UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXVIII

Combining evidence under dependence

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Bruno Scarpa

Co-supervisore: Prof. Aaditya Ramdas

Dottorando/a: Matteo Gasparin

1 December 2025

Abstract

The quantification of uncertainty is a central element of statistical inference, as it enables the assessment of the reliability of inferential conclusions. In practice, uncertainty may arise from multiple sources, which are often dependent on each other, posing new challenges both methodologically and in applications. An example of this occurs when different p-values, obtained from distinct statistical tests, are available for the same hypothesis.

This work is situated within this context, with the aim of contributing to the development of frequentist tools capable of handling complex scenarios where dependence among different measures of evidence cannot be neglected. In this perspective, new procedures are introduced for combining dependent p-values, with particular attention to the case in which they are exchangeable. Such situations arise, for instance, when p-values are generated through random splitting techniques, and therefore hold concrete relevance in many applications. In addition, methods are proposed for the combination of uncertainty sets, such as confidence regions or prediction sets, that may be arbitrarily dependent.

Finally, the results are extended and placed within the framework of conformal prediction, an approach that allows for the construction of *distribution-free* prediction set under mild assumption.

Sommario

La quantificazione dell'incertezza è un elemento centrale dell'inferenza statistica, poiché consente di valutare l'affidabilità delle conclusioni inferenziali. Nella pratica, l'incertezza può derivare da molteplici fonti, spesso tra loro dipendenti, ponendo nuove sfide sia sul piano metodologico sia su quello applicativo. Un esempio si verifica quando diversi p-value, ottenuti da test statistici distinti, sono disponibili per la stessa ipotesi.

Questo lavoro si inserisce in tale contesto, con l'obiettivo di contribuire allo sviluppo di strumenti frequentisti in grado di gestire scenari complessi, nei quali la dipendenza tra diverse misure di evidenza non può essere trascurata. In questa prospettiva, vengono introdotte nuove procedure per la combinazione di p-value dipendenti, con particolare attenzione al caso in cui essi siano scambiabili. Tale situazione si presenta, ad esempio, quando i p-value sono generati mediante tecniche di *random splitting*, e riveste quindi un interesse concreto in numerose applicazioni. Inoltre, vengono proposti metodi per la combinazione di insiemi di incertezza, come regioni di confidenza o insiemi predittivi, che possono essere arbitrariamente dipendenti.

Infine, i risultati vengono estesi e collocati all'interno del paradigma della conformal prediction, un approccio che consente di costruire insiemi predittivi *distribution-free* sotto ipotesi minimali.

Ai miei genitori

Acknowledgements

Since childhood, I pictured my future on a football pitch. However, over the years, I learned that dreams can change direction, much like a ball whose trajectory changes unexpectedly. That is how I ultimately found myself on a different path, one that has led me to this PhD in statistics. And, as at the end of every meaningful match, I feel it is time to pause, look back, and acknowledge those who have made this journey possible.

I would like to begin by expressing my gratitude to my supervisors, Bruno Scarpa and Aaditya Ramdas. Their mentorship has been invaluable, helping me grow both professionally and personally. Their guidance has shaped the path I have taken and is something I will carry with me throughout my life.

My deepest thanks go to my family, and especially to my parents. They were my first supporters and have accompanied me through all these years, giving me strength in difficult moments and sharing my joy in the happy ones. My greatest achievement is knowing that I have made you proud. A special thanks also goes to my cat, Leo, who has consistently brought joy to our home.

I would also like to begin by thanking my "PhD team": Daniela, Lorenzo, PG and Giovanni. Sharing this journey with you has made the difficult moments more manageable and the good ones even more meaningful. I am also grateful to all my friends who have accompanied me throughout these years, whether with a kind word, a laugh, or simply a shared spritz. In particular, I would like to sincerely thank Marco, Alessia, Riccardo, and Matteo, as well as the "Sezione Pionca" group, for their support. In addition, I would like to thank my "american friends": Rebecca, Lorenzo, Luca, Fede, Lucas, Yusuf, and Lasse, for making me feel at home even when thousands of kilometers apart.

Finally, I want to thank Meme, Mariangela Guidolin, Antonietta Mira and Elena Stanghellini for their advice and for supporting me, in various ways, throughout this path. A special thanks also goes to Aldo Solari and Simone Vantini for their valuable comments on this work.

Last but not least, I would like to thank F.C. Internazionale Milano, my team, which has supported me throughout this long journey. For 90 minutes each week, they gave me the chance to forget everything else and simply enjoy a passion.

In closing, I am deeply grateful to all who have accompanied me throughout this adventure.

Lugano, November 2025

Contents

List of Figures

List of Tables

Introduction

Overview

Quantifying uncertainty represents a cornerstone in statistics; indeed, assessing the reliability of inference and prediction is a fundamental task in many real-world applications. In particular, understanding and accurately representing uncertainty allows researchers and practitioners to make better informed decisions and interpret results appropriately.

In practice, it is often the case that multiple sources of uncertainty are available simultaneously, for example in the form several p-values for the same hypothesis, confidence intervals for a parameter of interest, or prediction sets for new observations. A natural question in such contexts is how to combine these multiple sources into a single summary, such as a combined p-value or a single confidence or prediction set, that it is able to capture the overall level of uncertainty while preserving the validity of the inference.

On the combination of p-values

The combination of p-values represents a fundamental task frequently encountered in statistical inference and its applications in the natural sciences. Within the realm of multiple testing, for instance, the focus lies in testing whether all individual null hypotheses are simultaneously true. This particular challenge, often referred to as global null testing, can be addressed by merging multiple p-values into a single p-value. Potential solutions, assuming the p-values are statistically independent, are provided in Fisher (1934); Pearson (1934); Simes (1986), with the latter also working under a certain notion of positive dependence (Sarkar, 1998; Benjamini and Yekutieli, 2001). See Owen (2009) for a review of the methods. Recently, harmonic mean p-values (Wilson, 2019) and methods under negative dependence (Chi *et al.*, 2025) have been developed in the context of multiple testing.

The assumption of independence, or positive/negative dependence, is often violated in many real-world applications (see e.g., Section 4.2 of Efron, 2010), and in certain scenarios, it may be preferable not to impose unverifiable conditions on the joint distribution of the p-values, beyond the minimum necessary assumption that each individual p-value is indeed marginally valid. Several methods are available for combining p-values with arbitrary dependencies; notably, the Bonferroni method is widely used, which involves multiplying the minimum of the p-values by the number of tests conducted. Other methods have been proposed in the literature, some based on order statistics (Rüger, 1978; Morgenstern, 1980; Hommel, 1983), while others rely on their arithmetic mean and other variants (Rüschendorf, 1982; Vovk and Wang, 2020; Vovk *et al.*, 2022b). Two prominent examples are that both 2 times the median of the p-values, and 2 times the average of the p-values, are valid combination rules, and the multiplicative factor of 2 cannot be reduced. Inevitably, these methods satisfying validity under arbitrary dependence come with a price to pay in terms of statistical power. Surprisingly, while many methods exist for combining dependent p-values, the case where the p-values are exchangeable has received little attention. An exception is the work of Choi and Kim (2023), which shows that the constant 2 in the "twice the mean" rule cannot be improved even under exchangeability.

While the combination of p-values is well established both in practice and in the statistical literature, a relatively unexplored issue arises when multiple uncertainty sets, such as confidence regions or prediction sets, need to be combined into a single set. In such cases, it becomes crucial to have practical and valid methods that can produce a single set, ensuring proper coverage while maintaining informativeness.

Conformal prediction

Nowadays, it is common to use black-box algorithms for prediction (Azzalini and Scarpa, 2012). Although these methods often provide strong predictive performance, quantifying their uncertainty and constructing accurate prediction sets is far from straightforward. In fact, the predictions are based on the patterns observed in the training data, but these predictions can be prone to errors due to various sources of uncertainty, such as noisy data, overfitting, or in general inaccurate model assumptions.

Conformal prediction has emerged as a general and versatile framework for constructing prediction sets in regression and classification tasks (Shafer and Vovk, 2008). Unlike traditional methods, which often depend on rigid distributional assumptions, conformal prediction transforms point predictions from any prediction, or black-box, algorithm into prediction sets that guarantee valid finite-sample marginal coverage. Originally

introduced by Saunders *et al.* (1999), it has become increasingly influential, with numerous methods and extensions being proposed since its introduction. Several influential contributions to the framework include the works of Lei *et al.* (2018), Romano *et al.* (2019), and Barber *et al.* (2021). While extensions and generalizations of the methods have been proposed by Kim *et al.* (2020), Ajroldi *et al.* (2023) and Gupta *et al.* (2022), among others. A different branch of the literature extends conformal prediction to settings where the standard assumptions may not hold, such as Tibshirani *et al.* (2019), Prinster *et al.* (2022), Barber *et al.* (2023) and Stutz *et al.* (2023).

Main contributions of the thesis

In this thesis, we aim to advance the development of novel frequentist methods for quantifying uncertainty under dependence. The first part addresses the combination of dependent p-values and uncertainty sets, while the second part focuses on conformal prediction. The contributions are summarized below:

Combining exchangeable p-values. As discussed in the previous section, the problem of combining p-values is both classical and fundamental, yet the standard assumption of independence is frequently violated or impossible to verify in many applications. Numerous well-known methods exist to combine a set of arbitrarily dependent p-values (for the same hypothesis) into a single p-value. We demonstrate that nearly all of these existing methods can be improved when the p-values are exchangeable or when external randomization is allowed. The main technical contribution is to show that all current combination rules can be derived by first calibrating the p-values into e-values (using an α-dependent calibrator), averaging these e-values, converting the result into a level-α test via Markov's inequality, and finally producing p-values by combining this family of tests. The improvements arise from recent randomized and exchangeable versions of Markov's inequality. *The results related to this part are based on Gasparin* et al. *(2025) and are presented in Chapter 2.*

Merging uncertainty sets via majority vote. We address the problem of combining arbitrarily dependent uncertainty sets into a single set in a black-box manner. We describe a simple majority vote procedure that produces a merged set with nearly the same error guarantee as the originals. Extensions include weighted averaging to incorporate prior information and a randomization technique that yields smaller sets without compromising coverage. *The results related to this part are based on Gasparin and Ramdas (2024b) and are presented in Chapters 3 and 4.*

Improving the statistical efficient of cross-conformal prediction. We introduce new variants of cross-conformal prediction that produce smaller prediction sets while maintaining the coverage guarantee. These methods build on recent advances in statistically efficient p-value combination, exploiting exchangeability and randomization. *The results related to this part are based on Gasparin and Ramdas (2025) and are presented in Chapter 5.*

Conformal online model aggregation. A relatively unexplored question in conformal prediction is model selection or aggregation: given the many available prediction methods, which should we conformalize? We propose a novel approach for conformal model aggregation in online settings, combining prediction sets from multiple algorithms via a voting scheme, with model weights adapted over time based on past performance. The algorithm is especially useful in decentralized settings, where the aggregator has access only to the prediction sets. *The results related to this part are based on Gasparin and Ramdas (2024a) and are presented in Chapter 6.*

Chapter 1

Statistical evidence: classical and modern perspectives

The objective of this chapter is to provide a brief roadmap of the main concepts of statistical evidence and to introduce the key tools that will be used in the subsequent chapters. We will give a concise overview of different approaches to quantifying and interpreting statistical evidence, with a particular focus on p-values, e-values, and their connections. In addition to discussing their individual properties, we will also describe techniques for aggregating p-values and e-values, which play a central role in the developments presented later.

1.1 P-values and testing

We begin with a sample space Ω equipped with a σ-algebra $\mathcal{F}$, and the set $\mathcal{M}$ of all probability measures (called distributions) on $(\Omega, \mathcal{F})$. Data are represented by Z, where Z is usually but not necessarily represented as $Z = (Z_1, \ldots, Z_n)$ where $Z_1, \ldots, Z_n$ are n independent and identically distributed (iid) random variables. In particular, our observed data are governed by some distribution $\mathbb{P}^\star \in \mathcal{M}$. Statistical inference arises from the fundamental fact that the distribution that generates the data is, either partially or entirely, unknown. As a result, the goal of statistical analysis is to reconstruct the underlying distribution based on the available observations.

In what follows, we focus our attention on hypothesis testing, whose aim is to assess whether the observed data are compatible with a given hypothesis. Here, a hypothesis is defined as a set of probability distributions on $\mathcal{M}$. A hypothesis is *simple* if it is a singleton, otherwise is called *composite*. Based on this definition, our goal is to test the null hypothesis H_0 that the data generating distribution lies in $\mathcal{P}$.

To measure evidence against a given hypothesis, we can use p-values. We now define a p-value in general terms, starting with the concept of p-variable.

Definition 1.1. A p-variable P for testing $\mathcal{P}$ is a $[0, \infty)$ random variable which satisfies

$$\mathbb{P}(P \leq \alpha) \leq \alpha, \quad \text{for all } \alpha \in (0, 1), \tag{1.1}$$

and $\mathbb{P} \in \mathcal{P}$. A p-variable P is exact if (1.1) holds with equality.

In words, a p-variable is random variable that is stochastically larger than a standard uniform under the null hypothesis. Typically, of course, P will only take values in $[0, 1]$, but nothing is lost by allowing the larger range above. A *p-value* in simply the value taken by a p-variable.

To evaluate if the observed data are compatible with the null hypothesis it is common to use a *test*.

Definition 1.2. A test is a measurable function $T : \Omega \to \{0, 1\}$. The *type-I error* of a test T for $\mathbb{P}$ is $\mathbb{E}_{\mathbb{P}}[T]$. A test has level $\alpha \in [0, 1]$ for $\mathcal{P}$ if its type-I error is at most α for every $\mathbb{P} \in \mathcal{P}$.

By combining Definitions 1.1 and 1.2, it is clear that if we have a p-variable P, we can obtain a level α test simply by thresholding P at α (i.e., $T = \mathbb{1}\{P \leq \alpha\}$). Ideally, a statistical test should reject the null hypothesis only when it is false, and refrain from rejecting it when it is true. In particular, the type-I error refers to the incorrect rejection of the null hypothesis when it is actually true.

The derivation of p-values or statistical tests tailored to a specific model lies beyond the scope of this work. However, a (non-exhaustive) list of commonly used approaches, both parametric and nonparametric, includes likelihood-based methods (Pace and Salvan, 1997), permutation-based procedures (Pesarin and Salmaso, 2010), and resampling techniques (Efron and Tibshirani, 1994; Politis *et al.*, 1999).

1.2 E-values: definition and basic properties

P-values are a cornerstone of statistical inference and have long played a central role in scientific research as a standard tool for assessing the evidence between data and a given hypothesis. Recently, a new line of research has introduced e-values as a complementary framework for hypothesis testing, offering different advantages in certain settings. We start by providing a formal definition.

Definition 1.3. An *e-variable* E for testing $\mathcal{P}$ is a non-negative extended random variable with $\mathbb{E}_{\mathbb{P}}[E] \leq 1$, for all $\mathbb{P} \in \mathcal{P}$.

Thus, whereas p-variables are subject to probabilistic constraints (i.e., they are super-uniform random variables under the null hypothesis), e-variables are constrained in expectation under the null hypothesis. An *e-value* is the value taken by an e-variable. In a sense, e-values have always been implicitly present in statistics. For example, the likelihood ratio in a simple versus simple hypothesis test is a valid e-value, as it satisfies the required expectation bound under the null hypothesis. The strength and intuitive appeal of e-values lie in their interpretation as outcomes of bets placed against the null hypothesis (Shafer, 2021).

At first glance, controlling the type-I error using e-values may seem difficult. However, it proves to be quite straightforward. In fact, by applying Markov's inequality, one obtains that

$$\mathbb{P}\left(E \geq \frac{1}{\alpha}\right) \leq \alpha \mathbb{E}[E] \leq \alpha, \tag{1.2}$$

which immediately yields a level α test by rejecting the null hypothesis whenever the e-value exceeds $1/\alpha$. It is clear that an important difference is that for p-values, smaller values suggest that the null hypothesis is false, but the opposite holds true for e-values where a large value is interpreted as evidence against H_0.

In the following, we highlight the crucial role of e-values in the context of combining dependent p-values. Beyond this setting, e-values may also be preferable to p-values in various other scenarios, including irregular models (Wasserman *et al.*, 2020), sequential inference (Ramdas *et al.*, 2023; Howard *et al.*, 2021), and multiple testing (Wang and Ramdas, 2022; Xu *et al.*, 2025). For a comprehensive treatment of e-values, we refer the interested reader to Ramdas and Wang (2025). As a final remark, it is evident from (1.2) that Markov's inequality plays a central role in this context. Extensions of this inequality, which are crucial for the developments presented in this work, will be discussed later in this chapter (Section 1.6).

1.3 Calibrators

Calibrators are functions that convert p-values to e-values (or viceversa). A first definition of calibrator is given in Vovk (1993). We now describe calibrators in both directions: from p-values to e-values (p-to-e) and from e-values to p-values (e-to-p). We begin by introducing the concept of *domination* between calibrators.

Definition 1.4. A calibrator f is said to dominate a calibrator g if $f \geq g$ (p-to-e) or $f \leq g$ (e-to-p). A calibrator is *admissible* if it is not strictly dominated by any other calibrator.

A *(p-to-e) calibrator* is a decreasing function $f : [0, \infty) \to [0, \infty]$ satisfying $f = 0$ on $(1, \infty)$ and $\int_0^1 f(p)\mathrm{d}p \leq 1$. Essentially, a calibrator transforms any p-variable to an e-variable. It is *admissible* if it is upper semicontinuous, $f(0) = \infty$, and $\int_0^1 f(p)\mathrm{d}p = 1$. This implies that it is not strictly dominated, in a natural sense, by any other calibrator (Proposition 2.1 and Proposition 2.2 in Vovk and Wang, 2021). Note that the domain $[0, \infty)$ can safely be restricted to $[0, 1]$, since p-values greater than one are irrelevant in hypothesis testing. Examples of admissible p-to-e calibrators include:

$$f(p) = \kappa p^{\kappa-1}, \quad \text{for } \kappa \in (0, 1),$$
$$f(p) = p^{-1/2} - 1.$$

In contrast, the transformation from e-values to p-values is much simpler. The only admissible *e-to-p calibrator*, which converts an e-variable into a p-variable, is the function

$$f(t) = \min\left\{\frac{1}{t}, 1\right\},$$

as proved in Vovk and Wang (2021). For this reason, we will omit the explicit mention of "p-to-e" when referring to calibrators in what follows. It is also worth noting that all calibrators are decreasing functions, reflecting the inverse relationship between the magnitude of p-values and e-values. Although calibrators are interesting tools in their own right, we will see that they play a key role in the combination of dependent p-values Vovk *et al.* (2022b).

1.4 Combining evidence: an overview

We now turn to methods for combining statistical evidence, specifically focusing on the aggregation of p-values or e-values that relate to the same statistical hypothesis. The combination of p-values (or e-values) serves as a fundamental tool in several statistical domains, such as multiple testing (Marcus *et al.*, 1976; Goeman and Solari, 2011), global null testing (Kost and McDermott, 2002), and conformal prediction (Vovk *et al.*, 2005; Vovk, 2015).

This section provides a brief overview of techniques, serving as a preparatory step for the results that follow. The presentation is not intended to be exhaustive, and some of

the methods discussed here will be revisited and expanded upon in a more formal way in subsequent chapters.

1.4.1 Combining p-values

Suppose we are testing the same hypothesis using $K \geq 2$ different statistical tests, resulting in p-values $p_1, \ldots, p_K$. A natural question is how to combine these into a single p-value. More formally, the goal is to identify a function that maps a vector of K p-variables into a valid p-variable. This problem has a long history in the statistical literature, with early solutions proposed by Fisher (1948) and Tippett (1941), among others. Their approaches rely on the assumption that the K p-values are statistically independent. However, in many practical settings, the assumption of independence may be unrealistic or overly restrictive. As a result, it is important to consider methods that remain valid under dependence.

Several methods for combining dependent p-values have been proposed in the literature. Among these, the most widely known approach that makes no assumptions on the dependence structure (i.e., remains valid under arbitrary dependence) is the Bonferroni method, which multiplies the smallest p-value by the number of p-values. An important extension was introduced by Rüger (1978), who showed that the λ-quantile of p-values $p_1, \ldots, p_K$ can be used as a valid combination rule if it is scaled by a factor of $1/\lambda$. For instance, multiplying the median by 2, often referred to as the "twice the median" rule, yields a valid combined p-value under arbitrary dependence. Interestingly, the arithmetic mean of the p-values does not itself yield a valid combined p-value in this setting. However, as shown by Rüschendorf (1982), it can be made valid by applying the same multiplicative correction factor of 2 as with the median.

Additional methods have also been developed for combining p-values under specific forms of dependence. A prominent example is the Simes combination rule (Simes, 1986), which is valid when the p-values satisfy the PDS (positive dependence through stochastic ordering) condition (Sarkar, 2008). In a certain sense, the Simes method can be viewed as selecting the most favorable (i.e., smallest) adjusted p-value from the Rüger family of combinations, but without incurring any additional correction. However, when p-values are arbitrarily dependent, the Simes procedure is no longer valid and requires an approximate multiplicative correction factor of $\log K$, as shown by Hommel (1983). More recently, Chi *et al.* (2025) investigated the behavior of the Simes combination under various notions of negative dependence, further clarifying its range of applicability and its robustness.

Combination Rule	Dependence Structure	Formula
Fisher	Independence	$F^{-1}\left(-2\sum_{k=1}^{K}\log p_k\right)$
Tippet	Independence	$1-(1-\min\{p_1,\ldots,p_K\})^K$
Simes	PDS	$\min_k \frac{K}{k}p_{(k)}$
Bonferroni	Arbitrary dependence	$K\cdot\min\{p_1,\ldots,p_K\}$
Twice the average	Arbitrary dependence	$2\frac{1}{K}\sum_{k=1}^{K}p_k$
Twice the median	Arbitrary dependence	$2\cdot\text{median}\{p_1,\ldots,p_K\}$
Hommel	Arbitrary dependence	$h_K\cdot\min_k \frac{K}{k}p_{(k)}$

TABLE 1.1: Some methods for combining p-values $p_1,\ldots,p_K$ into a single p-value. The second column indicates the dependence structure under which each combination rule is valid. Here, $p_{(k)}$ denotes the k-th smallest value among $p_1,\ldots,p_K$, and F denotes the cumulative distribution function (cdf) of a χ^2 distribution with $2K$ degrees of freedom. The term h_K is defined as $h_K=\sum_{k=1}^{K}\frac{1}{k}\approx\log K$.

Table 1.1 summarizes some of the p-value combination rules discussed above. A more in-depth discussion of methods that remain valid under arbitrary dependence is deferred to the next chapter. For comprehensive overviews of p-value combination techniques, we refer the reader to Owen (2009) and Cousins (2007).

1.4.2 Combining e-values

We now turn to the setting where, instead of p-values, we aim to combine K different e-values. We begin with the case of independent e-variables. When the K e-variables are independent, their product yields a valid e-variable. This follows directly from the properties of the expectation under independence and from the definition of e-variable (Definition 1.3).

When e-variables are dependent, the arithmetic average is a valid method for combining them, thanks to the linearity of expectation. Furthermore, as proved by Wang (2025) and Vovk and Wang (2021), the average is, in fact, the only admissible combination rule under arbitrary dependence. Thus, unlike in Section 1.4.1, the arithmetic mean provides a valid combination method without requiring any multiplicative factor. More generally, any convex combination of e-values retains validity in the dependent

setting. A more detailed discussion on the treatment of multiple e-values can be found in Chapter 8 of Ramdas and Wang (2025) and in Vovk and Wang (2021).

As a remark, when p-values or e-values corresponding to the same hypothesis are combined into a single aggregate value, the validity of the resulting inference depends on whether the assumptions underlying the chosen combination rule are satisfied. If these assumptions hold (for instance, if the p-values are independent and Fisher's method is applied), then the combined statistic preserves its nominal inferential guarantees, meaning that the resulting conclusions remain valid with respect to the underlying data-generating process.

1.5 Confidence and prediction sets

Up to this point, we have focused on hypothesis testing, where the primary objective is to assess whether the observed data provide sufficient evidence to reject a predefined null hypothesis. However, there are many settings in statistical inference where the goal is not to test a hypothesis, but rather to identify a set of elements of $\mathcal{M}$, corresponding to appropriate data generating distributions. In other contexts, the aim may be to construct a region where a future observation (usually Z_{n+1} if data are iid) is likely to fall with high probability. These tasks fall within broader categories such as confidence set constuction and predictive inference, both of which are central to statistical inference.

Suppose that c denotes a target of interest, for example an underlying functional or parameter of interest of the data distribution (e.g., the mean) or an outcome that we want to predict. As a result, c can be a fixed quantity or, in predictive settings, it can itself be a random variable. Then we say that the set $\mathcal{C}$ is a $(1-\alpha)$ confidence, or prediction, set if

$$\mathbb{P}\left(c \in \mathcal{C}\right) \geq 1-\alpha,$$

where $\alpha \in (0,1)$ is a user-chosen miscoverage rate. We say that the set $\mathcal{C}$ has *exact* coverage if $\mathbb{P}(c \in \mathcal{C}) = 1-\alpha$. The construction and the structure of $\mathcal{C}$ may vary depending on the problem at hand: it can be an interval, a multidimensional region, or a more general subset of the parameter or sample space.

There exists a duality between hypothesis testing and confidence sets. In fact, a confidence set with level $1-\alpha$ can be obtained by inverting a class of tests with level α; see, for example, Section 9.2 in Casella and Berger (2001). Similar results can be obtained, for example, in conformal prediction; where the construction of the prediction set can be rephrased in terms of conformal p-values (Lei *et al.*, 2018). We will discuss this setting in Chapter 4.

1.6 Markov's inequality and extensions

To conclude this chapter, we present a set of inequalities introduced in Ramdas and Manole (2024), which will play a central role in the discussions that follow. We begin by recalling Markov's inequality, which, as discussed in Section 1.2, is fundamental for controlling the type-I error in e-value based testing.

Theorem 1.5 (Markov's Inequality)**.** *Let X be a non-negative random variable, then, for any $a > 0$,*

$$\mathbb{P}(X \geq 1/a) \leq a\mathbb{E}[X].$$

The following inequalities can be viewed as an extension of Markov's inequality.

Theorem 1.6 (Exchangeable Markov Inequality)**.** *Let $X_1, X_2, \ldots$ form an exchangeable sequence of non-negative and integrable random variables. Then, for any $a > 0$,*

$$\mathbb{P}\left(\exists k \geq 1 : \frac{1}{k}\sum_{i=1}^{k} X_i \geq \frac{1}{a}\right) \leq a\mathbb{E}[X_1].$$

In addition, let $X_1, \ldots, X_K$ be exchangeable, non-negative and integrable random variables. Then, for any $a > 0$,

$$\mathbb{P}\left(\exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} X_i \geq \frac{1}{a}\right) \leq a\mathbb{E}[X_1].$$

The subsequent inequality is based on an external randomization of the threshold of Markov's inequality. Let $\mathcal{U}$ be the set of all uniform random variables on $[0, 1]$.

Theorem 1.7 (Uniformly-randomized Markov Inequality)**.** *Let X be a non-negative random variable independent of $U \in \mathcal{U}$. Then, for any $a > 0$,*

$$\mathbb{P}(X \geq U/a) \leq a\mathbb{E}[X]. \tag{1.3}$$

The results can also be expressed in additive terms; in fact (1.3) can be rephrased as

$$\mathbb{P}(X \geq 1/a - U') \leq a\mathbb{E}[X],$$

where U' is uniformly distributed in the interval $[0, 1/a]$. The equivalence is explained by the fact that $U' = U/a$ and $1 - U \overset{d}{=} U$. The third inequality combines the previous two theorems in the following way:

Theorem 1.8 (Exchangeable and Uniformly-randomized Markov Inequality)**.** *Let $X_1, \ldots, X_K$ be a set of exchangeable and non-negative random variables independent of $U \in \mathcal{U}$. Then, for any $a > 0$,*

$$\mathbb{P}\left(X_1 \geq U/a \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} X_i \geq \frac{1}{a}\right) \leq a\mathbb{E}[X_1].$$

These inequalities will serve as fundamental technical tools throughout the next chapters.

Chapter 2

Combining exchangeable p-values

The main objective of the chapter is to introduce and describe simple new valid merging methods valid under the assumption of exchangeability of the input p-values, which are more powerful than methods that assume arbitrary dependence. Implied by the relative strength of the dependence assumptions, the methods will be incomparable to methods assuming negative or positive dependence, and less powerful than methods assuming independence. However, specialized methods for handling exchangeable dependence are quite practically relevant. Such dependence is encountered, for example, in statistical testing via sample splitting, as we now elaborate.

