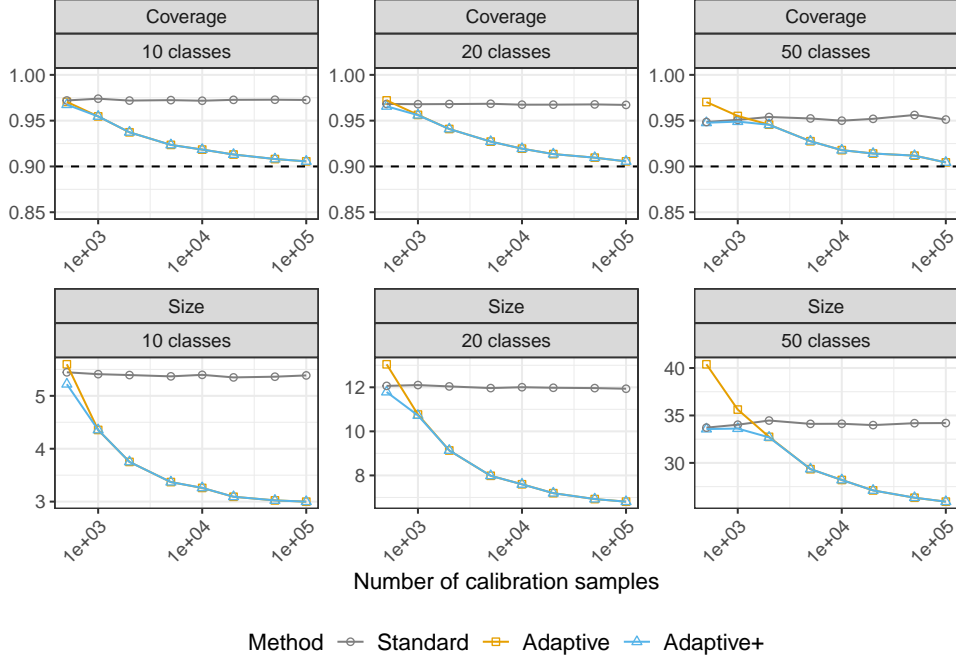The results of this comparison, presented in Figure 4, demonstrate that the advantage of our

Figure 3: Performances of conformal prediction methods on simulated data with contaminated labels. The contamination process is the two-level randomized response model with $\epsilon = 0.2$ and $\nu = 0.8$. The reported empirical coverage and average size of the prediction sets are stratified based on the number of possible labels K. Other details are as in Figure 1.

method, which focuses on achieving marginal rather than label-conditional coverage, becomes more pronounced in problems with many classes and low calibration sample sizes, where the label-conditional adaptive method fails to produce informative prediction sets.

Additional results with conclusions qualitatively similar to Figure 4 are presented in Appendix D.4. In particular, Figure A17 shows that conformal prediction sets with label-conditional coverage can become very large as the class imbalance increases, while our method remains informative even under extreme imbalance.

7 Applications to Real Data

We apply our method to the CIFAR-10H (Peterson et al., 2019) and BigEarthNet (Sumbul et al., 2019, BEN) data sets. In these examples, the availability of observations with both true and contaminated labels allows us to empirically estimate the contamination model matrix T. In particular, after randomly splitting each data set into training, calibration, and test subsets, we estimate T using the Y and $\tilde{Y}$ labels in the training data, as detailed below. Then, our method is applied to the calibration data without accessing the clean labels, reflecting a realistic setting where clean labels are unavailable at calibration time.

This setup is useful for assessing the robustness of our method when Assumption 1 may not hold exactly, yet a reasonable estimate of the contamination model is available. Beyond this setting, T can also be approximated from weaker forms of prior information, further extending the
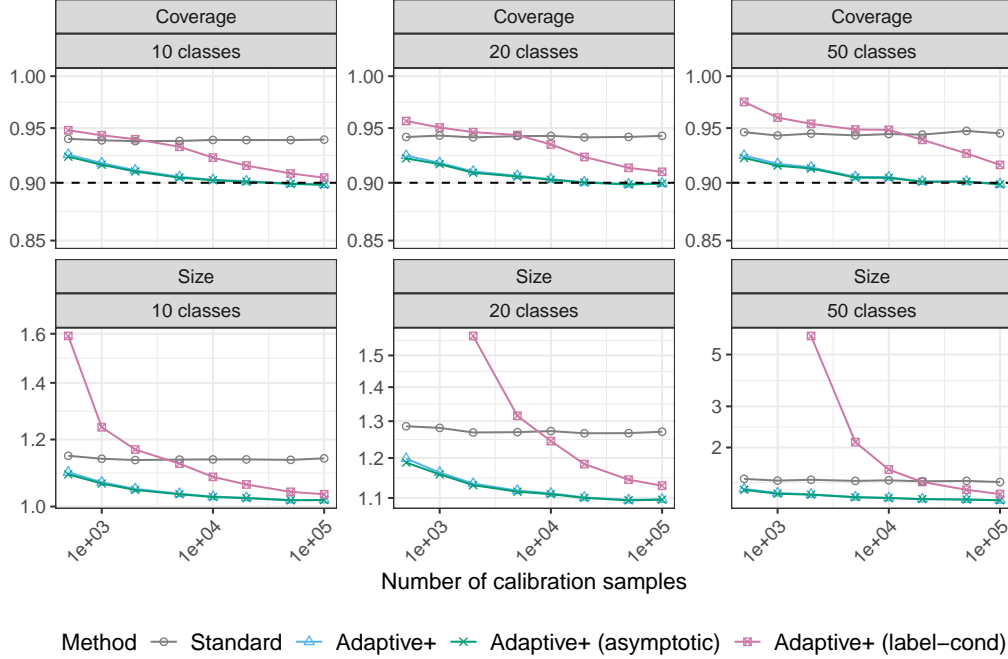
16

Figure 4: Performances of conformal prediction methods on simulated data with contaminated labels. The label contamination process is the two-level randomized response model with $\epsilon = 0.05$ and $\nu = 0.2$. The vertical axis for Size is truncated to highlight differences between the *Standard* and marginal *Adaptive+* methods, which would otherwise be obscured by the full scale required to accommodate the *Adaptive+ (label-cond)* method, especially for small samples. Other details are as in Figure 3.

applicability of our approach; see Section 8 for a more detailed discussion.

7.1 Analysis of CIFAR-10H Data

The CIFAR-10H data set contains 10,000 real-world images in 32x32 resolution, each annotated with both "ground-truth" labels (from the original CIFAR-10 data set) and imperfect labels provided by approximately 50 human annotators via Amazon Mechanical Turk. Variability in annotator opinions introduces label noise, which we model as a contamination process corrupting the true labels. Our objective is to construct conformal prediction sets for the true labels in a test subset of images using a calibration data set with noisy labels.

To simplify the analysis, we use a modified version of CIFAR-10H, where each image is assigned a single corrupted label $\tilde{Y}$, randomly drawn from a multinomial distribution based on the relative frequency of annotator-provided labels. In this setup, the true and noisy labels differ with a frequency of 0.046.

To compute the non-conformity scores, we use a pre-trained ResNet-18 convolutional neural network (Sesia et al., 2024). We then compare the prediction sets obtained with the *Standard* method with those produced by the optimistic *Adaptive* method from Section 6.4, separately applied using the finite-sample and asymptotic approximations of the ideal correction factor $\delta^*(n)$. As in Section 6.5, we include in the comparison the *Adaptive+ (label-cond)* method from Sesia et al.

(2024), which targets label-conditional coverage. For all adaptive methods the label contamination process is (approximately) modeled as a randomized response model whose noise parameter $\epsilon \in [0, 1)$ is estimated as the mean fraction of CIFAR-10H samples with mismatched clean and noisy labels.

Figure 5 presents results for the four conformal methods applied to a random test set of 500 images, with calibration sample sizes n ranging from 500 to 9,500. Each experiment is repeated 250 times using different random splits of CIFAR-10H into calibration and test sets. The methods are compared in terms of marginal coverage and average size of the prediction sets across varying calibration set sizes.

The results show that both the standard method and the label-conditional adaptive method are overly conservative, whereas our marginal adaptive method achieves valid coverage with smaller prediction sets, even with limited calibration set sizes. Targeting marginal coverage yields much more informative prediction sets than label-conditional coverage for small sample sizes (e.g., below 1,000). However, with very large calibration sets (e.g., around 10,000 samples), label-conditional coverage becomes more appealing, offering stronger guarantees with a more modest increase in prediction set size.
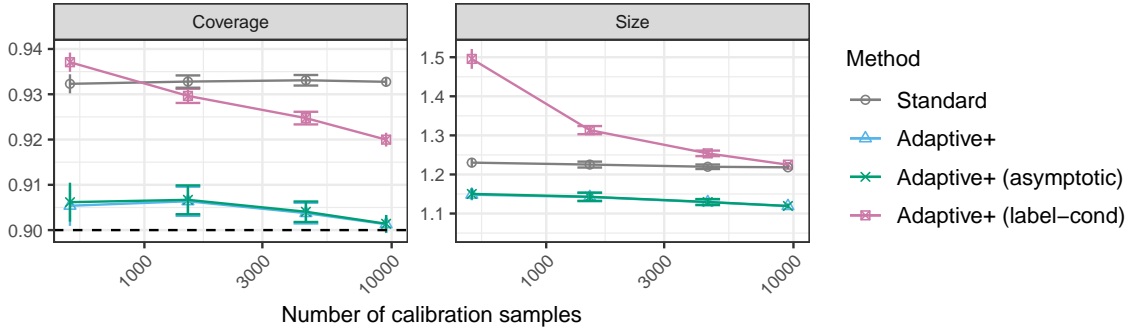

Figure 5: Performances of the conformal prediction methods on the CIFAR-10H data set with contaminated labels. The results are shown as a function of the number of calibration samples. The dashed line indicates the nominal 90% marginal coverage level.

7.2   Analysis of BigEarthNet Data

BigEarthNet is a large data set for remote sensing and Earth observation tasks. It consists of image patches derived from Sentinel-2 satellite imagery (Copernicus Program, European Space Agency), where each 120x120 pixel patch represents a 1.2 km x 1.2 km geographic area. Each patch is associated with multiple land cover classes from the 19-label CORINE Land Cover (CLC) database, reflecting the mix of cover types in the area.

The original annotations in this data set were later refined in the Refined BigEarthNet (REBEN) version (Clasen et al., 2024), which uses the updated CLC2018 database to correct some of the inaccurate labels present in the original data set. As a result, each patch is now associated with two sets of labels: one from the original BigEarthNet (BEN 1.0), which can be seen as being noisier, and another from REBEN, which is more reliable. We treat the BEN 1.0 labels as the corrupted version of the REBEN labels and use our method to construct adaptive prediction sets for the true REBEN labels.

To simplify the analysis, the 19 original labels are grouped into five broader and mutually exclusive classes: Coast, Waters and Wetlands, Arable Land, Agriculture, Vegetation, and Urban Fabric. Each patch is assigned a single label, and if not all its original labels from the CLC2018 database fall within the same broader category, the patch is labeled as Mixed. This leads to a classification problem with six classes.

The resulting classes are highly imbalanced, with relative frequencies ranging from approximately 0.69 to 0.001. In this setting, targeting label-conditional coverage would be completely impractical, as achieving nominal coverage for such rare classes would require an extremely large sample size. Marginal coverage, however, remains a feasible goal.

To compute the non-conformity scores, we train a ResNet-34 on 25,000 patches with noisy labels. The code for training this classifier is adapted from Pinto and St-Charles (2022), whose model was originally designed for multi-label classification on the BigEarthNet data set but is easily modified for our single-label classification task. The trained model computes non-conformity scores on a separate calibration set, drawn from a pool of approximately 270,000 patches, with the calibration set size varied as a control parameter.

We compare the prediction sets obtained using the standard conformal method with those produced by the optimistic implementation of our Adaptive method, as described in the previous section. Additionally, we include the *Adaptive+ (label-cond)* method from Sesia et al. (2024), targeting label-conditional coverage. The Adaptive methods use plug-in estimates of the transition matrix and the marginal label frequencies, both derived from the training set. Specifically, the marginal frequencies of the noisy labels are approximated by their relative empirical frequencies, and the transition matrix T is estimated as:

$$\hat{T}_{kl} = \frac{1}{|\{i \in \mathcal{D}^{\text{train}} : Y_i = l\}|} \sum_{i \in \mathcal{D}^{\text{train}} : Y_i = l} \mathbb{I}\left[\tilde{Y} = k\right].$$

As detailed in Appendix D.5, contamination primarily occurs in patches associated with agricultural activities or mixed land use types, which are often mislabeled as Urban Fabric.

Figure 6 compares the average coverage and size of prediction sets generated by the different calibration methods on a random test set of 500 patches, with calibration sample sizes ranging from 500 to 19,500. Each experiment is repeated 250 times with independent random splits of the data set into calibration and test sets. The results demonstrate that, while all adaptive methods correct the over-conservativeness of the standard method and reach valid coverage as the calibration sample size grows, only the marginal adaptive methods produce more informative prediction sets compared to the *Standard* method. In fact, the marginal adaptive prediction sets become increasingly more informative than the *Standard* prediction sets as n grows. The average size of the label-conditional prediction sets, on the contrary, is much larger than all others, especially for low n.

8   Discussion

This paper has introduced a novel method for split-conformal classification that can effectively handles random label noise in the calibration data. The key advantage of this method is its ability to generate efficient prediction sets with marginal coverage, even in challenging settings with numerous labels and potentially high class imbalance.

While the theoretical results in this paper rely on the simplifying premise of a contamination
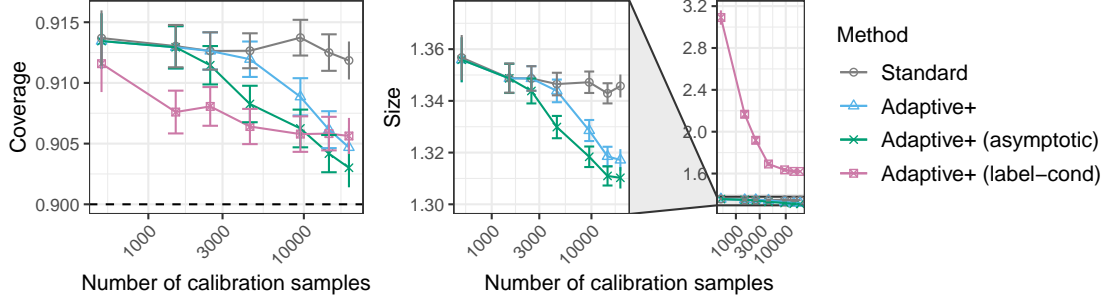
Figure 6: Performance of conformal prediction methods on the BigEarthNet data set with contaminated labels. In the middle panel, the average prediction set size is zoomed in to better illustrate differences between the *Standard* and marginal *Adaptive+* methods, which would otherwise be obscured by the full scale needed to accommodate the *Adaptive+ (label-cond)* method. The number of possible labels is $K = 6$. Other details are as in Figure 5.

model satisfying Assumption 1, with a known transition matrix T, the practical usefulness of our method extends beyond specific scenarios where these conditions hold exactly. As demonstrated by the numerical experiments with real data in Section 7, our method performs well in practice even when Assumption 1 does not hold and the true transition matrix T is replaced by an estimate.

For simplicity, in Section 7, the transition matrix T was estimated using an independent data set with both clean and contaminated labels, following the same distribution as the test and calibration data. However, our method is most valuable precisely in those scenarios where clean labels are unavailable. Fortunately, T can be estimated using only observations of Y and $\tilde{Y}$, without requiring X, at least under Assumption 1. This allows the contamination model to be inferred from a separate data set with a similar label noise process but potentially very different distributions of X and $Y \mid X$, making our approach very useful. For instance, in remote sensing, many data sets lack the refinement process applied to BEN (Koßmann et al., 2022; Schmitt et al., 2019) but share the same Sentinel-2 multispectral imaging technology and CLC2018-based labeling. In such cases, it may be reasonable to *transfer* an estimate of T obtained from the clean labels in BEN, effectively enabling the use of abundant low-quality data to calibrate accurate prediction sets with tight marginal coverage despite having seen no clean labels at all for the data of interest.

Moreover, although not explored here, theoretical guarantees for our method could be derived under weaker assumptions than those made in this paper. For example, our method could be extended to account for uncertainty in the estimated contamination model, effectively replacing a known T with a valid confidence band, using strategies similar to those proposed by Sesia et al. (2024). We leave this extension for future work.

Additional directions for future research could focus on deriving tighter finite-sample bounds for the correction factor, enabling prediction sets with nominal coverage even with smaller calibration sample sizes. Another intriguing direction involves extending our adaptive methodology to regression tasks, where the label space is continuous.

Acknowledgment of Generative AI Use

GPT-4o was used for language improvement.

Software Availability

An open-source software implementation of the methods described in this paper, along with the code needed to reproduce all numerical results, are available online at `https://github.com/tbortolotti/marginal-noise-adaptive-conformal`.

Acknowledgments

References

Barber, R. F., E. J. Candès, A. Ramdas, and R. J. Tibshirani (2023). Conformal prediction beyond exchangeability. *Ass. Stat. 51*(2), 816–845.

Clarkson, J., W. Xu, M. Cucuringu, and G. Reinert (2024). Split conformal prediction under data contamination. In *Proc. 13th Symp. Conformal Probabilistic Prediction Appl.*, Volume 230 of *Proceedings of Machine Learning Research*, pp. 5–27. PMLR.

Clasen, K. N., L. Hackel, T. Burgert, G. Sumbul, B. Demir, and V. Markl (2024). REBEN: Refined BigEarthNet dataset for remote sensing image analysis. *arXiv preprint arXiv:2407.03653*.

Ding, T., A. Angelopoulos, S. Bates, M. Jordan, and R. J. Tibshirani (2024). Class-conditional conformal prediction with many classes. *Adv. Neural Inf. Process. Syst. 36*.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–284. Springer.

Einbinder, B.-S., S. Feldman, S. Bates, A. N. Angelopoulos, A. Gendler, and Y. Romano (2024). Label Noise Robustness of Conformal Prediction. *J. Mach. Learn. Res. 25*(328), 1–66.

Koßmann, D., V. Brack, and T. Wilhelm (2022). Seasonet: A seasonal scene classification, segmentation and retrieval dataset for satellite imagery over germany. *IGARSS 2022 - IEEE Int. Geosci. Remote Sens. Symp.*, 243–246.

Lei, J., J. Robins, and L. Wasserman (2013). Distribution-free prediction sets. *J. Am . Stat. Assoc. 108*(501), 278–287.

Massart, P. (2000). Some applications of concentration inequalities to statistics. In *Ann. Fac. Sci. Toulouse Math.*, Volume 9, pp. 245–303.

Natalini, P. and B. Palumbo (2000). Inequalities for the incomplete gamma function. *Mathematical Inequalities & Applications 3*(1), 69–67.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res. 12*, 2825–2830.

Penso, C. and J. Goldberger (2024). A Conformal Prediction Score that is Robust to Label Noise. *arXiv preprint arXiv:2405.02648*.

Peterson, J. C., R. M. Battleday, T. L. Griffiths, and O. Russakovsky (2019). Human uncertainty makes classification more robust. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 9617–9626.

Pinto, J. and P.-L. St-Charles (2022). BigEarthNet. `https://github.com/jerpint/bigearthnet`.

Richardson, L. F. and R. T. Glazebrook (1911). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Phil. Trans. R. Soc. A 210*(459-470), 307–357.

Romano, Y., M. Sesia, and E. Candès (2020). Classification with valid and adaptive coverage. *Adv. in Neural Inf. Process. Syst. 33*, 3581–3591.

Schmitt, M., L. H. Hughes, C. Qiu, and X. X. Zhu (2019). SEN12MS – A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. IV-2/W7*, 153–160.

Sesia, M., Y. R. Wang, and X. Tong (2024). Adaptive conformal classification with noisy labels. *J. R. Stat. Soc. (B)*, qkae114.

Sumbul, G., M. Charfuelan, B. Demir, and V. Markl (2019). Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904. IEEE.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World.* Springer Science & Business Media, 2005.

Vovk, V., D. Lindsay, I. Nouretdinov, and A. Gammerman (2003). Mondrian confidence machine. *Technical Report*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Am . Stat. Assoc. 60*(309), 63–69.

A Review of Existing Methods

A.1 Standard Conformal Classification with Marginal Coverage

Algorithm A1: Standard conformal classification with marginal coverage.

Input: Data set $\{(X_i, \tilde{Y}_i)\}_{i=1}^n$ with observable labels $\tilde{Y}_i \in [K]$.

Unlabeled test point with features X_{n+1}.

Pre-trained K-class classification model $\hat{\pi}$. Prediction function $\mathcal{C}$.

Desired miscoverage probability $\alpha \in (0, 1)$.

1 Randomly split $[n]$ into two disjoint subsets, $\mathcal{D}^{\text{train}}$ and $\mathcal{D}$, defining $n = |\mathcal{D}|$.

2 Train the classifier $\mathcal{A}$ on the data in $\mathcal{D}^{\text{train}}$.

3 Compute $\hat{s}(X_i, \tilde{Y}_i)$ using (2), for all $i \in \mathcal{D}$.

4 Define $\hat{\tau}$ as the $\lceil (1+n) \cdot (1-\alpha) \rceil$ smallest value in $\{\hat{s}(X_i, \tilde{Y}_i)\}_{i \in \mathcal{D}}$.

5 Evaluate $\hat{C}(X_{n+1}) = \mathcal{C}(X_{n+1}, \hat{\tau}; \hat{\pi})$. **Output:** Conformal prediction set $\hat{C}(X_{n+1})$ for $\tilde{Y}_{n+1}$, satisfying (1).

Proposition A1 (e.g., from Lei et al. (2013) or Romano et al. (2020)). *If the data pairs $(X_i, \tilde{Y}_i)$, for all $i \in [n+1]$, are exchangeable random samples from some joint distribution, the prediction set $\hat{C}(X_{n+1})$ output by Algorithm A1 has marginal coverage (1) for the observable labels $\tilde{Y}$; i.e., $\mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1})\right] \geq 1-\alpha$. Further, if all scores $\hat{s}(X_i, \tilde{Y}_i)$ computed by Algorithm A1 are almost-surely distinct, $\mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1})\right] \leq 1 - \alpha + 1/(n+1)$, where $n = |\mathcal{D}|$ is the number of data points in the calibration set.*

A.2 Generalized Inverse Quantile Non-conformity Scores

We briefly review here the definition of the generalized inverse quantile non-conformity scores proposed by Romano et al. (2020), which are used in the empirical demonstrations presented in this paper. These non-conformity scores are designed to produce more flexible prediction sets that can account for possible heteroscedasticity in the distribution of $Y \mid X$.

For any $x \in \mathbb{R}^d$ and $t \in [0, 1]$, define

$$\hat{Q}(x, \hat{\pi}, t) = \min\{k \in \{1, \ldots, K\} : \hat{\pi}_{(1)}(x) + \hat{\pi}_{(2)}(x) + \ldots + \hat{\pi}_{(k)}(x) \geq t\}, \tag{A1}$$

where $\hat{\pi}_{(1)}(x) \geq \ldots \geq \hat{\pi}_{(K)}(x)$ are the descending order statistics of $\hat{\pi}(x, 1), \ldots, \hat{\pi}(x, K)$. Intuitively, $\hat{Q}(x, \hat{\pi}, \cdot)$ may be seen as a generalized quantile function. Similarly, let $\hat{r}(x, \hat{\pi}, k)$ denote the rank of $\hat{\pi}(x, k)$ among $\hat{\pi}(x, 1), \ldots, \hat{\pi}(x, K)$. With this notation, one can also define a corresponding generalized cumulative distribution function:

$$\hat{\Pi}(x, \hat{\pi}, k) = \hat{\pi}_{(1)}(x) + \hat{\pi}_{(2)}(x) + \ldots + \hat{\pi}_{(\hat{r}(x, \hat{\pi}, k))}(x).$$

Then, the function $\mathcal{C}$ proposed by Romano et al. (2020) can be written as:

$$\mathcal{C}(x; \hat{\pi}, \tau) = \{k \in [K] : \hat{r}(x, \hat{\pi}, k) \leq \hat{Q}(x, \hat{\pi}, \tau_k)\}, \tag{A2}$$

and the corresponding non-conformity scores defined in (2) can be evaluated efficiently by noting that $\hat{s}(x, k) = \hat{\Pi}(x, k)$; see Romano et al. (2020) for further details.

The prediction function defined in (A2) may be understood by noting that, if $\tau = (\tau_0, \ldots, \tau_0)$ for some $\tau_0 \in [0, 1]$, the output of $\mathcal{C}(x; \hat{\pi}, \tau)$ is the list of most likely classes according to $\hat{\pi}(x)$ up until the first label l for which $\hat{\Pi}(x, \hat{\pi}, l) \geq \tau_0$. Therefore, in the ideal case where $\hat{\pi}(x, k) = \mathbb{P}[Y = k \mid X = x]$, one can verify that $\mathcal{C}(x; \hat{\pi}, \tau)$ is the smallest possible (deterministic) prediction set for Y with perfect object-conditional coverage at level τ_0, i.e., satisfying $\mathbb{P}[Y \in \mathcal{C}(x; \hat{\pi}, \tau) \mid X = x] \geq \tau_0$. Note that Romano et al. (2020) also developed a more powerful randomized version of (A2) that enjoys similar theoretical properties while being able to produce even more informative prediction sets. The results of this paper can also seamlessly accommodate such additional randomness in $\mathcal{C}$, and indeed that is the practical approach followed in the empirical demonstrations presented in this paper, but we choose not to review such extension explicitly here to avoid making the notation too cumbersome.

B   Additional Methodological Details

B.1   The Conservativeness of Standard Conformal Predictions

Corollary A1. *Consider the same setting of Theorem 1 and assume Assumption 1 holds. Suppose also that the cumulative distribution functions of the scores (A11) satisfy*

$$\max_{l \neq k} F_k^l(t) \leq F_k^k(t), \tag{A3}$$

for all $t \in \mathbb{R}$ and $k \in [K]$. Then, $\Delta(\hat{\tau}) \geq 0$ almost-surely, and hence the predictions $\hat{C}(X_{n+1})$ of Algorithm A1 satisfy (1).

Intuitively, (A3) states that the scores $\hat{s}(X, k)$ assigned by the machine learning model tend to be smaller than any other score $\hat{s}(X, l)$ for $l \neq k$ among data points with true label $Y = k$. In other words, this could be interpreted as saying that the correct label is the most likely point prediction of the machine learning model, for each possible class $k \in [K]$.

B.2   Optimistic Adaptive Calibration

This section proposes an alternative method which can outperform both the standard conformal method and the adaptive method introduced above. We show that, under a mild stochastic dominance assumption, it is convenient to adopt a calibration method which optimistically chooses between the standard and the adaptive method depending on which approach leads to a lower calibrated threshold; that is, to a less conservative prediction set.

In the practice, the optimistic adaptive method consists in applying Algorithm 1 with the set $\hat{\mathcal{I}}$ replaced by

$$\hat{\mathcal{I}} := \left\{ i \in [n] : \frac{i}{n} \geq 1 - \alpha - \max\left\{ \hat{\Delta}(S_{(i)}) - \delta(n), -\frac{1-\alpha}{n} \right\} \right\}. \tag{A4}$$

It is possible to show that the prediction sets output by the optimistic approach have marginal coverage guarantee.

Let $U_1, \ldots, U_n$ be i.i.d. uniform random variables on $[0, 1]$, and denote their order statistics as $U_{(1)}, \ldots, U_{(n)}$; define

$$c(n) := \mathbb{E}\left[\sup_{i \in [n]} \left\{ \frac{i}{n} - U_{(i)} \right\} \right]. \tag{A5}$$

Note that a Monte Carlo simulation of n independent standard uniform random variables could be employed to get an estimate of the constant $c(n)$ with arbitrary precision. Therefore, the constant will be hereafter assumed to be known.

Proposition A2. *Under the setup of Theorem 2, assume also that $\inf_{t \in [0,1]} \Delta(t) \geq \delta(n) - (1-\alpha)/n$. If $\hat{C}(X_{n+1})$ is the prediction set output by Algorithm 1 applied with the set $\hat{\mathcal{I}}$ defined in (A4) instead of (11), and with a correction factor $\delta(n)$ satisfying $\delta(n) \geq \max\{\delta^*(n), c(n)\}$. Then, $\mathbb{P}[Y_{n+1} \in \hat{C}(X_{n+1})] \geq 1 - \alpha$.*

Note that the assumption $\inf_{t\in[0,1]} \Delta(t) \geq \delta(n) - (1-\alpha)/n$ in Proposition A2 is stronger than the stochastic dominance condition A3 in Corollary A1, but remains reasonable. In cases when the calibration set size is large enough to make the correction factor $\delta(n)$ small, the assumption aligns with (A3), which implies $\inf_{t\in[0,1]} \Delta(t) \geq 0$. It will be made clear in the following sections that, in the practice, the requirement $\delta(n) \geq \max\{\delta^*(n), c(n)\}$ is not more restrictive than $\delta(n) \geq \delta^*(n)$, and that it is easily met by all the practical adaptive methods proposed in this work.

B.3 Solving the Optimization Problem

The optimal choice of β in (14) depends on the structure of the contamination model through the inverse transition matrix W. Fortunately, the finite-sample correction term $\delta^{\mathrm{FS}}(n)$ in (14) can be practically computed by solving a convex optimization problem.

Let us define

$$\delta_1^{\mathrm{FS}}(n) := \inf_{\beta \in \mathbb{R}^{K+1}} \left\{ c(n)\left(\beta_0 + \frac{\sum_{k=1}^{K}\beta_k}{K}\right) + \frac{2}{\sqrt{n}} \max_{l\in[K]} \sum_{k} |\Omega_{kl}|\sqrt{\log(Kn+1)} \right\}, \qquad (A6)$$

$$\delta_2^{\mathrm{FS}}(n) := \inf_{\beta \in \mathbb{R}^{K+1}} \left\{ c(n)\left(\beta_0 + \frac{\sum_{k=1}^{K}\beta_k}{K}\right) \right.$$

$$\left. + \frac{2}{\sqrt{n}} 24 \max_{k,l\in[K]} |\Omega_{kl}| \frac{2\log K + 1}{2\log K - 1}\sqrt{2K\log K} \right\}. \qquad (A7)$$

Clearly, $\delta^{\mathrm{FS}}(n) \leq \min\{\delta_1^{\mathrm{FS}}(n), \delta_2^{\mathrm{FS}}(n)\}$. Moreover, the optimization problems in (A6) and (A7) are both convex. This suggests tackling the optimization problem in (14) by solving separately the two convex problems (A6) and (A7) and then taking the value of β corresponding to smallest of the two optimal values.

Intuitively, the more the contamination process resembles the randomized response model, the more similar the behavior of $\delta^{\mathrm{FS}}(n)$ is to that of $c(n)$. If the deviation of the contamination model from the randomized response model is non-negligible – that is, when the matrix Ω amplifies the second term in (14) – the worst case behaviour of the correction factor is $\min\{\mathcal{O}(\sqrt{(K\log K)/n}), \mathcal{O}(\sqrt{\log(Kn)/n})\}$, which still scales well with n, especially in those cases where K is considerably lower than n. We argue in Section B.4 that, for some special contamination models, the correction factors display improved scaling properties than that of the general case presented here; for the two-level randomized response model, for example, the finite-sample correction factor is proved to scale as $\mathcal{O}([1 + \sqrt{(2\log K)/K}]/\sqrt{n})$.

B.4 Simplified Methods for Special Contamination Models

B.4.1 Block-Randomized Response Model

Recall that $T = \mathbb{P}\left[\tilde{Y} = k \mid X, Y = l\right]$. The Block-Randomized response model with b blocks of dimension $m := K/b$ is identified by the matrix T in the form

$$T = (1 - \epsilon)I_K + \frac{\epsilon}{m} \cdot B_b, \tag{A8}$$

where B_b is a block-diagonal matrix with b constant blocks equal to J_m, i.e., the $(m \times m)$ matrix of ones.

It is possible to verify that matrix W corresponding to model (A8) reads

$$W = \frac{1}{1 - \epsilon}I_K - \frac{\epsilon}{m(1 - \epsilon)}B_b.$$

Consider for instance the case $K = 4$ and $b = 2$, so that $m = 2$. Then the matrix Ω is

$$\Omega = \begin{bmatrix} \frac{1}{1-\epsilon} - \frac{\epsilon}{2(1-\epsilon)} - \beta_0 - \frac{\beta_1}{4} & -\frac{\epsilon}{2(1-\epsilon)} - \frac{\beta_1}{4} & -\frac{\beta_1}{4} & -\frac{\beta_1}{4} \\ -\frac{\epsilon}{2(1-\epsilon)} - \frac{\beta_2}{4} & \frac{1}{1-\epsilon} - \frac{\epsilon}{2(1-\epsilon)} - \beta_0 - \frac{\beta_2}{4} & -\frac{\beta_2}{4} & -\frac{\beta_2}{4} \\ -\frac{\beta_3}{4} & -\frac{\beta_3}{4} & \frac{1}{1-\epsilon} - \frac{\epsilon}{2(1-\epsilon)} - \beta_0 - \frac{\beta_3}{4} & -\frac{\epsilon}{2(1-\epsilon)} - \frac{\beta_3}{4} \\ -\frac{\beta_4}{4} & -\frac{\beta_4}{4} & -\frac{\epsilon}{2(1-\epsilon)} - \frac{\beta_4}{4} & \frac{1}{1-\epsilon} - \frac{\epsilon}{2(1-\epsilon)} - \beta_0 - \frac{\beta_4}{4} \end{bmatrix}.$$

We are interested in studying the finite sample correction in (14) for this special case. Assume that the number b of blocks is fixed, and that $m = K/b$ is an integer. Set $\beta'_0 = \frac{1}{1-\epsilon}$ and $\beta'_k = -\frac{K\epsilon}{m(1-\epsilon)}$ for all k, and let $\beta' = (\beta'_0, \beta'_1, \ldots, \beta'_K)$. Note that $\beta'_0 + \frac{1}{K}\sum_k \beta'_k = (1 - b\epsilon)/(1 - \epsilon)$.

Consider the first term in (15), evaluated in β'. Excluding the multiplicative constant, the term reads

$$\max_{l \in [K]} \sum_k |\Omega_{kl}| = \frac{K - m}{K} \frac{K\epsilon}{m(1 - \epsilon)} = \frac{K - m}{m} \cdot \frac{\epsilon}{1 - \epsilon} = (b - 1) \cdot \frac{\epsilon}{1 - \epsilon}.$$

Now consider the second term in (15). For $\beta = \beta'$ one gets

$$\max_{k,l \in [K]} |\Omega_{kl}| = \frac{\epsilon}{m(1 - \epsilon)} = \frac{b\epsilon}{K(1 - \epsilon)}.$$

Putting all together, it follows from the definition of the correction term in (14) that

$$\delta^{\mathrm{FS}}(n, K) \leq \frac{1}{1 - \epsilon}(1 - b\epsilon) \cdot c(n) + \frac{1}{\sqrt{n}}B(K, n, \beta')$$

$$\leq \frac{1}{1 - \epsilon} \cdot c(n) + \frac{1}{\sqrt{n}}B(K, n, \beta'),$$

where

$$B(K, n, \beta') = 2\min\left\{(b - 1) \cdot \frac{\epsilon}{1 - \epsilon}\sqrt{\log(Kn + 1)}, 24 \cdot \frac{b\epsilon}{1 - \epsilon} \cdot \frac{2\log K + 1}{2\log K - 1}\sqrt{\frac{2\log K}{K}}\right\}. \tag{A9}$$
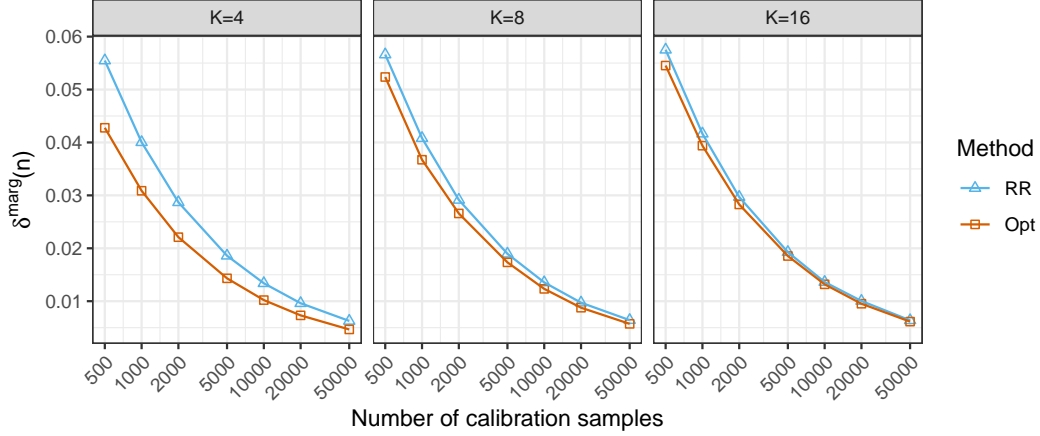
Figure A1: Finite-sample correction for the block randomized response model, estimated using the β which optimizes δ^{FS} for the randomized response model (RR) and the β resulting from the minimization problem in (14) (Opt). The estimates are displayed as function of the sample size, for different values of K. In these experiments, the number of blocks is $b = 2$ and $\epsilon = 0.1$.

This shows that, in the simplified case of the block-randomized response model, the finite sample correction scales at most as $\sqrt{\log(Kn)}$ with K.

Figures A1 and A2 display the comparison between the two estimates of the finite sample correction for the block randomized response model with two blocks, one obtained by using a basic definition of β (RR) and the other obtained by solving the optimization problem in (14) (Opt). The basic definition is set as the vector of β's which solve the minimization problem for the randomized response model described in Section 3.1, namely $\beta_0 = \frac{1}{1-\epsilon}$ and $\beta_k = -\frac{\epsilon}{1-\epsilon}$, $k = 1, \ldots, K$. The experiment is conducted by setting the number of blocks $b = 2$ and $\epsilon = 0.1$. In figure A1, the finite sample correction is displayed as a function of the sample size, for $K = 4, 8, 16$. As expected, both estimates of δ^{FS} decay as n increases and do not show any significant increase as the number of classes grows. Figure A2 displays δ^{FS} as a function of K, for $n = 100, 500, 1000$. When n is fixed and of the same order of K, the finite sample correction increases as K grows. When n is of greater order than K, the finite sample correction is almost insensitive to an increase in K. The difference between the optimal finite sample correction and that obtained with the naive beta accentuates for low values of K and of the sample size, and narrows as these two parameters increase.