There are at least two different reasons for which sample splitting is used: the first is to relax the assumptions needed to obtain theoretical guarantees, while the second is to reduce computational costs. Some examples of such procedures are Cox (1975); Wasserman and Roeder (2009); Banerjee *et al.* (2019); Wasserman *et al.* (2020); Shafer and Vovk (2008); Shekhar *et al.* (2022, 2023); Kim and Ramdas (2024). The drawback of these methods based on sample-splitting is that the obtained p-values are affected by the randomness of the split. Meinshausen *et al.* (2009) called this phenomenon as a *p-value lottery*. So one may instead repeat the same sample splitting procedure several times to obtain multiple p-values, which are exchangeable by design of the procedure.

One can of course combine such p-values by using the earlier mentioned rules (like twice the average) for arbitrarily dependent p-values (see Chapter 1). But we hope to do better by exploiting their exchangeability. In a paper that seemingly dampens that hope, Choi and Kim (2023) showed that in the rule of "twice the average", the constant factor of 2 cannot be improved even under exchangeability. However, their result does not imply that "twice the average" cannot be improved; it simply states that any such improvement cannot proceed by attempting to lower the constant of 2. Indeed, we will improve on this well known rule (and many others), but we proceed differently, not by

lowering the constant. Instead, we calculate twice the average of the first k p-values and take a minimum over all k (see Table 2.1).

We also show that no symmetric rule for merging arbitrarily dependent p-values (like the ones mentioned earlier) can be improved under exchangeability. To achieve any improvement, we must consider asymmetric rules that process the p-values in a particular order. This might initially appear paradoxical given the exchangeability of the p-values, but it is easily sorted out. In many practical settings, these exchangeable p-values can be generated one by one by repeating the same randomized procedure many times, generating a stream of p-values. In this case, our combination rules would simply process these p-values in the order that they are generated. This seems quite appropriate, and the main advantage of doing this is increased power over processing them symmetrically as a batch. In fact, as we discuss, we do not need to fix the number of p-values ahead of time, they can just be processed online, yielding a p-value whenever this procedure is stopped. This makes our merging rules particularly simple and practical.

The problem of combining such p-values from repeated sample splitting has been studied by other authors, such as DiCiccio *et al.* (2020), but the combination rules presented here are more powerful, and also more general and systematic because they apply more broadly. The same problem has also been studied in Guo and Shah (2024), where the authors propose a combination method based on subsampling to merge test statistics based on different random splits. Unlike our nonasymptotic guarantees that work directly with the p-values and are cheap to compute, their work provides only asymptotic guarantees under certain additional assumptions (like an asymptotic pivotal null distribution for their test statistics) and they require access to the full dataset on which to perform expensive subsampling-based recomputations. However, when their additional assumptions hold, their procedure can be expected to be more powerful than ours because it estimates and exploits the joint distribution of the p-values. With a similar aim, Ritzwoller and Romano (2023) investigates a method to enhance the reproducibility of statistical results obtained through sample-splitting. Specifically, their algorithm sequentially aggregates statistics across multiple sample splits until the variability induced by the different splits falls below a defined threshold.

It is perhaps interesting that all the aforementioned methods for merging under arbitrary dependence or under exchangeability are actually inadmissible when randomization is permitted. Randomization in the context of hypothesis testing is not new and is used, for example, in discrete tests; some examples are Fisher's exact test (Fisher, 1934) or the randomized test for a binomial proportion proposed in Stevens (1950). We

will see how the introduction of a simple external randomization (an independent uniform random variable and/or uniform permutation) can improve the existing merging rules for arbitrary dependence, as well as our new rules for exchangeable merging.

In terms of technical aspects, one of our main contributions is to point out explicitly how existing merging rules for arbitrary dependence are actually recovered in a unified manner: by transforming the p-values into e-values (Vovk and Wang, 2021; Wasserman *et al.*, 2020; Grünwald *et al.*, 2024) using different "calibrators", averaging the resulting e-values and finally applying Markov's inequality. This connection is particularly important, because then the improvements under exchangeability, or by randomization, are then achieved by invoking the "exchangeable Markov inequality" and "uniformly randomized Markov inequality".

The rest of the chapter is organized as follows. In Section 2.1, we introduce the notation and tools necessary for the chapter. In Section 2.2, the main results are presented in a general way, focusing on two distinct aspects: the first part addresses the case of exchangeable p-values, while the second part describes novel findings under the assumption of arbitrarily dependent p-values when randomization is allowed. Subsequent sections investigate the implications of these results across various p-merging functions commonly found in the literature. Specifically, Section 2.3 and Section 2.4 delve into the combination proposed by Rüger (1978) and Hommel (1983), respectively. Section 2.5 examines the case of arithmetic mean. The following two sections address two additional scenarios within the family of generalized means: namely, the harmonic mean (Section 2.6) and the geometric mean (Section 2.7). Section 2.8 presents some simulation results.

Before proceeding with the results, Table 2.1 presents some notable combination rules introduced in the literature and their corresponding exchangeable and randomized versions introduced in the following sections. These results are first derived in a general form and then discussed case-by-case.

2.1 Problem setup

Without loss of generality, let $(\Omega, \mathcal{F}, \mathbb{P})$ be an atomless probability space[1], and this is implicitly assumed in almost all papers in statistics; see Vovk and Wang (2021, Appendix D) for related results and discussions. Let $\mathcal{U}$ be the set of all uniform random variables on $[0,1]$ under $\mathbb{P}$. In the following, $K \geq 2$ is an integer. We use the shorthand notation

[1]A probability space is atomless if there exists a random variable on this space that is uniformly distributed on $[0,1]$.

Combination rule	Arbitrary dependence (known)	Exchangeability (new)	Arbitrary dependence, randomized (new)
Rüger combination	$\frac{K}{k} p_{(k)}$	$\frac{K}{k} \bigwedge_{m=1}^{K} p^m_{(\lambda_m)}$	$\frac{K}{k} p_{(\lceil Uk \rceil)}$
Arithmetic mean	$2A(\mathbf{p})$	$2 \bigwedge_{m=1}^{K} A(\mathbf{p}_m)$	$\frac{2}{2-U} A(\mathbf{p})$
Geometric mean	$eG(\mathbf{p})$	$e \bigwedge_{m=1}^{K} G(\mathbf{p}_m)$	$e^U G(\mathbf{p})$
Harmonic mean	$T_K^+ H(\mathbf{p})$	$T_K^+ \bigwedge_{m=1}^{K} H(\mathbf{p}_m)$	$(T_K U + 1) H(\mathbf{p})$

TABLE 2.1: Some combination rules for arbitrarily dependent p-values documented in literature, along with their exchangeable and randomized improvements. If randomization is permitted, one can also improve the existing rules for combining arbitrarily dependent p-values by using the exchangeable combination rule applied to a random permutation of the p-values (this is not presented as a separate column). Here, $\mathbf{p} = (p_1, \dots, p_K)$ denotes the vector of p-values, and $\mathbf{p}_m$ represents the vector containing the first m values of $\mathbf{p}$. In the table, $p_{(k)}$ is the k-th smallest value of $\mathbf{p}$, while $p^m_{(\lambda_m)}$ is the $\lambda_m = \lceil m \frac{k}{K} \rceil$ ordered value of $\mathbf{p}_m$. The random variable U is uniformly distributed in the interval $[0, 1]$. Additionally, A, G and H respectively denote the arithmetic mean, the geometric mean, and the harmonic mean. The values T_K and T_K^+ are given by $T_K = \log K + \log \log K + 1$ and $T_K^+ = T_K + 1$, for $K \geq 2$.

$x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$, $\bigvee_{k=1}^{K} x_k = \max\{x_1, \dots, x_K\}$, and $\bigwedge_{k=1}^{K} x_k = \min\{x_1, \dots, x_K\}$.

We fix $\mathcal{P} = \{\mathbb{P}\}$ throughout, and omit "for testing $\mathbb{P}$" when discussing p-variables and e-variables; we do not distinguish them from the commonly used terms "p-values" and "e-values", and this should create no confusion.

Our starting point is a collection of K p-variables $\mathbf{P} = (P_1, \dots, P_K)$ and we denote their observed (realized) values by $\mathbf{p} = (p_1, \dots, p_K)$. Borrowing terminology from Vovk *et al.* (2022b) and Vovk and Wang (2020), we have that a *p-merging function* is an increasing Borel function $F : [0, \infty)^{K+1} \to [0, \infty)$ such that $\mathbb{P}(F(\mathbf{P}) \leq \alpha) \leq \alpha$ whenever $P_1, \dots, P_K$ are p-variables. In other words, the function F, starting from K p-values, returns a p-value. A p-merging function is *symmetric* if $F(\mathbf{p})$ is invariant under any permutation of $\mathbf{p}$, and it is *homogeneous* if $F(\gamma \mathbf{p}) = \gamma F(\mathbf{p})$ for all $\mathbf{p}$ with $F(\mathbf{p}) \leq 1$ and $\gamma \in (0, 1]$. The class of homogeneous p-merging functions encompasses the *O-family* based on quantiles introduced in Rüger (1978), the Hommel's combination and the *M-family* introduced in Vovk and Wang (2020). We now introduce the notion of

domination in the context of p-merging functions.

Definition 2.1. A function F dominates (interpreted as better being smaller) a function G if

$$F(\mathbf{p}) \leq G(\mathbf{p}), \quad \text{for all } \mathbf{p},$$

and the domination is strict if $F(\mathbf{p}) < G(\mathbf{p})$, for at least one $\mathbf{p}$. A p-merging function F is *admissible* if it is not strictly dominated by any other p-merging function.

Although we defined p-values and e-values for a single probability measure $\mathbb{P}$, all results hold for composite hypotheses. More precisely, if $\mathbf{P}$ is a vector of p-variables for a composite hypothesis and F is a p-merging function, then $F(\mathbf{P})$ is a p-variable for the same composite hypothesis. See Vovk and Wang (2021) for precise definitions and related discussions.

For any function $F : [0, \infty)^K \to [0, \infty)$ and $\alpha \in (0, 1)$, let its rejection region at level α be given by

$$R_\alpha(F) := \left\{\mathbf{p} \in [0, \infty)^K : F(\mathbf{p}) \leq \alpha\right\}.$$

For any homogeneous F, $R_\alpha(F)$ for $\alpha \in (0, 1)$ takes the form $R_\alpha(F) = \alpha A$ for some $A \subseteq [0, \infty)^K$, where αA means the set $\{\alpha\, \mathbf{x} : \mathbf{x} \in A\}$.

Conversely, any increasing collection of Borel lower sets $\{R_\alpha \subseteq [0, \infty)^K : \alpha \in (0, 1)\}$ determines an increasing Borel function $F : [0, \infty)^K \to [0, 1]$ by the equation

$$F(\mathbf{p}) = \inf\{\alpha \in (0, 1) : \mathbf{p} \in R_\alpha\},$$

with the convention $\inf \varnothing = 1$ (throughout). It is immediate to see that F is a p-merging function if and only if $\mathbb{P}(\mathbf{P} \in R_\alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ and $\mathbf{P} \in \mathcal{U}^K$, where $\mathcal{U}^K$ is the set of all K-dimensional random vectors with components in $\mathcal{U}$.

Below, Δ_K is the standard K-simplex. Every admissible homogeneous p-merging function possesses a dual formulation expressed in terms of calibrators, as summarized below.

Theorem 2.2 (Vovk *et al.* (2022b); Theorem 5.1)**.** *For any admissible homogeneous p-merging function F, there exist $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and admissible calibrators $f_1, \dots, f_K$ such that*

$$R_\alpha(F) = \left\{\mathbf{p} \in [0, \infty)^K : \sum_{k=1}^{K} \lambda_k f_k\left(\frac{p_k}{\alpha}\right) \geq 1\right\} \qquad \text{for each } \alpha \in (0, 1). \tag{2.1}$$

Conversely, for any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and calibrators $f_1, \dots, f_K$, (2.1) determines a homogeneous p-merging function.

We will exploit this dual form to implement our "randomized" and "exchangeable" techniques, generating p-merging functions that consistently give smaller p-values (usually strictly) than those produced by their *original* counterparts. From (2.1), it is worth noting that $\sum_{k=1}^{K} \lambda_k f_k(P_k)$ is an e-value. We now present a useful lemma.

Lemma 2.3. *Let $f_1, \dots, f_K$ be K calibrators and $\mathbf{P} \in \mathcal{U}^K$. Then, for any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and $\alpha \in (0, 1]$,*

$$\frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\alpha} \right) \tag{2.2}$$

is an e-variable. If $f_1, \dots, f_K$ are admissible calibrators, then $\mathbb{E}\left[\frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\alpha}\right)\right] = 1$.

Proof. By definition, the quantity in (2.2) is non-negative. In addition, for any $\alpha \in (0, 1]$,

$$\begin{aligned} \mathbb{E}\left[\frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\alpha} \right) \right] &= \frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k \mathbb{E}\left[f_k \left(\frac{P_k}{\alpha} \right) \right] = \frac{1}{\alpha} \sum_{k=1}^{K} \lambda_k \int_0^{\alpha} f_k \left(\frac{p}{\alpha} \right) \mathrm{d}p \\ &= \sum_{k=1}^{K} \lambda_k \int_0^{1} f_k \left(p \right) \mathrm{d}p \leq 1. \end{aligned}$$

If the calibrators are admissible, one can see that the equality holds since $\int_0^1 f_k(p)\,\mathrm{d}p = 1$ for each k. □

Lemma 2.3 used the fact that a calibrator takes value 0 on $(1, \infty)$, which is not restrictive because the relevant range of p-values is $[0, 1]$. This condition is assumed when defining calibrators in Vovk *et al.* (2022b).

In particular, choosing $\lambda_1 = 1$ and $\lambda_k = 0$, for $k \geq 2$, we have that $(1/\alpha) f_1(P_1/\alpha)$ is an e-value, for all $\alpha \in (0, 1]$.

2.2 New results on merging p-values

2.2.1 Exchangeable p-values

Assuming iid p-values is often overly stringent in numerous practical applications. A more pragmatic and less restrictive assumption is exchangeability of p-values, indicating that the distribution of the p-values is unchanged under a permutation of the indices. Formally,

$$(P_1, \dots, P_K) \overset{d}{=} (P_{\sigma(1)}, \dots, P_{\sigma(K)}),$$

where $\overset{d}{=}$ represents equality in distribution while $\sigma : \{1, \dots, K\} \to \{1, \dots, K\}$ is any permutation of the indices. As discussed in the introduction of the chapter, this situation

is encountered in statistical testing using repeated sample splitting (repeated K times on the same data in an identical fashion). In this section, we assume that the sequence of p-variables $\mathbf{P} = (P_1, \ldots, P_K)$ is exchangeable and takes values in $[0,1]^K$.

Remark 2.4. The reader may note that exchangeability can be induced by processing the (potentially non-exchangeable) sequence of p-values $(P_1, \ldots, P_K)$ in a uniformly random order. As a consequence, this implies that if randomization is allowed, it is always possible to satisfy the exchangeability assumption even if the starting sequence has an arbitrary dependence. Stated alternatively, exchangeable combination rules can be safely applied to arbitrarily dependent p-values if the p-values are presented to the analyst in a random order.

We present an extension of the converse direction of Theorem 2.2, which is valid under exchangeability of the vector of p-values. In particular, to preserve exchangeability and use the result stated in Theorem 1.6, it is necessary that the same calibrator f is used for all p-values.

Theorem 2.5. *Let f be a calibrator, and $\mathbf{P} = (P_1, \ldots, P_K) \in \mathcal{U}^K$ be exchangeable. For each $\alpha \in (0,1)$, we have*

$$\mathbb{P}\left(\exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \geq 1\right) \leq \alpha.$$

Proof. The proof involves the use of the exchangeable Markov inequality (EMI) recalled in Theorem 1.6 for finite sequences:

$$\mathbb{P}\left(\exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \geq 1\right) \overset{(i)}{\leq} \mathbb{E}\left[f\left(\frac{P_1}{\alpha}\right)\right] = \alpha\mathbb{E}\left[\frac{1}{\alpha}f\left(\frac{P_1}{\alpha}\right)\right] \overset{(ii)}{\leq} \alpha,$$

where (i) is due to EMI while (ii) holds due to Lemma 2.3. □

We first give a formal definition of ex-p-merging function, which is a function that yields a p-value if its argument is a vector of exchangeable p-values.

Definition 2.6. An *ex-p-merging function* is an increasing Borel function $F : [0,1]^K \to [0,1]$ such that $\mathbb{P}(F(\mathbf{P}) \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ and $\mathbf{P} \in \mathcal{U}^K$ that is exchangeable. It is *homogeneous* if $F(\gamma\mathbf{p}) = \gamma F(\mathbf{p})$ for all $\gamma \in (0,1]$ and $\mathbf{p} \in [0,1]^K$. An ex-p-merging function F is *admissible* if for any ex-p-merging function G, $G \leq F$ implies $G = F$.

We now use the result given in Theorem 2.5 to derive better p-merging functions by exploiting the duality between rejection regions and p-merging functions. In particular, to derive an ex-p-merging the following steps are involved: Initially, the rejection regions

at level α based on a given calibrator are determined. As explained in Section 2.1, the ex-p-merging is established by choosing the smallest α for which the p-values $\mathbf{p}$ falls within the rejection region R_α. In the first step, the result stated in Theorem 2.5 helps to derive functions that dominate their counterpart valid under arbitrary dependence. To elaborate, starting from a calibrator f and $\alpha \in (0,1)$, we define the exchangeable rejection region

$$R_\alpha = \left\{ \mathbf{p} \in [0,1]^K : \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \text{ for some } k \leq K \right\}.$$

Using R_α, we can define the function $F : [0,1]^K \to [0,1]$ by

$$\begin{aligned} F(\mathbf{p}) &= \inf\{\alpha \in (0,1) : \mathbf{p} \in R_\alpha\} \\ &= \inf\left\{ \alpha \in (0,1) : \exists k \leq K \text{ s.t. } \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \right\} \\ &= \inf\left\{ \alpha \in (0,1) : \bigvee_{k=1}^{K} \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \right\}, \end{aligned} \tag{2.3}$$

where, and throughout, $\bigvee_{k=1}^{K} \frac{1}{k} \sum_{i=1}^{k} f(\frac{p_i}{\alpha})$ should be understood as $\bigvee_{k=1}^{K} (\frac{1}{k} \sum_{i=1}^{k} f(\frac{p_i}{\alpha}))$. Note that (2.3) is always smaller or equal than the p-merging function given by (2.1)

$$F'(\mathbf{p}) = \inf\left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} f\left(\frac{p_k}{\alpha}\right) \geq 1 \right\},$$

which is valid for p-values with an arbitrary dependence. This is particularly important since all admissible homogeneous and symmetric p-merging functions have the form F' for some admissible calibrator (a symmetric version of Theorem 2.2; see Vovk *et al.* (2022b)). This implies that the function defined in (2.3) dominates the function F'. In the following theorem we prove that (2.3) is a homogeneous ex-p-merging function.

Theorem 2.7. *If f is a calibrator and $\mathbf{P} \in \mathcal{U}^K$ is an exchangeable sequence, then F in* (2.3) *is a homogeneous ex-p-merging function.*

Proof. It is clear that F is increasing and Borel since R_α is a lower set. For an exchangeable sequence $\mathbf{P} \in \mathcal{U}^K$ and $\alpha \in (0,1)$, using Theorem 2.5 and the fact that $(R_\beta)_{\beta \in (0,1)}$

is nested, we have

$$\begin{aligned}
\mathbb{P}\left(F(\mathbf{P}) \leq \alpha\right) &= \mathbb{P}\big(\inf\{\beta \in (0,1) : \mathbf{P} \in R_\beta\} \leq \alpha\big) \\
&= \mathbb{P}\left(\inf\left\{\beta \in (0,1) : \exists k \leq K \text{ such that } \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1\right\} \leq \alpha\right) \\
&= \mathbb{P}\left(\bigcap_{\beta>\alpha}\left\{\exists k \leq K : \left(\frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right)\right) \geq 1\right\}\right) \\
&= \inf_{\beta>\alpha} \mathbb{P}\left(\exists k \leq K : \left(\frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right)\right) \geq 1\right) \leq \inf_{\beta>\alpha} \beta = \alpha.
\end{aligned}$$

Therefore F is a valid p-merging function. Homogeneity comes directly from the definition of (2.3). □

It is clear from (2.3) that the function depends on the order of the values in $\mathbf{p}$, and hence $F(\mathbf{p})$ is not a symmetric function. This is not a coincidence: in the next result, we show that any symmetric ex-p-merging function is actually valid under arbitrary dependence, and hence they cannot improve over the admissible p-merging functions studied by Vovk *et al.* (2022b). In particular, this implies that under exchangeability the multiplicative factor 2 for the arithmetic average cannot be improved, as earlier noted by Choi and Kim (2023), but it also extends their result to every other symmetric merging function.

Proposition 2.8. *A symmetric ex-p-merging function is necessarily a p-merging function. Hence, for an ex-p-merging function to strictly dominate an admissible p-merging function, it cannot be symmetric.*

Proof. Let $\mathbf{P} \in \mathcal{U}^K$, and let σ be a random permutation of $\{1, \dots, K\}$, uniformly drawn from all permutations of $\{1, \dots, K\}$ and independent of $\mathbf{P}$. Let $\mathbf{P}^\sigma = (P_{\sigma(1)}, \dots, P_{\sigma(K)})$. Note that $\mathbf{P}^\sigma$ is exchangeable by construction. If F is a symmetric ex-p-merging function, it must satisfy $F(\mathbf{P}^\sigma) = F(\mathbf{P})$. Because $F(\mathbf{P}^\sigma)$ is a p-variable, so is $F(\mathbf{P})$, showing that F is a p-merging function. □

Clearly, Proposition 2.8 implies that under symmetry, a function is a p-merging function if and only if it is an *ex-p-merging function*. Hence, to take advantage of the exchangeability of the p-values (over arbitrary dependence), one necessarily deviates from symmetric ways of merging p-values, as done in Theorem 2.7. More importantly, the proof of Proposition 2.8 illustrates the idea (mentioned in Remark 2.4) that for arbitrarily dependent p-values, we can first randomly permute them and then apply an

ex-p-merging function (not necessarily symmetric) such as the one in Theorem 2.7, to obtain a p-value.

The next result gives a simple condition on the calibrator f that guarantees that the probability of rejection using F in (2.3) is sharp for some $\mathbf{P}$.

Proposition 2.9. *Suppose that f is a convex admissible calibrator with $f(0+) \le K$ and $f(1) = 0$, and F is in (2.3). For $\alpha \in (0,1)$, there exists an exchangeable $\mathbf{P} \in \mathcal{U}^K$ such that $\mathbb{P}(F(\mathbf{P}) \le \alpha) = \alpha$.*

Proof. Take $U \in \mathcal{U}$ and an event A with $\mathbb{P}(A) = \alpha$ independent of U. Let $b = f(0+) \le K$. The condition on f guarantees that $f(U)$ is a random variable with support $[0, b]$, mean 1 and a decreasing density. The above conditions, using Theorem 3.2 of Wang and Wang (2016), guarantee that there exists $\mathbf{U} = (U_1, \dots, U_K) \in \mathcal{U}^K$ such that $\mathbb{P}(\sum_{i=1}^K f(U_i) = K) = 1$. We assume that $U, A, \mathbf{U}$ are mutually independent; this is possible as we are only concerned with distributions. Taking a uniformly drawn random permutation further allows us to assume that $\mathbf{U}$ is exchangeable. Let $P_i = \alpha U_i \mathbb{1}_A + (\alpha + (1-\alpha)U)\mathbb{1}_{A^c}$ for $i = 1, \dots, K$. It is clear that each $P_i \in \mathcal{U}$ and $\mathbf{P} = (P_1, \dots, P_K)$ is exchangeable. Moreover, by the definition of F,

$$\mathbb{P}(F(\mathbf{P}) \le \alpha) \ge \mathbb{P}\left(\frac{1}{K}\sum_{i=1}^K f(P_i/\alpha) \ge 1\right) = \mathbb{P}(A)\mathbb{P}\left(\sum_{i=1}^K f(U_i) = K\right) = \alpha.$$

The other inequality $\mathbb{P}(F(\mathbf{P}) \le \alpha) \le \alpha$ follows from Theorem 2.5. □

Remark 2.10. Proposition 2.9 is not sufficient to justify admissibility of F in (2.3). In general, admissibility of *ex-p-merging functions* remains unclear. For instance, take F in (2.3) with $f(p) = (2-2p)_+$, corresponding to the arithmetic average, as in Section 2.5 below. If $K = 2$, then F is not admissible as it is strictly dominated by the Bonferroni p-merging function given by $F_{\mathrm{Bonf}}(p_1, \dots, p_K) = K \min\{p_1, \dots, p_K\}$. For $K \ge 3$, F and F_{Bonf} are not comparable.

2.2.2 Homogeneity of p-merging functions

As seen from Theorem 2.7, the class of ex-p-merging functions we obtained in (2.3) are homogeneous. Indeed, all explicit p-merging functions in the literature are homogeneous; see Vovk *et al.* (2022b) for many examples. In the next result, we show a very special feature of homogeneous p-merging functions, justifying their relevance in applications.

Theorem 2.11. *Let F be a p-merging function. For any $\alpha \in (0,1)$, there exists a homogeneous p-merging function G such that $R_\alpha(F) \subseteq R_\alpha(G)$.*

Proof. We say that a set $R \subseteq [0,\infty)^K$ is a decreasing set if $\mathbf{x} \in R$ implies $\mathbf{y} \in R$ for all $\mathbf{y} \in [0,\infty)^K$ with $\mathbf{y} \le \mathbf{x}$ (componentwise).

Fix $\alpha \in (0,1)$, and note that $R_\alpha(F) = \{\mathbf{p} \in [0,\infty)^K : F(\mathbf{p}) \le \alpha\}$ is a decreasing set. Define

$$G(\mathbf{p}) = \inf\left\{\beta \in (0,1) : \mathbf{p} \in \frac{\beta}{\alpha} R_\alpha(F)\right\}, \quad \mathbf{p} \in [0,\infty)^K.$$

First, G is homogeneous, which follows from its definition. Second, $\mathbf{p} \in R_\alpha(F)$ implies $G(\mathbf{p}) \le \alpha$, and hence $R_\alpha(F) \subseteq R_\alpha(G)$. Third, G is increasing because $R_\alpha(F)$ is a decreasing set.

It remains to show that G is a p-merging function. The definition of G gives

$$G(\mathbf{p}) \le \beta \iff \mathbf{p} \in \bigcap_{\gamma > \beta} \frac{\gamma}{\alpha} R_\alpha(F).$$

If we can show

$$\mathbb{P}\left(\mathbf{P} \in \frac{\beta}{\alpha} R_\alpha(F)\right) \le \beta \quad \text{for all } \mathbf{P} \in \mathcal{U}^K \text{ and } \beta \in (0,1), \tag{2.4}$$

then

$$\mathbb{P}(G(\mathbf{P}) \le \beta) \le \inf_{\gamma > \beta} \mathbb{P}\left(\mathbf{P} \in \frac{\gamma}{\alpha} R_\alpha(F)\right) \le \inf_{\gamma > \beta} \gamma = \beta,$$

Lemma 2.12. *Let $R \subseteq [0,\infty)^K$ be a decreasing Borel set. For any $\beta \in (0,1)$, we have*

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \beta R) \ge \beta \iff \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R) = 1.$$

Proof of the lemma. Let $\mathcal{V}$ be the set of p-variables for $\mathbb{P}$. For any decreasing set L, we have

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in L) = \sup_{\mathbf{P} \in \mathcal{V}^K} \mathbb{P}(\mathbf{P} \in L). \tag{2.5}$$

This fact will be repeatedly used in the proof below.

We first prove the $\Leftarrow$ direction by contraposition. Suppose

$$\gamma := \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \beta R) < \beta.$$

Take an event A with probability β and any $\mathbf{P} \in \mathcal{U}^K$ independent of A. Define $\mathbf{P}^*$ by

$$\mathbf{P}^* = \beta \mathbf{P} \times \mathbb{1}_A + \mathbf{1} \times \mathbb{1}_{A^c}.$$

It is straightforward to check $\mathbf{P}^* \in \mathcal{V}^K$. Hence, by (2.5),

$$\beta \mathbb{P}\left(\mathbf{P} \in R\right) = \mathbb{P}(A)\mathbb{P}\left(\mathbf{P} \in R\right) \leq \mathbb{P}(\mathbf{P}^* \in \beta R) \leq \gamma,$$

and thus $\mathbb{P}(\mathbf{P} \in R) \leq \gamma/\beta$. Since $\gamma/\beta < 1$, this yields $\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R) < 1$ and completes the $\Leftarrow$ direction.