B.4.2 Two-Level Randomized Response Model

The model considered here describes a natural label contamination process involving two clearly defined groups of labels. Let ϵ and ν be the two parameters which identify the model, with $\epsilon \in [0, 1)$ and $\nu \in [0, 1]$. For simplicity, assume that the total number of possible labels K is even. The two-level randomized response model is identified by the transition matrix T, which is defined as a (2×2) block matrix in the form

$$T = \begin{pmatrix} (1 - \epsilon)I + \frac{\epsilon}{K}(1 + \nu)J & \frac{\epsilon}{K}(1 - \nu)J \\ \frac{\epsilon}{K}(1 - \nu)J & (1 - \epsilon)I + \frac{\epsilon}{K}(1 + \nu)J \end{pmatrix}.$$
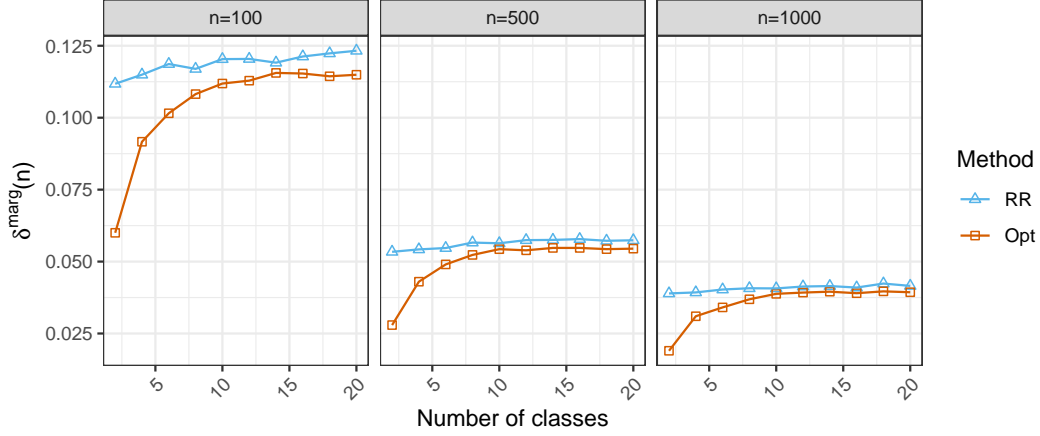
28

Figure A2: Finite-sample correction for the block randomized response model, estimated using the β which optimizes δ^{FS} for the randomized response model (RR) and the β resulting from the minimization problem in (14) (Opt). The estimates are displayed as function of the number of classes K, for different values of the sample size. In these experiments, the number of blocks is $b = 2$ and $\epsilon = 0.1$.

Above, I denotes the $K/2$-dimensional identity matrix, and J denotes the $K/2$-dimensional constant matrix of ones. For ease of notation, let

$$f = \epsilon(1 + \nu), \quad g = \epsilon(1 - \nu), \quad \text{and} \quad e = \frac{\epsilon(1 + \nu) - \epsilon^2(1 - \nu)}{1 - \epsilon + f/2}.$$

It is possible to verify that the inverse of T is the 2×2 block matrix W in the form

$$W = \begin{pmatrix} \frac{1}{1-\epsilon}I - \frac{p}{K}J & -\frac{h}{K}J \\ -\frac{h}{K}J & \frac{1}{1-\epsilon}I - \frac{p}{K}J \end{pmatrix},$$

where

$$p = \frac{1}{1 - \epsilon} \cdot \frac{e}{1 - \epsilon + e/2},$$

$$h = \frac{g}{(1 - \epsilon)^2} \cdot \left(1 - \frac{f}{2(1 - \epsilon + f/2)}\right)\left(1 - \frac{e}{2(1 - \epsilon + e/2)}\right).$$

If we consider for instance the case $K = 4$, then the matrix Ω is

$$\Omega = \begin{bmatrix} \frac{1}{1-\epsilon} - \beta_0 - \frac{1}{4}(p + \beta_1) & -\frac{1}{4}(p + \beta_1) & -\frac{1}{4}(h + \beta_1) & -\frac{1}{4}(h + \beta_1) \\ -\frac{1}{4}(p + \beta_2) & \frac{1}{1-\epsilon} - \beta_0 - \frac{1}{4}(p + \beta_2) & -\frac{1}{4}(h + \beta_2) & -\frac{1}{4}(h + \beta_2) \\ -\frac{1}{4}(h + \beta_3) & -\frac{1}{4}(h + \beta_3) & \frac{1}{1-\epsilon} - \beta_0 - \frac{1}{4}(p + \beta_3) & -\frac{1}{4}(p + \beta_3) \\ -\frac{1}{4}(h + \beta_4) & -\frac{1}{4}(h + \beta_4) & -\frac{1}{4}(p + \beta_4) & \frac{1}{1-\epsilon} - \beta_0 - \frac{1}{4}(p + \beta_4) \end{bmatrix}.$$

Set $\beta_0' = \frac{1}{1-\epsilon}$ and $\beta_k' = -p$ for all k, and let $\beta' = (\beta_0', \beta_1', \ldots, \beta_K')$. Note that

$$\beta_0' + \frac{1}{K}\sum_k \beta_k' = \frac{1}{1 - \epsilon}\left(1 - \frac{e}{1 - \epsilon + e/2}\right) \leq \frac{1}{1 - \epsilon}.$$

29

Then, the first term in (15) evaluated in correspondence of β' reads

$$\max_{l\in[K]}\sum_k |\Omega_{kl}| = \frac{K}{2}\frac{1}{K}|h-p| = \frac{1}{2}|h-p|,$$

while the second term reads

$$\max_{k,l\in[K]} |\Omega_{kl}| = \frac{1}{K}|h-p|.$$

Note that the quantity $|h-p|$ is independent of K and n, and that for any $\epsilon \in [0,1)$ and for any $\nu \in [0,1]$, $|h-p| < \infty$. In fact, one may verify that $h \geq 0$ and that it is a decreasing function of ν, while $p \geq \frac{\epsilon}{1-\epsilon} \geq h$ increases with ν. Then, $|h-p| \leq p(\nu=1) - h(\nu=1) = \frac{2\epsilon}{1-\epsilon}$. This implies that for every $\nu \in [0,1]$ and for every ϵ that is model-wise meaningful (e.g., ϵ in $[0,0.5]$), $|h-p|$ is bounded and small. For example, for $\epsilon = 0.2$, $|h-p| \leq 0.5$ for all ν.

Then, the finite sample correction in the specific case of the two-level randomized response model scales as $\frac{1}{\sqrt{n}}\left(1 + \sqrt{\frac{2\log K}{K}}\right)$ and reads

$$\delta^{\mathrm{FS}}(n,K) \leq c(n)\left(\beta_0' + \frac{\sum_{k=1}^K \beta_k'}{K}\right) + \frac{1}{\sqrt{n}}B(K,n,\beta')$$

$$\leq \frac{1}{1-\epsilon}\cdot c(n) + \frac{48}{\sqrt{n}}\cdot\frac{2\epsilon}{1-\epsilon}\cdot\frac{2\log K+1}{2\log K-1}\cdot\sqrt{\frac{2\log K}{K}}.$$

Figures A3–A4 display the comparison between the estimates of the finite sample correction δ^{FS} for the two-level randomized response model, obtained by using a basic definition of β (RR) and by solving the optimization problem in (14) (Opt). As in Section B.4.1, the basic definition is set as the vector of β's which solve the minimization problem for the randomized response model, namely $\beta_0 = \frac{1}{1-\epsilon}$ and $\beta_k = -\frac{\epsilon}{1-\epsilon}$, $k = 1,\ldots,K$. The comparison is done in two scenarios, identified by two values of ν. When $\nu = 0.2$, the two-level randomized response model is close to the randomized response model and the finite sample correction obtained with the naive β estimate is close to the optimal finite sample correction, for all $K \geq 4$ and for all n. When $\nu = 0.8$, the two-level randomized response model deviates from the randomized response model and the optimal finite sample correction improves the one obtained with the naive β, although the difference between the two estimates narrows both as K increases and as n increases. Figures A5–A6 display the behavior of the finite sample corrections obtained with the two methods, as the contamination model deviates from the randomized response model. The closer ν is to 1, the larger is the deviation from the randomized response model. As expected, the difference between the two finite sample corrections becomes more significant as ν increases, and is further magnified with an increase in ϵ, namely the parameter that quantifies the level of random noise (see Figure A6). These results show that calculating the optimized δ^{FS} proves more effective the more the model deviates from the randomized response model, especially when the number of classes is low and when the amount of random level noise is high. Again, an increase in sample size reduces the gap between the two finite sample corrections.

Overall, these experiments suggest a convenient practical rule. When the number of classes and the sample size are large, one can simplify the estimation routine by avoiding solving the optimization problem, and evaluating the finite sample correction at the β vector corresponding to the randomized response model. However, the optimal finite sample correction should be employed when the contamination model deviates from the randomized response model and the amount of label contamination is high.
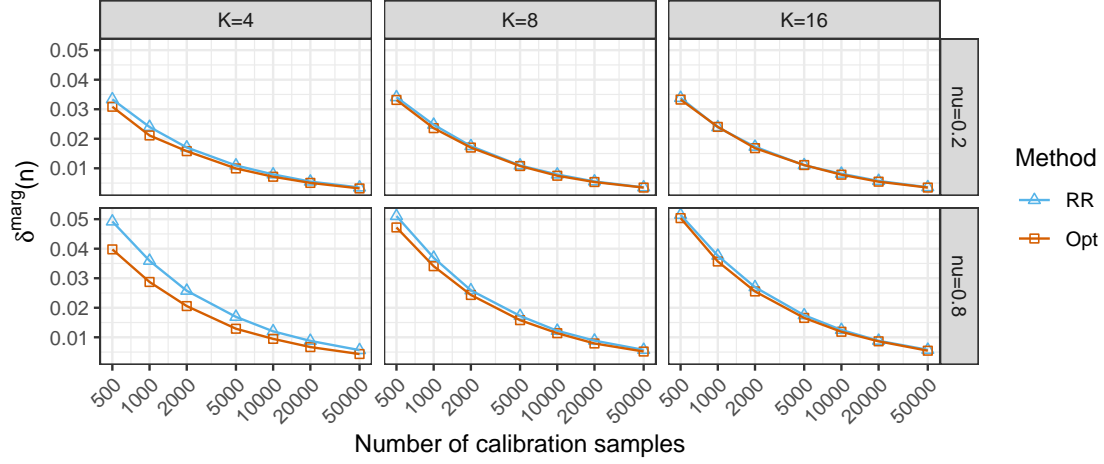
Figure A3: Finite-sample correction for the two-level randomized response model, estimated using the β which optimizes δ^{FS} for the randomized response model (RR) and the β resulting from the minimization problem in (14) (Opt). The estimates are displayed as function of the sample size, for different values of K and for $\nu \in \{0.2, 0.8\}$. Parameter ϵ is set to $\epsilon = 0.1$.
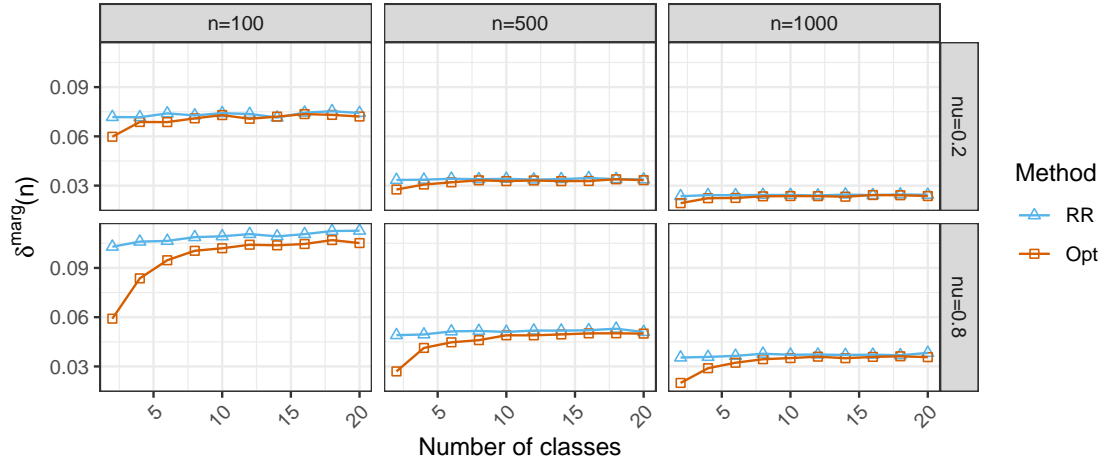

Figure A4: Finite-sample correction for the two-level randomized response model, estimated using the β which optimizes δ^{FS} for the randomized response model (RR) and the β resulting from the minimization problem in (14) (Opt). The estimates are displayed as function of the number of classes K, for different values of the sample size and for $\nu = 0.2, 0.8$. Parameter ϵ is set to $\epsilon = 0.1$.
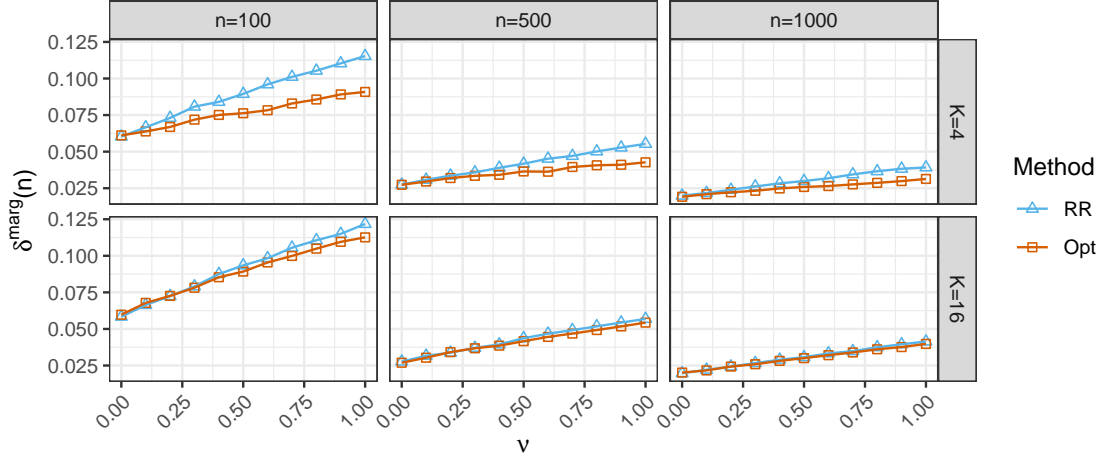
Figure A5: Finite-sample correction for the two-level randomized response model, estimated using the β which optimizes δ^{FS} for the randomized response model (RR) and the β resulting from the minimization problem in (14) (Opt). The estimates are displayed as function of ν, for different values of the sample size and for $K = 4, 16$. Parameter ϵ is set to $\epsilon = 0.1$.
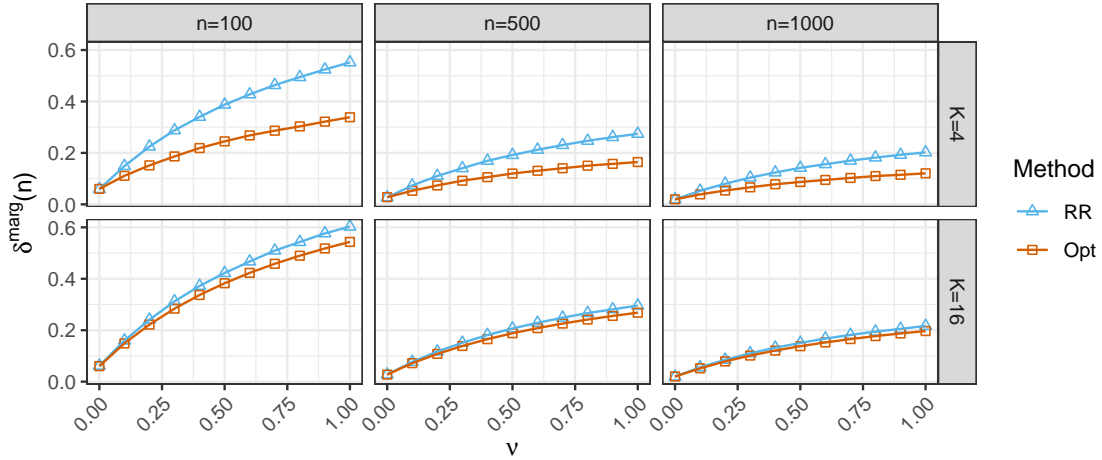

Figure A6: Finite-sample correction for the two-level randomized response model, estimated using the β which optimizes δ^{FS} for the randomized response model (RR) and the β resulting from the minimization problem in (14) (Opt). The estimates are displayed as function of ν, for different values of the sample size and for $K = 4, 16$. Parameter ϵ is set to $\epsilon = 0.5$.

B.5 Richardson Extrapolation

Richardson extrapolation (Richardson and Glazebrook, 1911) is a numerical technique which improves the order of accuracy of an estimate by leveraging a sequence of approximations obtained at different discretization steps and extrapolating to the zero step size. In our context, the method turns out to be particularly useful, as it allows us to reduce the computational burden which comes with calculating (17) for small values of h. By order of the Richardson extrapolation, we refer to the number of times that the extrapolation is applied recursively to improve the accuracy of the resulting estimate.

To show the effectiveness of estimating $\delta(n)$ via a Monte Carlo simulation coupled with Richarson extrapolation, we consider the Brownian bridge as an illustrative example. As the distribution of the supremum is known for the Brownian bridge, its expected value is a known target that enables us to assess how the estimation strategy behaves. These experiments additionally provide us with a convenient rule of thumb for setting the parameters h and M of the Monte Carlo simulation, together with the order of Richardson's extrapolation, which lead to an estimate of $\delta(n)$ that is sufficiently close to its true value.

B.5.1 Example: The F-Brownian Bridge

Let $X_1, \ldots, X_n$ be a random sample from a distribution function F on the real line. Let $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[X_i \leq t]$ and $F(t) = \mathbb{P}[X \leq t]$. From Donsker's theorem, we know that $\sqrt{n} \left(\hat{F}_n(t) - F(t) \right)$ converges to the F-Brownian bridge process, namely a zero mean Gaussian process with covariance function $c^B(t, s) = F(t \wedge s) - F(t)F(s)$. The supremum of the F-Brownian bridge process has Kolmogorov distribution, and its expected value is equal to the expected value of the sup of the uniform Brownian bridge; that is,

$$\mathbb{E}\left[\sup_t |B_F(t)|\right] = \mathbb{E}\left[\sup_t |B_{U[0,1]}(t)|\right] = \sqrt{\frac{\pi}{2}} \log 2. \tag{A10}$$

Figure A7 displays the estimates of (A10) obtained via Monte Carlo simulation with $M = 1,000,000$, as function of the number T of points of the grid. The standard method refers to the estimate obtained without employing Richardson extrapolation downstream of the MC simulation. Richardson(1) and Richardson(2) refer to the extrapolation method of order 1 and order 2, respectively. The dashed line marks the target (A10). This experiment demonstrates that, for a given value of h, the estimate obtained through Richardson extrapolation improves upon the estimate obtained solely from the Monte Carlo simulation performed on the grid corresponding to h. Furthermore, reducing h below a certain threshold does not lead to any significant improvement in the estimate obtained through Richardson extrapolation for higher values of h. Finally, no significant difference is observed between the estimates obtained using Richardson(1) or Richardson(2), suggesting that using one or the other method does not significantly impact the resulting estimate.

Figure A8 displays the variability of the estimates for each of the considered methods as function of M. The variability of the estimates is obtained by running the experiment $B = 20$ times, for $h = 1/1600$. These results suggest that choosing Richardson(1) over Richardson(2), although not providing any significant improvement in the estimates in mean, may result in more stable estimates for $M \geq 10,000$.
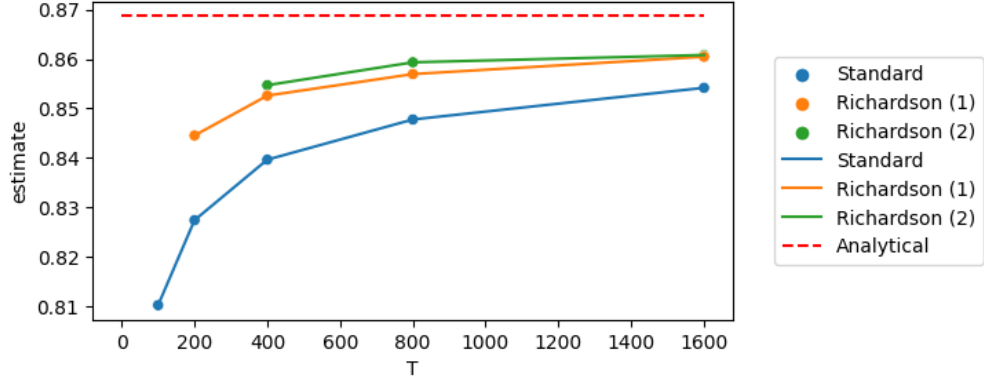
Figure A7: Estimate of the expected value of the supremum of the Brownian Bridge as function of $T = 1/h$. The estimates are obtained via a Monte Carlo simulation with $M = 1000000$. The standard method refers to the estimate obtained without employing Richardson extrapolation downstream of the MC simulation. Richardson(1) and Richardson(2) refer to the extrapolation method of order 1 and order 2, respectively. The dashed line marks the analytical value of the expectation of the supremum of the Brownian Bridge.
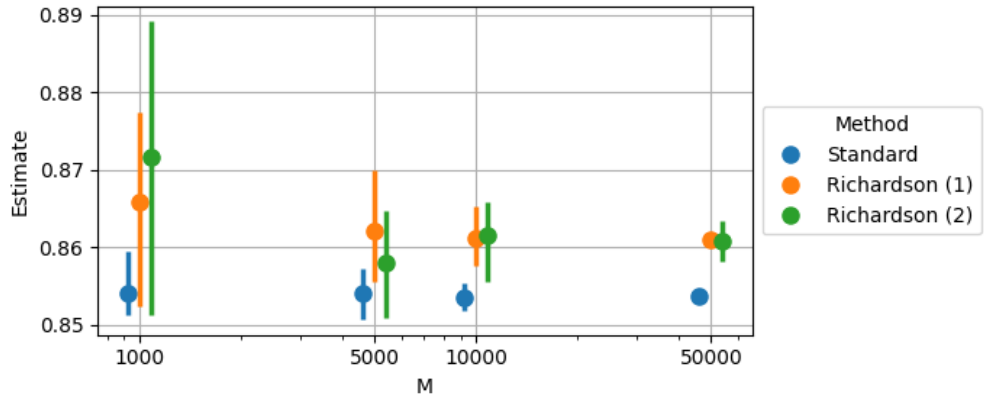

Figure A8: Variability of the estimates of the expected value of the supremum of the Brownian Bridge as function of M, obtained with $B = 20$ repeated experiments. The estimates are obtained via Monte Carlo simulation and for $h = 1/1600$. The dashed line marks the analytical value of the expectation of the supremum of the Brownian Bridge.

34

C Mathematical Proofs

C.1 Preliminaries

Proof of Theorem 1. For any $k, l \in [K]$ and $t \in [0, 1]$, define, as in Section 2.4,

$$F_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \leq t \mid Y = l\right], \qquad \tilde{F}_l^k(t) := \mathbb{P}\left[\hat{s}(X, k) \leq t \mid \tilde{Y} = l\right]. \tag{A11}$$

Above, $F_l^k(t)$ is the CDF of $\hat{s}(X, k)$, based on a fixed function $\hat{s}$ applied to a random X from the distribution of $X \mid Y = l$, while $\tilde{F}_l^k(t)$ is the corresponding CDF of $\hat{s}(X, k)$, with X conditioned on $\tilde{Y} = l$. Further, for any $k \in [K]$, define also $\tilde{\rho}_k := \mathbb{P}[Y = k]$ and $\tilde{\rho}_k := \mathbb{P}[Y = k]$, the marginal frequencies of the l-th class in the contaminated and clean labels, respectively.

By definition of the non-conformity score function in (2), $Y_{n+1} \in \hat{C}(X_{n+1})$ if and only if $\hat{s}(X_{n+1}, Y_{n+1}) \leq \hat{\tau}$. Therefore,

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right]$$
$$= \mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1})\right] + \left(\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] - \mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1})\right]\right).$$

The proof is completed by noting that the second term on the right-hand-side above is:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] - \mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1})\right]$$

$$= \sum_{k=1}^{K} \left(\rho_k \cdot \mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid Y_{n+1} = k\right] - \tilde{\rho}_k \cdot \mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1}) \mid \tilde{Y}_{n+1} = k\right]\right)$$

$$= \sum_{k=1}^{K} \left(\rho_k \cdot \mathbb{P}\left[\hat{s}(X_{n+1}, k) \leq \hat{\tau} \mid Y_{n+1} = k\right] - \tilde{\rho}_k \cdot \mathbb{P}\left[\hat{s}(X_{n+1}, k) \leq \hat{\tau} \mid \tilde{Y}_{n+1} = k\right]\right)$$

$$= \sum_{k=1}^{K} \left(\rho_k \cdot \mathbb{E}\left[\mathbb{P}\left[\hat{s}(X_{n+1}, k) \leq \hat{\tau} \mid Y_{n+1} = k, \mathcal{D}\right]\right]\right.$$
$$\left. - \tilde{\rho}_k \cdot \mathbb{E}\left[\mathbb{P}\left[\hat{s}(X_{n+1}, k) \leq \hat{\tau} \mid \tilde{Y}_{n+1} = k, \mathcal{D}\right]\right]\right)$$

$$= \mathbb{E}\left[\sum_{k=1}^{K} \left[\rho_k F_k^k(\hat{\tau}) - \tilde{\rho}_k \tilde{F}_k^k(\hat{\tau})\right]\right]$$

$$= \mathbb{E}\left[F(\hat{\tau}) - \tilde{F}(\hat{\tau})\right] = \mathbb{E}\left[\Delta(\hat{\tau})\right].$$

The expected value above is taken with respect to the randomness of the data in $\mathcal{D}$, hence including both the training data in $\mathcal{D}^{\text{train}}$ and the calibration data in $\mathcal{D}$. Equations (4) and (5) follow directly because

$$1 - \alpha \leq \mathbb{P}\left[\tilde{Y}_{n+1} \in \hat{C}(X_{n+1})\right] \leq 1 - \alpha + \frac{1}{n+1}$$

by Proposition A1, which applies here since $(X_i, \tilde{Y}_i)$ are i.i.d. random samples. □

Proof of Corollary A1. By Theorem 1, it suffices to prove $\mathbb{E}\left[\Delta(t)\right] \geq 0$ for all $t \in [0, 1]$.

To establish that, first define $M_{kl} := \mathbb{P}\left[Y = l \mid \tilde{Y} = k\right]$. Under Assumption 1, Proposition 1 in Sesia et al. (2024) connects the distribution of $X \mid \tilde{Y} = k$, namely $\tilde{P}_k$, and the distributions of $X \mid Y = l$, namely P_l, for all $k, l \in [K]$. That is,

$$\tilde{P}_k = \sum_{l=1}^{K} M_{kl} P_l. \tag{A12}$$

Note that combining (3) with (A12) gives

$$
\begin{aligned}
\Delta(t) &= F(t) - \tilde{F}(t) \\
&= \sum_{k=1}^{K} \left[\rho_k F_k^k(t) - \tilde{\rho}_k \tilde{F}_k^k(t) \right] \\
&= \sum_{k=1}^{K} \left[\rho_k F_k^k(t) - \sum_{l=1}^{K} M_{kl} \tilde{\rho}_k F_l^k(t) \right].
\end{aligned}
$$

By combining Bayes' theorem with Assumption 1, one gets $T_{kl} = M_{kl} \cdot \tilde{\rho}_k / \rho_l$. Then,

$$
\begin{aligned}
\Delta(t) &= \sum_{l=1}^{K} \rho_l F_l^l(t) - \sum_{l=1}^{K} \rho_l \sum_{k=1}^{K} T_{kl} F_l^k(t) \\
&= \sum_{l=1}^{K} \rho_l F_l^l(t) - \sum_{l=1}^{K} \rho_l \left[T_{ll} F_l^l(t) + \sum_{k \neq l} T_{kl} F_l^k(t) \right] \\
&\geq \sum_{l=1}^{K} \rho_l F_l^l(t) - \sum_{l=1}^{K} \rho_l \left[T_{ll} F_l^l(t) + \left(\sum_{k \neq l} T_{kl} \right) \max_{k \neq l} F_l^k(t) \right] \\
&= \sum_{l=1}^{K} \rho_l F_l^l(t) - \sum_{l=1}^{K} \rho_l \left[T_{ll} F_l^l(t) + \left(\sum_{k=1}^{K} T_{kl} \right) \max_{k \neq l} F_l^k(t) - T_{ll} \max_{k \neq l} F_l^k(t) \right] \\
&= \sum_{l=1}^{K} \rho_l F_l^l(t) - \sum_{l=1}^{K} \rho_l \left[T_{ll} F_l^l(t) + \max_{k \neq l} F_l^k(t) - T_{ll} \max_{k \neq l} F_l^k(t) \right] \\
&= \sum_{l=1}^{K} \rho_l (1 - T_{ll}) \left[F_l^l(t) - \max_{k \neq l} F_l^k(t) \right] \geq 0,
\end{aligned}
$$

where the last inequality is given by (A3). This implies that $\Delta(\hat{\tau}) \geq 0$ almost-surely. $\square$

C.2   Lower Bound on Coverage of Adaptive Prediction Sets

Proof of Theorem 2. The event $Y_{n+1} \in \hat{C}(X_{n+1})$ occurs if and only if $\hat{s}(X_{n+1}, Y_{n+1}) \leq \hat{\tau}$. Therefore, the probability of a miscoverage event conditional on the data in $\mathcal{D}$ can be decomposed as:

$$
\begin{aligned}
\mathbb{P}&\left[Y_{n+1} \notin \hat{C}(X_{n+1}) \mid \mathcal{D}\right] \\
&= \mathbb{P}\left[\tilde{Y}_{n+1} \notin \hat{C}(X_{n+1}) \mid \mathcal{D}\right] \\
&\quad + \left(\mathbb{P}\left[Y_{n+1} \notin \hat{C}(X_{n+1}) \mid \mathcal{D}\right] - \mathbb{P}\left[\tilde{Y}_{n+1} \notin \hat{C}(X_{n+1}) \mid \mathcal{D}\right]\right) \\
&= 1 - \tilde{F}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})}) \\
&= 1 - \hat{F}(S_{(\hat{i})}) - \hat{\Delta}(S_{(\hat{i})}) \\
&\quad + \left[\hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})})\right] + \left[\hat{\Delta}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})})\right] \\
&= \left\{1 - \frac{\hat{i}}{n} - \hat{\Delta}(S_{(\hat{i})}) + \delta(n)\right\} - \delta(n) + \sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left[\hat{\rho}_l \hat{F}_l^k(S_{(\hat{i})}) - \tilde{\rho}_l \tilde{F}_l^k(S_{(\hat{i})})\right] \\
&\leq \sup_{i \in \hat{\mathcal{I}}}\left\{1 - \frac{i}{n} - \hat{\Delta}(S_{(i)}) + \delta(n)\right\} - \delta(n) \\
&\quad + \sup_{t \in [0,1]}\left\{\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t)\right]\right\}.
\end{aligned}
$$

By definition of $\hat{\mathcal{I}}$, for all $i \in \hat{\mathcal{I}}$, $1 - i/n - \hat{\Delta}(S_{(i)}) + \delta(n) \leq \alpha$, which implies a.s.

$$
\sup_{i \in \hat{\mathcal{I}}}\left\{1 - \frac{i}{n} - \hat{\Delta}(S_{(i)}) + \delta(n)\right\} \leq \alpha.
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\left[Y_{n+1} \notin \hat{C}(X_{n+1})\right] &\leq \alpha + \mathbb{E}\left[\sup_{t \in [0,1]}\left\{\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t)\right]\right\}\right] - \delta(n) \\
&= \alpha + \delta^*(n) - \delta(n) \\
&\leq \alpha,
\end{aligned}
$$

where the last inequality above follows directly from the assumption that $\delta(n) \geq \delta^*(n)$. $\qquad\square$

Proof of Proposition A2. Proceeding as in the proof of Theorem 2, we obtain:

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}(X_{n+1}) \mid \mathcal{D}\right]$$

$$= 1 - \tilde{F}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})})$$

$$= 1 - \hat{F}(S_{(\hat{i})}) - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}$$

$$+ \hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})}) + \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}$$

$$= \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right]$$

$$+ \hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})}) + \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})}) - \delta(n), -\left(\Delta(S_{(\hat{i})}) + \frac{1-\alpha}{n}\right)\right\}$$

$$= \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right]$$

$$+ \hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})}) - \delta(n)$$

$$+ \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})}), -\left(\Delta(S_{(\hat{i})}) - \delta(n) + \frac{1-\alpha}{n}\right)\right\}$$

$$\leq \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right]$$

$$+ \hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})}) - \delta(n) + \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})}), 0\right\}.$$

The inequality follows from the assumption that

$$\inf_{t\in[0,1]}\Delta(t) \geq \delta(n) - (1-\alpha)/n. \tag{A13}$$

Then, we have found that

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}(X_{n+1}) \mid \mathcal{D}\right]$$

$$\leq \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right] - \delta(n)$$

$$+ \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})}) + \hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})}), \hat{F}(S_{(\hat{i})}) - \tilde{F}(S_{(\hat{i})})\right\}$$

$$\leq \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right] - \delta(n)$$

$$+ \sup_{t\in[0,1]} \max\left\{\hat{\psi}(t), \hat{F}(t) - \tilde{F}(t)\right\}$$

$$\leq \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right] - \delta(n)$$

$$+ \max\left\{\sup_{t\in[0,1]} \hat{\psi}(t), \sup_{t\in[0,1]}\left(\hat{F}(t) - \tilde{F}(t)\right)\right\}.$$

By taking the expected value, we get

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}(X_{n+1})\right] \le \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right] - \delta(n)$$

$$+ \max\left\{\mathbb{E}\left[\sup_{t \in [0,1]} \hat{\psi}(t)\right], \mathbb{E}\left[\sup_{t \in [0,1]} \left(\hat{F}(t) - \tilde{F}(t)\right)\right]\right\}$$

$$\le \left[1 - \frac{\hat{i}}{n} - \max\left\{\hat{\Delta}(S_{(\hat{i})}) - \delta(n), -\frac{1-\alpha}{n}\right\}\right] - \delta(n)$$

$$+ \max\left\{\delta^*(n), c(n)\right\}.$$

Then the proof is concluded by following the same steps of Theorem 2.

$\square$

C.3 Finite-Sample Correction Factor

Proof of Theorem 3. Consider $\delta^*(n)$ defined in (13). We want to show that

$$c(n) \ge \mathbb{E}\left[\sup_{t \in [0,1]} \hat{\psi}(t)\right],$$

in the case of the randomized response model.

To start, it is helpful to parametrize W in terms of coefficients $\beta_0, \beta_1, \ldots, \beta_K$; that is

$$W = T^{-1} = \frac{1}{1-\epsilon}I - \frac{\epsilon}{K(1-\epsilon)}J = \beta_0 I + \frac{\beta_k}{K}J,$$

where we define $\beta_0 := -\frac{1}{1-\epsilon}$ and $\beta_k := \frac{\epsilon}{1-\epsilon}$ for all $k \in [K]$.

Note that

$$\hat{\psi}(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}W_{kl}\left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l\right] - \tilde{\rho}_l \tilde{F}_l^k(t)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}(\beta_0 \mathbb{I}[k = l] + \beta_k/K)\left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l\right] - \tilde{\rho}_l \tilde{F}_l^k(t)\right]$$

$$= \beta_0 \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = k\right] - \tilde{\rho}_k \tilde{F}_k^k(t)\right]$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(\beta_k/K)\sum_{l=1}^{K}\left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l\right] - \tilde{\rho}_l \tilde{F}_l^k(t)\right]$$

$$= \beta_0 \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{I}\left[\hat{s}(X_i, \tilde{Y}_i) \le t\right] - \tilde{F}(t)\right]$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(\beta_k/K)\left[\mathbb{I}\left[\hat{s}(X_i, k) \le t\right] - \tilde{F}^k(t)\right]$$

$$= \hat{\psi}_1(t) + \hat{\psi}_2(t),$$

where

$$\hat{\psi}_1(t) := \beta_0 \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{I}\left[\hat{s}(X_i, \tilde{Y}_i) \le t \right] - \tilde{F}(t) \right],$$

$$\hat{\psi}_2(t) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (\beta_k/K) \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t \right] - \tilde{F}^k(t) \right],$$

$$\tilde{F}^k(t) := \mathbb{P}\left[\hat{s}(X_i, k) \le t \right],$$

$$\tilde{F}(t) := \mathbb{P}\left[\hat{s}(X_i, \tilde{Y}_i) \le t \right].$$

We now aim to bound the expected value of the supremum of $\hat{\psi}$, and we proceed by bounding separately the expected value of the suprema of $\hat{\psi}_1(t)$ and $\hat{\psi}_2(t)$.

We first start with $\hat{\psi}_1(t)$.
Let $U_1, \ldots, U_n$ be i.i.d uniform random variables on $[0,1]$, and denote their order statistics as $U_{(1)}, \ldots, U_{(n)}$. Then,

$$\sup_{t \in [0,1]} \left(\hat{F}(t) - \tilde{F}(t) \right) \overset{d}{=} \sup_{i \in [n]} \left\{ \frac{i}{n} - U_{(i)} \right\}.$$

Following from (A5), this implies that

$$\mathbb{E}\left[\sup_{t \in [0,1]} \left(\hat{F}(t) - \tilde{F}(t) \right) \right] = \mathbb{E}\left[\sup_{i \in [n]} \left\{ \frac{i}{n} - U_{(i)} \right\} \right] = c(n).$$

Therefore,

$$\mathbb{E}\left[\sup_{t \in [0,1]} \hat{\psi}_1(t) \right] = \beta_0 c(n).$$

Similarly, we can deal with $\hat{\psi}_2(t)$:

$$\mathbb{E}\left[\sup_{t \in [0,1]} \hat{\psi}_2(t) \right] \le \sum_{k=1}^{K} \frac{\beta_k}{K} \mathbb{E}\left[\sup_{t \in [0,1]} \left(\hat{F}^k(t) - \tilde{F}^k(t) \right) \right]$$

$$\le \frac{\sum_{k=1}^{K} \beta_k}{K} c(n).$$

The result follows by noting that $\beta_0 + \frac{1}{K} \sum_{k=1}^{K} \beta_k = 1$. $\qquad\square$

Proof of Theorem 4. Let us define the empirical process $\hat{\psi}(t)$ as follows:

$$\hat{\psi}(t) := \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right].$$

Our goal is equivalent to that of proving

$$\mathbb{E}\left[\sup_{t \in [0,1]} \hat{\psi}(t) \right] \le \delta^{\mathrm{FS}}(n).$$

40

Recall that the matrix $\bar{W}$ is defined such that $\bar{W}_{kl} = \beta_0 \mathbb{I}\,[k = l] + \beta_k/K$, for some $\beta_0, (\beta_k)_k$. The intuition behind the proof is as follows. We will derive a bound as a function of the parameters β; then, the result can be obtained by taking the minimum over β.