Next we show the $\Rightarrow$ direction. Suppose $\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \beta R) \geq \beta$. For any $\epsilon \in (0, \beta)$, there exists $\mathbf{P} = (P_1, \dots, P_K) \in \mathcal{U}^K$ such that $\mathbb{P}(\mathbf{P} \in \beta R) > \beta - \epsilon$. Let $A = \{\mathbf{P} \in \beta R\}$, $\gamma = \mathbb{P}(A)$, and B be an event containing A with $\mathbb{P}(B) = \beta \vee \gamma$. Let $\mathbf{P}^* = (P_1^*, \dots, P_K^*)$ follow the conditional distribution of $\mathbf{P}/\beta$ given B. We have

$$\mathbb{P}(\mathbf{P}^* \in R) = \mathbb{P}(\mathbf{P} \in \beta R \mid B) = \mathbb{P}(A \mid B) = \frac{\gamma}{\beta \vee \gamma}.$$

Note that for $k \in \{1, \dots, K\}$,

$$\mathbb{P}(P_k^* \leq p) = \mathbb{P}(P_k/\beta \leq p \mid B) \leq \frac{\mathbb{P}(P_k \leq \beta p)}{\mathbb{P}(B)} = \frac{\beta p}{\beta \vee \gamma} \leq p,$$

and hence $\mathbf{P}^* \in \mathcal{V}^K$. Since $\gamma > \beta - \epsilon$ and $\epsilon \in (0, \beta)$ is arbitrary, we can conclude $\sup_{\mathbf{P} \in \mathcal{V}^K} \mathbb{P}(\mathbf{P} \in R) = 1$, yielding $\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R) = 1$ via (2.5). □

Now we resume the proof of Theorem 2.11. Fix $\beta \in (0, 1)$. Take any $\lambda > 1$ such that $\lambda(\beta \vee \alpha) < 1$. Lemma 2.12 yields, for any decreasing set R,

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\beta R) < \lambda\beta \iff \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\alpha R) < \lambda\alpha. \tag{2.6}$$

Let $R = R_\alpha(F)/(\alpha\lambda)$, which is a decreasing set. Note that

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\alpha R) = \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in R_\alpha(F)) \leq \alpha < \lambda\alpha$$

and this leads to, by using (2.6),

$$\sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}\left(\mathbf{P} \in \frac{\beta}{\alpha} R_\alpha(F)\right) = \sup_{\mathbf{P} \in \mathcal{U}^K} \mathbb{P}(\mathbf{P} \in \lambda\beta R) < \lambda\beta.$$

Since $\lambda > 1$ can be arbitrarily close to 1, we conclude that (2.4) holds true, and this completes the proof. □

As a consequence of Theorem 2.11, if the level α is determined before choosing the p-merging function, then it suffices to consider homogeneous ones, since their rejection

sets are at least as larger as those of other p-merging functions. Since all admissible homogeneous p-merging functions have the form in Theorem 2.2, the class of our ex-p-merging functions in (2.3), sharing a similar form to (2.1), is quite broad. Note that Theorem 2.11 does not imply that there exists a homogeneous p-merging function G dominating F in general, because the construction of G depends on the given α.

2.2.3 Sequentially combining a stream of exchangeable p-values

In Section 2.2.1 we have seen that, if our starting vector of p-values $\mathbf{P} = (P_1, \dots, P_K)$ is exchangeable, then it is possible to derive new combination rules exploiting their exchangeability. The technique for developing these new rules involves converting the initial p-values into e-values through an α-dependent calibrator. Following this transformation, the exchangeable Markov inequality (Theorem 1.6) plays a crucial role in formulating these rules. In particular, notice that the inequality in Theorem 1.6 is uniformly valid; therefore, the exchangeable sequence of p-values need not be limited to a set with cardinality K, but it is possible to continue to add p-values and stop when the procedure is stable.

To provide a concrete example, in the case where p-values are obtained by applying a sample-splitting procedure to the same dataset, a researcher can obtain one p-value at a time simply by performing a new data split. The researcher would then aim to combine these p-values sequentially and potentially stop when the procedure appears to stabilize.

We first define the following lemma:

Lemma 2.13. *Let* $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2) \in [0,1]^K$ *be a vector with* $\mathbf{p}_1 \in [0,1]^{K_1}, \mathbf{p}_2 \in [0,1]^{K_2}$ *and* $K = K_1 + K_2$. *In addition, let* F *be the function defined in* (2.3). *Then*

$$F(\mathbf{p}) = F(\mathbf{p}_1, \mathbf{p}_2) \leq F(\mathbf{p}_1).$$

Proof. Fix any $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2) \in [0,1]^K$ and suppose that $F(\mathbf{p}_1) = \alpha$. By definition, we have

$$\exists k \leq K_1 : \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \implies \exists k \leq K_1 + K_2 : \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1.$$

This implies

$$F(\mathbf{p}) = \inf\left\{\beta \in (0,1) : \exists k \leq K \text{ s.t. } \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\beta}\right) \geq 1\right\} \leq \alpha,$$

due to the fact that f is increasing in β. □

The above result implies that the function F in (2.3) is non-increasing as more parameters (i.e., p-values) are added. In addition, the following holds.

Theorem 2.14. *Let $P_1, P_2, ... \in \mathcal{U}^\infty$ be an infinite exchangeable sequence and let F be the function defined in* (2.3)*. Then,*

$$\mathbb{P}(\exists k \geq 1 : F(\mathbf{P}_k) \leq \alpha) \leq \alpha,$$

where $\mathbf{P}_k = (P_1, \ldots, P_k)$.

Proof. From Lemma 2.13, we have that

$$\exists k \leq K : F(\mathbf{p}_k) \leq \alpha \iff F(\mathbf{p}_K) \leq \alpha,$$

where $\mathbf{p}_k = (p_1, \ldots, p_k)$ is the sequence containing the first k values of $\mathbf{p}$. Then we can write,

$$\begin{aligned}
\mathbb{P}\left(\exists k \geq 1 : F(\mathbf{P}_k) \leq \alpha\right) &= \mathbb{P}\left(\bigcup_{K \in \mathbb{N}} \{\exists k \leq K : F(\mathbf{P}_k) \leq \alpha\}\right) \\
&= \lim_{K \to \infty} \mathbb{P}\left(\exists k \leq K : F(\mathbf{P}_k) \leq \alpha\right) \\
&= \lim_{K \to \infty} \mathbb{P}\left(F(\mathbf{P}_K) \leq \alpha\right) \leq \alpha,
\end{aligned}$$

where the last inequality is due to Theorem 2.7. □

The result allows the analyst to either continue collecting new (exchangeable) p-values or to cease based on the outcome. In particular, in the example described at the beginning of the section, a researcher can stop the procedure when the result seems to stabilize. However, there are some issues to consider when applying such a procedure. In particular, some calibrators f depend on the number K of p-values. For example, this is the case for the Hommel combination (Section 2.4) and for the harmonic mean (Section 2.6). As a final caveat, exchangeability becomes more stringent when the number K grows; for instance, in general, for a given K-dimensional exchangeable vector $(P_1, \ldots, P_K)$, there may not exist P_{K+1} such that $(P_1, \ldots, P_{K+1})$ is exchangeable. Luckily, in many practical situations, the procedure that produced the K exchangeable p-values in the first place could also produce more of them; for example, this happens when the p-value was produced by sample splitting.

2.2.4 Randomized p-merging functions

In this subsection, we start with a collection of arbitrarily dependent p-values and we will show how it is possible to enhance existing merging rules using a simple randomization trick. In this case, we denote

$$\mathcal{U}^K \otimes \mathcal{U} = \{(\mathbf{P}, U) \in \mathcal{U}^K \times \mathcal{U} : U \text{ and } \mathbf{P} \text{ are independent}\},$$

and we state a randomized version of the converse direction of Theorem 2.2, by changing the constant 1 in (2.1) to a uniform random variable U.

Theorem 2.15. *Let $f_1, \dots, f_K$ be calibrators and $(P_1, \dots, P_K, U) \in \mathcal{U}^K \otimes \mathcal{U}$. For each $\alpha \in (0,1)$ and $(\lambda_1, \dots, \lambda_K) \in \Delta_K$, we have*

$$\mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right) \geq U\right) \leq \alpha.$$

If $f_1, \dots, f_K$ are admissible calibrators and $\mathbb{P}(\sum_{k=1}^{K} \lambda_k f_k(P_k/\alpha) \leq 1) = 1$, then equality holds

$$\mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\beta}\right) \geq U\right) = \beta \quad \text{for all } \beta \in (0, \alpha]. \tag{2.7}$$

Proof. From direct calculation and using Lemma 2.3,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right) \geq U\right) &= \mathbb{E}\left[\mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right) \geq U\right) \mid \mathbf{P}\right] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right) \wedge 1\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right] = \alpha \mathbb{E}\left[\frac{1}{\alpha}\sum_{k=1}^{K} \lambda_k f_k\left(\frac{P_k}{\alpha}\right)\right] \leq \alpha,
\end{aligned}$$

The equality for $\beta = \alpha$ follows because $\sum_{k=1}^{K} \lambda_k f_k(P_k/\alpha) \leq 1$ and $\int_0^1 f_k(p)\mathrm{d}p = 1$ for each k guarantee the inequalities in the above set of equations are equalities. For $\beta < \alpha$, it suffices to notice that $\sum_{k=1}^{K} \lambda_k f_k(P_k/\alpha)$ is increasing in α. □

The result in Theorem 2.15 is a direct consequence of the uniformly randomized Markov inequality (UMI) introduced by Ramdas and Manole (2024); see Theorem 1.7.

Definition 2.16. A *randomized p-merging function* is an increasing Borel function $F : [0,1]^{K+1} \to [0,1]$ such that $\mathbb{P}(F(\mathbf{P}, U) \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ and $(\mathbf{P}, U) \in \mathcal{U}^K \otimes \mathcal{U}$. It is *homogeneous* if $F(\gamma \mathbf{p}, u) = \gamma F(\mathbf{p}, u)$ for all $\gamma \in (0,1]$ and $(\mathbf{p}, u) \in [0,1]^{K+1}$. A

randomized p-merging function F is admissible if for any randomized p-merging function G, $G \leq F$ implies $G = F$.

Let $f_1, \ldots, f_K$ be calibrators and $(\lambda_1, \ldots, \lambda_K) \in \Delta_K$. For $\alpha \in (0,1)$, define the randomized rejection region by

$$R_\alpha = \left\{ (\mathbf{p}, u) \in [0,1]^{K+1} : \sum_{k=1}^{K} \lambda_k f_k \left(\frac{p_k}{\alpha} \right) \geq u \right\}$$

where we set $f_k(p_k/u) = 0$ if $u = 0$. Using R_α, we can define the function $F : [0,1]^{K+1} \to [0,1]$ by

$$\begin{aligned} F(\mathbf{p}, u) &= \inf\{\alpha \in (0,1) : (\mathbf{p}, u) \in R_\alpha\} \\ &= \inf \left\{ \alpha \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k \left(\frac{p_k}{\alpha} \right) \geq u \right\}, \end{aligned} \tag{2.8}$$

with the convention $0 \times \infty = \infty$ (this guarantees $F(\mathbf{p}, u) = 0$ when any component of $(\mathbf{p}, u)$ is 0).

Theorem 2.17. *If $f_1, \ldots, f_K$ are calibrators and $(\lambda_1, \ldots, \lambda_K) \in \Delta_K$, then F in (2.8) is a homogeneous randomized p-merging function. Moreover, F is lower semicontinuous.*

Proof. It is clear that F is increasing and Borel since R_α is a lower set. For $(\mathbf{P}, U) = (P_1, \ldots, P_K, U) \in \mathcal{U}^K \otimes \mathcal{U}$ and $\alpha \in (0,1)$, using Theorem 2.15 and the fact that $(R_\beta)_{\beta \in (0,1)}$ is nested, we have

$$\begin{aligned} \mathbb{P}(F(\mathbf{P}, U) \leq \alpha) &= \mathbb{P}\left(\inf \left\{ \beta \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\beta} \right) \geq U \right\} \leq \alpha \right) \\ &= \mathbb{P}\left(\bigcap_{\beta > \alpha} \left\{ \sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\beta} \right) \geq U \right\} \right) \\ &= \inf_{\beta > \alpha} \mathbb{P}\left(\sum_{k=1}^{K} \lambda_k f_k \left(\frac{P_k}{\beta} \right) \geq U \right) \leq \inf_{\beta > \alpha} \beta = \alpha. \end{aligned}$$

Therefore, F is a randomized p-merging function. Homogeneity of F follows from (2.8).

Since F is homogeneous and increasing, it is continuous in $\mathbf{p}$. Moreover, for fixed $\mathbf{p} \in [0,1]^K$, since

$$\bigcap_{v < u} \left\{ \alpha \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k \left(\frac{p_k}{\alpha} \right) \geq v \right\} = \left\{ \alpha \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k \left(\frac{p_k}{\alpha} \right) \geq u \right\},$$

we have

$$\lim_{v\uparrow u} F(\mathbf{p}, v) = \liminf_{v\uparrow u} \left\{ \alpha \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k\left(\frac{p_k}{\alpha}\right) \geq v \right\}$$
$$= \inf \bigcap_{v<u} \left\{ \alpha \in (0,1) : \sum_{k=1}^{K} \lambda_k f_k\left(\frac{p_k}{\alpha}\right) \geq v \right\} = F(\mathbf{p}, u).$$

Therefore, $u \mapsto F(\mathbf{p}, u)$ is lower semi-continuous. □

In case of symmetric p-merging functions (i.e., $F(\mathbf{p}, u) = F(\mathbf{q}, u)$ for any permutation $\mathbf{q}$ of $\mathbf{p}$), we have the following corollary, which directly follows from Theorem 2.17.

Corollary 2.18. *For any calibrator f,*

$$F(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} f\left(\frac{p_k}{\alpha}\right) \geq u \right\},$$

is a homogeneous, symmetric randomized p-merging function.

A simple observation is that replacing $f(p)$ with $f(p) \wedge K$ does not change the function F. This observation allows us to only consider calibrators that are bounded above by K, which can improve some existing p-merging functions. This is similar to what was done in Vovk *et al.* (2022b) in the context of deterministic p-merging functions.

Remark 2.19 (P-hacking via repeated derandomization). Randomized methods may not cause a lack of reproducibility if they are part of a standard automated data analysis pipeline without a human in the loop. However, if a human is actively involved in the analysis, then there is a risk of "p-hacking", where a (malevolent) researcher re-runs the randomized method many times in order to obtain a desirable result according to their own utility. In our setting, the issue translates into sampling different U and picking the smallest one (which would be closer to 0 the more times the procedure is re-run). Clearly, this procedure is not valid and can be particularly problematic in the context of confirmatory analysis, where the end product is a binary decision.

Remark 2.20 (Internal randomization). The use of *internal* (as opposed to *external*) randomization can be particularly useful to prevent and mitigate the risk of p-hacking mentioned above. We mention two such strategies below. First, note that F in (2.8) is increasing in each of its arguments. If one has prior information that one of the p-values, say P_1, is independent of the rest (but the rest can be arbitrarily dependent), then one can use P_1 for randomization and apply F (with one less input dimension for $\mathbf{p}$) to $(\mathbf{p}, u) = (P_2, \ldots, P_K, U)$ with $U = P_1$ to obtain a p-value that does not depend

on external randomization. Monotonicity of $u \mapsto F(\mathbf{p}, u)$ guarantees two things. First, a p-variable may be stochastically larger than a standard uniform one, so increasing monotonicity is needed for validity. Second, if P_1 is indeed very small, i.e., it carries signal against the null, then the combined p-value will benefit from this signal. This form of internal randomization has been discussed in Wang (2024, Section B.1).

An alternative method of internal randomization through data (instead of p-values) is discussed by Ramdas and Manole (2024, Section 10.6). To understand their proposal, assume that each p-value is calculated using a function of only the order statistics of iid data; for example, it could be based only on sums, like the t-statistic. In this case, the rank of the first data point (amongst all the data points) is a discrete uniform random variable, and can be used in place of U. This way, if the dataset is itself public (posted by a previous research paper, for example), the ordering itself is not in the hands of the researcher analyzing that data, reducing the risk of p-hacking. We refer interested readers to Ramdas and Manole (2024, Section 10) or Lei and Sudijono (2024) for further discussions.

It is feasible to combine the results presented in Subections 2.2.1 and 2.2.4 through the formulation of novel p-merging functions that exploit both the properties of exchangeability and randomization. These results are presented in Appendix A.1 and are based on the exchangeable and uniformly randomized Markov inequality presented in Theorem 1.8.

2.2.5 Instantiating the above ideas

The ideas above were admittedly somewhat abstract, but provide us with the general tools to improve specific combination rules. The following sections do this for several rules, one by one. We now recall the rules already introduced in Section 1.4.1 that are valid under the assumption of arbitrary dependence. In detail, one of the most commonly employed p-merging functions is the Bonferroni method:

$$F_{\text{Bonf}}(\mathbf{p}) = Kp_{(1)},$$

where $p_{(1)}$ is the minimum of observed p-values. Rüger (1978) extended the aforementioned rule in a more general sense. In particular, it is possible to prove that

$$F_{\text{R}}(\mathbf{p}) := \frac{K}{k} p_{(k)}, \quad k \in \{1, \dots, K\}, \tag{2.9}$$

is a p-value, where $p_{(k)}$ represents the k-th smallest p-value among $(p_1, \dots, p_K)$. In other words, the λ-quantile $p_{(\lceil \lambda K \rceil)}$ is a p-value if multiplied by the factor $1/\lambda$. In particular, the robust and widely used combination rule, twice the median, is part of this class. The next section improves on this combination rule.

Vovk and Wang (2020) introduced the class of p-merging functions based on the generalized mean, also called *M-family*. This general class takes the form

$$a_{r,K} \left(\frac{p_1^r + \cdots + p_K^r}{K} \right)^{1/r}, \tag{2.10}$$

where $r \in \mathbb{R} \setminus \{0\}$ and $a_{r,K}$ is the smallest constant making (2.10) a p-merging function. This class encompasses numerous well-known cases, each distinguished by different values of the parameter r. In particular, if $r = 1$ then (2.10) reduces to the simple average introduced in Rüschendorf (1982) and the value $a_{1,K} = 2$. Another important case is the harmonic mean obtained with $r = -1$. Among this class, the harmonic mean combination rule should be used when substantial dependence among the p-values is suspected. If the dependence is really strong, the arithmetic mean might be a safer option. In the following sections, we demonstrate that if p-values exhibit exchangeability or if randomization is allowed, then it becomes feasible to enhance most of these combination rules.

Before continuing, we provide a few simple examples to highlight the benefits of the proposed findings and demonstrate how our methods can enhance the existing approaches. In the examples we will use some rules proposed in Table 2.1, foreshadowing many results to come.

Example 2.1. *Suppose that the vector* $\mathbf{P}$ *of* 3 *p-values is generated as follows. With* 0.9 *probability,* $\mathbf{P} = (P_1, P_2, P_3)$ *where* P_1, P_2, P_3 *are independent, and with* 0.1 *probability* $\mathbf{P} = (P_4, P_4, P_4)$*. We assume* $P_i \in \mathcal{U}$ *under the null, while under the alternative each* P_i *is distributed as* $\text{Beta}(0.2, 1)$*. The Beta distribution* $(a, 1)$ *with small* $a > 0$ *is a typical model for p-values under the alternative hypothesis; see Sellke* et al. *(2001). Clearly, the p-values are exchangeable under the null. Suppose that one want to use the rule* $(3/2)p_{(2)}$ *to combine the p-values. Then we can check numerically that, fixing the threshold* α *to* 0.05*, the probability of rejection is* 0.5101 *under the alternative. If we use the ex-p-merging derived from the median (see Theorem 2.21 below) the probability of rejection increases to* 0.6207*.*

Example 2.2. *Suppose that we want to test an hypothesis and we have that under the null* $P_1 \in \mathcal{U}$ *while under the alternative* $P_1 \sim \text{Beta}(0.2, 1)$*. The vector of p-values is generated as follows* $\mathbf{P} := (P_1, P_2) = (P_1, 1 - P_1)$*, where* P_2 *is an exact p-value and* $\mathbf{P}$

is exchangeable under the null. In this scenario, the commonly applied twice the mean, even though controls the type-I error, has no power under the alternative hypothesis because the result is always equal to 1*. On the other hand, using the exchangeable rule derived from the arithmetic average (see Theorem 2.27), the probability of rejection is* $\mathbb{P}(P_1 \leq \alpha/2) \approx 0.48$ *under the alternative for* $\alpha = 0.05$*.*

2.2.6 Combining asymptotic p-values

Before proceeding with the remainder of the chapter and introducing new merging rules based on the results presented in the preceding sections, we want to examine the scenario wherein the p-values are asymptotically valid. Many of the results obtained in the literature rely on uniform or super-uniform p-values (see Section 2.1); however, in statistical applications, p-values are often asymptotic, and they are not necessarily valid p-values in finite sample. See, for example, Severini (2000) for an introduction to p-values obtained using likelihood-based methods.

All methods work also for asymptotic p-values, i.e., those that converge in distribution to p-values. Suppose that $(\mathbf{P}_n)_{n\in\mathbb{N}}$ is a sequence of nonnegative random vectors that converges to a vector $\mathbf{P}$ of p-values in distribution. With an upper semicontinuous calibrator f (recall that all admissible calibrators are upper semicontinuous), for each $\alpha \in (0,1)$ and $u \in (0,1)$, the rejection sets R_α given by

$$R_\alpha = \left\{ \mathbf{p} \in [0,\infty)^K : \bigvee_{k\leq K} \frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \right\}$$

or

$$R_\alpha = \left\{ \mathbf{p} \in [0,\infty)^K : \frac{1}{K} \sum_{k=1}^{K} f\left(\frac{p_k}{\alpha}\right) \geq u \right\}$$

are closed. As a consequence, by the Portmanteau Theorem,

$$\limsup_{n\to\infty} \mathbb{P}(\mathbf{P}_n \in R_\alpha) \leq \mathbb{P}(\mathbf{P} \in R_\alpha).$$

Therefore, any methods that produce a p-value for the vector $\mathbf{P}$ of p-values (exchangeable or arbitrarily dependent) produce an asymptotic p-value for any $(\mathbf{P}_n)_{n\in\mathbb{N}}$ that converges to $\mathbf{P}$ in distribution.

2.3 Improving Rüger's combination rule

Vovk *et al.* (2022b) showed that the p-merging function defined in (2.9), with a trivial modification (i.e., return 0 if $p_{(1)} = 0$; see Theorem 7.3 of Vovk *et al.* (2022b)) is admissible for $k \neq K$, and it is admissible among symmetric p-merging functions when $k = K$. In particular, the corresponding calibrator that induces (2.9) is

$$f(p) = \frac{K}{k}\mathbb{1}\{p \in (0, k/K]\} + \infty\mathbb{1}\{p = 0\},$$

and this implies that we can exploit directly the duality between rejection regions and p-merging functions.

2.3.1 An exchangeable Rüger combination rule

If the exchangeability condition is satisfied, then we can obtain something better than the combination rule in (2.9). First, it is clear that $f(p) = \frac{K}{k}\mathbb{1}\{p \in (0, k/K]\} + \infty\mathbb{1}\{p = 0\}$ is an admissible calibrator. We now define the function

$$F_{\mathrm{ER}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{K}{k}\mathbb{1}\left\{\frac{p_i}{\alpha} \leq \frac{k}{K}\right\} \geq 1\right\}.$$

Below, for fixed $k \in \{1, \ldots, K\}$, we let $p^{\ell}_{(\lambda_\ell)}$ denote the $\lceil \ell \frac{k}{K} \rceil$-th ordered value obtained using the first ℓ values of $\mathbf{p}$. Essentially, $p^{\ell}_{(\lambda_\ell)}$ is the upper quantile of order k/K obtained using the first ℓ p-values.

Theorem 2.21. *For any fixed $k \in \{1, \ldots, K\}$ the function F_{ER} satisfies*

$$F_{\mathrm{ER}}(\mathbf{p}) = \left(\frac{K}{k}\bigwedge_{\ell=1}^{K} p^{\ell}_{(\lambda_\ell)}\right)\mathbb{1}\{p_{(1)} > 0\} \text{ for } \mathbf{p} \in [0,1]^K,$$

where $\lambda_\ell := \lceil \ell \frac{k}{K} \rceil$, and it is an ex-p-merging function that dominates F_{R} in (2.9).

Proof. According to Theorem 2.7, it follows that the function F_{ER} is a valid ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. Note that $F_{\mathrm{ER}}(\mathbf{p}) \leq \alpha$ if and only if

$$\exists \ell \leq K : \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{K}{k}\mathbb{1}\left\{\frac{p_i}{\alpha} \leq \frac{k}{K}\right\} \geq 1 \implies \exists \ell \leq K : \sum_{i=1}^{\ell} \mathbb{1}\left\{p_i \leq \alpha\frac{k}{K}\right\} \geq \left\lceil \ell\frac{k}{K}\right\rceil.$$

Where rounding up is due to the fact that the summation takes values in positive integers. This holds true if and only if

$$\exists \ell \leq K : p^{\ell}_{(\lambda_\ell)} \leq \alpha \frac{k}{K},$$

where $p^{\ell}_{(\lambda_\ell)}$ is the $\lceil \ell \frac{k}{K} \rceil$-th ordered value of $(p_1, \ldots, p_\ell)$. Rearranging the terms, we obtain that it is verified when

$$\frac{K}{k} \bigwedge_{\ell=1}^{K} p^{\ell}_{(\lambda_\ell)} \leq \alpha,$$

which complete the first part of the proof. For the second statement, it is possible to note that the element $p^{\ell}_{(\lambda_\ell)}$ in the sequence coincides with $p_{(k)}$ when $\ell = K$. □

It may be useful to note that the Bonferroni rule is not improved using this method. Indeed, fixing $k = 1$, we find that $p^{\ell}_{(\lambda_\ell)}$ reduces to the minimum of the first ℓ p-values subsequently taking the minimum of the obtained sequences coincides with the overall minimum. In addition, Rüger can be sharp, for some exchangeable $\mathbf{P}$, i.e., satisfying $F_{\mathrm{R}}(\mathbf{P}) \in \mathcal{U}$, and so is our proposal.

2.3.2 A randomized Rüger combination rule

In this part, we prove that if randomization is allowed, it becomes feasible to enhance the combination introduced by Rüger (1978), even if the sequence of p-values presents an arbitrary dependence and the obtained result has nice properties in terms of interpretability. As before, we define the merging function

$$F_{\mathrm{UR}}(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{i=1}^{K} \frac{K}{k} \mathbb{1} \left\{ \frac{p_i}{\alpha} \leq \frac{k}{K} \right\} \geq u \right\}.$$

Theorem 2.22. *For any fixed $k \in \{1, \ldots, K\}$, the function F_{UR} satisfies*

$$F_{\mathrm{UR}}(\mathbf{p}, u) = \frac{K}{k} p_{(\lceil uk \rceil)} \mathbb{1}\{p_{(1)} > 0\},$$

and it is a randomized p-merging function that dominates F_{R} in (2.9).

Proof. According to Corollary 2.18, it follows that the function F_{UR} is a valid randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p}, u) \in (0,1]^{K+1}$, then it is possible to note that $F(\mathbf{p}, u) \leq \alpha$ if and only if

$$\frac{1}{K} \sum_{i=1}^{K} \frac{K}{k} \mathbb{1} \left\{ \frac{p_i}{\alpha} \leq \frac{k}{K} \right\} \geq u \implies \sum_{i=1}^{K} \mathbb{1} \left\{ p_i \leq \alpha \frac{k}{K} \right\} \geq \lceil uk \rceil.$$

Rearranging the terms this holds true only if

$$p_{(\lceil uk \rceil)} \leq \alpha \frac{k}{K} \implies \frac{K}{k} p_{(\lceil uk \rceil)} \leq \alpha,$$

which concludes the claim. Since $u \leq 1$ almost surely, we have $p_{(\lceil uk \rceil)} \leq p_{(k)}$. □

The above theorem implies in particular that the Rüger combination rule is inadmissible if external randomization is allowed, despite it being admissible if randomization is not allowed (Vovk *et al.*, 2022b). The fact that $p_{(\lceil UK \rceil)}$ is a p-value is particularly interesting. It has a simple interpretation: sort the p-values and pick the one at a uniformly random index. In addition, for the latter combination rule (2.7) holds for all $\beta \in (0, 1]$.

It is worth noting that when $k = 1$ the Rüger combination rule reduces to the Bonferroni method; however, the introduction of a randomization does not yield any practical benefit since $\lceil U \rceil = 1$.

2.4 Improving Hommel's combination rule

The method proposed in Section 2.3 requires us to choose the value of k in advance; a solution that solves the problem is proposed by Hommel (1983). Hommel's combination rule is given by

$$F'_{\mathrm{Hom}}(\mathbf{p}) := h_K \bigwedge_{k=1}^{K} F_{\mathrm{R}}(\mathbf{p}; k) = \left(\sum_{k=1}^{K} \frac{1}{k} \right) \bigwedge_{k=1}^{K} \frac{K}{k} p_{(k)}, \tag{2.11}$$

with $h_K = \sum_{k=1}^{K} \frac{1}{k}$. This function allows selecting the minimum derived from the combinations based on ordered statistics with a multiplicative cost of $h_K \approx \log K$.