Note that

$$\hat{\psi}(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(W_{kl} - \bar{W}_{kl} \right) \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(\beta_0 \mathbb{I}\,[k = l] + \beta_k/K \right) \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(W_{kl} - \bar{W}_{kl} \right) \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$+ \beta_0 \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = k \right] - \tilde{\rho}_k \tilde{F}_k^k(t) \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (\beta_k/K) \sum_{l=1}^{K} \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(W_{kl} - \bar{W}_{kl} \right) \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$+ \beta_0 \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{I}\left[\hat{s}(X_i, \tilde{Y}_i) \le t \right] - \tilde{F}(t) \right]$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (\beta_k/K) \left[\mathbb{I}\,[\hat{s}(X_i, k) \le t] - \tilde{F}^k(t) \right]$$

$$= \hat{\psi}_1(t) + \hat{\psi}_2(t) + \hat{\psi}_3(t),$$

where

$$\hat{\psi}_1(t) := \beta_0 \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{I}\left[\hat{s}(X_i, \tilde{Y}_i) \le t \right] - \tilde{F}(t) \right],$$

$$\hat{\psi}_2(t) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (\beta_k/K) \left[\mathbb{I}\,[\hat{s}(X_i, k) \le t] - \tilde{F}^k(t) \right]$$

$$\hat{\psi}_3(t) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} (W_{kl} - \bar{W}_{kl}) \left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right],$$

$$\tilde{F}^k(t) := \mathbb{P}\left[\hat{s}(X_i, k) \le t \right],$$

$$\tilde{F}(t) := \mathbb{P}\left[\hat{s}(X_i, \tilde{Y}_i) \le t \right].$$

41

We now bound the suprema of these three processes separately, since

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}(t)\right] \leq \mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_1(t)\right] + \mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_2(t)\right] + \mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_3(t)\right].$$

For $\hat{\psi}_1(t)$ and $\hat{\psi}_2(t)$ we proceed as in the proof of Theorem 3. Let's start from $\hat{\psi}_1(t)$. First, note that

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_1(t)\right] = \beta_0 \mathbb{E}\left[\sup_{t\in[0,1]} \left(\hat{F}(t) - \tilde{F}(t)\right)\right].$$

Let $U_1, \ldots, U_n$ be i.i.d uniform random variables on $[0,1]$, and denote their order statistics as $U_{(1)}, \ldots, U_{(n)}$. Then,

$$\sup_{t\in[0,1]} \left(\hat{F}(t) - \tilde{F}(t)\right) \overset{d}{=} \sup_{i\in[n]} \left\{\frac{i}{n} - U_{(i)}\right\}.$$

Following from (A5), this implies that

$$\mathbb{E}\left[\sup_{t\in[0,1]} \left(\hat{F}(t) - \tilde{F}(t)\right)\right] = \mathbb{E}\left[\sup_{i\in[n]} \left\{\frac{i}{n} - U_{(i)}\right\}\right] = c(n).$$

Therefore,

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_1(t)\right] = \beta_0 c(n).$$

Similarly, we can deal with $\hat{\psi}_2(t)$:

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_2(t)\right] = \mathbb{E}\left[\sup_{t\in[0,1]} \sum_{k=1}^{K} \frac{\beta_k}{K} \left(\hat{F}^k(t) - \tilde{F}^k(t)\right)\right]$$

$$\leq \sum_{k=1}^{K} \frac{\beta_k}{K} \mathbb{E}\left[\sup_{t\in[0,1]} \left(\hat{F}^k(t) - \tilde{F}^k(t)\right)\right]$$

$$\leq \frac{\sum_{k=1}^{K} \beta_k}{K} c(n).$$

What remains to be bound is $\hat{\psi}_3(t)$. For simplicity, define $\Omega_{kl} := W_{kl} - \bar{W}_{kl}$, so that

$$\hat{\psi}_3(t) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \Omega_{kl} \left[\mathbb{I}\left[\hat{s}(X_i, k) \leq t, \tilde{Y}_i = l\right] - \tilde{\rho}_l \tilde{F}_l^k(t)\right].$$

The bound on $\hat{\psi}_3(t)$ is found following two alternative approaches leading to two different bounds, which we combine by taking the minimum of the two. Both approaches focus on bounding the supremum of $|\hat{\psi}_3(t)|$. They both rely on symmetrization and on bounding of the Rademacher dimension of a specific set of functions, and differ in the way the bound on the Rademacher dimension is found. One approach employs Massart's inequality, the other employs the chaining technique. The result is gathered in Lemma A1 and reads

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}_3(t)\right] \leq \mathbb{E}\left[\sup_{t\in[0,1]} \left|\hat{\psi}_3(t)\right|\right] \leq \frac{1}{\sqrt{n}} B(K, n, \beta),$$

42

where $B(K, n, \beta)$ is defined in (15).

Putting everything together, we find that, for any $\beta \in [0,1]^{K+1}$,

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}(t)\right] \leq c(n)\left(\beta_0 + \frac{\sum_{k=1}^K \beta_k}{K}\right) + \frac{1}{\sqrt{n}}B(K,n,\beta).$$

Therefore,

$$\mathbb{E}\left[\sup_{t\in[0,1]} \hat{\psi}(t)\right] \leq \inf_\beta \left\{c(n)\left(\beta_0 + \frac{\sum_{k=1}^K \beta_k}{K}\right) + \frac{1}{\sqrt{n}}B(K,n,\beta)\right\}$$
$$=: \delta^{\mathrm{FS}}(n).$$

$\square$

Lemma A1. *Let $(X_i, \tilde{Y}_i)_{i=1}^n \sim P$ be an independent and identically distributed random sample. For all i, let $A_i(t) = \sum_{k=1}^K \sum_{l=1}^K \Omega_{kl} \mathbb{I}\left[s(X_i, k) \leq t, \tilde{Y}_i = l\right]$. Then,*

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n A_i(t) - \mathbb{E}\left[A(t)\right]\right|\right] \leq \frac{2}{\sqrt{n}}\min\left\{\max_{l\in[K]}\sum_k |\Omega_{kl}|\sqrt{\log(Kn+1)}, \right.$$
$$\left. 24\max_{k,l\in[K]}|\Omega_{kl}|\frac{2\log K + 1}{2\log K - 1}\sqrt{2K\log K}\right\}. \quad \text{(A14)}$$

Proof of Lemma A1. For ease of notation, let $Z_i = (X_i, \tilde{Y}_i)$ and $Z_1^n = (Z_1, \ldots, Z_n)$. In the following, we denote by $\mathbb{E}\left[\cdot\right]$ the expected value of a random quantity taken with respect to the joint distribution of $Z_1, \ldots, Z_n$.

We study the quantity

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n A_i(t) - \mathbb{E}\left[A(t)\right]\right|\right] \quad \text{(A15)}$$

by using a symmetrization argument which links (A15) with the Rademacher complexity of a properly defined set of vectors. We then get the result by finding a bound for the Rademacher complexity.

First, we define $\mathcal{H}$ as the set of random functions

$$\mathcal{H} = \left\{\sum_{k=1}^K \sum_{l=1}^K \Omega_{kl} \mathbb{I}\left[\tilde{Y} = l\right] \mathbb{I}\left[\hat{s}(X, k) \leq t\right] : t \in [0,1]\right\} = \{h(Z)\}.$$

It is worth emphasizing that, even though we omit the pedix t for ease of notation, any $h \in \mathcal{H}$ is a function of the random variable Z and is indexed by t. Then, it should be clear that any $h(Z)$ is a scalar random variable.

Note that

$$\sup_{t\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n A_i(t) - \mathbb{E}\left[A(t)\right]\right| = \sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}\left[h(Z)\right]\right|. \quad \text{(A16)}$$

We denote by $\mathcal{H}(Z_1^n)$ the set of vectors in $[0,1]^n$ that can be realized by applying any function $h \in \mathcal{H}$ to the collection Z_1^n, i.e., $\{(h(Z_1), \ldots, h(Z_n))\}_{h \in \mathcal{H}}$. In addition, we denote by $\mathcal{R}(\mathcal{H}, Z_1^n)$ the Rademacher complexity of $\mathcal{H}(Z_1^n)$. That is,

$$\mathcal{R}(\mathcal{H}, Z_1^n) = \mathbb{E}_{\eta_1^n} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \eta_i h(Z_i) \right| \right],$$

where $\eta_1^n = (\eta_1, \ldots, \eta_n)$ is a sequence of independent Rademacher random variables.

Following from the symmetrization lemma (e.g., Wainwright (2019), p. 106), which we recall in Lemma A2, it holds that

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}\left[h(Z)\right] \right| \right] \leq \mathbb{E}\left[2\mathcal{R}(\mathcal{H}, Z_1^n)\right]. \tag{A17}$$

We now follow two alternative approaches to bound the Rademacher complexity of $\mathcal{H}(Z_1^n)$. In one approach, the straightforward application of Massart's lemma leads to the first bound. In the second approach, a chaining technique is employed to find a convenient covering number for $\mathcal{H}(Z_1^n)$, which is then used to bound the Rademacher complexity of $\mathcal{H}(Z_1^n)$ via Dudley's entropy integral. The final bound for (A15) is obtained by taking the minimum of the two bounds.

Bound via direct use of Massart's lemma.
First, we show that $\mathcal{H}(Z_1^n)$ is bounded and with finite cardinality.

Concerning cardinality, one easily checks that $|\mathcal{H}(Z_1^n)| \leq Kn + 1$. Indeed, it is sufficient to note that a realization Z_1^n is associated to Kn scores $\hat{s}(X_i, k)$, with $i = 1, \ldots, n$ and $k = 1, \ldots, K$. The ordered scores $\hat{S}_{(1)}, \ldots, \hat{S}_{(Kn)}$ partition the real line in $Kn + 1$ intervals. The result on cardinality follows by noting that a component of the vectors in $\mathcal{H}(Z_1^n)$ changes value whenever t shifts between two intervals.

Concerning boundedness, it suffices to observe that

$$\max_{a \in \mathcal{H}(Z_1^n)} ||a|| = \max_{j \in [Kn+1]} ||a_j|| = \max_{j \in [Kn+1]} \sqrt{\sum_{i=1}^n a_{ji}^2}$$

$$= \max_{s \in \{-\infty, \hat{S}_{(1)}, \ldots, \hat{S}_{(Kn)}\}} \sqrt{\sum_{i=1}^n \left(\sum_{k,l} \Omega_{kl} \mathbb{I}\left[\tilde{Y}_i = l\right] \mathbb{I}\left[\hat{s}(X_i, k) \leq s\right] \right)^2}$$

$$= \max_{s \in \{-\infty, \hat{S}_{(1)}, \ldots, \hat{S}_{(Kn)}\}} \sqrt{\sum_{i=1}^n \left(\sum_{k} \Omega_{k\tilde{Y}_i} \mathbb{I}\left[\hat{s}(X_i, k) \leq s\right] \right)^2}$$

$$\leq \sqrt{\sum_{i=1}^n \max_{l \in [K]} \sum_{k} (\Omega_{kl})^2} = \sqrt{n} \max_{l \in [K]} \sum_{k} |\Omega_{kl}|.$$

By applying Massart's lemma (Massart (2000), Lemma 5.2), we find:

$$\mathcal{R}(\mathcal{H}, Z_1^n)) \leq \frac{1}{n} \sqrt{n} \max_{l \in [K]} \sum_k |\Omega_{kl}| \sqrt{2 \log(Kn + 1)}$$

$$= \max_{l \in [K]} \sum_k |\Omega_{kl}| \sqrt{\frac{2 \log(Kn + 1)}{n}}. \tag{A18}$$

By combining (A18) with (A17) and (A16), one immediately has that

$$\mathbb{E} \left[\sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n A_i(t) - \mathbb{E}[A(t)] \right| \right] \leq 2 \max_{l \in [K]} \sum_k |\Omega_{kl}| \sqrt{\frac{2 \log(Kn + 1)}{n}}. \tag{A19}$$

Finding an ϵ-covering number for $\mathcal{H}$.
First, note that $h \in \mathcal{H}$ may be expressed as

$$h(Z) = \sum_{k=1}^K \Omega_{k\tilde{Y}} \mathbb{I}\left[\hat{s}(X, k) \leq t\right].$$

In the following, let $h(Z_1^n)$ denote the vector $(h(Z_1), \ldots, h(Z_n))$ in $\mathbb{R}^n$. Additionally, let us denote $L_2(P)$-norm the norm $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n h(Z_i)^2\right]$ where the expectation is taken with respect to the joint distribution of $Z_1, \ldots, Z_n$. By definition of ϵ-cover, an ϵ-cover of $\mathcal{H}$ with respect to the $L_2(P)$-norm is a set of functions $\mathcal{G}$ such that for all $h \in \mathcal{H}$ there exists $g \in \mathcal{G} : \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (h(Z_i) - g(Z_i))^2\right] \leq \epsilon^2$. Recall that the ϵ-covering number of $\mathcal{H}$ is defined as the cardinality of its smallest ϵ-cover and is denoted as $N(\epsilon, \mathcal{H}, L_2(P))$ (e.g., Wainwright (2019), Definition 5.1).

For a fixed ϵ, consider the set of functions

$$\mathcal{G}(\epsilon) = \left\{ g(Z) = \sum_{k=1}^K \Omega_{k\tilde{Y}} \mathbb{I}\left[\hat{s}(X, k) \leq a_{kj_k}\right] : j_k = 1, \ldots, b, \ b := \frac{K^2}{\epsilon^2} \right\}.$$

Recall that $\tilde{F}^k(t) := \mathbb{P}\left[\hat{s}(X_i, k) \leq t\right]$. For each $k \in [K]$, the parameters $\{a_{kj_k}\}_{j_k}$ are set such that $a_{kj_k} = (\tilde{F}^k)^{-1}\left(\frac{j_k}{b}\right)$. This implies partitioning the $[0, 1]$ interval in b sub-intervals.

In other words, each function g in $\mathcal{G}(\epsilon)$ is identified by a K-dimensional vector $a = (a_{1j_1}, \ldots, a_{Kj_K})$. As each component of a can take b values, $|\mathcal{G}(\epsilon)| = b^K = \left(\frac{K^2}{\epsilon^2}\right)^K$.

Now, we show that $\mathcal{G}(\epsilon)$ is a η-cover of $\mathcal{H}$, with $\eta = \epsilon \frac{\max_{kl} |\Omega_{kl}|}{\sqrt{2}}$. Consider any function $h \in \mathcal{H}$.

Then,

$$\frac{1}{n}\sum_{i=1}^{n}\left(h(Z_i)-g(Z_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\Omega_{k\tilde{Y}_i}\left(\mathbb{I}\left[\hat{s}(X_i,k)\leq t\right]-\mathbb{I}\left[\hat{s}(X_i,k)\leq a_{kj_k}\right]\right)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{k,k'=1}^{K}\Omega_{k\tilde{Y}_i}\Omega_{k'\tilde{Y}_i}\left(\mathbb{I}\left[\hat{s}(X_i,k)\leq t\right]-\mathbb{I}\left[\hat{s}(X_i,k)\leq a_{kj_k}\right]\right)$$

$$\cdot\left(\mathbb{I}\left[\hat{s}(X_i,k')\leq t\right]-\mathbb{I}\left[\hat{s}(X_i,k')\leq a_{kj_k}\right]\right)$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{k,k'=1}^{K}\max_{k,l\in[K]}(\Omega_{kl})^2\left(\mathbb{I}\left[\hat{s}(X_i,k)\leq t\right]-\mathbb{I}\left[\hat{s}(X_i,k)\leq a_{kj_k}\right]\right)$$

$$\leq \frac{1}{n}\max_{k,l\in[K]}(\Omega_{kl})^2\sum_{k,k'=1}^{K}\sum_{i=1}^{n}\left(\mathbb{I}\left[\hat{s}(X_i,k)\leq t\right]-\mathbb{I}\left[\hat{s}(X_i,k)\leq a_{kj_k}\right]\right).$$

Then,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(h(Z_i)-g(Z_i)\right)^2\right] \leq \frac{1}{n}\max_{k,l\in[K]}(\Omega_{kl})^2\sum_{k,k'=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbb{I}\left[\hat{s}(X_i,k)\leq t\right]-\mathbb{I}\left[\hat{s}(X_i,k)\leq a_{kj_k}\right]\right)\right]$$

$$\leq \frac{1}{n}\max_{k,l\in[K]}(\Omega_{kl})^2\sum_{k,k'=1}^{K}\sum_{i=1}^{n}\left(\tilde{F}^k(t)-\tilde{F}^k(a_{kj_k})\right)$$

$$\leq \frac{1}{n}\max_{k,l\in[K]}(\Omega_{kl})^2\sum_{k,k'=1}^{K}\sum_{i=1}^{n}\frac{1}{2b} = \frac{1}{n}\max_{k,l\in[K]}(\Omega_{kl})^2\frac{K^2n}{2b}$$

$$= \frac{1}{2}\max_{k,l\in[K]}(\Omega_{kl})^2\epsilon^2 = \eta^2.$$

The third inequality descends directly from how the parameters a_{kj_k} are selected. In fact, note that for each $t\in[0,1]$ (i.e., for each $h\in\mathcal{H}$) there exists a vector $a=(a_{1j_1},\ldots,a_{Kj_K})$ (i.e., a function $g\in\mathcal{G}$) such that $\tilde{F}^k(a_{kj_k})-\tilde{F}^k(t)\leq\frac{1}{2b}$, for all $k\in[K]$. Then $\mathcal{G}(\epsilon)$ is a η-cover of $\mathcal{H}$. The cardinality of the cover is $|\mathcal{G}(\epsilon)|=\left(\frac{K}{\epsilon}\right)^{2K}$.

Then, by Dudley's entropy integral (e.g., Wainwright (2019), Eq. (5.48)), we have

$$\mathbb{E}\left[\mathcal{R}(\mathcal{H},Z_1^n)\right] \leq \frac{24}{\sqrt{n}}\int_0^\infty\sqrt{\log N(\eta,\mathcal{H},L_2(P))}d\eta$$

$$\leq \frac{24}{\sqrt{n}}\int_0^\infty\sqrt{\log(|\mathcal{G}(\epsilon)|)}d\eta$$

$$\leq \frac{24}{\sqrt{n}}\max_{k,l\in[K]}|\Omega_{kl}|\int_0^1\sqrt{2K\log\left(\frac{K}{\epsilon}\right)}d\epsilon.$$

By solving the integral, one finds that

$$\mathbb{E}\left[\mathcal{R}(\mathcal{H},Z_1^n)\right] < 24\max_{k,l\in[K]}|\Omega_{kl}|\frac{2\log K+1}{2\log K-1}\sqrt{\frac{2K\log K}{n}}, \tag{A20}$$

46

from which it directly follows that

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}A_i(t)-\mathbb{E}\left[A(t)\right]\right|\right] < 48\max_{k,l\in[K]}|\Omega_{kl}|\frac{2\log K+1}{2\log K-1}\sqrt{\frac{2K\log K}{n}}. \qquad (A21)$$

For the sake of completeness, we here report the calculations that lead to (A20).

$$\begin{aligned}
\int_0^1\sqrt{2K\log\left(\frac{K}{\epsilon}\right)}d\epsilon &= \sqrt{2K}\int_{-\infty}^0\sqrt{\log(K)-u}\ e^u du\\
&= \sqrt{2K}\int_{-\infty}^0\sqrt{\log(K)-u}\ e^u du\\
&= \sqrt{2K}\int_{\log K}^{\infty}\sqrt{v}e^{\log K-v}dv\\
&= \sqrt{2}K^{3/2}\int_{\log K}^{\infty}\sqrt{v}e^{-v}dv\\
&= \sqrt{2}K^{3/2}\Gamma\left(\frac{3}{2},\log K\right)\\
&< \sqrt{2}K^{3/2}\frac{2\log K+1}{2\log K-1}\sqrt{\log K}\frac{1}{K}\\
&= \frac{2\log K+1}{2\log K-1}\sqrt{2K\log K}.
\end{aligned}$$

The last inequality follows from a known upper bound for the incomplete Gamma function (Natalini and Palumbo, 2000). $\qquad\square$

Lemma A2. *Let $Z_1^n = (Z_1,\ldots,Z_n)$ be a collection of independent random samples. Then,*

$$\mathbb{E}_{Z_1^n}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i)-\mathbb{E}\left[h(Z)\right]\right|\right] \leq \mathbb{E}_{Z_1^n}\left[2\mathcal{R}(\mathcal{H},Z_1^n)\right],$$

where $\mathcal{H}$ is the set defined in the proof of Lemma A1 and $\mathcal{R}(\mathcal{H},Z_1^n)$ denotes the Rademacher complexity of $\mathcal{H}(Z_1^n)$.