It is possible to prove that the Hommel combination rule is not admissible and is dominated by the *grid harmonic merging function* introduced in Vovk *et al.* (2022b). For completeness, we state here a useful lemma.

Lemma 2.23. *Let f be a function defined by*

$$f(p) = \frac{K \mathbb{1}\{h_K p \leq 1\}}{\lceil K h_K p \rceil}. \tag{2.12}$$

Then f is an admissible calibrator. Moreover, the p-merging function induced by f is

$$F_{\mathrm{Hom}}(\mathbf{p}) := \inf \left\{ \alpha \in (0, 1) : \sum_{k=1}^{K} \frac{\mathbb{1}\{h_K p_k / \alpha \leq 1\}}{\lceil K h_K p_k / \alpha \rceil} \geq 1 \right\},$$

is valid and it dominates the Hommel combination rule.

Proof. It is simple to see that f is decreasing and it is upper semicontinuous (the function f has discontinuity points in $i/(Kh_K),\ i = 1, \dots, K$, but it is simple to prove that $\lim_{x \to x_0} f(x) \leq f(x_0)$). In addition,

$$\int_0^1 f(p)\mathrm{d}p = \int_0^1 \frac{K\mathbb{1}\{h_K p \leq 1\}}{\lceil Kh_K p \rceil}\mathrm{d}p = \sum_{j=1}^{K} \frac{K}{j}\frac{1}{Kh_K} = \frac{1}{h_K}\sum_{j=1}^{K}\frac{1}{j} = 1.$$

Due to Theorem 2.2 we have that F_{Hom} is admissible.

To prove the last part, we see that for any $\mathbf{p} \in (0,1]^K$ and $\alpha \in (0,1)$ we have that $F_{\text{Hom}}(\mathbf{p}) \leq \alpha$, if and only if

$$\bigwedge_{k=1}^{K} \frac{K}{k} p_{(k)} \leq \frac{\alpha}{h_K}.$$

This implies that, for some $m \in \{1, \dots, K\}$, we have that

$$\sum_{j=1}^{K} \mathbb{1}\left\{\frac{K}{m} h_K p_j \leq \alpha\right\} \geq m.$$

We now define this chain of inequalities,

$$\begin{aligned} 1 &\leq \sum_{j=1}^{K} \frac{1}{m}\mathbb{1}\left\{\frac{K}{m} h_K p_j \leq \alpha\right\} \overset{(i)}{\leq} \sum_{j=1}^{K} \frac{1}{\lceil Kh_K p_j/\alpha \rceil}\mathbb{1}\left\{\frac{K}{m} h_K p_j \leq \alpha\right\} \\ &\overset{(ii)}{\leq} \sum_{j=1}^{K} \frac{1}{\lceil Kh_K p_j/\alpha \rceil}\mathbb{1}\left\{h_K p_j \leq \alpha\right\}, \end{aligned} \tag{2.13}$$

where (i) is a consequence of $(1/m)\mathbb{1}\{Kp_j h_K \leq \alpha m\} \leq (1/\lceil Kh_K p_j/\alpha \rceil)\mathbb{1}\{Kp_j h_K \leq \alpha m\}$, for all $j = 1, \dots, K$, while (ii) to the fact that $K/m \geq 1$. This concludes the proof. □

The calibrator in (2.12) coincides, with a slight adjustment, with the BY calibrator introduced in Xu *et al.* (2024). An interesting fact is that the function F_{Hom} is always admissible in the class of symmetric p-merging functions; while it is admissible in the class of p-merging function (not necessarily symmetric) if K is not a prime number (Vovk *et al.*, 2022b, Theorem 7.1). The Hommel function allows for the selection of the minimum among the K possible different quantiles of $\mathbf{p}$.

In reality, one can choose to select only certain quantiles among K (e.g., one can select the minimum between K times the minimum, 2 times the median and the maximum),

hoping to pay a price less than $\log K$. We treat this problem in Appendix A.2, where we introduce a generalization of the Hommel combination rule.

2.4.1 An exchangeable Hommel's combination rule

Starting from the results in the previous sections, we can introduce a merging function, which improves Hommel's combination if the input p-values are exchangeable. In particular, we define the function

$$F_{\mathrm{EHom}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{K\mathbb{1}\{h_K p_i/\alpha \leq 1\}}{\lceil K h_K p_i/\alpha \rceil} \geq 1\right\}. \tag{2.14}$$

Theorem 2.24. *The function F_{EHom} is an ex-p-merging function and it strictly dominates the function F_{Hom}.*

Proof. According to Theorem 2.7 and Lemma 2.23, it follows that F_{EHom} is a ex-p-merging function. It is simple to see that $F_{\mathrm{EHom}} \leq F_{\mathrm{Hom}}$. □

The computation of a closed form for (2.14) is not straightforward, a possible solution to calculate the value of F_{EHom} is by using Algorithm 5 defined in Appendix A.6.

2.4.2 A randomized Hommel's combination rule

The *randomized* version of the function F_{Hom} has been proposed in Xu and Ramdas (2023, Appendix E) and takes the following form:

$$F_{\mathrm{UHom}}(\mathbf{p}, u) := \inf\left\{\alpha \in (0,1) : \sum_{k=1}^{K} \frac{\mathbb{1}\{h_K p_k/\alpha \leq 1\}}{\lceil K h_K p_k/\alpha \rceil} \geq u\right\}.$$

For completeness, we report here the following theorem.

Theorem 2.25. *The function F_{UHom} is a randomized p-merging function and it strictly dominates the function F_{Hom}.*

Proof. According to Corollary 2.18 and Lemma 2.23, it follows that F_{UHom} is a randomized p-merging function. It is simple to see that $F_{\mathrm{UHom}} \leq F_{\mathrm{Hom}}$. □

The value of the function F_{UHom} can be computed using Algorithm 1 in Vovk *et al.* (2022b), substituting the value of 1 by u.

2.5 Improving the "twice the average" combination rule

We now study the case of $r = 1$ in (2.10), which corresponds to the arithmetic mean. The general case, which allows $r \in \mathbb{R} \setminus \{0\}$, is considered in Appendix A.3. In the following, let $A(\mathbf{p})$ denote the arithmetic average of any vector $\mathbf{p}$, let $\mathbf{p}_m := (p_1, \dots, p_m)$ denote the vector containing the first m p-values, and let $\mathbf{p}_{(m)}$ denote the vector containing the smallest m elements of $\mathbf{p}$: $\mathbf{p}_{(m)} = (p_{(1)}, \dots, p_{(m)})$ such that $p_{(1)} \leq \cdots \leq p_{(m)}$. In addition, we denote by $\mathbf{p}^\ell_{(m)} = (p^\ell_{(1)}, \dots, p^\ell_{(m)})$, $m \in \{1, \dots, \ell\}$, the vector containing the smallest m elements of $(p_1, \dots, p_\ell)$. First, let us introduce a lemma that will be instrumental in subsequent discussions.

Lemma 2.26. *Let $f(p) = (2-2p)_+ \mathbb{1}\{p \in (0,1]\} + \infty \mathbb{1}\{p = 0\}$. Then, f is an admissible calibrator.*

Proof. It is easy to see that f is decreasing and $\int_0^1 (2 - 2p)\mathrm{d}p = 1$. In addition, it is upper semi-continuous and $f(0) = \infty$. □

In particular, we have that the calibrator defined in Lemma 2.26 is larger or equal than $f'(p) := (2 - 2p)$, which is the function inducing the average combination rule

$$F'_{\mathrm{A}}(\mathbf{p}) := 2A(\mathbf{p}), \tag{2.15}$$

which is a valid p-merging function. Note that f' can take negative values (its arguments p/α can be larger than 1), so it is technically not a calibrator in the sense of our definitions.

2.5.1 Exchangeable average combination rule

We now define ex-p-merging functions

$$F_{\mathrm{EA}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(2 - 2\frac{p_i}{\alpha}\right)_+ \geq 1\right\};$$
$$F'_{\mathrm{EA}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(2 - 2\frac{p_i}{\alpha}\right) \geq 1\right\}.$$

Theorem 2.27. *The dominations among ex-p-merging functions* $F_{\mathrm{EA}} \le F'_{\mathrm{EA}} \le F'_{\mathrm{A}}$ *are strict. Moreover,*

$$F'_{\mathrm{EA}}(\mathbf{p}) = 2\left\{\bigwedge_{m=1}^{K} A(\mathbf{p}_m)\right\};$$
$$F_{\mathrm{EA}}(\mathbf{p}) = \left(\bigwedge_{\ell=1}^{K}\bigwedge_{m=1}^{\ell} \frac{2A(\mathbf{p}^\ell_{(m)})}{(2-\ell/m)_+}\right) \mathbb{1}\{p_{(1)} > 0\}.$$

Proof. According to Theorem 2.7 and Lemma 2.26, it follows that the function F_{EA} is an ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. We note that $F_{\mathrm{EA}}(\mathbf{p}) \le \alpha$ if and only if, for some $\ell \in \{1,\dots,K\}$, we have

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(2 - \frac{2p_i}{\alpha}\right)_+ \ge 1. \tag{2.16}$$

This implies that there exists $\ell \le K$ such that

$$\frac{1}{\ell}\sum_{j=1}^{m}\left(2 - \frac{2p^\ell_{(j)}}{\alpha}\right) \ge 1 \quad \text{for some } m \in \{1,\dots,\ell\},$$

where we recall that $p^\ell_{(j)}$ is the j-th ordered value of the vector $(p_1,\dots,p_\ell)$. This is due to the fact that the contribution of p_i in the left-hand side of (2.16) vanishes for large values of p_i. Rearranging the terms, it is possible to obtain that exists $\ell \le K$ such that

$$\frac{2A(\mathbf{p}^\ell_{(m)})}{2-\ell/m} \le \alpha \quad \text{for some } m \in \{1,\dots,\ell\}.$$

Taking an infimum over m yields

$$\left(\bigwedge_{m=1}^{\ell} \frac{2A(\mathbf{p}^\ell_{(m)})}{(2-\ell/m)_+}\right) \le \alpha \quad \text{for some } \ell \in \{1,\dots,K\}.$$

Actually, the index m can start from $\lceil \ell/2 \rceil$ since the first $\lceil \ell/2 \rceil - 1$ terms in the denominator are smaller than zero. Taking an infimum also over ℓ gives the desired result.

The statements on domination in Theorems 2.27, 2.28, and other similar results are straightforward from definitions and we omit the proof. □

Despite being strictly dominated, F'_{EA} is very interpretable: it is just the minimum (over m) of "twice the average" of the first m p-values.

2.5.2 Randomized average combination rule

We now derive an improvement for the "twice the average" rule using a simple randomization trick. In this case, we do not require exchangeability but we allow for an arbitrary dependence among the p-values. We define the randomized p-merging function F_{UA} by

$$F_{\mathrm{UA}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(2 - \frac{2p_k}{\alpha}\right)_+ \geq u\right\}.$$

Clearly, $F_{\mathrm{UA}}(\mathbf{p}, u) \leq F'_{\mathrm{UA}}(\mathbf{p}, u)$, where F'_{UA} is defined by

$$F'_{\mathrm{UA}}(\mathbf{p}, u) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(2 - \frac{2p_k}{\alpha}\right) \geq u\right\}.$$

In particular, if randomization is not allowed then u is replaced by 1 and $F'_{\mathrm{UA}}(\mathbf{p}, 1)$ coincides with F'_{A} in (2.15). A p-merging function (such as F'_{A}) can also be seen as a randomized p-merging function, for which the argument u does not affect its value.

Theorem 2.28. *The dominations among randomized p-merging functions $F_{\mathrm{UA}} \leq F'_{\mathrm{UA}} \leq F'_{\mathrm{A}}$ are strict. Moreover,*

$$F'_{\mathrm{UA}}(\mathbf{p}, u) = \frac{2A(\mathbf{p})}{2-u};$$
$$F_{\mathrm{UA}}(\mathbf{p}, u) = \left(\bigwedge_{m=1}^{K} \frac{2A(\mathbf{p}_{(m)})}{(2 - Ku/m)_+}\right) \mathbb{1}\{p_{(1)} > 0\}.$$

Proof. According to Corollary 2.18 and Lemma 2.26, it follows that the function F_{UA} is a randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p}, u) \in (0,1]^{K+1}$. Note that $F_{\mathrm{UA}}(\mathbf{p}, u) \leq \alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{m}\left(2 - \frac{2p_{(k)}}{\alpha}\right) \geq u \quad \text{for some } m \in \{1, \ldots, K\}.$$

Rearranging terms, it is

$$\frac{\sum_{k=1}^{m} 2p_{(k)}}{2m - Ku} \leq \alpha \quad \text{for some } m \in \{1, \ldots, K\}.$$

Taking an infimum over m yields the desired formula. □

A method that can be directly compared with Theorem 2.28 is to use $F^*_{\mathrm{UA}}(\mathbf{p}, u) := A(\mathbf{p})/(2-2u)$ proposed by Wang (2024, Section B.2). This function F^*_{UA} is also a randomized p-merging function. One can see that F^*_{UA} does not dominate and is not

dominated by any of F_{A}, F_{UA} and F'_{UA}. Moreover, there is a simple relationship: $F'_{\mathrm{UA}}(\mathbf{p}, u) \leq F^*_{\mathrm{UA}}(\mathbf{p}, u)$ if and only if $u \geq 2/3$ for every $\mathbf{p}$ that is not the zero vector.

2.6 Improving the harmonic mean combination rule

The harmonic mean p-value was studied by Wilson (2019). In our context, it corresponds to the merging function in (2.10) when $r = -1$. We first state a lemma on a calibrator that we will use later.

Lemma 2.29. *Define the function*

$$f(p) = \min\left\{\frac{1}{T_K p} - \frac{1}{T_K}, K\right\} \mathbb{1}\{p \in [0,1]\},$$

with $T_K \geq 1$. Then f is a calibrator if T_K satisfies $KT_K + 1 - e^{T_K} \leq 0$, and in particular, f is a calibrator if $T_K = \log K + \log\log K + 1$.

Proof. Let $K \geq 2$ and $T_K \geq 1$. Define the function $f : [0, \infty) \to [0, \infty]$ as

$$p \mapsto \min\left\{\frac{1}{T_K p} - \frac{1}{T_K}, K\right\} \mathbb{1}\{p \in [0,1]\},$$

that is decreasing in p. Then,

$$\begin{aligned}
\int_0^1 f(p)\mathrm{d}p &= \int_0^1 \min\left\{\frac{1}{T_K p} - \frac{1}{T_K}, K\right\} \mathbb{1}\{p \in [0,1]\}\mathrm{d}p \\
&= \int_0^{(T_K K+1)^{-1}} K\mathrm{d}p + \int_{(T_K K+1)^{-1}}^1 \left(\frac{1}{T_K p} - \frac{1}{T_K}\right)\mathrm{d}p \\
&= \frac{K}{T_K K + 1} - \frac{K}{T_K K + 1} + \frac{\log(T_K K + 1)}{T_K} = \frac{\log(K T_K + 1)}{T_K}.
\end{aligned}$$

This implies that $\int_0^1 f(p)\mathrm{d}p \leq 1$ if and only if

$$KT_K + 1 - e_K^T \leq 0. \tag{2.17}$$

We would like to choose T_K as small as possible. One possible candidate is $T_K = \log K + \log\log K + 1$. Indeed, plugging $T_K = \log K + \log\log K + 1$ into the left-hand side of (2.17) we find that it is verified when

$$\log\log K + 1 + \frac{1}{K} \leq (e-1)\log K,$$

and this holds if $K \geq 2$ by checking $K = 2, 3$ and using the derivative of both sides for $K \geq 4$. □

In the following, we fix $T_K = \log K + \log\log K + 1$ and denote by $H(\mathbf{p}) = K(\sum_{k=1}^{K} 1/p_k)^{-1}$ the harmonic mean of the vector $\mathbf{p}$ where K is the number of elements contained in $\mathbf{p}$. We begin with a result that is new even under arbitrary dependence.

Proposition 2.30. $F'_{\mathrm{H}}(\mathbf{p}) := (T_K + 1)H(\mathbf{p})$ *is a p-merging function.*

Proof. It is straightforward to see that F is an increasing function. By direct calculation,

$$\begin{aligned}
\mathbb{P}(F(\mathbf{P}) \leq \alpha) &= \mathbb{P}((T_K+1)H(\mathbf{P}) \leq \alpha) \\
&= \mathbb{P}\left((T_K+1)K\left(\sum_{k=1}^{K}\frac{1}{P_k}\right)^{-1} \leq \alpha\right) \\
&= \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\alpha}{T_K P_k} - \frac{1}{T_K}\right) \geq 1\right) \\
&\overset{(i)}{\leq} \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right)\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\} \geq 1\right) \\
&= \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}\min\left\{\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right), K\right\}\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\} \geq 1\right) \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\min\left\{\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right), K\right\}\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\}\right] \\
&= \alpha\mathbb{E}\left[\frac{1}{\alpha}\frac{1}{K}\sum_{k=1}^{K}\min\left\{\left(\frac{1}{T_K P_k/\alpha} - \frac{1}{T_K}\right), K\right\}\mathbb{1}\left\{\frac{P_k}{\alpha} \in [0,1]\right\}\right] \leq \alpha,
\end{aligned}$$

where (i) is due to the fact that $(1/(T_K x) - 1/T_K)$ is negative for $x > 1$, and (ii) holds due to Markov's inequality. The last inequality is a consequence of Lemma 2.3 and Lemma 2.29. □

The above result differs from the formulation given in Vovk and Wang (2020, Proposition 9), which states that $e\log KH(\mathbf{p})$ is a p-merging function, thus sharpening their result for $K \geq 4$. Below we will further improve this result for exchangeable p-values. Moreover, a correction factor in the order of $\log K$ as $K \to \infty$ (although smaller than T_K) is needed even for independent p-values (see Proposition 6 of Chen *et al.* (2025)). The harmonic mean has some advantages over many other rules under certain dependence conditions; see Gui *et al.* (2025). It performs similarly to the Hommel combination; see Chen *et al.* (2023).

2.6.1 Exchangeable harmonic mean combination rule

We define homogeneous ex-p-merging functions as follows

$$F_{\mathrm{EH}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{\alpha}{T_K p_i} - \frac{1}{T_K}\right)_+ \ge 1\right\};$$
$$F'_{\mathrm{EH}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \le K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{\alpha}{T_K p_i} - \frac{1}{T_K}\right) \ge 1\right\}.$$

Theorem 2.31. *The dominations among ex-p-merging functions $F_{\mathrm{EH}} \le F'_{\mathrm{EH}} \le F'_{\mathrm{H}}$ are strict. Moreover,*

$$F'_{\mathrm{EH}}(\mathbf{p}) = \bigwedge_{m=1}^{K} (T_K + 1) H(\mathbf{p}_m);$$
$$F_{\mathrm{EH}}(\mathbf{p}) = \bigwedge_{\ell=1}^{K}\left(\bigwedge_{m=1}^{\ell}\left(\frac{\ell\, T_K}{m} + 1\right) H(\mathbf{p}^{\ell}_{(m)})\right).$$

Proof. According to Theorem 2.7 and Lemma 2.29, it follows that F_{EH} is an ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. We note that $F_{\mathrm{EH}}(\mathbf{p}) \le \alpha$ if and only if, for some $\ell \in \{1,\dots,K\}$, we have

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{\alpha}{T_K p_i} - \frac{1}{T_K}\right)_+ \ge 1.$$

This implies that exists $\ell \le K$ such that

$$\frac{1}{\ell}\sum_{j=1}^{m}\left(\frac{\alpha}{T_K p^{\ell}_{(j)}} - \frac{1}{T_K}\right) \ge 1 \quad \text{for some } m \in \{1,\dots,\ell\},$$

where we recall that $p^{\ell}_{(j)}$ is the j-th ordered value of the vector $(p_1,\dots,p_\ell)$. Rearranging the terms it is possible to obtain that exist $\ell \le K$ such that

$$\left(\frac{\ell\, T_K}{m} + 1\right) H(\mathbf{p}^{\ell}_{(m)}) \le \alpha \quad \text{for some } m \in \{1,\dots,\ell\},$$

Taking an infimum over m, we get

$$\bigwedge_{m=1}^{\ell}\left(\frac{\ell\, T_K}{m} + 1\right) H(\mathbf{p}^{\ell}_{(m)}) \le \alpha \quad \text{for some } \ell \in \{1,\dots,K\}.$$

Taking an infimum over ℓ yields the desired result. □

2.6.2 Randomized harmonic mean combination rule

Similarly to Section 2.5, we derive an improvement for the harmonic mean using a randomization trick in the case of arbitrarily dependent p-values. Define

$$F_{\mathrm{UH}}(\mathbf{p},u)=\inf\left\{\alpha\in(0,1):\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\alpha}{T_Kp_k}-\frac{1}{T_K}\right)_+\geq u\right\};$$
$$F'_{\mathrm{UH}}(\mathbf{p},u)=\inf\left\{\alpha\in(0,1):\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\alpha}{T_Kp_k}-\frac{1}{T_K}\right)\geq u\right\}.$$

Theorem 2.32. *The dominations among randomized p-merging functions* $F_{\mathrm{UH}}\leq F'_{\mathrm{UH}}\leq F'_{\mathrm{H}}$ *are strict. Moreover,*

$$F'_{\mathrm{UH}}(\mathbf{p},u)=(T_Ku+1)H(\mathbf{p});$$
$$F_{\mathrm{UH}}(\mathbf{p},u)=\bigwedge_{m=1}^{K}\left(\frac{uKT_K}{m}+1\right)H(\mathbf{p}_{(m)}).$$

Proof. According to Corollary 2.18 and Lemma 2.29, it follows that F_{UH} is a randomized p-merging function. Fix any $\alpha\in(0,1)$ and $(\mathbf{p},u)\in(0,1]^{K+1}$. Then $F_{\mathrm{UH}}(\mathbf{p},u)\leq\alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{m}\left(\frac{\alpha}{T_Kp_{(k)}}-\frac{1}{T_K}\right)\geq u\quad\text{ for some } m\in\{1,\dots,K\}.$$

Rearranging the terms, it is

$$(uKT_K+m)\left(\sum_{k=1}^{m}\frac{1}{p_{(k)}}\right)^{-1}\leq\alpha\quad\text{for some } m\in\{1,\dots,K\}.$$

Taking a minimum over m yields the desired formula. □

A non-randomized improvement of F'_{H} can be achieved fixing $u=1$ in F_{UH}. This coincides with the function $F_{\mathrm{H}}(\mathbf{p})=\bigwedge_{m=1}^{K}((KT_K)/m+1)H(\mathbf{p}_{(m)})$.

2.7 Improving the geometric mean combination rule

We now derive some new combination based on the geometric mean, a special case of (2.10) when $r\to 0$. Let $G(\mathbf{p})=(\prod_{k=1}^{K}p_k)^{1/K}$ denote the geometric mean of the vector $\mathbf{p}$. The calibrator, in this case, is given by $f(p)=(-\log p)_+$, which is an admissible calibrator. Actually, a slightly improved calibrator is $f(p)=(-(\log p)/T)_+\wedge K$ for

some $T < 1$ satisfying $\int_0^1 f(p)\mathrm{d}p \leq 1$. This condition is verified when $1 - e^{-KT} \leq T$, which makes T very close to 1 for K moderately large (see Section 3.2 in Vovk and Wang, 2020). In the sequel, we denote by

$$F'_{\mathrm{G}}(\mathbf{p}) := eG(\mathbf{p}), \tag{2.18}$$

which is a valid p-merging function studied by Vovk and Wang (2020).

2.7.1 Exchangeable geometric mean combination rule

Following the same approach as in the preceding sections, we define

$$F_{\mathrm{EG}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\log\frac{\alpha}{p_i}\right)_+ \geq 1\right\};$$
$$F'_{\mathrm{EG}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell}\log\frac{\alpha}{p_i} \geq 1\right\}.$$

Theorem 2.33. *The dominations among ex-p-merging functions $F_{\mathrm{EG}} \leq F'_{\mathrm{EG}} \leq F'_{\mathrm{G}}$ are strict. Moreover,*

$$F'_{\mathrm{EG}}(\mathbf{p}) = e\left\{\bigwedge_{m=1}^{K} G(\mathbf{p}_m)\right\};$$
$$F_{\mathrm{EG}}(\mathbf{p}) = \bigwedge_{\ell=1}^{K}\left(\bigwedge_{m=1}^{\ell} e^{\ell/m}G(\mathbf{p}^{\ell}_{(m)})\right).$$

Proof. According to Theorem 2.7, it follows that the function F_{EG} is an ex-p-merging function. Fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$. Then $F_{\mathrm{EG}}(\mathbf{p}) \leq \alpha$, if and only if exists $\ell \in \{1,\ldots,K\}$ such that

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(-\log p_i + \log\alpha)_+ \geq 1.$$

This is verified when exists $\ell \leq K$ such that

$$\frac{1}{\ell}\sum_{j=1}^{m}(-\log p^{\ell}_{(j)} + \log\alpha) \geq 1 \quad \text{for some } m \in \{1,\ldots,\ell\},$$

where we recall that $p^{\ell}_{(j)}$ is the j-th ordered value of the vector $(p_1, \dots, p_\ell)$. Rearranging the terms it is possible to obtain that exists $\ell \leq K$ such that

$$e^{\ell/m} G(\mathbf{p}^{\ell}_{(m)}) \leq \alpha \quad \text{for some } m \in \{1, \dots, \ell\}.$$

Taking an infimum over m, we get

$$\bigwedge_{m=1}^{\ell} \left(e^{\ell/m} G(\mathbf{p}^{\ell}_{(m)}) \right) \leq \alpha \quad \text{for some } \ell \in \{1, \dots, K\}.$$

Taking an infimum over ℓ yields the desired result. □

2.7.2 Randomized geometric mean combination rule

As in the previous sections, we define the randomized p-merging functions as follows:

$$F_{\mathrm{UG}}(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} \left(\log \frac{\alpha}{p_k} \right)_+ \geq u \right\};$$
$$F'_{\mathrm{UG}}(\mathbf{p}, u) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} \log \frac{\alpha}{p_k} \geq u \right\}.$$

Theorem 2.34. *The dominations among randomized p-merging functions $F_{\mathrm{UG}} \leq F'_{\mathrm{UG}} \leq F'_{\mathrm{G}}$ are strict. Moreover,*

$$F'_{\mathrm{UG}}(\mathbf{p}, u) = e^u G(\mathbf{p});$$
$$F_{\mathrm{UG}}(\mathbf{p}, u) = \bigwedge_{m=1}^{K} \left(e^{u\frac{K}{m}} G(\mathbf{p}_{(m)}) \right).$$

Proof. According to Corollary 2.18, it follows that F_{UG} is a randomized p-merging function. Fix any $\alpha \in (0,1)$ and $(\mathbf{p}, u) \in (0,1]^{K+1}$. We have that $F_{\mathrm{UG}}(\mathbf{p}, u) \leq \alpha$ if and only if

$$\frac{1}{K} \sum_{k=1}^{m} (-\log p_{(k)} + \log \alpha) \geq u \quad \text{for some } m \in \{1, \dots, K\}.$$

Rearranging the terms, it is

$$e^{u\frac{K}{m}} G(\mathbf{p}_{(m)}) \leq \alpha \quad \text{for some } m \in \{1, \dots, K\}.$$

Taking a minimum over m yields the desired formula. □

A non-randomized improvement of the combination in (2.18) can be obtained fixing $u = 1$ in F_{UG}. This gives the combination rule $F_{\mathrm{G}}(\mathbf{p}) = \bigwedge_{m=1}^{K} e^{K/m} G(\mathbf{p}_{(m)})$.

2.8 Simulation study

In the previous sections, new p-merging functions have been introduced. These new rules are obtained using a randomization trick or they rely on exchangeability of p-values. Specifically, the introduced rules have been shown to dominate their original counterparts by utilizing randomness or exchangeability (or both). In this section, our aim is to investigate their performance using simulated data.

We consider the example described in Vovk and Wang (2020, Section 6) (a similar example is proposed in Chen *et al.* (2023)), where p-values are generated in the following way:

$$X_k = \rho Z + \sqrt{1-\rho^2} Z_k - \mu, \quad P_k = \Phi(X_k), \tag{2.19}$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, $Z, Z_1, \dots, Z_K \overset{iid}{\sim} \mathcal{N}(0,1)$, and $\mu \geq 0$ and $\rho \in [0,1]$ are constants. It is simple to prove that $P_1, \dots, P_K$ are exchangeable and their marginal distribution does not depend on ρ. In addition, if $\rho = 0$ then $P_1, \dots, P_K$ are independent while if $\rho = 1$ then $P_1 = \cdots = P_K$. The value p_k, in this case, can be interpreted as the p-value resulting from a one-side z-test of the null hypothesis $\mu = 0$ against the alternative $\mu > 0$ from the statistic $X_k \sim \mathcal{N}(-\mu, 1)$ with unknown μ. We let the parameter μ vary in the interval $[0, 3]$ (if $\mu = 0$ then H_0 is true) and fix the upper bound of the type-I error to the nominal level 0.05.

We compare the different ex-p-merging functions introduced in the previous sections, with the addition of the Bonferroni method. The parameter k for the Rüger combination rule is set to $K/2$ (twice the median). The parameter ρ is set to the values $\rho = \{0.1, 0.9\}$, corresponding to weak and strong dependence among p-values. Each simulation is repeated for a total of $B = 10,000$ replications, and we report the observed empirical average. In Figure 2.1, we can notice that the error level is controlled at the nominal level 0.05 for all the proposed methods. Variations in terms of power are observed depending on whether the parameter ρ is set to 0.9 or 0.1. Specifically, ex-p-merging functions that exhibit strong performance in the left plot tend to show reduced effectiveness in the right plot, and the opposite is also true. As expected, the Bonferroni method shows a higher power near independence where also the ex-Hommel combination rule defined in (2.14) seems to perform well. A simulation study comparing the different randomized p-merging functions is reported in Appendix A.7.

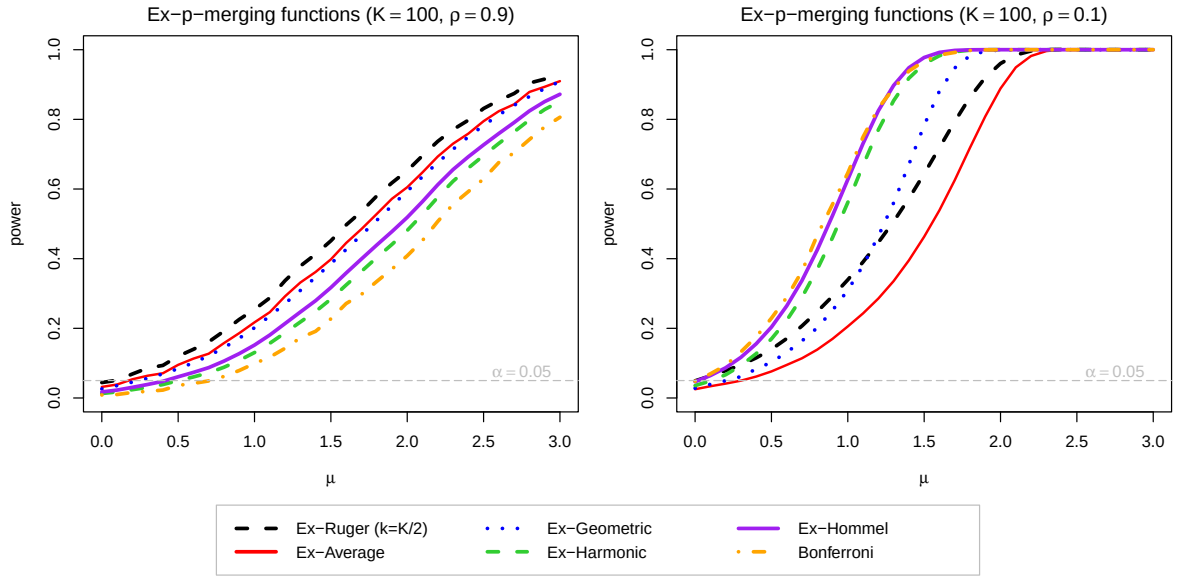

FIGURE 2.1: Combination of p-values using different ex-p-merging functions under high (left) and low (right) dependence. The performance of the different ex-p-merging functions is almost reversed in the two situations.

In this last part, we aim to explore the scenario where p-values are exchangeable under the null hypothesis but not under the alternative. Indeed, the p-values generated as in (2.19) are exchangeable under both the null and the alternative hypotheses; however, for the results in Section 2.2, it is only necessary for the p-values to be exchangeable under H_0. Specifically, if the p-values are not exchangeable under the alternative, they could be arranged in a particular way to obtain a more powerful procedure. In other words, if it is suspected that some p-values are smaller under the alternative hypothesis, they could be placed at the beginning of the vector to enhance the power of the procedure since our rules that are valid under exchangeability process the vector of p-values sequentially. Clearly, the procedure of ordering p-values in a particular way must preserve the exchangeabilty under the null hypothesis (i.e., data-dependent ordering is usually not allowed since it violates exchangeability). In the simulation setting, suppose to have K independent studies, each with observations X_{ij}, $i = 1, \dots, K$, $j = 1, \dots, n_i$, that are iid from a normal distribution with mean μ and variance 1. In addition, $X_{0j}, j = 1, \dots, n_0$, is assumed to be an additional sample from the same population and common for all studies. We define the quantity $\bar{X}_i$ as

$$\bar{X}_i = \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} X_{ij}, \quad i = 0, 1, \dots, K,$$

that is distributed as $\mathcal{N}(\sqrt{n_i}\mu, 1)$. The interest is to test the hypothesis $\mu = 0$ under the alternative $\mu \neq 0$ and the test statistic used is

$$T_k := \frac{\bar{X}_k + \bar{X}_0}{\sqrt{2}}, \quad k = 1, \ldots, K,$$

where it is possible to see that each study use the common sample in the "same way". Under the null hypothesis, the test statistics have the same marginal distribution and the model $(T_1, \ldots, T_K)$ is exchangeable. Under the alternative, the mean of the test statistics depends on the sample size, so the model cannot be exchangeable. In particular, studies with a higher number of observations are more powerful. The p-values to test the null hypothesis are given by

$$P_k = 2\Phi(-|T_k|), \quad k = 1, \ldots, K.$$

Specifically, in the simulated scenario, $K = 10$, n_i, $i = 1, \ldots, K$, take values $10, 20, \ldots, 100$ and $n_0 = 25$. The number of replications is $B = 10,000$ and the ex-p-merging functions used are F_{EA} ("twice the mean") and F_{ER} with $k = K/2$ ("twice the median"). We let the parameter μ varies in the interval $[0, 0.5]$ and three different solutions are compared: (i) the p-values are ordered in increasing order with respect to the sample size, (ii) the p-values are ordered in decreasing order with respect to the sample size, and (iii) the p-values are randomly ordered. In addition, we compare the ex-p-merging functions with the "standard" rules valid under arbitrary dependence F_{A} and F_{R}. The results are reported in Figure 2.2, where we see that when the p-values are ordered in decreasing order with respect to the number of observations, the combined tests are more powerful. This is because the power of individual p-values increases with the sample size. Overall, the proposed ex-p-merging are more powerful than the rules valid under arbitrary dependence.

Results under other setups are reported in Appendix A.7, with similar qualitative conclusions.

2.9 On the effect of the randomization

Before concluding, we make a final observation on the role of randomization. Randomization based methods for combining p-values are introduced to address the problem of aggregating dependent p-values. Two forms of randomization are employed: (i) a random permutation, corresponding to an independent draw from a discrete distribution, and (ii) an independent draw from a continuous uniform distribution.

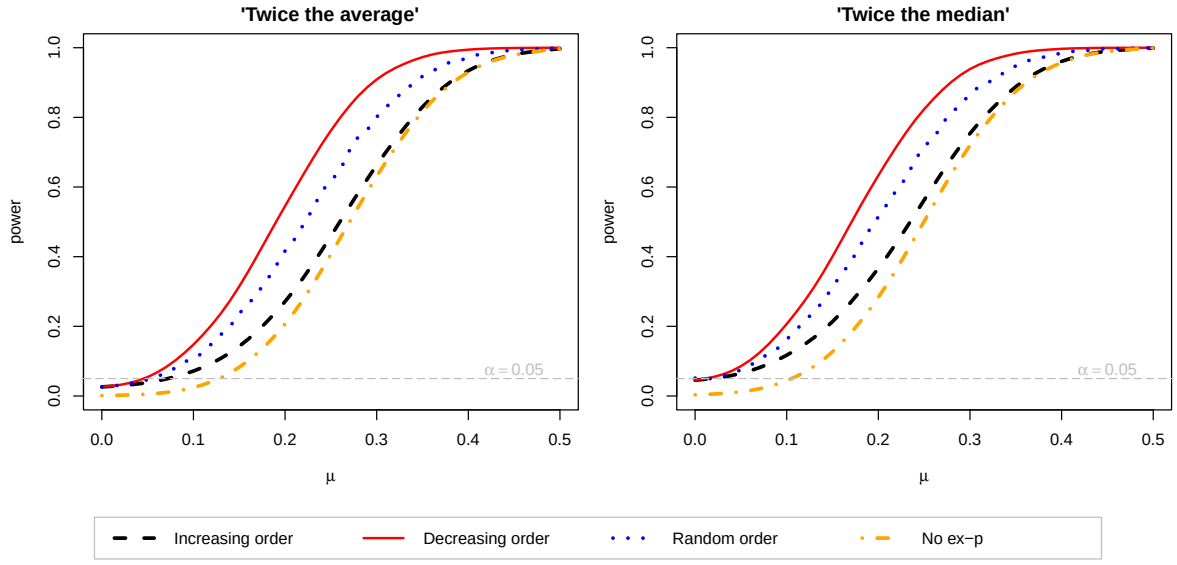

FIGURE 2.2: Combination of p-values using different ex-p-merging functions and different ordering based on the sample size. Non ex-p-merging functions valid under arbitrary dependence are added for comparison. The ex-p-merging rules are more powerful if p-values are ordered in decreasing order with respect to the sample size.

To make the role of randomization explicit, consider a fixed significance level $\alpha \in (0, 1)$. Let $\mathbf{P}$ denote the vector of p-variables, and let V represent the randomization element (we avoid the notation U since it does not necessarily follow a uniform distribution, but may instead represent a random permutation). Following Lehmann *et al.* (1986), we define the function

$$\phi_\alpha(\mathbf{P}, V) = \mathbb{1}\{F(\mathbf{P}, V) \leq \alpha\}.$$

In the previous section, we established that, under the null. $\mathbb{E}[\phi_\alpha(\mathbf{P}, V)] \leq \alpha$. However, it remains to clarify how randomization affects the decision once the p-values have been observed. In this case, the relevant quantity is $\mathbb{E}[\phi_\alpha(\mathbf{p}, V)]$, which represents the expected decision under randomization given the observed p-values $\mathbf{p}$.

If this expected value equals one, the null hypothesis would be rejected even without randomization. Conversely, there are situations in which the expected value is zero, indicating that randomization provides no practical advantage at the given level. In general, evaluating or estimating this expectation can be informative, as it reveals the extent to which randomization refines, or fails to refine, the testing rule. We present an example.

Example 2.3. *Let* $\mathbf{p} = (0.01, 0.02, 0.03, 0.06)$ *be the vector of observed p-values. The rule* $2A(\mathbf{p})$ *does not reject the null hypothesis at the significance level* $\alpha = 0.05$*. However,*

the randomized "twice the average" rule (Table 2.1) can reject the null hypothesis, and this occurs with positive probability. In particular,

$$\mathbb{E}[\phi_\alpha(\mathbf{p}, U)] = \mathbb{P}\left(\frac{2}{2-U}A(\mathbf{p}) \leq \alpha\right) = \frac{4}{5}.$$

Similarly, it is possible to compute $\mathbb{E}[\phi_\alpha(\mathbf{p}, V)]$ *in the case where* V *is a random permutation. In this case, the computation simply consists of evaluating all the ex-p-merging functions corresponding to the possible permutations of* $\mathbf{p}$ *and counting how many of these values are less than or equal to* α*. The resulting probability in this case is* $2/3$*.*

2.10 Discussion

In this chapter, we describe novel p-merging functions for the scenario where the p-values are exchangeable. These new rules are demonstrated to dominate their original counterparts derived under the assumption of arbitrary dependence. Furthermore, we illustrate how a simple randomization trick (introduction of a uniform random variable or uniform permutation) can also be employed in the case of arbitrarily dependent p-values to yield more powerful rules than existing ones. These results are proposed in a fully general form, addressing the relationship between p-merging functions and e-values introduced by their respective calibrators. In particular, once the corresponding e-value is obtained, it becomes feasible to utilize Markov's inequality or its randomized and exchangeable generalizations from Ramdas and Manole (2024).

As a practical recommendation, we suggest the exchangeable improvement of the Hommel combination if we have no apriori idea how strong the dependence is likely to be, but to use the exchangeable improvement of "twice the median" if the p-values are thought to be strongly dependent.

As an open question, it remains unclear whether our methods are further improvable, stemming from the unknown admissibility or tightness of exchangeable Markov's inequality beyond extreme cases.

Chapter 3

Merging uncertainty sets via majority vote

In the previous chapter, we focused on methods for combining dependent p-values. A related problem arises when the quantities of interest are uncertainty sets rather than p-values. This chapter develops methods for combining K uncertainty sets (e.g., prediction or confidence sets) that are arbitrarily dependent (perhaps due to shared data) in order to obtain a single set with nearly the same coverage. As one motivation, consider K "agents" that process some private and public data in different ways in order to define K uncertainty sets. Their use of the public data in unknown ways causes the sets to be dependent. Our work studies ways to combine these sets, despite unknown dependence.

Formally, we start with a collection of K different sets $\mathcal{C}_k$ (one from each agent), each having a confidence level $1-\alpha$ for some $\alpha \in (0,1)$:

$$\mathbb{P}\big(c \in \mathcal{C}_k\big) \geq 1-\alpha, \quad k = 1, \ldots, K, \tag{3.1}$$

where c denotes our target (e.g., an outcome that we want to predict, or some underlying functional of the data distribution). As in Section 1.5, we say that $\mathcal{C}_k$ has *exact* coverage if $\mathbb{P}(c \in \mathcal{C}_k) = 1-\alpha$. Since the sets $\mathcal{C}_k$ are based on data, they are random quantities by definition, but c can be either fixed or random; for example, in the case of confidence sets for a target functional of a distribution it is fixed, but it is random in the case of prediction sets for an outcome (e.g., conformal prediction). The proposed method will be agnostic to such details.

Our objective as the "aggregator" of uncertainty is to combine the sets in a black-box manner in order to create a new set that exhibits favorable properties in both coverage and size. A first (trivial) solution is to define the set $\mathcal{C}^J$ as the union of the others: $\mathcal{C}^J = \bigcup_{k=1}^{K} \mathcal{C}_k$. Clearly, $\mathcal{C}^J$ respects the property defined in (3.1), but the

resulting set is typically too large and has significantly inflated coverage. On the other hand, the set resulting from the intersection $\mathcal{C}^I = \bigcap_{k=1}^K \mathcal{C}_k$ is narrower, but typically has inadequate coverage: it guarantees at least $1-K\alpha$ coverage by the Bonferroni inequality (Bonferroni, 1936), but this is uninformative when K is large.

This work addresses the setting where only a single interval is known from each agent. In fact, if the aggregator knows the $(1-\alpha)$-confidence intervals for *every* $\alpha \in (0,1)$, they can combine these using well-known rules to combine dependent p-values as the ones described in Chapter 1 and in Chapter 2. A major motivation is aesthetic: while the literature on "black-box" combination of p-values (that is, without access to the underlying data) is rich — dating from Fisher's famous combination rule to modern treatments (Vovk *et al.*, 2022b) — analogous results for uncertainty sets like confidence or prediction intervals are less studied or known. A second motivation is practical — while in some situations it is possible to obtain the p-values for every possible target value of c, there are many cases where this is intractable, and it is easier to simply obtain the confidence intervals at a fixed $1-\alpha$ level. Some examples of the latter type include procedures that are tuned to a particular level α, such as HulC (Kuchibhotla *et al.*, 2024). In yet other cases, for example in differential privacy or federated learning, the aggregator may only have access to the K different intervals and not the p-values or the raw data due to privacy constraints (Humbert *et al.*, 2023; Waudby-Smith *et al.*, 2023). In the following, we will define new aggregation schemes based on the simple concept of voting, which can be used to merge uncertainty sets.

3.1 Majority vote: definition, extensions and properties

We now study a versatile method for combining uncertainty sets that is evidently broadly applicable. The key idea is based on voting: each agent's interval gets to vote, and each point in the space of interest (where the target c lies) will be part of the final set if it is vouched for by more than half (or more generally, some fraction) of the agents. Majority voting is popular in other contexts, like ensemble methods for prediction; see Breiman (1996) and Kuncheva *et al.* (2003); Kuncheva (2014). The idea has been proposed within the context of combining conformal prediction intervals by Cherubin (2019) and Solari and Djordjilović (2022).

This section compiles the relevant results in a succinct manner, and building on these, we extend the method in multiple directions. Specifically, we allow for the incorporation of a priori information, and additionally, we are able to achieve smaller sets through the

use of a simple randomization or permutation technique without altering the coverage properties.

The majority vote procedure for sets can be seen as dual to the results in Rüger (1978), who presented a method for combining K different p-values for testing a null hypothesis based on their order statistics; see Appendix B.1. These results are used, for example, in multi-split inference where the single agent wants to reduce the randomness induced by sample-splitting by performing many random splits and combining the results; see DiCiccio *et al.* (2020). Recently, Guo and Shah (2024) introduced a different subsampling-based method to conduct inference in the case of multiple splits. Their results assume exchangeability of the underlying sets or p-values. We handle both the exchangeable and non-exchangeable cases, but the more important distinction is that our method is "black-box" (needing to know no details of how the original sets were constructed) while theirs is not. Further, their method needs more distributional assumptions for their asymptotics to hold, while ours is nonasymptotically valid under no assumptions. Finally, their methods are computationally intensive and sophisticated to understand and use, while ours are simple in comparison. In turn, one may expect that when one has access to the full data to perform subsampling, and their distributional assumptions are true, their method may be more efficient than ours.

3.1.1 The majority vote procedure

Let the observed data $z = (z_1, \ldots, z_n)$ be a realization of the random vector $Z = (Z_1, \ldots, Z_n)$. In particular, $z = (z_1, \ldots, z_n)$ is a point in a sample space $\mathcal{Z}$, while our target c is a point in a measurable space $(\mathcal{S}, \mathcal{F}, \nu)$, where $\mathcal{F}$ is a σ-algebra on $\mathcal{S}$ and ν is a measure on $\mathcal{S}$. As mentioned earlier, it is important to note that c can itself be a random variable. The sets $\mathcal{C}_k = \mathcal{C}_k(z) \subseteq \mathcal{S}$, $k = 1, \ldots, K$, based on the observed data, follow the property (3.1), where the probability refers to the joint distribution (Z, c) (or only Z if c is fixed). Naturally, the different sets may have only been constructed using different subsets of z (as different public and private data may be available to each of the agents). Let us define a new set $\mathcal{C}^M$ that includes all points *voted* by at least half of the sets:

$$\mathcal{C}^M := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} > \frac{1}{2} \right\}. \tag{3.2}$$

The following result stems from Kuncheva *et al.* (2003) and Cherubin (2019), and again later by Solari and Djordjilović (2022), but we provide a short proof to be self-contained.

Theorem 3.1. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be $K \geq 2$ different sets satisfying property* (3.1)*. Then, the set $\mathcal{C}^M$ defined in* (3.2) *has coverage of at least $1 - 2\alpha$:*

$$\mathbb{P}\big(c \in \mathcal{C}^M\big) \geq 1 - 2\alpha. \tag{3.3}$$

The proof essentially relies on defining e-values $E_k = \phi_k/\alpha$, combining them by averaging, and applying Markov's inequality.

Proof. Let $\phi_k = \phi_k(Z, c) = \mathbb{1}\{c \notin \mathcal{C}_k\}$ be a Bernoulli random variable such that $\mathbb{E}[\phi_k] \leq \alpha$, $k = 1, \ldots, K$. We have by Markov's inequality,

$$\mathbb{P}(c \notin \mathcal{C}^M) = \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}\phi_k \geq \frac{1}{2}\right) \leq 2\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\phi_k\right] = \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\phi_k\right] \leq 2\alpha,$$

which concludes the proof. □

The merged set could be the empty set — however, it cannot be empty with probability more than 2α, because it would otherwise violate the coverage guarantee. In practice, it occurs with much smaller probability, almost never occurring in our simulations.

Remark 3.2. Actually, a slightly tighter bound can be obtained if K is odd. In this case, for a point to be contained in $\mathcal{C}^M$, it must be voted for by at least $\lceil K/2 \rceil$ of the other sets. This implies that, with the same arguments used in Theorem 3.1, the probability of miscoverage is equal to $\alpha K/\lceil K/2 \rceil = 2\alpha K/(K+1)$, approaching (3.3) for large K.

This result is known to be tight in a worst-case sense; a simple example from Kuncheva *et al.* (2003) shows that if K is odd and if the sets have a particular joint distribution, then the error will equal $(\alpha K)/\lceil K/2 \rceil$. This worst-case distribution allows for only two types of cases: either all agents provide the same set that contains c (so majority vote is correct), or $\lfloor K/2 \rfloor$ sets contain c but the others do not (so majority vote is incorrect). Each of the latter cases happens with some probability p, so the probability that majority vote makes an error is $\binom{K}{\lfloor K/2 \rfloor + 1}p$. The probability that any particular agent makes an error is $\binom{K-1}{\lfloor K/2 \rfloor}p$, which we set as our choice of α, and then we see that the probability of error for majority vote simplifies to $\alpha K/\lceil K/2 \rceil$. *Despite the apparent tightness of majority vote in the worst-case, we will develop several ways to improve this procedure in non-worst-case instances, while retaining the same worst-case performance.*

Remark 3.3 (When does majority vote overcover and when does it undercover?). While the worst case theoretical guarantee for majority vote is a coverage level of $1 - 2\alpha$, sometimes it will get close to the desired $1 - \alpha$ coverage, and sometimes it may even

overcover, achieving coverage closer to one. Here, we provide some intuition for when to expect each type of behavior in practice assuming $\alpha < 1/2$, foreshadowing many results to come. If the sets are actually independent (or nearly so), we should expect the method to have coverage more than $1 - \alpha$. This can be seen via an application of Hoeffding's inequality in place of Markov's inequality in the proof of Theorem 3.1: since each ϕ_k has expectation (at most) α, we should expect $\frac{1}{K}\sum_{k=1}^{K}\phi_k$ to concentrate around α, and the probability that this average exceeds $1/2$ is exponentially small (as opposed to 2α), being at most $\exp(-2K(1/2-\alpha)^2)$ by Hoeffding's inequality. In contrast, if the sets are identical (the opposite extreme of independence), clearly the method has coverage $1 - \alpha$. As argued in the previous remark, there is a worst case dependence structure that forces majority to vote to have an error of (essentially) 2α. Finally, if the sets are exchangeable, it appears more likely that the coverage will be closer to $1 - \alpha$ than $1 - 2\alpha$. While one informal reason may be that exchangeability connects the two extremes of independence and being identical (with coverages close to 1 and $1 - \alpha$), a slightly more formal reason is that under exchangeability, we will later in this section actually devise a strictly tighter set $\mathcal{C}^E$ than $\mathcal{C}^M$ which also achieves the same coverage guarantee of $1-2\alpha$, thus making $\mathcal{C}^M$ itself likely to have a substantially higher coverage.

3.1.2 Other thresholds and upper bounds

The above method can be easily generalized beyond the threshold of $1/2$. We record it as a result for easier reference. For any $\tau \in [0, 1)$, let

$$\mathcal{C}^{\tau} := \left\{ s \in \mathcal{S} : \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} > \tau \right\}. \tag{3.4}$$

Theorem 3.4. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be $K \geq 2$ different sets satisfying property* (3.1)*. Then,*

$$\mathbb{P}\big(c \in \mathcal{C}^{\tau}\big) \geq 1 - \alpha/(1-\tau).$$

The proof follows the same lines as the original results outlined in Theorem 3.1 and is thus omitted. As expected, it can be noted that the obtained bound decreases as τ increases. In fact, for larger values of τ, smaller sets will be obtained. One can check that this result also yields the right bound for the intersection ($\tau = 1 - 1/K$) and the union ($\tau = 0$). In certain situations, it is possible to identify an upper bound to the coverage of the set resulting from the majority vote.

Theorem 3.5. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be $K \geq 2$ different sets having exact coverage $1-\alpha$. Then,*

$$\mathbb{P}(c \in \mathcal{C}^M) \leq 1 - \frac{K\alpha - \left\lceil \frac{K}{2} \right\rceil + 1}{K - \left\lceil \frac{K}{2} \right\rceil + 1}. \tag{3.5}$$

Proof. Let $r := \left\lceil \frac{K}{2} \right\rceil$, $\phi_k = \mathbb{1}\{c \notin \mathcal{C}_k\}$ be a Bernoulli random variable such that $\mathbb{E}[\phi_k] = \alpha$, $k = 1, \ldots, K$, and $S_K = \sum_{k=1}^{K} \phi_k$ taking values in $\{0, 1, \ldots, K\}$. By definition, we know that

$$\mathbb{E}[S_K] = \sum_{k=1}^{K} \mathbb{E}[\phi_k] = K\alpha.$$

Let us define $\rho_j = \mathbb{P}(S_K = j)$. Now we can write

$$\begin{aligned}
\mathbb{E}[S_K] &= \sum_{j=0}^{K} j\rho_j = \sum_{j=0}^{r-1} j\rho_j + \sum_{j=r}^{K} j\rho_j \pm (r-1)\sum_{j=0}^{r-1}\rho_j \pm K\sum_{j=r}^{K}\rho_j \\
&= (r-1)\sum_{j=0}^{r-1}\rho_j + K\sum_{j=r}^{K}\rho_j - \sum_{j=0}^{r-1}(r-1-j)\rho_j - \sum_{j=r}^{K}(K-j)\rho_j \\
&= (r-1)\left(1 - \mathbb{P}\left(S_K \geq r\right)\right) + K\mathbb{P}\left(S_K \geq r\right) - m.
\end{aligned}$$

Since $m \geq 0$, then

$$K\alpha \leq (r-1)\left(1 - \mathbb{P}\left(S_K \geq \left\lceil \frac{K}{2} \right\rceil\right)\right) + K\mathbb{P}\left(S_K \geq \left\lceil \frac{K}{2} \right\rceil\right)$$

which implies

$$\mathbb{P}\left(S_K \geq \left\lceil \frac{K}{2} \right\rceil\right) \geq \frac{K\alpha - r + 1}{K - r + 1}.$$

From Theorem 3.1 we know that

$$\mathbb{P}(c \in \mathcal{C}^M) = 1 - \mathbb{P}(c \notin \mathcal{C}^M) = 1 - \mathbb{P}\left(S_K \geq \left\lceil \frac{K}{2} \right\rceil\right) \leq 1 - \frac{K\alpha - r + 1}{K - r + 1},$$

which concludes the proof. □

For typically employed values of α, this upper bound is useful only for small K. When $K = 2$, it can be seen that (3.2) coincides with the intersection between the two sets; this correctly implies that the coverage in this case lies in $[1-2\alpha, 1-\alpha]$.

3.1.3 On the computation of the majority vote set

Even if the input sets are intervals, the majority vote set may be a union of intervals. In Appendix B.2, we describe a simple aggregation algorithm to find this set quickly

by sorting the endpoints of the intervals and checking some simple conditions. But in practice, we find that it is indeed a non-empty interval in the vast majority of our simulations, suggesting that it may be possible to identify some sufficient conditions under which this happens. One such condition is given below, and is represented in Figure 3.1.

Lemma 3.6. *If $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are one-dimensional intervals and $\cap_{k=1}^K \mathcal{C}_k \neq \varnothing$, then $\mathcal{C}^\tau$ is an interval for any τ, and, in particular, $\mathcal{C}^M$ is an interval.*

Proof. We provide a "visual" proof of this fact. Consider a "histogram" view of the voting procedure. Every point on the real line is assigned a score between 0 and 1, which is the fraction of sets that contained it. The resulting score curve is almost a kernel density estimate, except that it is not normalized: the "density" is given by $f(s) = \frac{1}{K}\sum_{k=1}^K \mathbb{1}\{s \in \mathcal{C}_k\}$. If this density is unimodal, then every level set will be an interval (and the level sets just correspond to $\mathcal{C}^\tau$ for various τ). Now we argue that the input sets being intervals and $\cap_{k=1}^K \mathcal{C}_k \neq \varnothing$ is sufficient for this unimodality. Take $\tau = 1/2$ in what follows for simplicity, to focus on $\mathcal{C}^M$. First note that if $\cap_{k=1}^K \mathcal{C}_k \neq \varnothing$, then $\cap_{k=1}^K \mathcal{C}_k$ is itself an interval (being the intersection of convex sets, it must be convex), and it must be contained in $\mathcal{C}^M$. Starting from this interval, when we attempt to grow the interval by checking points just to the right (or left) of the current set, one will observe that the score of points can only decrease as we move further out. This is because our starting interval $\cap_{k=1}^K \mathcal{C}_k$ lies "inside" every single input interval, and so as we move outwards from it, we can only hit closing endpoints of these intervals, slowly starting to exclude them one by one, but we can never hit an opening endpoint of an interval at a later point. This concludes the proof of unimodality, and hence of the lemma. □

A typical example of a non-empty intersection arises when all intervals contain the target c, and, as mentioned earlier, this probability is at least $1 - K\alpha$. Clearly, this is an underestimate of the true probability, and it is non-trivial only for moderate values of K. One can avoid the problem of having a union of intervals by taking the convex hull of set (the smallest interval containing the set) and this solution is used, for example, in Gupta *et al.* (2022) in the framework of leave-one-out conformal prediction.