Proof of Lemma A2. Let $Z_1',\ldots,Z_n'$ be distributed as $Z_1,\ldots,Z_n$ but independent of $Z_1,\ldots,Z_n$. Let $\eta_1^n = (\eta_1,\ldots,\eta_n)$ be an i.i.d sequence of Rademacher variables, independent of $Z_1,\ldots,Z_n$ and $Z_1',\ldots,Z_n'$. This implies that the variables $h(Z_i)-h(Z_i')$ are independent, symmetric

and distributed as $\eta_i(h(Z_i) - h(Z_i'))$. Then,

$$
\mathbb{E}_{Z_1^n}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}_Z[h(Z)]\right|\right] = \mathbb{E}_{Z_1^n}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}_{Z_1}[h(Z')]\right|\right]
$$

$$
\leq \mathbb{E}_{Z_1^n}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n (h(Z_i) - h(Z_i'))\right|\right]
$$

$$
= \mathbb{E}_{Z_1^n,\eta_1^n}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \eta_i\left(h(Z_i) - h(Z_i')\right)\right|\right]
$$

$$
\leq 2\mathbb{E}_{Z_1^n,\eta_1^n}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \eta_i h(Z_i)\right|\right]
$$

$$
= \mathbb{E}_{Z_1^n}\left[2\mathbb{E}_{\eta_1^n}\left[\left|\frac{1}{n}\sum_{i=1}^n \eta_i h(Z_i)\right|\right]\right]
$$

$$
= \mathbb{E}_{Z_1^n}\left[2\mathcal{R}(\mathcal{H}, Z_1^n)\right].
$$

Above, the first inequality follows from the application of Jensen's inequality. $\qquad\square$

Lemma A3. *Let $U_1, \ldots, U_n$ be i.i.d uniform random variables on $[0,1]$, and denote their order statistics as $U_{(1)}, \ldots, U_{(n)}$. Then*

$$
\mathbb{E}\left[\sup_{i\in[n]}\left\{\frac{i}{n} - U_{(i)}\right\}\right] \leq \sqrt{\frac{\pi}{2n}}. \tag{A22}
$$

Proof of Lemma A3. Let $\hat{F}(t)$ denote the empirical distribution function based on $U_1, \ldots, U_n$. Then,

$$
\frac{i}{n} - U_{(i)} \overset{d}{=} \hat{F}(t) - t.
$$

This implies that

$$
\mathbb{E}\left[\sup_{i\in[n]}\left\{\frac{i}{n} - U_{(i)}\right\}\right] = \mathbb{E}\left[\sup_{t\in[0,1]}\left(\hat{F}(t) - t\right)\right]
$$

$$
\leq \mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{F}(t) - t\right|\right]
$$

$$
= \int_0^\infty \mathbb{P}\left[\sup_{t\in[0,1]}\left|\hat{F}(t) - t\right| > z\right] dz
$$

$$
\leq 2\int_0^\infty e^{-2nz^2} dz
$$

$$
= \sqrt{\frac{\pi}{2n}}.
$$

Above, the second inequality follows from a direct application of the Dvoretzky-Kiefer-Wolfowitz inequality. $\qquad\square$

C.4 Asymptotic Approximation of the Correction Factor

Proof of Theorem 5. Consider the definition of the empirical process $\hat{\psi}(t)$:

$$\hat{\psi}(t) = \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left[\mathbb{I} \left[\hat{s}(X_i, k) \leq t, \tilde{Y}_i = l \right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right].$$

Note that $\hat{\psi}(t)$ is an empirical process with mean 0. The convergence of $\hat{\psi}$ to a Gaussian limit can be established via Donsker's theorem; e.g., see Van der Vaart (2000, Ch. 19).

Let $\mathcal{F}$ be the following class of functions of $Z = (X, \tilde{Y})$, indexed by $t \in [0, 1]$,

$$\mathcal{F} = \left\{ \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \mathbb{I} \left[\hat{s}(X, k) \leq t, \tilde{Y} = l \right] : t \in [0, 1] \right\}. \tag{A23}$$

Let f_t be the element of $\mathcal{F}$ corresponding to t, and note that it may be expressed as

$$f_t(Z) = \sum_{k=1}^{K} W_{k\tilde{Y}} \mathbb{I} \left[\hat{s}(X, k) \leq t \right].$$

Now let $Z_i = (X_i, \tilde{Y}_i)$, $i = 1, \ldots, n$, be independent and identically distributed realizations from some unknown distribution P. Now denote

$$\mathbb{P}_n f_t = \frac{1}{n} \sum_{i=1}^{n} f_t(Z_i), \qquad\qquad P f_t = \mathbb{E} \left[\mathbb{P}_n f_t \right].$$

Note that $\hat{\psi}(t) = \mathbb{P}_n f_t - P f_t$.

Let us denote the $L_2(P)$ be the set of measurable functions whose second powers are P-integrable, i.e., $\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} g(Z_i)^2 \right] < \infty$. Given two functions l and u, the bracket $[l, u]$ is the set of all functions g with $l \leq g \leq u$. An ϵ-bracket in $L_2(P)$ is a bracket $[l, u]$ with

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} (u(Z_i) - l(Z_i))^2 \right] < \epsilon^2.$$

The bracketing number $N(\epsilon, \mathcal{F}, L_2(P))$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$; e.g., see Van der Vaart (2000, p. 270). The bracketing integral (Van der Vaart, 2000, p. 270) is defined starting from the bracketing number as

$$J(\delta, \mathcal{F}, L_2(P)) := \int_0^{\delta} \sqrt{\log N(\epsilon, \mathcal{F}, L_2(P))} d\epsilon.$$

Proceeding in a similar way as for finding the ϵ-covering number of $\mathcal{H}$ in the proof of Lemma A1, it is possible to show that the bracketing integral of $\mathcal{F}$ is finite.

In fact, let us define the following set of functions

$$\mathcal{M}(\epsilon) = \left\{ m(Z) = \sum_{k=1}^{K} W_{k\tilde{Y}} \mathbb{I}[\hat{s}(X, k) \leq c_{kj_k}], \ j_k = 0, 1, \ldots, d, \ d := \frac{K^2}{\epsilon^2} \right\},$$

where $c_{kj_k} := \left(\tilde{F}^k \right)^{-1} \left(\frac{j_k}{d} \right)$ for all $k \in [K]$. Recall that $\left(\tilde{F}^k \right)^{-1}(0) = 0$ and $\left(\tilde{F}^k \right)^{-1}(1) = 1$ because $\tilde{F}^k$ is defined in $[0, 1]$ for all k. We now show that, for each element $f_t \in \mathcal{F}$, there exist $u_t, l_t \in \mathcal{M}(\epsilon)$ such that $[l_t, u_t]$ is an ϵ-bracket for f.

Fix any $f_t \in \mathcal{F}$, and note that

$$f_t(Z) = \sum_{k=1}^{K} W_{k\tilde{Y}} \mathbb{I}[\hat{s}(X, k) \leq t]$$

$$= \min_{k,l} W_{kl} \sum_{k=1}^{K} \mathbb{I}[\hat{s}(X, k) \leq t] + \sum_{k=1}^{K} Q_{k\tilde{Y}} \mathbb{I}[\hat{s}(X, k) \leq t],$$

where $Q_{kl} = W_{kl} - \min_{k,l} W_{kl}$. Note that $Q_{kl} \geq 0$ for all $k, l \in [K]$. Now consider the two functions $l_t, u_t \in \mathcal{M}(\epsilon)$ defined as

$$l_t(Z) = \min_{k,l} W_{kl} \sum_{k=1}^{K} \mathbb{I}[\hat{s}(X, k) \leq t] + \sum_{k=1}^{K} Q_{k\tilde{Y}} \mathbb{I}\left[\hat{s}(X, k) \leq c_{k(j_k - 1)}\right],$$

$$u_t(Z) = \min_{k,l} W_{kl} \sum_{k=1}^{K} \mathbb{I}[\hat{s}(X, k) \leq t] + \sum_{k=1}^{K} Q_{k\tilde{Y}} \mathbb{I}\left[\hat{s}(X, k) \leq c_{kj_k}\right],$$

where, for each $k \in [K]$, the index j_k is set so that $c_{k(j_k-1)} \leq t \leq c_{kj_k}$. It is trivial to see that $l_t \leq f_t \leq u_t$, which implies that $[l_t, u_t]$ is a bracket of f_t. We now show that $[l_t, u_t]$ is in fact an η-bracket in $L_2(P)$, with $\eta = \epsilon \max_{k,l} Q_{kl}$.

Indeed,

$$\frac{1}{n} \sum_{i=1}^{n} (u_t(Z_i) - l_t(Z_i))^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} Q_{k\tilde{Y}} \mathbb{I}[\hat{s}(X_i, k) \leq c_{kj_k}] - \mathbb{I}\left[\hat{s}(X_i, k) \leq c_{k(j_k-1)}\right] \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k,k'=1}^{K} Q_{k\tilde{Y}} Q_{k'\tilde{Y}} \left(\mathbb{I}[\hat{s}(X_i, k) \leq c_{kj_k}] - \mathbb{I}\left[\hat{s}(X_i, k) \leq c_{k(j_k-1)}\right] \right)$$

$$\cdot \left(\mathbb{I}\left[\hat{s}(X_i, k') \leq c_{k'j_{k'}}\right] - \mathbb{I}\left[\hat{s}(X_i, k') \leq c_{k'(j_{k'}-1)}\right] \right)$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{k,k'=1}^{K} \max_{k,l \in [K]} (Q_{kl})^2 \left(\mathbb{I}[\hat{s}(X_i, k) \leq c_{kj_k}] - \mathbb{I}\left[\hat{s}(X_i, k) \leq c_{k(j_k-1)}\right] \right)$$

$$\leq \frac{1}{n} \max_{k,l \in [K]} (Q_{kl})^2 \sum_{k,k'=1}^{K} \sum_{i=1}^{n} \left(\mathbb{I}[\hat{s}(X_i, k) \leq c_{kj_k}] - \mathbb{I}\left[\hat{s}(X_i, k) \leq c_{k(j_k-1)}\right] \right).$$

50

Then,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(u_t(Z_i) - l_t(Z_i))^2\right]$$

$$\leq \frac{1}{n}\max_{k,l\in[K]}(Q_{kl})^2 \sum_{k,k'=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\left(\mathbb{I}\left[\hat{s}(X_i,k)\leq c_{kj_k}\right] - \mathbb{I}\left[\hat{s}(X_i,k)\leq c_{k(j_k-1)}\right]\right)\right]$$

$$\leq \frac{1}{n}\max_{k,l\in[K]}(Q_{kl})^2 \sum_{k,k'=1}^{K}\sum_{i=1}^{n}\left(\tilde{F}^k(c_{kj_k}) - \tilde{F}^k(c_{k(j_k-1)})\right)$$

$$\leq \frac{1}{n}\max_{k,l\in[K]}(Q_{kl})^2 \sum_{k,k'=1}^{K}\sum_{i=1}^{n}\frac{1}{d}$$

$$= \frac{1}{n}\max_{k,l\in[K]}(Q_{kl})^2\frac{K^2 n}{d}$$

$$= \max_{k,l\in[K]}(Q_{kl})^2\epsilon^2$$

$$= \eta^2.$$

With the same argument, one may find an η-bucket in $\mathcal{M}(\epsilon)$ for any $f_t \in \mathcal{F}$. Then, $N(\eta, \mathcal{H}, L_2(P)) \leq |\mathcal{M}(\epsilon)|$. Note that each function m in $\mathcal{M}(\epsilon)$ is identified by a K-dimensional vector $c = (c_{1j_1}, \ldots, c_{Kj_K})$. As each component of c can take $d+1$ values, this implies that $|\mathcal{M}(\epsilon)| = (d+1)^K = \left(\frac{K^2}{\epsilon^2} + 1\right)^K \leq \left(\frac{K}{\epsilon} + 1\right)^{2K}$.

It is then possible to find an upper bound for the bracketing number of $\mathcal{F}$ with $\delta = 1$, which reads

$$J(1, \mathcal{F}, L_2(P)) \leq \int_0^1 \sqrt{\log\left(\frac{K}{\epsilon} + 1\right)^{2K}}d\epsilon$$

$$\leq \int_0^1 \sqrt{2K\log\left(\frac{K}{\epsilon} + 1\right)}d\epsilon$$

$$\leq \int_0^1 \sqrt{2K\log\left(\frac{K+1}{\epsilon}\right)}d\epsilon.$$

The fact that is integral is finite can be proven by repeating similar calculations as those shown at the end of the proof of Lemma A1.

As $J(1, \mathcal{F}, L_2(P))$ is finite, it follows from Donsker's theorem (e.g., Van der Vaart (2000), Ch. 19, Theorem 19.4) that $\mathcal{F}$ is a P-Donsker class of functions. Consequently, $\sqrt{n}\left(\mathbb{P}_n f_t - P f_t\right)$ converges in distribution to a centered Gaussian process with covariance function

$$G(t_1, t_2) = \mathbb{E}\left[f_{t_1} f_{t_2}\right] - \mathbb{E}\left[f_{t_1}\right]\mathbb{E}\left[f_{t_2}\right], \ \forall t_1, t_2 \in [0, 1].$$

Let GBB denote the limiting Gaussian process. Then, for n that tends to infinity and for all $t \in [0, 1]$

$$\sqrt{n}\cdot\hat{\psi}(t) \xrightarrow{d} \text{GBB}(t).$$

$\square$

C.5 Theoretical Upper Bound on Coverage

Proof of Theorem 6. Define the events $\mathcal{A}_1 = \{\hat{\mathcal{I}} = \emptyset\}$ and $\mathcal{A}_2 = \{\hat{i} = 1\}$. Then,

$$
\begin{aligned}
\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] &\leq \mathbb{E}\left[\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]\mathbb{I}\left[\mathcal{A}_1\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]\mathbb{I}\left[\mathcal{A}_2\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]\mathbb{I}\left[\mathcal{A}_1^c \cap \mathcal{A}_2^c\right]\right] \\
&= \mathbb{P}\left[\mathcal{A}_1\right] + \mathbb{E}\left[\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]\mathbb{I}\left[\mathcal{A}_2\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]\mathbb{I}\left[\mathcal{A}_1^c \cap \mathcal{A}_2^c\right]\right].
\end{aligned}
\tag{A24}
$$

We will now separately bound the three terms on the right-hand-side of (A24). The following notation will be useful for this purpose. For all $i \in [n]$, let $U_i \sim \text{Uniform}(0, 1)$ be independent and identically distributed uniform random variables, and denote their order statistics as $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$.

- We show below that the probability of the event $\mathcal{A}_1$ can be bound from above as:

$$
\mathbb{P}\left[\hat{\mathcal{I}} = \emptyset\right] \leq \frac{1}{\sqrt[4]{n}} + \frac{1}{n}.
\tag{A25}
$$

First, note that

$$
\mathbb{P}\left[\hat{\mathcal{I}} \neq \emptyset\right] \geq \mathbb{P}\left[n \in \mathcal{I}\right] = \mathbb{P}\left[1 - \alpha - \hat{\Delta}(S_{(n)}) + \delta(n) \leq 1\right].
\tag{A26}
$$

We show now that

$$
\mathbb{P}\left[1 - \alpha - \hat{\Delta}(S_{(n)}) + \delta(n) \leq 1\right] \geq 1 - \frac{1}{\sqrt[4]{n}} - \frac{1}{n}.
\tag{A27}
$$

Indeed, for any $i \in [n]$,

$$
\begin{aligned}
&1 - \alpha - \hat{\Delta}(S_{(i)}) + \delta(n) \\
&\quad = 1 - \alpha - \Delta(S_{(i)}) + \delta(n) + \Delta(S_{(i)}) - \hat{\Delta}(S_{(i)}) \\
&\quad \leq 1 - \alpha - \Delta(S_{(i)}) + \delta(n) \\
&\qquad + \sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\tilde{\rho}_l \tilde{F}_l^k(S_{(i)}) - \hat{\rho}_l \hat{F}_l^k(S_{(i)})\right) - \left(\tilde{F}(S_{(i)}) - \hat{F}(S_{(i)})\right) \\
&\quad \leq 1 - \alpha - \Delta(S_{(i)}) + \delta(n) \\
&\qquad + \sup_{t \in [0,1]}\left|\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t)\right)\right| + \sup_{t \in [0,1]}\left|\hat{F}(t) - \tilde{F}(t)\right|.
\end{aligned}
$$

By applying the Dvoretzky–Kiefer–Wolfowitz inequality on the last term above, we have that, with probability at most $1/n$,

$$\sup_{t \in [0,1]} \left| \hat{F}(t) - \tilde{F}(t) \right| > \sqrt{\frac{\log(2n)}{2n}}.$$

Further, we know from Markov's inequality that, with probability at most $1/\sqrt[4]{n}$,

$$\sup_{t \in [0,1]} \left| \sum_{k=1}^{K} \sum_{l=1}^{K} W_{kl} \left(\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t) \right) \right| > \sqrt[4]{n} \delta^{**}(n).$$

Combining these two results we have that, with probability at least $1 - \frac{1}{n} - \frac{1}{\sqrt[4]{n}}$,

$$1 - \alpha - \hat{\Delta}(S_{(i)}) + \delta(n) \leq 1 - \alpha - \Delta(S_{(i)}) + \delta(n) + \sqrt{\frac{\log(2n)}{2n}} + \sqrt[4]{n} \delta^{**}(n).$$

Now let

$$d(n) := \inf_{\beta} \left\{ \frac{1}{\sqrt[4]{n}} \sqrt{\frac{\pi}{2}} \left(|\beta_0| + \frac{\sum_{k=1}^{K} |\beta_k|}{K} \right) + \frac{1}{\sqrt[4]{n}} B(K, n, \beta) \right\}.$$

From Theorem A1 we have that

$$1 - \alpha - \hat{\Delta}(S_{(i)}) + \delta(n) \leq 1 - \alpha - \Delta(S_{(i)}) + \delta(n) + \sqrt{\frac{\log(2n)}{2n}} + d(n).$$

Then, it follows from Assumption 6 that, with probability at least $1 - \frac{1}{n} - \frac{1}{\sqrt[4]{n}}$,

$$1 - \alpha - \hat{\Delta}(S_{(i)}) + \delta(n) \leq 1.$$

- Under $\mathcal{A}_2$, the second term on the right-hand-side of (A24) can be written as:

$$\begin{aligned}
\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D} \right] &= \mathbb{P}\left[\hat{s}(X_{n+1}, Y_{n+1}) \leq S_{(1)} \mid \mathcal{D} \right] \\
&= F(S_{(1)}) \\
&= \sum_{k=1}^{K} \rho_k F_k^k(S_{(1)}) \\
&= \sum_{k=1}^{K} \rho_k \sum_{l=1}^{K} V_{kl} \tilde{F}_l^k(S_{(1)}).
\end{aligned}$$