3.1.4 Equal-width intervals and the "Median of Midpoints"

A special case for the majority vote set in (3.2) occurs if the input sets are (or can be bounded by) intervals of the same width. In this case, each interval can be represented as its midpoint plus/minus a half-width, and the following rule results in a more

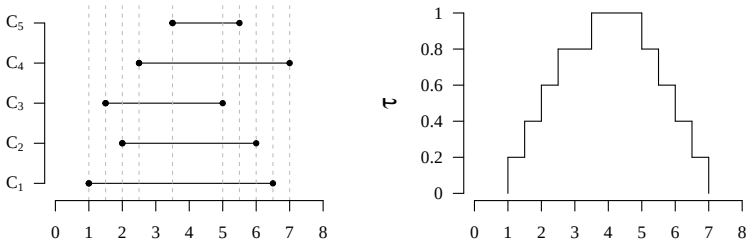

FIGURE 3.1: Visual representation of the majority vote procedure.

computationally efficient procedure. We call this method the "Median of Midpoints".

Theorem 3.7. *Suppose the input sets $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are intervals having the same width, and let $\mathcal{C}^M$ be their majority vote set. Let the midpoint of $\mathcal{C}_k$ be denoted by c_k. If K is odd, let $c_{(\lceil K/2 \rceil)}$ denote their median, and let $\mathcal{C}^{(K/2)}$ denote the interval whose midpoint is $c_{(\lceil K/2 \rceil)}$. If K is even, let $\mathcal{C}^{(K/2)}$ be defined by the intersection of the sets whose midpoints are $c_{(K/2)}$ and $c_{(1+K/2)}$. Then, $\mathcal{C}^{(K/2)} \supseteq \mathcal{C}^M$ and is hence also a $1-2\alpha$ uncertainty set. Further, $\mathcal{C}^{(K/2)}$ has at most the same width as the input sets. If $\cap_{k=1}^K \mathcal{C}_k \neq \varnothing$, then $\mathcal{C}^{(K/2)} = \mathcal{C}^M$.*

Proof. Assume for simplicity that all intervals (and hence midpoints) are distinct. Sort the intervals by their midpoints: let $\mathcal{C}_{(k)}$ denote the k-th ordered set if its midpoint is $c_{(k)}$. We first note that if the intervals have the same width, then the intervals that contain the target c must form a contiguous set according to this ordering. To elaborate, it is apparent that the covering intervals, if there are any, must be $\mathcal{C}_{(a)}, \mathcal{C}_{(a+1)}, \ldots, \mathcal{C}_{(b)}$ for some $1 \leq a \leq b \leq K$; indeed, it is not possible for $\mathcal{C}_{(a)}$ and $\mathcal{C}_{(a+2)}$ to cover without $\mathcal{C}_{(a+1)}$ also covering (because they have the same width). The above observation implies the following: if $> K/2$ intervals cover the target c, then so does the median interval. In particular, if K is even then c must be contained both in the intervals with midpoints $c_{(K/2)}$ and $c_{(1+K/2)}$; in fact, if c is exclusively contained within one of the two intervals, then it is contained at most in $K/2$ of the intervals (and not $K/2+1$). If E denotes the event that $> K/2$ intervals contain the target c, we argued earlier that $\mathbb{P}(E) \geq 1-2\alpha$. More generally, if it happens to be the case that $> K/2$ intervals include an arbitrary point s, then the "median interval" $\mathcal{C}^{(K/2)}$ must also contain s. Thus, $\mathcal{C}^{(K/2)} \supseteq \mathcal{C}^M$ as claimed.

To prove the final claim, let us assume for simplicity that all midpoints are distinct and all intervals are closed. Let $c_{(1)}, \ldots, c_{(K)}$ be the ordered midpoints of $\mathcal{C}_1, \ldots, \mathcal{C}_K$ and δ denote the width of the intervals. The boundaries of the intervals can be defined as

$$a_{(k)} := c_{(k)} - \delta/2, \quad b_{(k)} := c_{(k)} + \delta/2, \quad k = 1, \ldots, K.$$

By definition, we have $a_{(1)} < \cdots < a_{(K)}$ and $b_{(1)} < \cdots < b_{(K)}$. In addition, since $\cap_{k=1}^{K} \mathcal{C}_k \neq \varnothing$ then $\cap_{k=1}^{K} \mathcal{C}_k$ must coincide with $[a_{(K)}, b_{(1)}]$. This implies that

$$a_{(1)} < \cdots < a_{(K)} < b_{(1)} < \cdots < b_{(K)},$$

and if $s \in [a_{(K)}, b_{(1)}]$ then it is *voted* by all the intervals, in symbols, $\sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} = K$. With similar arguments, we find that if $s \in [a_{(K-1)}, a_{(K)})$ or $s \in (b_{(1)}, b_{(2)}]$ then $\sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} = K - 1$; since s will belong to all intervals except $\mathcal{C}_K$ in the first case, while s will belong to all intervals except $\mathcal{C}_1$ in the second case. In general, we see that if $s \in [a_{(K-j)}, a_{(K-j+1)})$ or $s \in (b_{(j)}, b_{(j+1)}]$ then $\sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} = K - j$, for $j = 1, \ldots, K-1$. This implies that the majority set $\mathcal{C}^M$ is defined as

$$\mathcal{C}^M = \begin{cases} \left(\cup_{j=\lceil K/2 \rceil}^{K-1} [a_{(j)}, a_{(j+1)})\right) \bigcup [a_{(K)}, b_{(1)}] \bigcup \left(\cup_{j=1}^{\lceil K/2 \rceil - 1} (b_{(j)}, b_{(j+1)}]\right) = [a_{\lceil K/2 \rceil}, b_{\lceil K/2 \rceil}], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } K \text{ is odd;} \\ \left(\cup_{j=K/2+1}^{K-1} [a_{(j)}, a_{(j+1)})\right) \bigcup [a_{(K)}, b_{(1)}] \bigcup \left(\cup_{j=1}^{K/2-1} (b_{(j)}, b_{(j+1)}]\right) = [a_{K/2+1}, b_{K/2}], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } K \text{ is even;} \end{cases}$$

which concludes the claim. □

3.1.5 How large is the majority vote set?

One naive way to combine the K sets is to randomly select one of them as the final set; this method clearly has coverage $1 - \alpha$, and its length is in between their union and intersection, so it seems reasonable to ask how it compares to majority vote. Surprisingly, majority vote is not always strictly better than this approach in terms of the expected length of the set: consider, for example, three nested intervals $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ of width $10, 8$ and 3, respectively. The majority vote set is $\mathcal{C}_2$, with a length of 8, but randomly selecting an interval results in an average length of 7. However, we show next that the majority vote set cannot be more than twice as large, and for interval inputs, never wider than the largest interval.

Theorem 3.8. *Let $\nu(\mathcal{C}^\tau)$ be the measure (size) of $\mathcal{C}^\tau$ in* (3.4)*. Then, for all $\tau \in [0, 1)$,*

$$\nu(\mathcal{C}^\tau) \leq \frac{1}{K\tau} \sum_{k=1}^{K} \nu(\mathcal{C}_k). \tag{3.6}$$

If the input sets are K one-dimensional intervals, for all $\tau \in [\frac{1}{2}, 1)$, we have

$$\nu(\mathcal{C}^\tau) \leq \max_k \nu(\mathcal{C}_k). \tag{3.7}$$

Proof. The first part involves the observation that $\mathbb{1}\{y > 1\} \leq y$ for $y \geq 0$:

$$\begin{aligned}\nu(\mathcal{C}^\tau) &= \int_{\mathcal{S}} \mathbb{1}\left\{\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{x \in \mathcal{C}_k\} > \tau\right\} d\nu(x) \\ &\leq \int_{\mathcal{S}} \frac{1}{K\tau}\sum_{k=1}^{K} \mathbb{1}\{x \in \mathcal{C}_k\} d\nu(x) = \frac{1}{K\tau}\sum_{k=1}^{K} \nu(\mathcal{C}_k),\end{aligned}$$

where integration is with respect to the measure ν.

For the second part, let the endpoints of $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be denoted by $(a_1, b_1), \ldots, (a_K, b_K)$. Let $\bar{m}_j := (a_j + b_j)/2$ be the midpoint of the interval $\mathcal{C}_j$, $j = 1, \ldots, K$. Without loss of generality, let us suppose that the length of the largest interval is $\max_k(b_k - a_k) = (b_1 - a_1) =: \delta$. Let us define a new collection of intervals $\mathcal{C}_1, \mathcal{C}_2^\delta, \ldots, \mathcal{C}_K^\delta$, where the interval $\mathcal{C}_j^\delta$ has endpoints

$$a_j^\delta = \bar{m}_j - \frac{\delta}{2}, \quad b_j^\delta = \bar{m}_j + \frac{\delta}{2},$$

for $j = 2, \ldots, K$. This implies that $\mathcal{C}_j \subseteq \mathcal{C}_j^\delta$ and intervals $\mathcal{C}_1, \mathcal{C}_2^\delta, \ldots, \mathcal{C}_K^\delta$ have the same width. Now, we can prove that $\mathcal{C}^M(\mathcal{C}_1, \ldots, \mathcal{C}_K) \subseteq \mathcal{C}^M(\mathcal{C}_1, \mathcal{C}_2^\delta, \ldots, \mathcal{C}_K^\delta)$. Indeed, if the point $s \in \mathcal{C}^M(\mathcal{C}_1, \ldots, \mathcal{C}_K)$ then it is contained in at least more than $K/2$ of the initial intervals, but by definition s must be included in the same set of the new intervals since $\mathcal{C}_j \subseteq \mathcal{C}_j^\delta$. From Theorem 3.7, we have $\mathcal{C}^M(\mathcal{C}_1, \mathcal{C}_2^\delta, \ldots, \mathcal{C}_K^\delta) \subseteq \mathcal{C}^{(K/2)}$. So, if K is odd, then $\mathcal{C}^{(K/2)}$ is an interval with width δ, while if K is even it corresponds to the intersection of two equal-sized intervals, so its width is $\leq \delta$. Taking advantage of the fact that $\mathcal{C}^\tau$ is decreasing in τ (i.e., $\mathcal{C}^{\tau_2} \subseteq \mathcal{C}^{\tau_1}$ if $\tau_1 \leq \tau_2$) then $\nu(\mathcal{C}^\tau) \leq \max_k \nu(\mathcal{C}_k)$ continues to hold for all $\tau \in [\frac{1}{2}, 1)$. □

In (3.6), ν could be the Lebesgue measure (for intervals), or the counting measure (for discrete, categorical sets), for example. The proof of (3.7) uses the fact that the majority vote set is elementwise monotonic in its input sets, meaning that if any of the input sets gets larger, the majority vote set can never get smaller. From (3.6) we have that if $\tau = 1/2$, then the measure of the majority vote set is never larger than 2 times the average of the measure of the initial sets. This result is essentially tight as can be seen in the following example involving one-dimensional intervals. For odd K, let $(K+1)/2$ intervals have a length L, while the rest have a length of nearly 0. The average length is

then $(K+1)L/(2K)$, and the majority vote has length L, whose ratio approaches $1/2$ for large K. Clearly, this is just an example pointing out the tightness of the obtained bound; in practice, we usually observe a less adversarial behavior. In addition, (3.6) gives the right bound for the intersection ($\tau \uparrow 1$) and for the union ($\tau \uparrow 1/K$).

Also the bound in (3.7) is tight as can be seen from the following example. Consider a scenario similar to that of the last example. Suppose we have $K > 2$ confidence intervals, with odd K, such that $\mathcal{C}_1 = \cdots = \mathcal{C}_{\lceil K/2 \rceil} = (-\epsilon, 1)$ and $\mathcal{C}_{\lceil K/2 \rceil + 1} = \cdots = \mathcal{C}_K = (0, 1+\epsilon)$, for some $\epsilon > 0$. By definition, if $\tau = (K-2)/(2K) < 1/2$, then the set $\mathcal{C}^\tau$ coincides with the union of the initial sets, which is larger than the input intervals and has width $1+2\epsilon$. On the other hand, choosing $\tau = 1/2$, as a consequence of Theorem 3.7 we obtain a set whose size is $1+\epsilon$, matching the width of the input sets. This fact provides a practical justification for choosing $\tau = 1/2$: it is the smallest τ for which the combined set cannot be larger than the input intervals. In particular, a simple majority vote seems to offer a good compromise between coverage and size.

3.1.6 Combining independent or nested confidence sets

When $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are independent, then a more tailored combination rule can yield a smaller set with exactly $1-\alpha$ coverage. The rule is similar to (3.2), albeit with a different threshold, which is related to the quantile of a binomial distribution with K trials and parameter $1-\alpha$. We define $Q_K(\alpha)$ as the α-quantile of a Binom$(K, 1-\alpha)$:

$$Q_K(\alpha) := \max\{x \in \{0, \ldots, K\} : F(x) \leq \alpha\}, \tag{3.8}$$

where $F(\cdot)$ is the cumulative distribution function of a Binom$(K, 1-\alpha)$.

Proposition 3.9. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be $K \geq 2$ different independent sets satisfying* (3.1) *and for a fixed parameter of interest c. Then, the set*

$$\mathcal{C}^M = \left\{ s \in \mathcal{S} : \sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} > Q_K(\alpha) \right\}$$

has coverage equal to $1-\alpha$.

The next lemma will be needed to prove Proposition 3.9. In the following, we denote $X \sim PB(p_1, \ldots, p_K)$ as a Poisson binomial random variable distributed as the sum of K independent Bernoulli random variables with parameters $p_1, \ldots, p_K$.

Lemma 3.10 (Boland *et al.* (2002); Tang and Tang (2023))**.** *Let* $X \sim PB(p_1, \ldots, p_K)$ *and* $Y \sim \text{Binom}(K, p)$ *then* X *is stochastically larger than* Y*,* $X \geq_{st} Y$*, if*

$$p^K \leq \prod_{k=1}^{K} p_k.$$

The proof is given in Boland *et al.* (2002) and discussed in Tang and Tang (2023).

Proof of Proposition 3.9. Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be a collection of independent confidence sets for the parameter c. Then $\phi_k = \mathbb{1}\{c \in \mathcal{C}_k\}$ is a Bernoulli random variable with parameter $p_k \geq 1 - \alpha$. In addition, $\phi_1, \ldots, \phi_K$ are independent (transformation of independent quantities). Suppose that $p_k = 1 - \alpha$, for all $k = 1, \ldots, K$, then

$$\mathbb{P}(c \in \mathcal{C}^M) = \mathbb{P}\left(\sum_{k=1}^{K} \mathbb{1}\{c \in \mathcal{C}_k\} > Q_K(\alpha)\right) = \mathbb{P}(S_K > Q_K(\alpha)) \geq 1 - \alpha,$$

where $S_K = \sum_{k=1}^{K} \phi_k \sim \text{Binom}(K, 1-\alpha)$ and $Q_K(\alpha)$ defined in (3.8). If $p_k \geq 1-\alpha$, $k = 1, \ldots, K$, then S_K is distributed as a Poisson binomial with parameters $p_1, \ldots, p_K$ and $\prod_{k=1}^{K} p_k \geq (1-\alpha)^K$. This implies that S_K is stochastically larger than a $\text{Binom}(K, 1-\alpha)$ due to Lemma 3.10. □

We emphasize that this result requires that c be a fixed quantity. If c were to be random, the independence between the events $\mathbb{1}\{c \in \mathcal{C}_k\}$ and $\mathbb{1}\{c \in \mathcal{C}_l\}$, with $k \neq l$, would be compromised even if the sets were based on independent observations.

Another (trivial) special case where it is possible to achieve a coverage of level $1 - \alpha$ appears when the sets are almost surely nested (not necessarily independent). Let us suppose that $\mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_K$ holds almost surely and we obtain the set $\mathcal{C}^M$ as in (3.2). By definition, all the points contained in $\mathcal{C}_1$ will be part of the set $\mathcal{C}^M$, which implies that $\mathcal{C}^M$ is a set with confidence level equal to $1 - \alpha$. But of course, in that case, $\mathcal{C}_1$ is itself a smaller and valid combination. If some, but not all, the sets are almost surely nested, the natural way to merge them is to pick the smallest one of the nested ones, and combine it with the others via majority vote.

3.1.7 Combining exchangeable uncertainty sets

In many practical applications, the independence of sets is often violated. Surprisingly, when $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are not independent, but are exchangeable, something better than a naive majority vote can be accomplished. To describe the method, denote the set (3.2) as $\mathcal{C}^M(1:K)$ to highlight that it is based on the majority vote of sets $\mathcal{C}_1, \ldots, \mathcal{C}_K$. Now

define

$$\mathcal{C}^E := \bigcap_{k=1}^{K} \mathcal{C}^M(1:k),$$

which can be equivalently represented as

$$\mathcal{C}^E = \left\{ s \in \mathcal{S} : \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}\{s \in \mathcal{C}_j\} > \frac{1}{2} \text{ for all } k \leq K \right\}. \tag{3.9}$$

Essentially, $\mathcal{C}^E$ is formed by the intersection of sets obtained through sequential processing of the sets derived from the majority vote.

Theorem 3.11. *If $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are $K \geq 2$ exchangeable sets having coverage $1-\alpha$, then $\mathcal{C}^E$ is a $1-2\alpha$ uncertainty set, and it is never worse than majority vote ($\mathcal{C}^E \subseteq \mathcal{C}^M$).*

Proof. Let $\phi_k = \mathbb{1}\{c \notin \mathcal{C}_k\}$ be Bernoulli random variable such that $\mathbb{E}[\phi_k] \leq \alpha$, $k = 1, \ldots, K$. Then, if the sets are exchangeable (jointly with c, if random), we have that $\phi_1, \ldots, \phi_K$ are exchangeable and

$$\mathbb{P}(c \notin \mathcal{C}^E) = \mathbb{P}\left(\exists k \leq K : c \notin \mathcal{C}^M(1:k)\right) = \mathbb{P}\left(\exists k \leq K : \frac{1}{k}\sum_{j=1}^{k} \phi_j \geq \frac{1}{2}\right) \leq 2\mathbb{E}\left[\phi_1\right] \leq 2\alpha,$$

where the first inequality holds due to the exchangeable Markov inequality in Theorem 1.6. It is simple to see that $\mathcal{C}^E \subseteq \mathcal{C}^M$, since $\mathcal{C}^M(1:K)$ coincides with $\mathcal{C}^M$. □

We note that $\mathcal{C}^M(1:2)$ is the intersection of $\mathcal{C}^1$ and $\mathcal{C}^2$, so we can omit $\mathcal{C}^M(1:1) = \mathcal{C}^1$ from the intersection defining $\mathcal{C}^E$, to observe that $\mathcal{C}^E = \bigcap_{k=2}^{K} \mathcal{C}^M(1:k)$. Since the set $\mathcal{C}^E$ depends on the order of arrival of the different sets, one may worry that $\mathcal{C}^E$ is less stable than $\mathcal{C}^M$ (that treats the sets symmetrically). However, the exchangeable method allows the user to combine the obtained sets sequentially and to stop when the procedure stabilizes.

Despite the previous result working only for exchangeable sets, it points out a simple way at improving majority vote for arbitrarily dependent sets: process them in a random order. To elaborate, let π be a uniformly random permutation of $\{1, 2, \ldots, K\}$ that is independent of the K sets, and define

$$\mathcal{C}^\pi := \bigcap_{k=1}^{K} \mathcal{C}^M(\pi(1):\pi(k)). \tag{3.10}$$

Since $\mathcal{C}^M(\pi(1):\pi(K)) = \mathcal{C}^M(1:K)$, $\mathcal{C}^\pi$ is also never worse than majority vote despite satisfying the same coverage guarantee:

Corollary 3.12. *If $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are $K \geq 2$ arbitrarily dependent uncertainty sets having coverage $1-\alpha$, and π is a uniformly random permutation independent of them, then $\mathcal{C}^\pi$ is a $1-2\alpha$ uncertainty set, and it is never worse than majority vote ($\mathcal{C}^\pi \subseteq \mathcal{C}^M$).*

The proof follows as a direct corollary of Theorem 3.11 by noting that the random permutation π induces exchangeability of the sets (the joint distribution of every permutation of sets is the same, due to the random permutation). Of course, if the sets were already "randomly labeled" (for example, to make sure there was no special significance to the labels), then the aggregator does not need to perform an extra random permutation.

3.1.8 Improving majority vote via random thresholding

Moving in a different direction below, we demonstrate that the majority vote can be improved with the aim of achieving a tighter set through the use of independent randomization, while maintaining the same coverage.

Let U be an independent random variable distributed uniformly on $[0,1]$ (i.e., $U \in \mathcal{U}$). We then define a new set $\mathcal{C}^R$ as:

$$\mathcal{C}^R := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} > \frac{1}{2} + U/2 \right\}. \tag{3.11}$$

As a small variant, define

$$\mathcal{C}^U := \left\{ s \in \mathcal{S} : \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} > U \right\}. \tag{3.12}$$

Theorem 3.13. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be $K \geq 2$ different sets satisfying the property* (3.1)*. Then, the set $\mathcal{C}^R$ has coverage at least $1-2\alpha$ and is never larger than $\mathcal{C}^M$, while $\mathcal{C}^U$ has coverage at least $1-\alpha$ and is never smaller than $\mathcal{C}^R$.*

The proof follows as a special case of the next subsection's result and is thus omitted. Even though $\mathcal{C}^U$ does not improve on $\mathcal{C}^M$, we include it since it involves random thresholding and delivers the same coverage as the input sets, a feature that we do not know how to obtain without randomization (unless in one of the cases described in Section 3.1.6).

There may of course be certain concerns with using randomization when a human is actively involved with data analysis, due to fears of a form of "p-hacking" caused by repeatedly running the above procedure many times and picking the most suitable merged set for the story they wish to tell. There is no fool-proof way to avoid this. We

recommend using randomization in automated pipelines because it is more statistically efficient to do so, but we suggest using the non-randomized procedures when humans are involved.

3.1.9 Weighted majority vote

It is not unusual for each interval to be assigned distinct "weights" (importances) in the voting procedure. This can occur, for instance, when prior studies empirically demonstrate that specific methods for constructing uncertainty sets consistently outperform others. Alternatively, a researcher might assign varying weights to the sets based on their own prior insights. Assume as before that the sets $\mathcal{C}_1, \ldots, \mathcal{C}_K$ based on the observed data follow the property (3.1), and let $w = (w_1, \ldots, w_K)$ be a set of weights, such that

$$w_k \in [0, 1] \quad \text{and} \quad \sum_{k=1}^{K} w_k = 1. \tag{3.13}$$

These weights can be interpreted as the aggregator's prior belief in the quality of the received sets. A higher weight signifies that we attribute greater importance to that specific interval. As before, let U be an independent random variable that is distributed uniformly on $[0, 1]$. We then define a new set $\mathcal{C}^W$ as:

$$\mathcal{C}^W := \left\{ s \in \mathcal{S} : \sum_{k=1}^{K} w_k \mathbb{1}\{s \in \mathcal{C}_k\} > \frac{1}{2} + U/2 \right\}. \tag{3.14}$$

Theorem 3.14. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be $K \geq 2$ different sets satisfying property* (3.1)*. Then, the set $\mathcal{C}^W$ defined in* (3.14) *has coverage of at least $1 - 2\alpha$:*

$$\mathbb{P}\big(c \in \mathcal{C}^W\big) \geq 1 - 2\alpha. \tag{3.15}$$

In addition, let $\nu(\mathcal{C}^W)$ be the measure associated with the set $\mathcal{C}^W$, then

$$\nu(\mathcal{C}^W) \leq 2 \sum_{k=1}^{K} w_k \nu(\mathcal{C}_k). \tag{3.16}$$

Proof. Let $\phi_k = \mathbb{1}\{c \notin \mathcal{C}_k\}$ be a Bernoulli random variable such that $\mathbb{E}[\phi_k] \leq \alpha$, $k = 1, \ldots, K$. Then using Uniformly-randomized Markov inequality (Theorem 1.7),

$$\begin{aligned}\mathbb{P}(c \notin \mathcal{C}^W) &= \mathbb{P}\left(\sum_{k=1}^{K} w_k (1 - \phi_k) \leq \frac{1}{2} + U/2\right) = \mathbb{P}\left(\sum_{k=1}^{K} w_k \phi_k \geq \frac{1}{2} - U/2\right) \\ &\leq 2\mathbb{E}\left[\sum_{k=1}^{K} w_k \phi_k\right] = 2\sum_{k=1}^{K} w_k \mathbb{E}\left[\phi_k\right] \leq 2\alpha \sum_{k=1}^{K} w_k = 2\alpha,\end{aligned}$$

which proves (3.15).

In order to prove (3.16), we follow the same lines as in Theorem 3.8. In particular,

$$\begin{aligned}\nu(\mathcal{C}^W) &= \int_{\mathcal{S}} \mathbb{1}\left\{\sum_{k=1}^{K} w_k \mathbb{1}\{x \in \mathcal{C}_k\} > \frac{1}{2} + \frac{u}{2}\right\} d\nu(x) \leq \int_{\mathcal{S}} \mathbb{1}\left\{\sum_{k=1}^{K} w_k \mathbb{1}\{x \in \mathcal{C}_k\} > \frac{1}{2}\right\} d\nu(x) \\ &\leq \int_{\mathcal{S}} 2\sum_{k=1}^{K} w_k \mathbb{1}\{x \in \mathcal{C}_k\} d\nu(x) = 2\sum_{k=1}^{K} w_k \nu(\mathcal{C}_k),\end{aligned}$$

which concludes the proof. □

If the weights are equal to $w_k = \frac{1}{K}$, for all k, then the set $\mathcal{C}^W$ coincides with the set $\mathcal{C}^R$ defined in (3.11) and it is a subset of that in (3.2). This means that in the case of a democratic (equal-weighted) vote $\mathcal{C}^R \subseteq \mathcal{C}^M$, since $\mathcal{C}^M$ is obtained by choosing $U = 0$. Furthermore, (3.16) says that the measure of the set obtained using the weighted majority method cannot be more than twice the average measure obtained by randomly selecting one of the intervals with probabilities proportional to w. When there are only two sets and randomization is avoided, the weighted majority vote set will correspond to the set with the greater weight. In Appendix B.3 we extend the result to the case where the number of the sets can be uncountable.

3.1.10 Different coverage levels and asymptotic coverage

It may happen that the property (3.1) is not met, meaning that the sets may have different coverage. The agents may intend to provide a $1 - \alpha$ confidence set, but may unintentionally overcover or undercover, or some agents could be malevolent. As examples of the former case, we know that under regularity conditions confidence intervals constructed using likelihood methods have an asymptotic coverage of level $1 - \alpha$ Pace and Salvan (1997), but the asymptotics may not yet have kicked in or the regularity conditions may not hold. Another example appears in the conformal prediction framework when the exchangeability assumption is not satisfied but we do not know the amount of

deviation from exchangeability, as considered in Barber *et al.* (2023). As another example in conformal prediction, the jackknife+ method by Barber *et al.* (2021) when run at level α may deliver coverage anywhere between $1-2\alpha$ and 1, even under exchangeability. As a last example, a Bayesian agent may provide a credible interval, which may not be a valid confidence set in the frequentist sense. The set in (3.14) still delivers a sensible guarantee.

Proposition 3.15. *If $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are $K \geq 2$ different sets having coverage $1-\alpha_1, \ldots, 1-\alpha_K$ (possibly unknown), then the set $\mathcal{C}^W$ defined in* (3.14) *has coverage*

$$\mathbb{P}(c \in \mathcal{C}^W) \geq 1 - 2\sum_{k=1}^{K} w_k \alpha_k.$$

In particular, this implies that the majority vote of asymptotic $(1-\alpha)$ intervals has asymptotic coverage at least $(1-2\alpha)$.

The proof is identical to that of Theorem 3.14, with the exception that the expected value for the variables ϕ_k is equal to α_k, and is thus omitted. If the α_k levels are known (which they may not be, unless the agents report it and are accurate), and if one in particular wishes to achieve a target level $1-\alpha$, then it is always possible to find weights $(w_1, \ldots, w_K)$ that achieve this as long as $\alpha/2$ is in the convex hull of $(\alpha_1, \ldots, \alpha_K)$. Since it is desirable to have as small an interval as possible if coverage (3.1) is respected, we would like to assign a higher weight to intervals of smaller size. However, the weights must be assigned before seeing the sets.