Then, using Assumption 5, we obtain:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]$$

$$\leq \sum_{k=1}^{K} \rho_k \left(V_{kk} + \sum_{l \neq k} |V_{kl}|\right) \tilde{F}_k^k(S_{(1)})$$

$$= \sum_{k=1}^{K} \frac{\rho_k}{\tilde{\rho}_k} \left(V_{kk} + \sum_{l \neq k} |V_{kl}|\right) \tilde{\rho}_k \tilde{F}_k^k(S_{(1)})$$

$$\leq \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^{K} |V_{kl}|\right)\right] \sum_{k=1}^{K} \tilde{\rho}_k \tilde{F}_k^k(S_{(1)})$$

$$= \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^{K} |V_{kl}|\right)\right] \tilde{F}(S_{(1)}).$$

Therefore,

$$\mathbb{E}\left[\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right] \mathbb{I}\left[\mathcal{A}_2\right]\right] \leq \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^{K} |V_{kl}|\right)\right] \mathbb{E}\left[\tilde{F}(S_{(1)})\right]$$

$$= \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^{K} |V_{kl}|\right)\right] \mathbb{E}\left[U_{(1)}\right] \tag{A28}$$

$$= \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^{K} |V_{kl}|\right)\right] \frac{1}{n+1}.$$

- Under $\mathcal{A}_1^c \cap \mathcal{A}_2^c$, by definition of $\hat{\mathcal{I}}$, for any $i \leq \hat{i} - 1$,

$$\frac{i}{n} < 1 - \left(\alpha + \hat{\Delta}(S_{(i)}) - \delta(n)\right).$$

Therefore, choosing $i = \hat{i} - 1$, we get:

$$\frac{\hat{i}}{n} < 1 - \left(\alpha + \hat{\Delta}(S_{(\hat{i}-1)}) - \delta(n)\right) + \frac{1}{n}.$$

As in the proof of Theorem 2, the probability of coverage conditional on the labeled data in $\mathcal{D}$ can be written as:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}\right]$$

$$= \tilde{F}(S_{(\hat{i})}) + \Delta(S_{(\hat{i})})$$

$$= \hat{F}(S_{(\hat{i})}) + \hat{\Delta}(S_{(\hat{i})}) + \sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\tilde{\rho}_l \tilde{F}_l^k(S_{(\hat{i})}) - \hat{\rho}_l \hat{F}_l^k(S_{(\hat{i})})\right)$$

$$= \frac{\hat{i}}{n} + \hat{\Delta}(S_{(\hat{i})}) + \sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\tilde{\rho}_l \tilde{F}_l^k(S_{(\hat{i})}) - \hat{\rho}_l \hat{F}_l^k(S_{(\hat{i})})\right)$$

$$< 1 - \left(\alpha + \hat{\Delta}(S_{(\hat{i}-1)}) - \delta(n)\right) + \frac{1}{n}$$
$$+ \hat{\Delta}(S_{(\hat{i})}) + \sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\tilde{\rho}_l \tilde{F}_l^k(S_{(\hat{i})}) - \hat{\rho}_l \hat{F}_l^k(S_{(\hat{i})})\right)$$

$$= 1 - \alpha + \delta(n) + \frac{1}{n}$$
$$+ \left(\hat{\Delta}(S_{(\hat{i})}) - \hat{\Delta}(S_{(\hat{i}-1)})\right) + \sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\tilde{\rho}_l \tilde{F}_l^k(S_{(\hat{i})}) - \hat{\rho}_l \hat{F}_l^k(S_{(\hat{i})})\right)$$

$$\leq 1 - \alpha + \delta(n) + \frac{1}{n}$$
$$+ \left(\hat{\Delta}(S_{(\hat{i})}) - \hat{\Delta}(S_{(\hat{i}-1)})\right) + \sup_{t\in[0,1]}\left|\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\left(\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t)\right)\right|.$$

The expected value of the last term above is $\delta^{**}(n)$.

The term $\hat{\Delta}(S_{(\hat{i})}) - \hat{\Delta}(S_{(\hat{i}-1)})$ above is given by

$$\hat{\Delta}(S_{(\hat{i})}) - \hat{\Delta}(S_{(\hat{i}-1)})$$
$$= \left[\hat{\Delta}(S_{(\hat{i})}) - \Delta(S_{(\hat{i})})\right] + \left[\Delta(S_{(\hat{i})}) - \Delta(S_{(\hat{i}-1)})\right] + \left[\Delta(S_{(\hat{i}-1)}) - \hat{\Delta}(S_{(\hat{i}-1)})\right]$$
$$\leq 2 \sup_{t\in[0,1]}\left|\hat{\Delta}(t) - \Delta(t)\right| + \left|\Delta(S_{(\hat{i})}) - \Delta(S_{(\hat{i}-1)})\right|.$$

We already know how to bound the expected value of the first term on the right-hand-side above. Indeed,

$$\sup_{t\in[0,1]}\left|\hat{\Delta}(t) - \Delta(t)\right|$$
$$\leq \sup_{t\in[0,1]}\left|\sum_{kl} W_{kl}\left(\hat{\rho}_l \hat{F}_l^k(t) - \tilde{\rho}_l \tilde{F}_l^k(t)\right)\right| + \sup_{t\in[0,1]}\left|\hat{F}(t) - \tilde{F}(t)\right|,$$

so that by making use of the DKW inequality we get

$$2\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{\Delta}(t) - \Delta(t)\right|\right] \leq 2\delta^{**}(n) + \sqrt{\frac{2\pi}{n}}.$$

Concerning the second term, let $U_{(1)}, \ldots, U_{(n)}$ be order statistics of n i.i.d. uniform random variables on $[0, 1]$. Then observe that

$$\Delta(S_{(\hat{i})}) - \Delta(S_{(\hat{i}-1)}) = \Delta(F^{-1}(F(S_{(\hat{i})}))) - \Delta(F^{-1}(F(S_{(\hat{i}-1)})))$$
$$= \Delta(F^{-1}(U_{(\hat{i})})) - \Delta(F^{-1}(U_{(\hat{i}-1)})).$$

From Assumption 4 it follows that, as F and $\tilde{F}$ are Lipschitz with constants $f_{\max}$ and $\tilde{f}_{\max}$ respectively, then Δ and F^{-1} are Lipschitz with constants $f_{\max} + \tilde{f}_{\max}$ and $1/f_{\min}$ respectively. Consequently, $\Delta(F^{-1}(\cdot))$ is Lipschitz with constant $\frac{f_{\max} + \tilde{f}_{\max}}{f_{\min}}$. It follows that

$$\left| \Delta(F^{-1}(U_{(\hat{i})})) - \Delta(F^{-1}(U_{(\hat{i}-1)})) \right| \leq \frac{f_{\max} + \tilde{f}_{\max}}{f_{\min}} \left(U_{(\hat{i})} - U_{(\hat{i}-1)} \right)$$
$$\leq \frac{f_{\max} + \tilde{f}_{\max}}{f_{\min}} \max_{2 \leq i \leq n} \left(U_{(\hat{i})} - U_{(\hat{i}-1)} \right)$$
$$\leq \frac{f_{\max} + \tilde{f}_{\max}}{f_{\min}} \max_{1 \leq i \leq n+1} D_i,$$

where $D_1 = U_{(1)}$, $D_i = U_{(i)} - U_{(i-1)}$ for $i = 2, \ldots, n$, and $D_{n+1} = 1 - U_{n_{\text{cal}}}$. By a standard result on maximum uniform spacing,

$$\mathbb{E} \left[\max_{1 \leq i \leq n+1} D_i \right] = \frac{1}{n+1} \sum_{j=1}^{n+1} \frac{1}{j}.$$

At this point, we have proven that

$$\mathbb{E} \left[\mathbb{P} \left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D} \right] \mathbb{I} \left[\mathcal{A}_1^c \cap \mathcal{A}_2^c \right] \right]$$
$$\leq 1 - \alpha + \delta(n) + \frac{1}{n} + 3\delta^{**}(n) + \sqrt{\frac{2\pi}{n}} + \frac{1}{n+1} \sum_{j=1}^{n+1} \frac{1}{j}. \tag{A29}$$

Finally, combining (A24) with (A25), (A28), and (A29) leads to the desired result:

$$\mathbb{P} \left[Y_{n+1} \in \hat{C}(X_{n+1}) \right]$$
$$\leq \mathbb{P} \left[\mathcal{A}_1 \right] + \mathbb{E} \left[\mathbb{P} \left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D} \right] \mathbb{I} \left[\mathcal{A}_2 \right] \right]$$
$$+ \mathbb{E} \left[\mathbb{P} \left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D} \right] \mathbb{I} \left[\mathcal{A}_1^c \cap \mathcal{A}_2^c \right] \right]$$
$$\leq 1 - \alpha + \delta(n) + \frac{2}{n} + 3\delta^{**}(n) + \frac{1}{\sqrt[4]{n}} + \frac{1}{n+1} \cdot \left[\max_{k \in [K]} \left(\frac{\rho_k}{\tilde{\rho}_k} \sum_{l=1}^{K} |V_{kl}| \right) - 1 \right]$$
$$= 1 - \alpha + \delta(n) + \varphi(n).$$

$\square$

Theorem A1. *Under the assumptions of Theorem 2, consider the empirical process $\hat{\psi}(t)$ defined in (12). Then,*

$$\delta^{**}(n) \leq \inf_{\beta} \left\{ \frac{1}{\sqrt{n}} \sqrt{\frac{\pi}{2}} \left(|\beta_0| + \frac{\sum_{k=1}^{K} |\beta_k|}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\},$$

where

$$B(K, n, \beta) := 2 \min \left\{ \max_{l \in [K]} \sum_{k} |\Omega_{kl}| \sqrt{\log(Kn+1)}, \right.$$

$$\left. 24 \max_{k,l \in [K]} |\Omega_{kl}| \frac{2 \log K + 1}{2 \log K - 1} \sqrt{2K \log K} \right\}. \tag{A30}$$

Proof of Theorem A1. Our goal is that of proving

$$\mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{\psi}(t) \right| \right] \leq \inf_{\beta} \left\{ \frac{1}{\sqrt{n}} \sqrt{\frac{\pi}{2}} \left(|\beta_0| + \frac{\sum_{k=1}^{K} |\beta_k|}{K} \right) + \frac{1}{\sqrt{n}} B(K, n, \beta) \right\}.$$

We proceed similarly to the proof of Theorem 4. Recall that the matrix $\bar{W}$ is defined such that $\bar{W}_{kl} = \beta_0 \mathbb{I}[k = l] + \beta_k/K$, for some $\beta_0, (\beta_k)_k$. As in the proof of Theorem 4, we will derive a bound as a function of the parameters β; then, the result can be obtained by taking the minimum over β.

Recall that $\hat{\psi}(t) = \hat{\psi}_1(t) + \hat{\psi}_2(t) + \hat{\psi}_3(t)$, where

$$\hat{\psi}_1(t) := \beta_0 \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{I}\left[\hat{s}(X_i, \tilde{Y}_i) \leq t\right] - \tilde{F}(t) \right],$$

$$\hat{\psi}_2(t) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (\beta_k/K) \left[\mathbb{I}\left[\hat{s}(X_i, k) \leq t\right] - \tilde{F}^k(t) \right]$$

$$\hat{\psi}_3(t) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} (W_{kl} - \bar{W}_{kl}) \left[\mathbb{I}\left[\hat{s}(X_i, k) \leq t, \tilde{Y}_i = l\right] - \tilde{\rho}_l \tilde{F}_l^k(t) \right],$$

$$\tilde{F}^k(t) := \mathbb{P}\left[\hat{s}(X_i, k) \leq t\right],$$

$$\tilde{F}(t) := \mathbb{P}\left[\hat{s}(X_i, \tilde{Y}_i) \leq t\right].$$

Then, it follows that

$$\left| \hat{\psi}(t) \right| \leq \left| \hat{\psi}_1(t) \right| + \left| \hat{\psi}_2(t) \right| + \left| \hat{\psi}_3(t) \right|.$$

We now bound the suprema of these three processes separately, since

$$\mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{\psi}(t) \right| \right] \leq \mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{\psi}_1(t) \right| \right] + \mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{\psi}_2(t) \right| \right] + \mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{\psi}_3(t) \right| \right].$$

Let's start from $\hat{\psi}_1(t)$. First, note that

$$\mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{\psi}_1(t) \right| \right] \leq |\beta_0| \mathbb{E} \left[\sup_{t \in [0,1]} \left| \hat{F}(t) - \tilde{F}(t) \right| \right].$$

Then, it follows from the DKW inequality that, for any $\eta > 0$

$$\mathbb{P} \left[\sup_{t \in [0,1]} \left| \hat{F}(t) - \tilde{F}(t) \right| > \eta \right] \leq \exp\left\{ -2n\eta^2 \right\}.$$

This directly implies that

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{\psi}_1(t)\right|\right] \le |\beta_0|\sqrt{\frac{\pi}{2n}}.$$

Similarly, we can deal with $\hat{\psi}_2(t)$:

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{\psi}_2(t)\right|\right] = \mathbb{E}\left[\sup_{t\in[0,1]}\left|\sum_{k=1}^{K}\frac{\beta_k}{K}\left(\hat{F}^k(t) - \tilde{F}^k(t)\right)\right|\right]$$

$$\le \sum_{k=1}^{K}\frac{|\beta_k|}{K}\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{F}^k(t) - \tilde{F}^k(t)\right|\right]$$

$$\le \frac{\sum_{k=1}^{K}|\beta_k|}{K}\sqrt{\frac{\pi}{2n}}.$$

A bound for $\hat{\psi}_3(t)$ is identified by defining $\Omega_{kl} := W_{kl} - \bar{W}_{kl}$, so that

$$\hat{\psi}_3(t) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\Omega_{kl}\left[\mathbb{I}\left[\hat{s}(X_i, k) \le t, \tilde{Y}_i = l\right] - \tilde{\rho}_l\tilde{F}_l^k(t)\right].$$

Then, it follows from Lemma A1 that

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{\psi}_3(t)\right|\right] \le \frac{1}{\sqrt{n}}B(K, n, \beta),$$

where $B(K, n, \beta)$ is defined in (A30).

Putting everything together, we find that, for any $\beta \in [0, 1]^{K+1}$,

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{\psi}(t)\right|\right] \le \frac{1}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\left(|\beta_0| + \frac{\sum_{k=1}^{K}|\beta_k|}{K}\right) + \frac{1}{\sqrt{n}}B(K, n, \beta).$$

Therefore,

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\hat{\psi}(t)\right|\right] \le \inf_{\beta}\left\{\frac{1}{\sqrt{n}}\sqrt{\frac{\pi}{2}}\left(|\beta_0| + \frac{\sum_{k=1}^{K}|\beta_k|}{K}\right) + \frac{1}{\sqrt{n}}B(K, n, \beta)\right\}.$$

$\square$

D   Additional Numerical Results

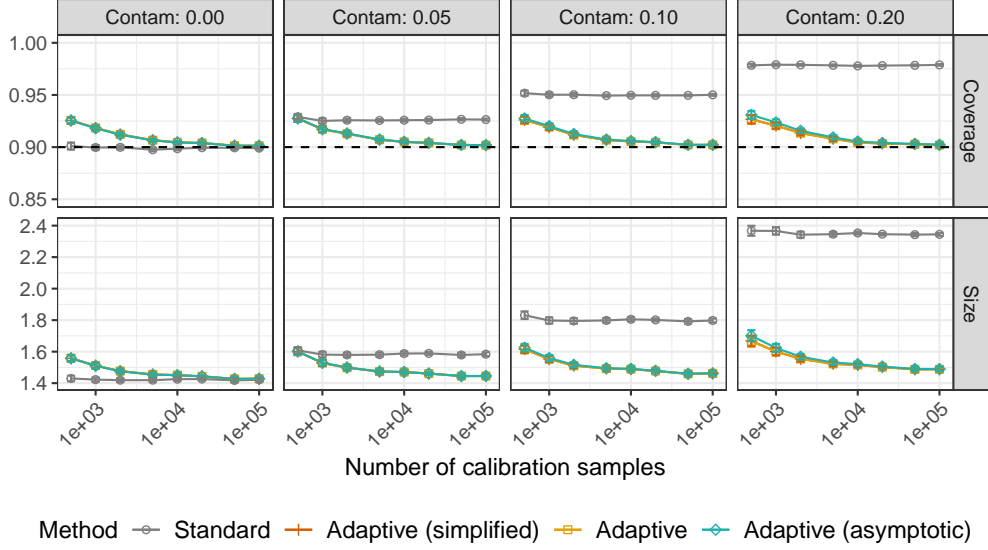## D.1   The Impact of the Label Contamination Strength

Figure A9: Performances of different conformal methods on simulated data with labels contaminated by a randomized response model, as function of the number of calibration samples. The empirical coverage and average size of the prediction sets are stratified based on the strength ϵ of the label contamination process. The number of classes is $K = 4$.
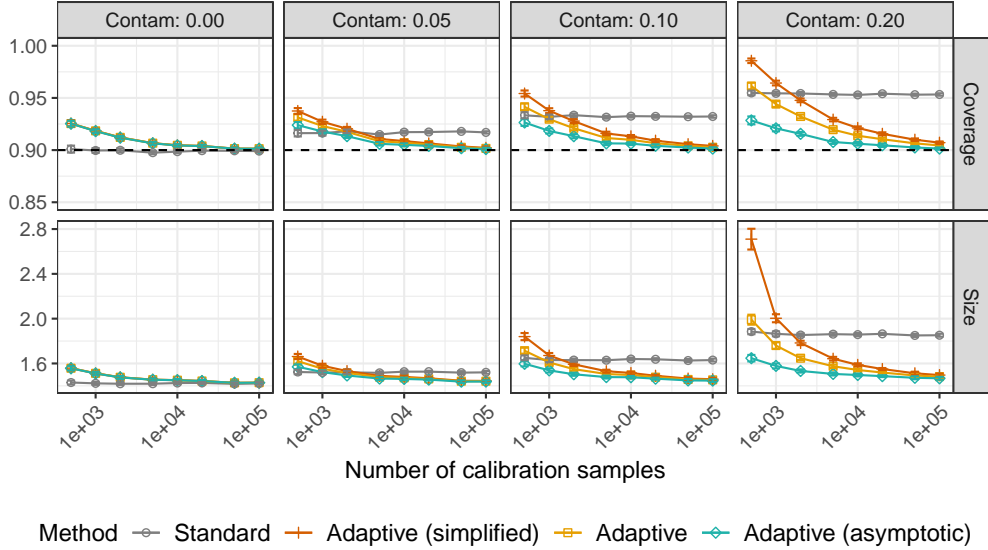

Figure A10: Performances of different conformal methods on simulated data with labels contaminated by a block randomized response model, as function of the number of calibration samples. The empirical coverage and average size of the prediction sets are stratified based on the strength ϵ of the label contamination process. The number of classes is $K = 4$.

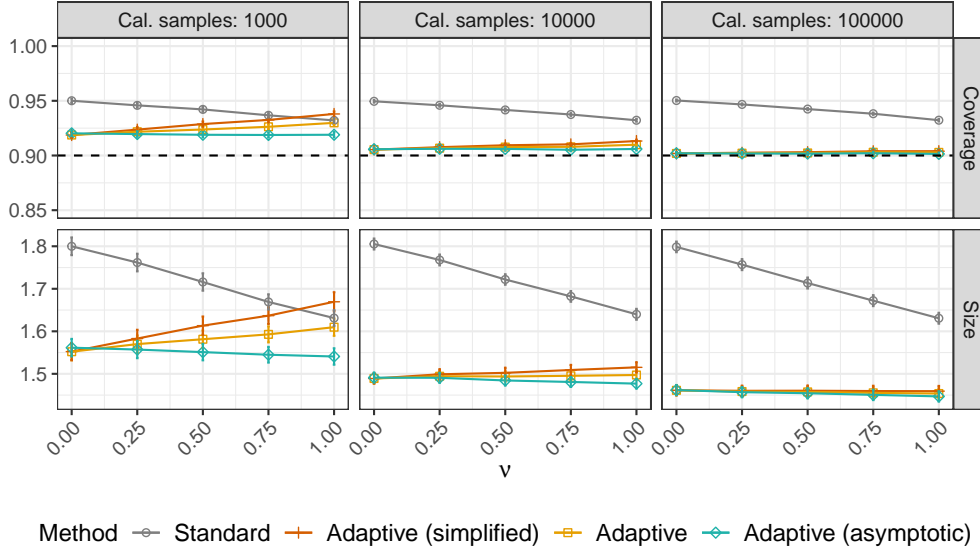## D.2 The Impact of the Label Contamination Model

Figure A11: Performances of different conformal methods on simulated data with labels contaminated by a two-level randomized response model, as function of ν. The reported empirical coverage and average size of the prediction sets are stratified based on the number of calibration samples. The strength parameter of the label contamination process is $\epsilon = 0.1$. The number of labels is $K = 4$.