3.1.11 Sequentially combining uncertainty sets

We show simple extensions of the results to two different sequential settings:

Sequential data: Imagine that we observe data sequentially one at a time $Z_1, Z_2, \ldots$ and wish to estimate some parameter c with increasing accuracy as we observe more samples, and wish to continuously monitor the resulting confidence intervals as they get tighter over time. A $(1-\alpha)$-confidence sequence $(\mathcal{C}^{(t)})_{t\geq 1}$ for a parameter of interest c is a time-uniform confidence interval:

$$\mathbb{P}(\forall t \geq 1 : c \in \mathcal{C}^{(t)}) \geq 1 - \alpha.$$

Here t indexes sample size that is used to calculate a single agent's confidence interval $\mathcal{C}^{(t)}$ (meaning that $\mathcal{C}^{(t)}$ is based on $(Z_1, \ldots, Z_t)$). Suppose now that we have K different

confidence sequences for a parameter that need to be combined into a single confidence sequence. For this setting we show a simple result:

Proposition 3.16. *Given K different $1-\alpha$ confidence sequences for the same parameter that are being tracked in parallel, their majority vote set is a $1-2\alpha$ confidence sequence.*

It may not be initially apparent how to deal with the time-uniformity in the definitions. The proof proceeds by first observing that an equivalent definition of a confidence sequence is a confidence interval that is valid at any arbitrary stopping time τ (here the underlying filtration is implicitly that generated by the data itself). Howard *et al.* (2021) proved that $(\mathcal{C}_k^{(t)})_{t\geq 1}$ is a confidence sequence if and only if for every stopping time τ, $\mathbb{P}(c \in \mathcal{C}_k^{(\tau)}) \geq 1-\alpha$, for all k. Now, the proof follows by applying earlier results.

Sequential set arrival: Now, let the data be fixed, but consider the setting where an unknown number of confidence sets arrive one at a time in a random order, and need to be combined on the fly. Now, we propose to simply take a majority vote of the sequences we have seen thus far. Borrowing terminology from earlier, denote

$$\mathcal{C}^E(1:t) := \bigcap_{i=1}^{t} \mathcal{C}^M(1:i). \tag{3.17}$$

We claim that the above sequence of sets is actually a $1-2\alpha$ confidence *sequence* for c:

Theorem 3.17. *Given an exchangeable sequence of confidence sets $\mathcal{C}_1, \mathcal{C}_2, \ldots$ (or confidence sets arriving in a uniformly random order), the sequence of sets formed by their "running majority vote" $(\mathcal{C}^E(1:t))_{t\geq 1}$ is a $1-2\alpha$ confidence sequence:*

$$\mathbb{P}\left(\exists t \geq 1 : c \notin \mathcal{C}^E(1:t)\right) \leq 2\alpha.$$

Proof. By direct calculation

$$\begin{aligned}\mathbb{P}(\exists t \geq 1 : c \notin \mathcal{C}^E(1:t)) &= \mathbb{P}\left(\cup_{T\geq 1}\{\exists t \leq T : c \notin \mathcal{C}^E(1:t)\}\right) \\ &= \lim_{T\to\infty} \mathbb{P}(c \notin \mathcal{C}^E(1:T)) \leq 2\alpha,\end{aligned}$$

where the last equality is due to the fact that $\mathcal{C}^E(1:t)$ shrinks when t increases, while the last inequality is due to Theorem 3.11. □

3.2 Simulation study: Private multi-agent CI

As a case study, we employ the majority voting method in a situation where certain public data may be available to all agents, and certain private data may only be available

to one (or a few) but not all agents. Consider a scenario involving K distinct agents, each providing a *locally differentially private* confidence interval for a common parameter of interest. As opposed to the centralized model of differential privacy in which the aggregator is trusted, local differential privacy is a stronger notion that does not assume a trusted aggregator, and privacy is guaranteed at an individual level at the source of the data. Further details about the definition of local privacy are not important for understanding this example; interested readers may consult Dwork and Roth (2014).

For $k = 1, \ldots, K$, suppose that the k-th agent has data about n individuals $(X_{1,k}, \ldots, X_{n,k})$ that they wish to keep locally private (we assume each agent has the same amount of data for simplicity). They construct their "locally private interval" based on the data $(Z_{1,k}, \ldots, Z_{n,k})$, which represents privatized views of the original data. Suppose that an unknown fraction of the observations may be shared among agents, indicating that the reported confidence sets are not independent. An example of such a scenario could be a medical study, where each patient represents an observation, and a significant but unknown number of patients may be shared among different research institutions, or some amount of public data may be employed. Consequently, the confidence intervals generated by various research centers (agents) are not independent.

In the following, we refer to the scenario described in Waudby-Smith *et al.* (2023), where the data $(X_{1,k}, .., X_{n,k}) \sim P$, and P is any $[0, 1]$-valued distribution with mean θ^*. Data $(Z_{1,k}, .., Z_{n,k})$ are ε-locally differentially private (ε-LDP) views of the original data obtained using the nonparametric randomized response mechanism. The mechanism requires one additional parameter, G, which we set to a value of 1 for simplicity (the mechanism stochastically rounds the data onto a grid of size $G + 1$, which is two in our case: the boundary points of 0 and 1).

A possible solution to construct a (locally private) confidence interval for the mean parameter θ^* is to use the locally private Hoeffding inequality as proposed in Waudby-Smith *et al.* (2023). In particular, let $\hat{\theta}_k$ be the adjusted sample mean for agent k:

$$\hat{\theta}_k := \frac{\sum_{i=1}^{n} Z_{i,k} - n(1-r)/2}{nr},$$

where $r := (\exp\{\varepsilon\} - 1)/(\exp\{\varepsilon\} + 1)$. Then the interval

$$\mathcal{C}_k = \left[\hat{\theta}_k - \sqrt{\frac{-\log(\alpha/2)}{2nr^2}}, \hat{\theta}_k + \sqrt{\frac{-\log(\alpha/2)}{2nr^2}}\right]$$

is a valid $(1 - \alpha)$-confidence interval for the mean θ^*. The width of the confidence interval depends solely on the number of observations n, coverage α, and privacy ε.

Scenario	p	$\mathcal{C}_k$	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	$\mathcal{C}^\pi$
(i)	-	0.3214	0.3058	0.2282	0.3212	0.3210
		(0.9880)	(1.0000)	(0.9752)	(0.9892)	(0.9892)
(ii)	0.5	0.3214	0.3062	0.2302	0.3218	0.3215
		(0.9877)	(1.0000)	(0.9749)	(0.9879)	(0.9879)
(ii)	0.75	0.3214	0.3062	0.2302	0.3218	0.3215
		(0.9877)	(1.0000)	(0.9749)	(0.9879)	(0.9879)
(ii)	0.9	0.3214	0.3063	0.2297	0.3223	0.2319
		(0.9877)	(1.0000)	(0.9764)	(0.9870)	(0.9775)

TABLE 3.1: Empirical average length of intervals and corresponding average coverage (within brackets) for the two simulation scenarios. In the second scenario, the percentage of shared observations is denoted as p. The α-level is set to 0.1 — the first column shows that the employed confidence interval is conservative (but tighter ones are more tedious to describe). The majority vote set is smaller than the individual ones, but it overcovers (despite the theoretically guarantee being one that permits some undercoverage), which is an intriguing phenomenon. The randomized majority vote method produces the smallest sets than the others while maintaining good coverage. Randomized voting is not very different from the original intervals.

Once the various agents have provided their confidence sets, a non-trivial challenge may arise in merging them to obtain a unique interval for the parameter of interest. One possible solution is to use the majority-vote procedure described in the previous sections. We conducted a simulation study within this framework. In the first scenario, at each iteration, $n \times (K/2)$ observations were generated from a standard uniform random variable, and each agent was randomly assigned n observations. In the second scenario, the first agent had n observations generated from a uniform random variable. For all agents with $k > 1$, a percentage p of their observations was shared with the preceding agent, while the remaining portion was generated from Unif$(0, 1)$. In both scenarios, the number of agents is equal to 10, the number of observations (n) is common among the agents and equal to 100, while the privacy parameter ε is set to 2 (which is an appropriate value for local privacy, indeed Apple uses a value of 4 to collect data from iPhones; see Apple Inc. (2022)). The number of replications for each scenario is 10 000.

As can be seen in Table 3.1, the length of the intervals constructed by the agents remains constant throughout the simulations, since the values of n and ε remain unchanged. In contrast, the intervals formed by the majority and randomized majority methods are smaller, compared to those constructed by individual agents. The coverage level achieved by individual agents' intervals (first column) significantly exceeds the threshold of $1 - \alpha$, but this is expected since the intervals are nonasymptotically valid and conservative. The coverage derived from the majority method is notably high, approaching 1. The incorporation of a randomization greatly reduces the length of the sets

while maintaining coverage at a slightly lower level than that of single agent-based intervals. The use of the randomized union introduced in (3.12) produces sets with nearly the same lenght and coverage as the ones produced by single agents. If the aggregator had access to the $(1-\alpha)$-confidence intervals level for all possible α, it would be able to derive the confidence distribution. The actual values of the length and coverage should not be given too much attention: there are other, more sophisticated, intervals derived in the aforementioned paper (empirical-Bernstein, or asymptotic) and these would have shorter lengths and less conservative coverage, but they take more effort to describe here in self-contained manner and were thus omitted.

3.3 Derandomizing statistical procedures

One application of the presented bounds is in the derandomization of existing randomized methods that are based in some way on data splitting. We explore different such methods here, such as the median-of-means (Devroye *et al.*, 2016) and HulC (Kuchibhotla *et al.*, 2024). The first of these produces a point estimator, while the second produces an interval, but both can be derandomized using the same set of ideas.

Since the chapter has so far focused on uncertainty sets, we first describe a general result that derandomizes point estimators "directly".

Theorem 3.18. *Suppose $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are K univariate point estimators of θ that are each built using n data points and satisfy a high probability concentration bound:*

$$\mathbb{P}(|\hat{\theta}_k - \theta| \leq w(n, \alpha)) \geq 1 - \alpha,$$

for some function w. Then, their median $\theta_{(\lceil K/2 \rceil)}$ satisfies

$$\mathbb{P}(|\hat{\theta}_{(\lceil K/2 \rceil)} - \theta| \leq w(n, \alpha)) \geq 1 - 2\alpha.$$

Further, if $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K, \hat{\theta}_{K+1} \ldots$ are exchangeable, then

$$\mathbb{P}(\forall K \geq 1 : |\hat{\theta}_{(\lceil K/2 \rceil)} - \theta| \leq w(n, \alpha)) \geq 1 - 2\alpha.$$

An analogous statement also holds if we instead had $\mathbb{P}(\ell(n, \alpha) \leq \hat{\theta}_k - \theta \leq u(n, \alpha)) \geq 1 - \alpha$, meaning that the two tails had different behaviors.

Proof. We provide two proofs: direct and indirect, starting with the latter. For each $k = 1, \ldots, K$, note that $\hat{\theta}_k \pm w(n, \alpha)$ is a $1 - \alpha$ confidence interval for θ. (If w involves other unknown nuisance parameters, we can pretend that these are being constructed

by an oracle). Now, Theorem 3.7 directly implies the result, by noting that the median of the midpoints of these K intervals is exactly $\hat{\theta}_{(\lceil K/2 \rceil)}$. For a direct proof, note that for $\hat{\theta}_{(\lceil K/2 \rceil)}$ to be more than $w(n, \alpha)$ away from θ, it would have to be the case that at least $\lceil K/2 \rceil$ of the individual $\hat{\theta}_k$ estimators would have to be more than $w(n, \alpha)$ away from θ, but the latter event happens with probability at most 2α, following the proof of the majority vote procedure. The exchangeability claim follows by an argument identical to Theorem 3.11. □

In the particular context of "double machine learning", Chernozhukov *et al.* (2018a, Corollary 3.3) also proposes repeating their sample-splitting based procedure several times and taking either the median or the mean of the resulting estimates, but their arguments justifying this choice are asymptotic. Further, these asymptotic arguments do not distinguish between the median and mean combination rules, but the authors recommend the median rule because it is more robust to outliers. Our justification above is nonasymptotic and rather different in flavor, and it applies in broader situations, beyond the ones considered in the above work. Indeed, derandomization can be applied in other contexts involving data splitting; some examples are Decrouez and Hall (2014); Banerjee *et al.* (2019); Kim and Ramdas (2024).

3.3.1 The Median-of-Median-of-Means (MoMoM) procedure

The aim of the celebrated median-of-means procedure is to produce estimators of the mean that have subGaussian tails, despite the underlying unbounded data having only a finite variance. It stems back to, at least, a book by Nemirovskij and Yudin (1983), but some modern references include Devroye *et al.* (2016) and Lugosi and Mendelson (2019).

The setup assumes n iid data points from an unknown univariate distribution P with unknown finite variance, whose mean μ we seek to estimate. The method works by randomly dividing the n points into B buckets of roughly n/B points each. One takes the mean within each bucket, and then calculates the median $\hat{\mu}^{\mathrm{MoM}}$ of those numbers.

The method does not produce confidence intervals for the mean; indeed, one needs to know a bound on the variance σ^2 for any nonasymptotic CI to exist. But the point estimator $\hat{\mu}^{\mathrm{MoM}}$ satisfies the following subGaussian tail bound: for any $t \geq 0$,

$$\mathbb{P}(|\hat{\mu}^{\mathrm{MoM}} - \mu| \geq C\sigma\sqrt{t/n}) \leq 2e^{-t},$$

for a constant $C = \sqrt{\pi} + o(1)$, where the $o(1)$ term vanishes as $B, n/B \to \infty$. This is a much faster rate than the $1/t^2$ on the right hand side obtained for the sample mean via

Chebyshev's inequality. However, one drawback of the method is that it is randomized, meaning that it depends on the random split of the data into buckets. Minsker (2023) proposed a derandomized variant that even improves the constant C to $\sqrt{2} + o(1)$. However, it is computationally intensive: it involves taking the median of *all possible sets* of size n/B.

Here, we point out that Theorem 3.18 yields a simple way to derandomize $\hat{\mu}^{\mathrm{MoM}}$. We simply repeat the median-of-means procedure independently K times to obtain exchangeable estimators $\{\hat{\mu}_k^{\mathrm{MoM}}\}_{k=1}^K$, and then report the "median of median of means" (MoMoM):

$$\hat{\mu}_K^{\mathrm{MoMoM}} := \hat{\mu}_{\lceil K/2 \rceil}^{\mathrm{MoM}}.$$

Clearly, $\{\hat{\mu}_k^{\mathrm{MoM}}\}_{k=1}^K$ form an exchangeable set, whose empirical distribution will stabilize as $K \to \infty$, and thus whose median will also be stabilize for large K.

Corollary 3.19. *Under the setup described above, the* median of median of means *satisfies a nearly identical subGaussian behavior as the original:*

$$\mathbb{P}\left(|\hat{\mu}_K^{\mathrm{MoMoM}} - \mu| \geq C\sigma\sqrt{t/n}\right) \leq 4e^{-t},$$

for any $t \geq 0$ *and* $C = \sqrt{\pi} + o(1)$*, as before. In fact,*

$$\mathbb{P}\left(\exists K \geq 1 : |\hat{\mu}_K^{\mathrm{MoMoM}} - \mu| \geq C\sigma\sqrt{t/n}\right) \leq 4e^{-t}.$$

The proof is an immediate consequence of Theorem 3.18 and is thus omitted. The simultaneity over K allows us to choose any sequence of constants $k_1 \leq k_2 \leq \ldots$ and produce the sequence of estimators $\hat{\mu}_{k_1}^{\mathrm{MoMoM}}, \hat{\mu}_{k_2}^{\mathrm{MoMoM}}, \ldots$, and plot this sequence as it is being generated. We can then stop the derandomization whenever the plot appears to stabilize, with the guarantee now holding for whatever data-dependent random K we stopped at. Clearly, one can use the same technique to derandomize other random estimators by taking their median while "importing" their nonasymptotic bounds.

3.3.1.1 Simulation study

As previously mentioned, a potential method to derandomize the MoM is to repeat the procedure K times and subsequently take the median of the estimators. The natural question that arises is: how large should K be in practice to achieve a derandomized result? Indeed, considering computational and time resources, one would prefer a small value for K. We conducted a simulation study to study the impact of K.

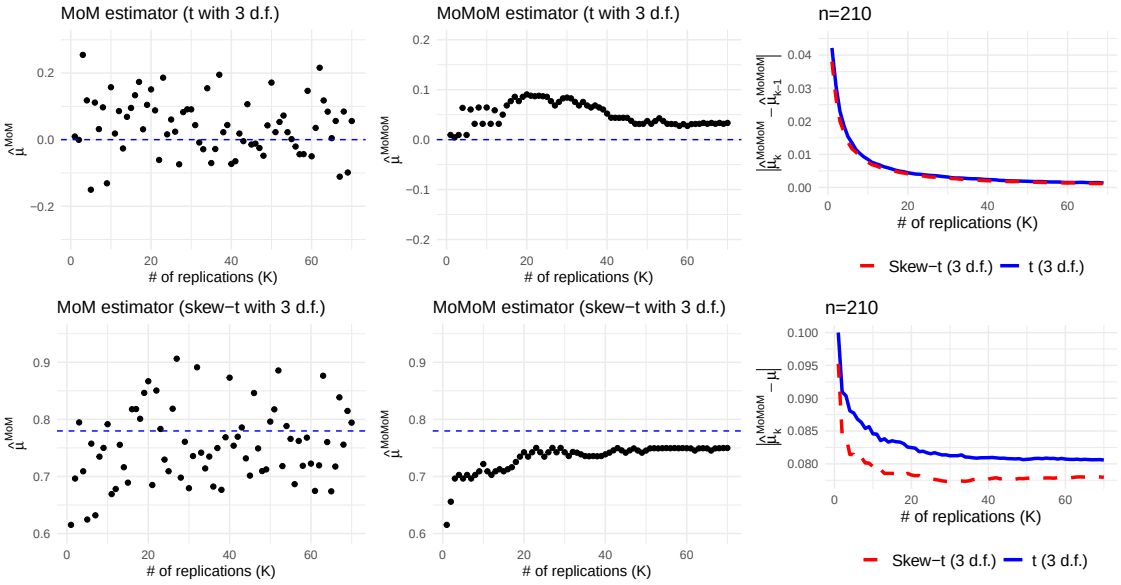

FIGURE 3.2: First column: MoM estimator obtained during various replications in a *a single run of the procedure* using data generated from the t-distribution (above) and the skew-t distribution (below). The dashed line is the true mean μ. Second column: MoMoM estimator obtained during the replications; both series stabilize for $k > 40$. Third column: average over 1000 runs of the absolute difference $|\hat{\mu}_k^{\text{MoMoM}} - \hat{\mu}_{k-1}^{\text{MoMoM}}|$ against k (above) and average over 1000 runs of the absolute difference $|\hat{\mu}_k^{\text{MoMoM}} - \mu|$ against K. Note that the third column is only available to us in the simulation, but plots like the second column can be constructed on the fly as K increases, and it can be adaptively tracked and stopped, while retaining the same statistical guarantee at the stopped K.

We simulated observations from a standard Student's t-distribution with 3 degrees of freedom and from a skew-t distribution with 3 degrees of freedom and 1 as a skewness parameter (Azzalini and Capitanio, 2003). The number of batches is equal to $B = 21$ and the number of data points generated are $n = 210$. For each iteration, we computed $\hat{\mu}_1^{\text{MoMoM}}, \ldots, \hat{\mu}_K^{\text{MoMoM}}$ with $K = 70$ and considered the absolute value of the difference $\hat{\mu}_k^{\text{MoMoM}} - \hat{\mu}_{k-1}^{\text{MoMoM}}$ as a measure of stability. In the first two columns of Figure 3.2, we present two examples of the procedure, where it can be observed that in the first case, $\hat{\mu}^{\text{MoMoM}}$ tends to stabilize after 50 iterations, while in the second case, where the data are generated from a skew-t distribution, 25 iterations are sufficient. In the third column of Figure 3.2, we report the empirical average over 1000 replications of $|\hat{\mu}_k^{\text{MoMoM}} - \hat{\mu}_{k-1}^{\text{MoMoM}}|$. We observe an inflection point around 25 for both distributions, and it seems that once this threshold is reached, the gain becomes negligible.

3.3.2 Derandomizing HulC

Recently, Kuchibhotla *et al.* (2024) proposed a rather general inference procedure for deriving confidence intervals for a target functional θ, using any point estimator $\hat{\theta}$. Their HulC procedure starts off in a similar fashion to the median-of-means: it divides the n iid data points into B buckets of n/B, and computes the point estimator $\hat{\theta}_b$ in each bucket (this could be the sample mean, as in the previous subsection). It then reports a confidence interval by using certain quantiles of $\{\hat{\theta}_b\}_{b=1}^B$ (as opposed to a point estimator given by their median). The miscoverage rate is not exactly α, but is determined by the "median bias" of the underlying estimator, a quantity that must vanish asymptotically for any nonasymptotic confidence interval to exist. It is important to remark that the final output of the HulC is a confidence set for the functional of interest. Clearly, given a level α, one could test the hypothesis that θ equals a certain value simply by checking if that value falls within the level $1-\alpha$ interval. But, calculating the p-value for a given hypothesis is not straightforward, as there is no guarantee that the sets produced with different coverage levels are nested as α decreases due to the randomness of the procedure (this is due to the fact that the number of buckets depends on α). This makes our method particularly useful in this setting, since it simply requires the sets and not the underlying p-values.

As with MoM, the HulC confidence interval depends on the random split of the data. Our majority vote procedure effectively derandomizes the procedure while retaining the optimal rate of shrinkage (in settings where HulC possesses this property). We remark that the similarity between the first steps of HulC and MoM has not been previously explicitly noted, but it is interesting because HulC effectively provides a confidence interval for the MoM procedure. The coverage is not exactly $1-\alpha$ due to the unknown median bias, but whatever the resulting coverage is, a similar coverage is retained by our various majority vote sets. (Indeed a nonasymptotically valid confidence interval is impossible without any apriori bound on the variance.)

Simulation study. We conducted a simulation study wherein we generated data from a Student's t-distribution with 3 dof. The HulC method was used to obtain a confidence interval at a level $1-\alpha$, and the Median of Means (MoM) was used as the estimator in this context. First, we define the median bias of the estimator $\hat{\mu}^{\mathrm{MoM}}$ as

$$\text{Med-Bias}_{\mu^*}(\hat{\mu}^{\mathrm{MoM}}) = \left(\frac{1}{2} - \min\left\{\mathbb{P}(\hat{\mu}^{\mathrm{MoM}} \geq \mu^*), \mathbb{P}(\hat{\mu}^{\mathrm{MoM}} \leq \mu^*)\right\}\right)_+,$$

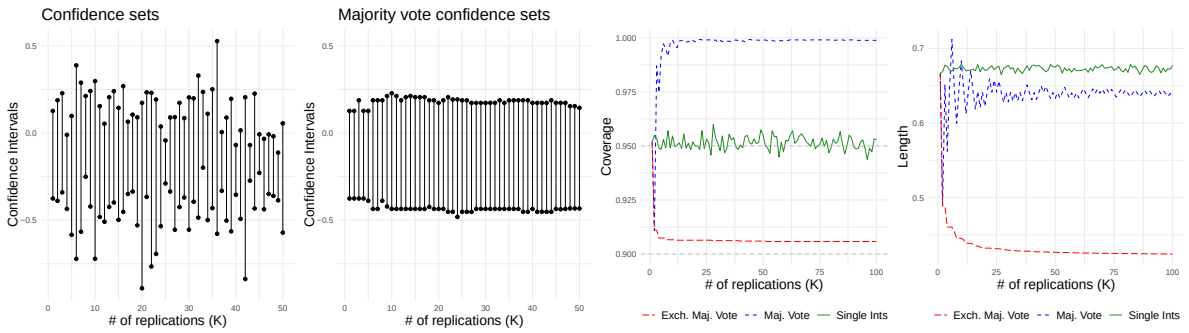

FIGURE 3.3: First two columns: example of *a single run of the procedure* using 210 observations generated from the t-distribution. Left: confidence intervals obtained for different random splits, Right: the merged sets. Last two columns: average over 5000 runs of the empirical coverage (left) and length (right) of $\mathcal{C}^M$ and $\mathcal{C}^E$ against K.

where μ^* denotes the true mean of the distribution. In this case, the empirical means used as estimators in each batch are median-unbiased, i.e. $\mathbb{P}(\hat{\mu}_j \geq \mu^*) = \mathbb{P}(\hat{\mu}_j \leq \mu^*) = \frac{1}{2}$, since the data are independent and identically distributed from a symmetric location family (Kuchibhotla *et al.*, 2023). We define B_1 and B_2 as the number of splits used for the MoM and HulC procedure, respectively. If B_1 is odd, then $\mathbb{P}(\hat{\mu}^{\text{MoM}} \geq \mu^*)$ implies that at least $\lceil B_1/2 \rceil$ of the estimators $\hat{\mu}_j$ are greater than μ^*. Since the $\hat{\mu}_j$ are independent, this corresponds to the probability that a $\text{Binom}(B_1, 1/2)$ is greater than $\lceil B_1/2 \rceil - 1$ which is equal to half. The same result is obtained with $\mathbb{P}(\hat{\mu}^{\text{MoM}} \leq \mu^*)$. This implies that if B_1 is odd, then $\hat{\mu}^{\text{MoMoM}}$ is a median unbiased estimator for the mean. The number of data splits B_2 to reach an interval with miscoverage equal to $\alpha = 0.05$ is 6 (see eq. 4 in Kuchibhotla *et al.* (2024)) while we fix $n = 210$ and $B_1 = 7$. Since intervals are constructed using different data splits one can use $\mathcal{C}^E$ to build a valid $1-2\alpha$. From Figure 3.3, we see that the intervals $\mathcal{C}^M$ are stable if the number of replications is greater than 5, and $\mathcal{C}^E$ is simply given by the running intersection of these intervals. In the third column of Figure 3.3, we see that $\mathcal{C}^M$ tends to produce intervals with an higher coverage, while $\mathcal{C}^E$ stabilizes around $1 - 2\alpha$.

3.4 Discussion

The chapter presents a method to address the question of merging dependent sets in an efficient manner, where efficiency is measured in both coverage and size. The approach can be seen as the confidence set analog of Rüger (1978) which combines p-values through quantiles. The randomized version yields better results in terms of both coverage and width, without altering the theoretical properties of the method. The

improvements for exchangeable sets are particularly striking and are achieved using recently proposed extensions of Markov's inequality. The proposed method can be used to derandomize statistical procedures that are based on data-splitting. In both real and simulated examples, the method achieves good results in terms of coverage and size. The method is versatile and is clearly applicable in more scenarios than we have explored here.

Chapter 4

Conformal prediction

In previous chapters, we studied methods for combining p-values and uncertainty sets. These approaches are primarily focused on ensuring valid probabilistic statements in the presence of dependence. We now apply these ideas to the conformal prediction framework, a relatively recent approach for constructing predictive sets for machine learning models. This setting is particularly relevant, as black-box algorithms are increasingly used to generate predictions across a wide range of real-world applications, including medical studies, economics, and social sciences. Although such algorithms often achieve strong predictive performance, a major challenge lies in accurately quantifying the uncertainty associated with their predictions. This is especially critical in high-risk contexts, such as medical diagnosis or autonomous driving, where incorrect predictions can have serious consequences.

In this chapter, we briefly outline the general ideas of conformal prediction and describe some of its main approaches, such as split conformal prediction and full conformal prediction. While at first glance this topic may seem disconnected from the previous chapters, we will see that, in a broad sense, the results can be recast as a hypothesis testing problem. We refer the reader to Fontana *et al.* (2023), Angelopoulos and Bates (2023), and Angelopoulos *et al.* (2025) for deeper and comprehensive treatments of conformal prediction and its extensions.

4.1 Problem setup and review of the literature

Assume we have independent and identically distributed (iid) training samples $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, drawn from a probability distribution Q, where $\mathcal{X}$ represents the feature space and $\mathcal{Y}$ the response space. Using these training data, our goal is to obtain a prediction set for the response variable Y_{n+1} based on the covariates

X_{n+1}, under the assumption that the test pair (X_{n+1}, Y_{n+1}) is independently sampled from the same distribution Q. Actually, the results will be shown to hold more generally under the assumptions of exchangeability of the $n+1$ data points with the iid assumption as a special case. A typical scenario involves applying a prediction algorithm to the training data in order to find a prediction for the response value. In particular, let $\hat{\mu} : \mathcal{X} \to \mathcal{Y}'$ be a regression function obtained by applying an algorithm $\mathcal{A}$ to the training points, where $\mathcal{Y}'$ is the prediction space (in regression problems we usually have $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathcal{Y}' = \mathbb{R}$). Formally, $\mathcal{A}$ is a mapping from $\cup_{d\geq 1}(\mathcal{X} \times \mathcal{Y})^d$ (the set of all possible training datasets of any size $d \geq 1$), to the space of functions $\mathcal{X} \to \mathcal{Y}'$. Starting from the regression function $\hat{\mu}$, we aim to construct a prediction set $\mathcal{C}(X_{n+1})$ that contains the point Y_{n+1} with high probability. Since no assumptions are made about the distribution Q, the method is said to be *distribution-free*.