D.3 The Impact of the Number of Classes

Figure A12 suggests that all implementations of our adaptive method remain similarly effective even when the number K of possible labels becomes very large. However, while this is often true, it is not always the case, depending on the structure of the contamination model and on the label contamination strength ϵ. The experiments in Figure A12 use a randomized response model, namely the setting for which the finite-sample implementations are optimized. In contrast, Figure A13 examines the case where the contamination model is a two-level randomized response model with $\nu = 0.8$ and $\epsilon = 0.1$, corresponding to a significant deviation from the simple randomized response model. Under this scenario, the finite-sample evaluation of the correction term becomes increasingly conservative relative to the more efficient *Asymptotic* implementation as K grows. Figure A14 shows that this effect gets more pronounced as ϵ increases. Analogous results are obtained for the block-randomized response model with two blocks, with $\epsilon = 0.1$ and $\epsilon = 0.2$, as shown in Figures A15 and A16.
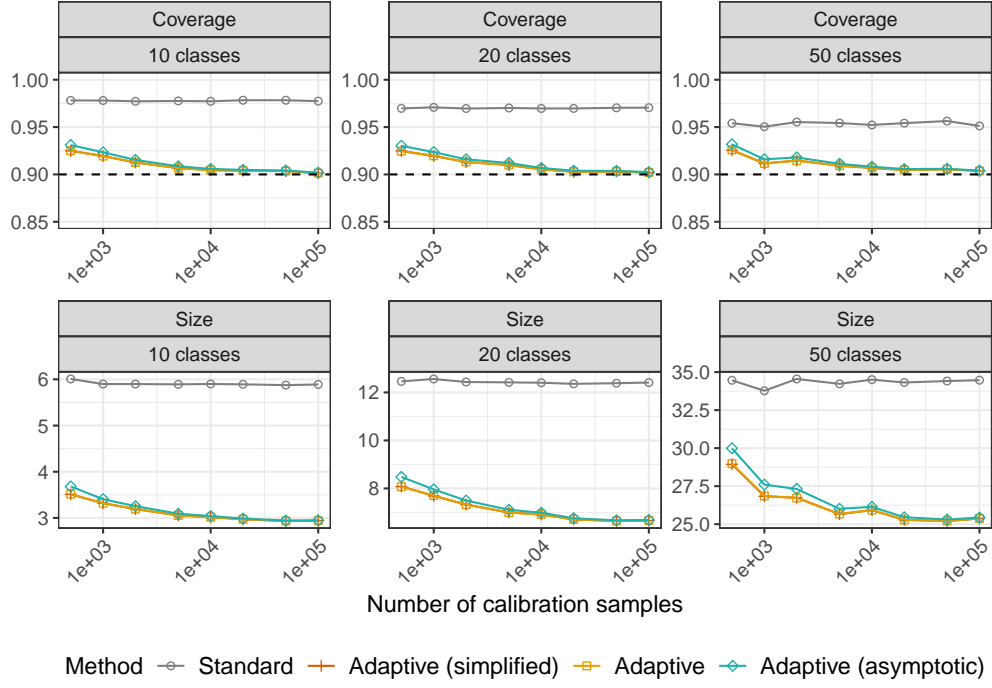
Randomized response model.

Figure A12: Performances of different conformal prediction methods on simulated data with contaminated labels. The label contamination process is the randomized response model with $\epsilon = 0.2$. The reported empirical coverage and average size of the prediction sets are stratified based on the number of labels. The dashed horizontal line indicates the 90% nominal marginal coverage.
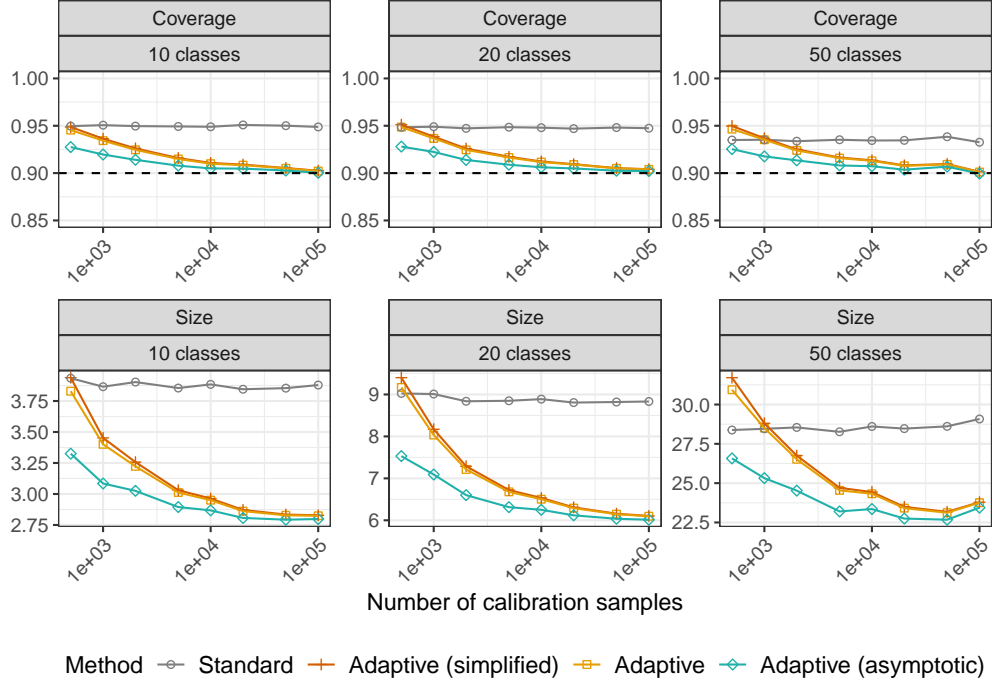
Two-level randomized response model with $\nu = 0.8$.

Figure A13: Performances of different conformal prediction methods on simulated data with contaminated labels. The label contamination process is the two-level randomized response model described in Section B.4.2, with $\epsilon = 0.1$ and $\nu = 0.8$. Other details are as in Figure A12.
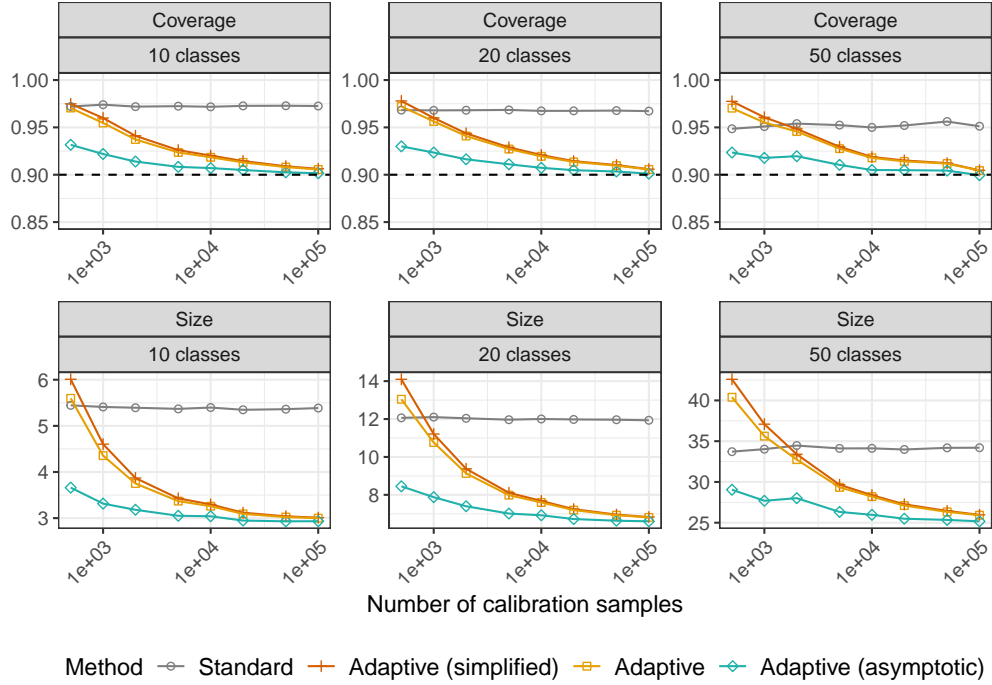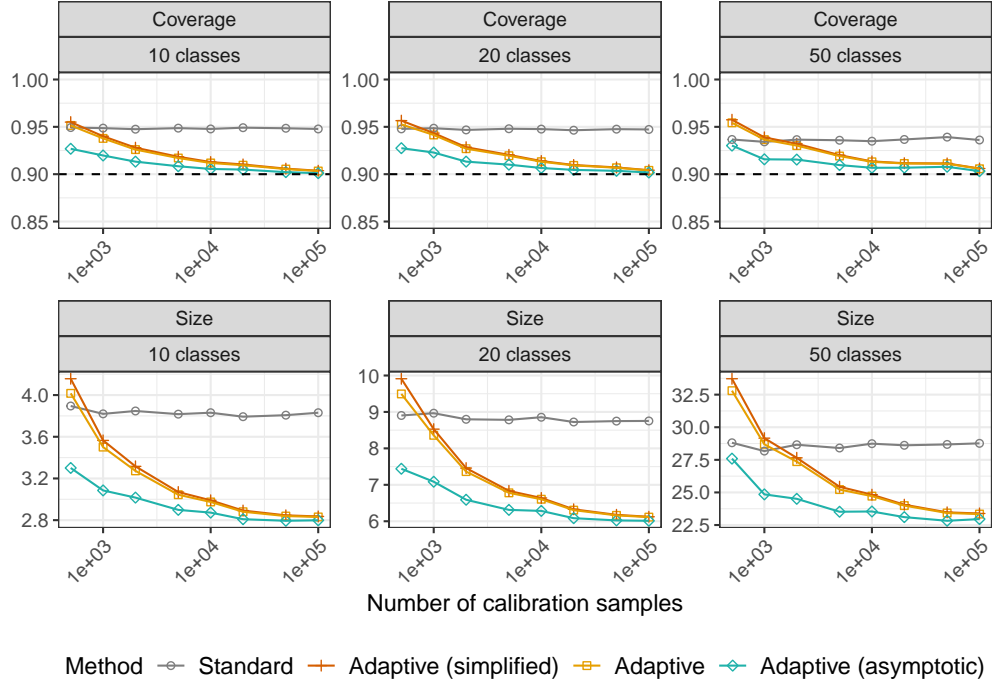

Figure A14: Performances of different conformal prediction methods on simulated data with contaminated labels. The label contamination process is the two-level randomized response model described in Section B.4.2, with $\epsilon = 0.2$ and $\nu = 0.8$. Other details are as in Figure A12.

Block-randomized response model.

Figure A15: Performances of different conformal prediction methods on simulated data with contaminated labels. The label contamination process is the block-randomized response model, with two blocks, with $\epsilon = 0.1$. Other details are as in Figure A12.
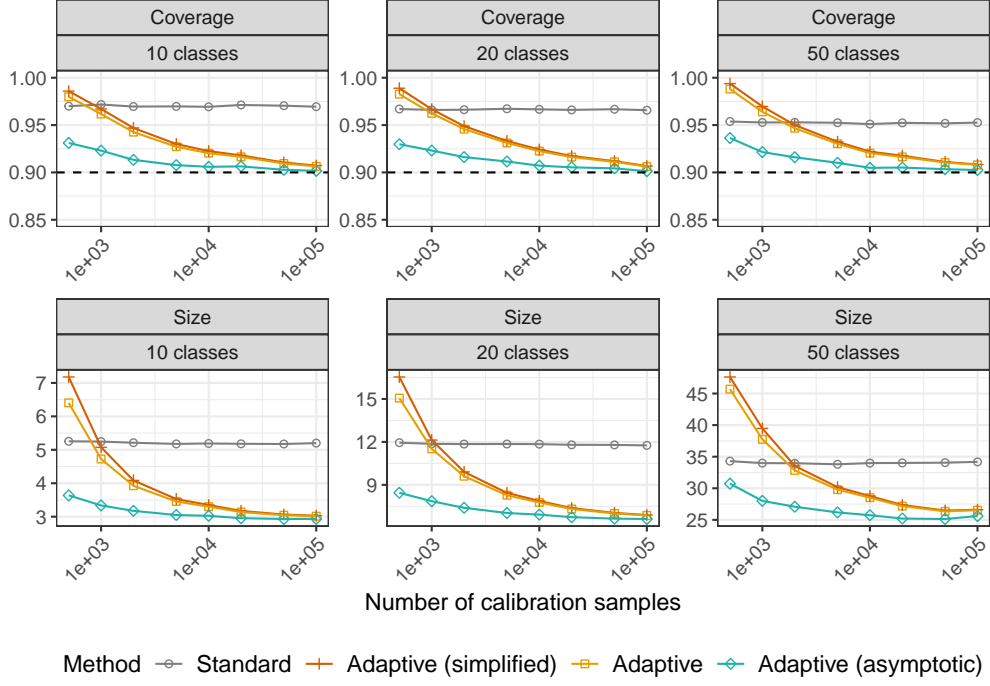
Figure A16: Performances of different conformal prediction methods on simulated data with contaminated labels. The label contamination process is the block-randomized response model, with two blocks, with $\epsilon = 0.2$. Other details are as in Figure A12.

D.4 The Advantage of Targeting Marginal Coverage

In these experiments, class imbalance is introduced and controlled through an imbalance indicator μ. The imbalance indicator adjusts the class distribution using an exponential decay function, where higher values of μ result in greater disparity among class proportions. These ratios are applied as a post-sampling step, where the sampled data is re-weighted or re-sampled to align with the desired class balance. This ensures precise control over the degree of imbalance across different experimental settings. Figure A17 compares the performances of our optimistic adaptive methods for marginal coverage with the optimistic adaptive method for label-conditional coverage of Sesia et al. (2024). The comparison is made in terms of marginal coverage and average size of the prediction sets output by the methods, and the results are stratified based on μ. As expected, the results show that the label-conditional method produces increasingly large prediction sets, as the class imbalance gets more severe. The performance of the method for marginal coverage, on the other hand, is not impacted by class imbalance, and the prediction sets output by the method remain informative also for high values of μ.
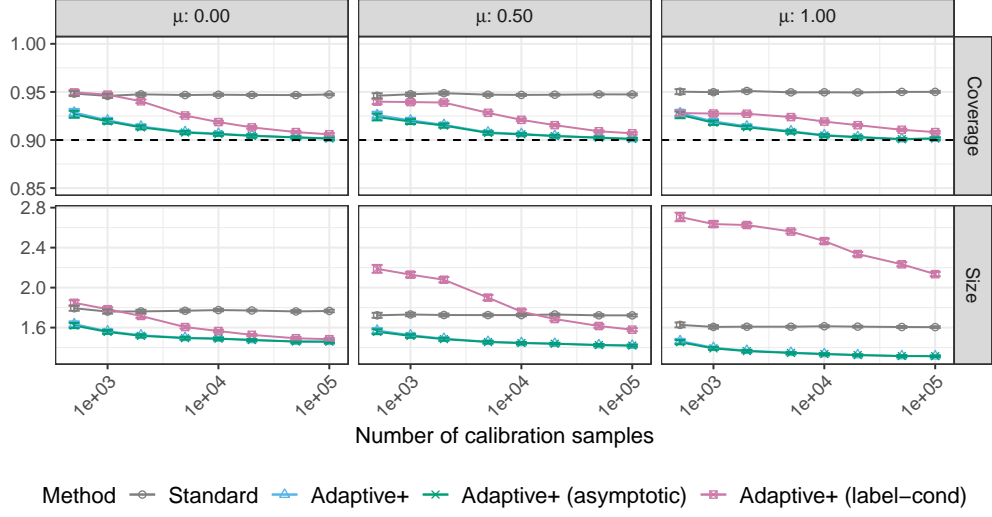
Figure A17: Performances of different conformal methods on simulated data with labels contaminated by a block-randomized response model with two blocks, as function of the number of samples. The reported empirical coverage and average size of the prediction sets are stratified based on the class imbalance, controlled by the parameter μ. The number of classes is $K = 4$, the strength parameter of the label contamination process is $\epsilon = 0.1$, and $\nu = 0.2$. The dashed horizontal line indicates the 90% nominal marginal coverage.

D.5   Additional Results on the BigEarthNet Data

The contamination model matrix T is estimated on the training set with

$$\hat{T}_{kl} = \frac{1}{|\{i : Y_i = l\}|} \sum_{i:Y_i=l} \mathbb{I}\left[\tilde{Y} = k\right].$$

The estimated transition matrix is

$$T = \begin{pmatrix} 0.986 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.988 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.956 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.994 & 0 & 0.001 \\ 0.013 & 0.008 & 0.042 & 0.006 & 1 & 0.017 \\ 0.001 & 0.004 & 0.002 & 0 & 0 & 0.982 \end{pmatrix},$$

where the rows and columns refer to Coast, waters and wetlands, Arable land, Agriculture, Vegetation, Urban fabric and Mixed, in this specific order. The form of T indicates that the contamination mainly manifests in the mislabelling of patches related to agricultural activities, or characterized by mixtures of land use types, as patches uniquely composed of urban fabric.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

Bortolotti, T.; Wang, Y. X. R.; Tong X.; Menafoglio, A.; Vantini, S.; Sesia, M.
Noise-Adaptive Conformal Classification with Marginal Coverage

18/2025 Antonietti, P.F.; Corti, M.; Gómez, S.; Perugia, I.
A structure-preserving LDG discretization of the Fisher-Kolmogorov equation for modeling neurodegenerative diseases

16/2025 Radisic, I.; Regazzoni, F.; Bucelli, M.; Pagani, S.; Dede', L.; Quarteroni, A.
Influence of cellular mechano-calcium feedback in numerical models of cardiac electromechanics

17/2025 Botti, M.; Mascotto, L.
Sobolev--Poincaré inequalities for piecewise $W^{1,p}$ functions over general polytopic meshes

15/2025 Fois, M.; de Falco, C.; Formaggia L.
Efficient particle generation for depth-averaged and fully 3D MPM using TIFF image data

13/2025 Scimone, R.; Menafoglio, A.; Secchi, P.
Estimating Non-Stationarity in Spatial Processes: an approach based on Random Domain Decomposition

14/2025 Nicolussi, F.; Masci, C.
Stratified Multilevel Graphical Models: Examining Gender Dynamics in Education

12/2025 Alessandro Andrea Zecchi, Claudio Sanavio, Simona Perotto e Sauro Succi
Improved amplitude amplification strategies for the quantum simulation of classical transport problems

11/2025 Tonini, A.; Dede', L.
Enhanced uncertainty quantification variational autoencoders for the solution of Bayesian inverse problems

10/2025 Botti, M.; Mascotto, L.; Vacca, G.; Visinoni, M.
Stability and interpolation estimates of Hellinger-Reissner virtual element spaces