Before proceeding with the remainder of the chapter, we define the score function $s = s((x, y); \mathcal{D})$, which quantifies the non-conformity of a point in the sample space with respect to the dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^d$ used to train the prediction model. In particular, we assume that the score function s adheres to a symmetry property:

$$s\big((x, y); \mathcal{D}\big) = s\big((x, y); \mathcal{D}^\pi\big), \tag{4.1}$$

where π is any permutation of the indices $[d] := \{1, \ldots, d\}$ and $\mathcal{D}^\pi$ refer to the dataset whose elements are permuted by π. For example, when considering residual scores $|y - \hat{\mu}(x)|$ in a regression problem, the symmetry of the score function is satisfied if the prediction algorithm is symmetric, which means that $\mathcal{A}(\mathcal{D}) = \mathcal{A}(\mathcal{D}^\pi)$. In addition, we denote the dataset containing the observations in the set I as $\mathcal{D}_I = (Z_i : i \in I)$.

4.1.1 Review of the literature

Conformal prediction was first introduced and formalized by Saunders *et al.* (1999) and Vovk *et al.* (2005). From a methodological point of view, influential contributions include the works of Lei *et al.* (2018), Romano *et al.* (2019), and Barber *et al.* (2021). In particular, full conformal prediction by Vovk *et al.* (2005), demonstrates favorable properties regarding the coverage and the size of the prediction set. However, these advantages are counterbalanced by a substantial computational cost, which limits its practical application. In fact, we will see that the method requires one to train the model for every possible value of the response, and this procedure is usually computationally burdensome. To alleviate this problem, split conformal prediction (Papadopoulos *et al.*, 2002; Lei *et al.*, 2018) has been proposed as a solution. The procedure involves a random

partition of the data into two subsets: the first subset is used to train the prediction algorithm, while the remaining part is used to calibrate the predictions and to obtain the prediction interval. Although this variant proves to be computationally efficient, it suffers from reduced efficiency in terms of the width of the resulting prediction set; this is due to the fact that only a fraction of the data is used to train the model.

Several "hybrid" solutions have been proposed in the literature, which can be considered between split conformal prediction and full conformal prediction. Examples include cross-conformal prediction (Vovk, 2015; Vovk *et al.*, 2018), multi-split conformal prediction (Solari and Djordjilović, 2022), the jackknife+ (Barber *et al.*, 2021) and out-of-bag conformal prediction (Linusson *et al.*, 2020; Gupta *et al.*, 2022). These techniques generally result in smaller prediction intervals compared to split conformal prediction and involve less computational effort than full conformal prediction. However, one of the main drawbacks of these methods is the reduced marginal coverage guarantee, which is less than the usual $1-\alpha$ level.

Other studies extend conformal prediction to scenarios where the standard assumptions might not hold, examples include Tibshirani *et al.* (2019), Barber *et al.* (2023), Gibbs and Candès (2021), and Gibbs and Candès (2024).

In the following, we present the main methods in conformal prediction and extend the application of the majority vote approach within this framework.

4.2 Full conformal prediction

We start by describing the full (or transductive) conformal prediction introduced by Vovk *et al.* (2005). In particular, let $\mathcal{D}_{[n]} = (Z_1, \ldots, Z_n)$ denote the training set, while $\mathcal{D}_{[n+1]} = \mathcal{D}_{[n]} \cup (X_{n+1}, Y_{n+1})$ denotes training set with the true test point. In practice, we never have access to the true test point Y_{n+1}; however, we can hypothesize that the response takes the value $y \in \mathcal{Y}$ (i.e., assume $Y_{n+1} = y$). We therefore define $\mathcal{D}^{y}_{[n+1]} = \mathcal{D} \cup (X_{n+1}, y)$ as the augmented set with the hypothesized test point.

The full conformal prediction method is outlined in Algorithm 1. Roughly speaking, the procedure assesses whether a hypothesized point $y \in \mathcal{Y}$ conforms well with the observed data. As can be seen, the method requires retraining the model for every candidate $y \in \mathcal{Y}$, which makes it computationally expensive. The full conformal prediction set possesses the following guarantee:

Theorem 4.1 (Vovk *et al.* (2005))**.** *If data are exchangeable then*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{full}}(X_{n+1})\right) \geq 1-\alpha. \tag{4.2}$$

Data: Dataset $\mathcal{D}_n$, test point X_{n+1}, α-level, algorithm $\mathcal{A}$, score function s, $\mathcal{Y} = \{y', y'', \ldots, y^{n_{\text{grid}}}\}$.
for $y \in \mathcal{Y}$ **do**
  Train the algorithm $\mathcal{A}$ to $\mathcal{D}^y_{[n+1]}$;
  Compute the conformal score $S^y_i := s((X_i, Y_i); \mathcal{D}^y_{[n+1]})$, $i \in 1, \ldots, n$;
  Compute the conformal score for the hypothesized test point $S^y_{n+1} := s((X_{n+1}, y); \mathcal{D}^y_{[n+1]})$;
  Compute the quantile $\hat{q}^y := \text{quantile}\,(S^y_1, \ldots, S^y_n; (1-\alpha)(1+1/n))$
end
Result: $\mathcal{C}^{\text{full}}(X_{n+1}) = \{y \in \mathcal{Y} : S^y_{n+1} \leq \hat{q}^y\}$
Algorithm 1: Full conformal prediction algorithm. The parameter n_{grid} represents the number of different possible y values.

It is important to note, that the probability is marginal and it is computed with respect to the training data and the test point (X_{n+1}, Y_{n+1}). A possible way to prove the theorem is to reformulate the procedure in terms of conformal p-values. In fact, the set $\mathcal{C}^{\text{full}}(X_{n+1})$ can be expressed as

$$\mathcal{C}^{\text{full}}(X_{n+1}) = \{y \in \mathcal{Y} : P(y) > \alpha\},$$

where

$$P(y) = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\{s((X_{n+1}, y); \mathcal{D}^y_{[n+1]}) \leq s((X_i, Y_i); \mathcal{D}^y_{[n+1]})\}}{n+1}. \tag{4.3}$$

Since the data points $Z_i, i \in [n+1]$, are exchangeable, the corresponding scores are also exchangeable. This follows because the score function is permutation invariant with respect to the dataset, as stated in (4.1). In particular, fixing $y = Y_{n+1}$, we obtain

$$\mathbb{P}\big(P(Y_{n+1}) \leq \alpha\big) = \mathbb{P}\left(\sum_{i=1}^{n+1} \mathbb{1}\{s((X_{n+1}, Y_{n+1}); \mathcal{D}_{[n+1]}) \leq s((X_i, Y_i); \mathcal{D}_{[n+1]})\} \leq \alpha(n+1)\right) \leq \alpha,$$

which holds because, under exchangeability, the ordering of the scores is not relevant (see Kuchibhotla (2020)). This argument is sufficient to establish the coverage property in (4.2).

In addition, if the scores are almost surely distinct, it is possible to derive an upper bound on the inclusion probability of the form

$$\mathbb{P}\big(Y_{n+1} \in \mathcal{C}^{\text{full}}(X_{n+1})\big) \leq 1 - \alpha + \frac{1}{n+1}.$$

A proof of this result can be found in Kuchibhotla (2020) or Angelopoulos *et al.* (2025).

4.3 Split conformal prediction

The appealing theoretical properties of full conformal prediction are contrasted by very high computational cost, which hinders its practical application. A solution is offered in the form by split conformal prediction (Papadopoulos *et al.*, 2002; Lei *et al.*, 2018). We assume that we are in the setup described in Section 4.1 and the goal is to obtain a prediction set for the response value Y_{n+1} given the training data and covariates in X_{n+1}. In this case, data are divided into two disjoint subsets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$. The algorithm is trained using the data points in $\mathcal{D}_{\text{train}}$, while the scores $S_i := s\left((X_i, Y_i); \mathcal{D}_{\text{train}}\right), i \in \mathcal{D}_{\text{cal}}$, are obtained from the observations in the calibration set. The split conformal prediction set is simply defined as

$$\mathcal{C}^{\text{split}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s\left((X_{n+1}, y); \mathcal{D}_{\text{train}}\right) \leq \hat{q} \right\}, \tag{4.4}$$

where $\hat{q} := \text{quantile}\left(S_1, \ldots, S_{|\mathcal{D}_{\text{cal}}|}; (1-\alpha)(1 + 1/|\mathcal{D}_{\text{cal}}|)\right)$[1]. The set in (4.4) can be re-written as

$$\mathcal{C}^{\text{split}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : P(y) > \alpha \right\},$$

where

$$P(y) = \frac{1 + \sum_{i \in \mathcal{D}_{\text{cal}}} \mathbb{1}\left\{ s\left((X_{n+1}, y); \mathcal{D}_{\text{train}}\right) \leq s\left((X_i, Y_i); \mathcal{D}_{\text{train}}\right) \right\}}{|\mathcal{D}_{\text{cal}}| + 1},$$

that is a p-value when calculated in Y_{n+1} similar to the one defined in (4.3). This implies that if the data are iid or exchangeable, the marginal coverage of the set $\mathcal{C}^{\text{split}}_{n,\alpha}(X_{n+1})$ is at least $1 - \alpha$; in symbols:

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{split}}_{n,\alpha}(X_{n+1})\right) \geq 1 - \alpha.$$

In addition, if the residuals have no ties (they have a continuous joint distribution) then

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{split}}_{n,\alpha}(X_{n+1})\right) \leq 1 - \alpha + \frac{1}{|\mathcal{D}_{\text{cal}}| + 1}. \tag{4.5}$$

The proof of the results can be found in Lei *et al.* (2018), and the result in (4.5) states that the marginal coverage is essentially $1 - \alpha$ when the number of observations in the calibration set is sufficiently large.

One of the attractive properties of split conformal prediction is that the computational cost of the procedure is low compared to that of full (or transductive) conformal prediction. In fact, the model only needs to be trained once, and the predictions are

[1] We define $\text{quantile}(z; \gamma) = \inf\{a : n^{-1} \sum_{i=1}^{n} \mathbb{1}\{z_i \leq a\} \geq \gamma\}$, for any $z \in \mathbb{R}^n$.

then calibrated using the data points in $\mathcal{D}_{\text{cal}}$.

4.4 Majority vote and conformal prediction

We now extend the results of Chapter 3 to the framework of conformal prediction. Suppose we have access to K conformal prediction sets with level $1-\alpha$, i.e.:

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_k(X_{n+1})\right) \geq 1-\alpha, \quad k = 1, \ldots, K, \tag{4.6}$$

where $\alpha \in (0,1)$. The K different intervals can differ due to the algorithm used to obtain the predictions or the variant of conformal prediction employed (such as split or full conformal prediction). Theorems 3.1 and 3.14 are specialized (in the conformal case) to obtain the following result:

Corollary 4.2. *Let $\mathcal{C}_1(X_{n+1}), \ldots, \mathcal{C}_K(X_{n+1})$ be $K \geq 2$ different conformal prediction sets, $X_{n+1} \in \mathcal{X}$ and $w = (w_1, \ldots, w_k)$ defined as in (3.13). Then,*

$$\mathcal{C}^M(X_{n+1}) = \left\{ y \in \mathcal{Y} : \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{y \in \mathcal{C}_k(X_{n+1})\} > \frac{1}{2} \right\},$$
$$\mathcal{C}^W(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} w_k \mathbb{1}\{y \in \mathcal{C}_k(X_{n+1})\} > 1/2 + U/2 \right\},$$

where $U \in \mathcal{U}$, are valid conformal prediction sets with level $1-2\alpha$.

In other words, Corollary 4.2 says that the majority vote approach still holds when we are dealing with K different conformal prediction sets. As explained in Section 3.1.7, if the sets are exchangeable then it is possible to obtain better results than using a simple majority vote procedure. Specifically, Theorem 3.11 can be specialized in the context of conformal prediction.

Corollary 4.3. *Let $\mathcal{C}_1(X_{n+1}), \ldots, \mathcal{C}_K(X_{n+1})$ be $K \geq 2$ exchangeable conformal prediction intervals having coverage $1-\alpha$, then*

$$\mathcal{C}^E(X_{n+1}) = \left\{ y \in \mathcal{Y} : \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}\{y \in \mathcal{C}_j(X_{n+1})\} > \frac{1}{2} \ \forall k \leq K \right\}, \tag{4.7}$$

is a valid $1-2\alpha$ conformal prediction set.

If the sets are non-exchangeable then exchangeability can be achieved through a random permutation π of the indexes $\{1, \ldots, K\}$, in order to obtain the set $\mathcal{C}^{\pi}(X_{n+1})$ described in (3.10).

In the following, we will study the properties of the method through a simulation study and an application to real data. One consistent phenomenon that we seem to observe empirically is that $\mathcal{C}^M$ actually has coverage $1-\alpha$ (better than $1-2\alpha$ as promised by the theorem), and the smaller sets $\mathcal{C}^R$ and $\mathcal{C}^\pi$ have coverage between $1-\alpha$ and $1-2\alpha$.

4.4.1 Simulation study

We carried out a simulation study in order to investigate the performance of the proposed majority vote approach. We apply the majority vote procedure on simulated high-dimensional data with $n = 100$ observations and $p = 120$ regressors. Specifically, we simulate the design matrix $\mathbf{X}_{n\times p}$, where each column is independent of the others and contains standard normal entries. The outcome vector is equal to $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where β is a sparse vector with only the first $m = 10$ elements different from 0 (generated independently from a $\mathcal{N}(0, 4)$) while $\varepsilon \sim \mathcal{N}_n(0, I_n)$. A test point (X_{n+1}, Y_{n+1}) is generated with the same data-generating mechanism. At each iteration we estimate the regression function using the lasso algorithm (Tibshirani, 1996) with penalty parameter λ varying over a fixed sequence of values $K = 20$ and then construct a conformal prediction interval for each λ in X_{n+1} using the split conformal method presented in package R `ConformalInference`. We run 10 000 iterations at $\alpha = 0.05$.

We then merge the K different sets using the method described in Corollary 4.2 with $w_k = \frac{1}{K}, k = [K]$ (an additional simulation to study the impact of w is reported in Appendix C.2). These weights can be interpreted, from a Bayesian perspective, as a discrete uniform prior on λ. From an alternative perspective, each agent represents a value of the penalty parameter, and the *aggregator* equally weighs the various intervals constructed by the various agents. An example of the result is shown in Figure 4.1, where the value of U is fixed to $1/2$ for standardization. Clearly, this is just an instance of the possible sets that can be obtained; however we know from the theory that the set is always smaller than $\mathcal{C}^M$ and larger than the intersection of the set. The empirical coverages of the intervals $\mathcal{C}^M(X_{n+1})$ and $\mathcal{C}^R(X_{n+1})$ are $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{Y_{n+1}^b \in \mathcal{C}_b^M(X_{n+1})\} = 0.97$ and $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{Y_{n+1}^b \in \mathcal{C}_b^R(X_{n+1})\} = 0.92$. By definition, the second method produces narrower intervals while maintaining the coverage level $1-2\alpha$. As explained in the previous sections, by inducing exchangeability through permutation, it may be possible to enhance the majority vote results. In fact, the empirical coverage of the sets $\mathcal{C}^\pi$ is 0.93 while the sets are smaller than the ones produced by the simple majority vote. Furthermore, we tested the sets $\mathcal{C}^U(X_{n+1})$ defined in (3.12) and obtained an empirical coverage equal to 0.96, which is very close to the nominal level $1-\alpha$. In all five cases,

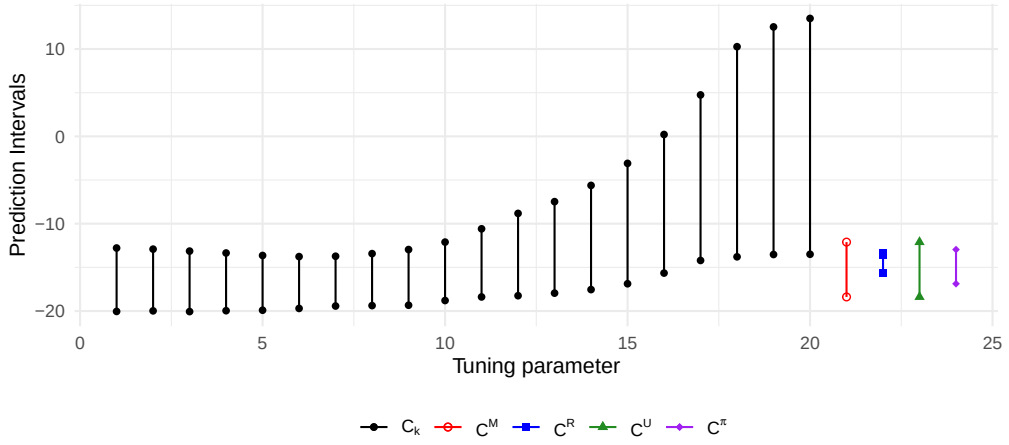

FIGURE 4.1: Intervals obtained using different values of λ, $\mathcal{C}^M(X_{n+1})$, $\mathcal{C}^R(X_{n+1}), \mathcal{C}^U(X_{n+1})$ and $\mathcal{C}^\pi(X_{n+1})$. For standardization, the value of U in the randomized thresholds is set to $1/2$. The smallest set $\mathcal{C}^R$. Since $U = 1/2$, the sets $\mathcal{C}^M$ and $\mathcal{C}^U$ coincides.

the occurrence of obtaining an union of intervals or the empty set as output is very low, specifically less than 1% of the iterations.

4.4.2 Real data example

We used the proposed methods in a real dataset regarding Parkinson's disease (Tsanas and Little, 2009). The goal is to predict the total UPDRS (Unified Parkinson's Disease Rating Scale) score using a range of biomedical voice measurements from people suffering early-stage Parkinson's disease. We used split conformal prediction and $K = 4$ different algorithms (linear model, lasso, random forest, neural net) to obtain the conformal prediction sets. In particular, we choose $n = 5000$ observations to construct our intervals and the others $n_0 = 875$ observations as test points. Prior weights also in this case are uniform over the K models. If previous studies had been carried out, one could, for example, put more weight on methods with better performance. Otherwise, one can assign a higher weight to more flexible algorithms such as random forest or neural net.

The results are reported in Table 4.1 where it is possible to note that all merging procedures obtain good results in terms of length and coverage. In addition, also the randomized vote ($\mathcal{C}^U$) obtains good results in terms of coverage, with an empirical length that is slightly larger than the one obtained by the neural net. The percentage of times that a union of intervals or an empty set is outputted is nearly zero for all methods. In this situation, the intervals produced by the random forest outperform the others in terms of size of the sets; as a consequence, one may wish to put more weight into the

Methods	LM	Lasso	RF	NN	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	$\mathcal{C}^\pi$
Coverage	0.958	0.960	0.949	0.961	0.951	0.923	0.961	0.918
Size	40.143	40.150	13.286	32.533	29.508	20.620	32.544	20.710

TABLE 4.1: Empirical coverage and length of the methods for the Parkinson's dataset.

method, which results in smaller intervals on average.

4.5 Multi-split conformal prediction

As discussed earlier, full conformal prediction method originally introduced by Vovk *et al.* (2005), exhibits good properties in terms of coverage and size of the prediction set; however, these are counterbalanced by a notable computational cost, which makes its practical application challenging. To address this issue, a potential solution is the adoption of *split* conformal prediction (Papadopoulos *et al.*, 2002; Lei *et al.*, 2018), which involves the use of a random split of the data into two parts. Although this variant proves to be highly computational efficient, it introduces an additional layer of randomness that stems from the randomness of the data split. In addition, only a fraction of the data is used to train the model, leading to a reduced efficiency of the prediction set in terms of width. Several works aim to mitigate this problem by proposing various ways to combine the intervals obtained from different splits (Solari and Djordjilović, 2022; Barber *et al.*, 2021; Vovk, 2015).

In particular, the multi-split conformal prediction method introduced in Solari and Djordjilović (2022) involves the construction of K distinct intervals of level $1-\alpha(1-\tau)$, each originating from a different random split. Subsequently, these intervals are merged using the mechanism described in (3.4). Although the final interval achieves a coverage of $1-\alpha$, it is possible to enhance the procedure by exploiting the exchangeability of sets and using the results introduced in Corollary 4.3 with the threshold set to τ. A simple possible solution is to fix K in advance, construct the K different sets in parallel, and subsequently merging them using the set $\mathcal{C}^E$ described in (4.7) (with a threshold different than $1/2$ and set to τ). Another method is to not fix K in advance and instead merge the sets online, through $\mathcal{C}^E(1:t)$ described in (3.17) and with a threshold equal to τ rather than half. In this case, the coverage $1-\alpha$ is guaranteed uniformly over the number of splits.

We conducted a simulation study in which the data were generated using the same generative mechanism described in Section 4.4.1. Also in this case, the algorithm employed was Lasso with $\lambda = 1$, while the intervals varied due to the data split on which they were constructed. The number of splits used was set at $\{1, 5, 10, 20, 30, 50\}$, while

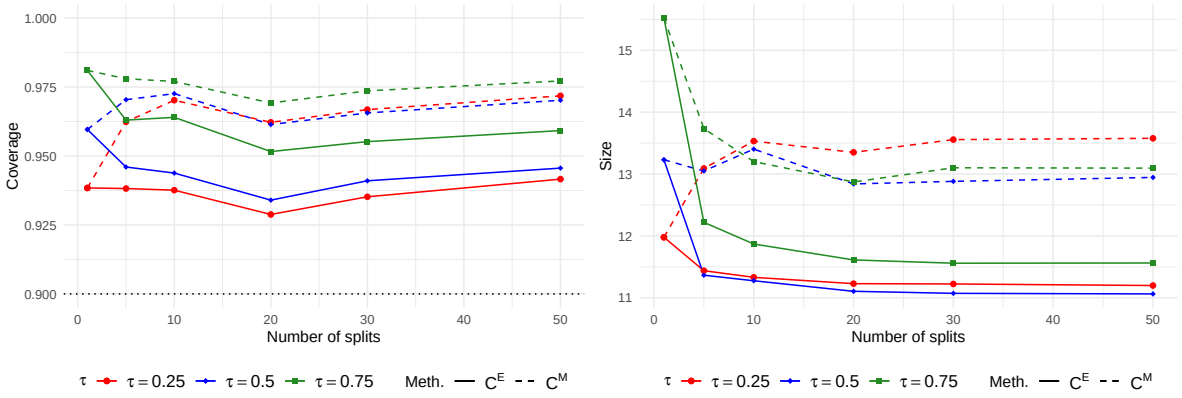

FIGURE 4.2: Comparison between the *simple* majority vote ($\mathcal{C}^M$) and exchangeable majority vote ($\mathcal{C}^E$) with different τ. The size of the $\mathcal{C}^E$ is smaller than $\mathcal{C}^M$.

for the parameter τ, three values were selected, namely $\{0.25, 0.50, 0.75\}$. The error rate α is set to 0.1 and the procedure was repeated 5000 times.

As evident from Figure 4.2, the methods lead to a greater coverage than the specified level $1-\alpha$. The coverage is higher for the *simple* majority vote compared to that based on the exchangeability of the sets; while, the size of the resulting sets is significantly smaller when using the method based on the exchangeability of sets. The number of splits does not appear to have a significant impact on the size of the sets. Once more, $\tau = 0.5$ seems to offer good performances in terms of width.

4.6 Discussion

This chapter provides an introduction to conformal prediction methods, presenting the key ideas, principles, and main approaches that underpin this framework. Commonly used variants, such as split conformal prediction and full conformal prediction, are introduced and described.

In the latter part of the chapter, the results presented in Chapter 3 are extended, showing how the techniques for combining uncertainty sets can be adapted and applied within the conformal prediction setting.

Chapter 5

Improving the statistical efficiency of cross-conformal prediction

One limitation of split conformal prediction, as discussed in Section 4.3, is that only a portion of the dataset is used to train the predictive model, which usually translates into reduced efficiency. To address this, Vovk (2015) proposed cross-conformal prediction, a modification designed to improve the size of prediction sets. However, when trained with a miscoverage rate of α, the method guarantees a marginal coverage of at least $1-2\alpha-2(1-\alpha)(K-1)/(n+K)$, where n is the number of observations and K denotes the number of folds. A simple modification of the method achieves coverage of at least $1-2\alpha$. In this chapter, we focus the attention on cross-conformal prediction and we show that the method can be improved without altering the coverage guarantee. In other words, it is possible to obtain smaller prediction sets while ensuring the same (worst-case) miscoverage rate. Starting from a modification of the method (Vovk *et al.*, 2018; Barber *et al.*, 2021), the new results are obtained using the combination of dependent p-values described in Chapter 2. Importantly, these results are obtained in a fully general manner, and do not need any specific prediction model or ensemble method to be used.

We consider the same setting introduced in Section 4.1, where the objective is to obtain a prediction set for Y_{n+1} given its covariate vector and data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.

5.1 (Modified) Cross-conformal prediction

This section will recap two methods: cross-conformal prediction, and modified cross-conformal prediction. We let K denote the number of folds, and we focus on the (practical) case when K is small like $K = 5$ or $K = 10$. We will always assume that

$m = n/K$ is an integer, which is achievable by only throwing away less than K points from the original dataset. However, we point out in Appendix C.4 that both methods have guarantees without this assumption.

5.1.1 Cross-conformal prediction

Cross-conformal prediction, introduced by Vovk (2015), is a method to obtain distribution-free prediction intervals. It can be considered as a combination of split conformal prediction and cross-validation. It works as follows: data are divided into K disjoint subsets (or folds) $I_1, \ldots, I_K$ of size $m = n/K$. The (cross-validation) scores are defined as:

$$S_i^{\mathrm{CV}} = s\left((X_i, Y_i); \mathcal{D}_{[n]\setminus I_{k(i)}}\right) \quad i = 1, \ldots, n, \tag{5.1}$$

where $I_{k(i)}$ is the subset containing the i-th data point. The cross-conformal prediction set is simply defined as

$$\mathcal{C}^{\mathrm{cross}}_{n,K,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{ s\left((X_{n+1}, y); \mathcal{D}_{[n]\setminus I_{k(i)}}\right) \leq S_i^{\mathrm{CV}} \right\}}{n+1} > \alpha \right\}. \tag{5.2}$$

Vovk *et al.* (2018) proves that the interval in (5.2) is such that

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\mathrm{cross}}_{n,K,\alpha}(X_{n+1}) \right) \geq 1 - 2\alpha - 2(1-\alpha)\frac{1 - 1/K}{n/K + 1}, \tag{5.3}$$

where the probability is marginal and is computed with respect to $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$. In particular, when K is small compared to n (that is, $n \gg K$), the additional term is negligible and the coverage is essentially at least $1 - 2\alpha$. To prove the result in (5.3), it is useful to define for each subset, $k \in [K]$, the quantity

$$P_k(y) = \frac{1 + \sum_{i \in I_k} \mathbb{1}\left\{ s\left((X_{n+1}, y); \mathcal{D}_{[n]\setminus I_k}\right) \leq S_i^{\mathrm{CV}} \right\}}{m+1}, \tag{5.4}$$

that is a p-value if computed using the response test value Y_{n+1} and if data $(X_i, Y_i), i \in [n+1]$, are iid or at least exchangeable (i.e., $\mathbb{P}(P_k(Y_{n+1}) \leq \alpha) \leq \alpha$). This is due to the fact that the scores in $I_k \cup \{n+1\}$ are exchangeable, since the prediction algorithm is trained only on the training points in $[n] \setminus I_k$, so (5.4) can be seen as a rank-based p-value. It is possible to relate the set defined in (5.2) with the cross-conformal p-values

in (5.4). In particular, a point y is included in $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ if and only if

$$\frac{1}{K}\sum_{k=1}^{K} P_k(y) > \alpha + (1-\alpha)\frac{K-1}{K+n}. \tag{5.5}$$

The multiplicative factor of two in the coverage statement in (5.3) arises from the fact that the average of arbitrarily dependent p-values remains a p-value up to a factor of 2:

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} P_k(Y_{n+1}) \leq \alpha\right) \leq 2\alpha. \tag{5.6}$$

This implies that the statement in (5.3) can be proved by combining the results in (5.5) and (5.6). For a detailed discussion on cross-conformal prediction see, for example, Chapter 4.4 of Vovk *et al.* (2022a).

Remark 5.1. The coverage statement in (5.3) is meaningless when K is large. In fact, Barber *et al.* (2021) proves a different bound for the miscoverage rate valid for large K. However, in practical applications, the number of splits is usually small if compared with the number of observations (e.g., $K = 5$ or $K = 10$) and the bound in (5.3) is the one that applies. We discuss the two different bounds and the connection with the CV+ method by Barber *et al.* (2021) in Appendix C.3.

Remark 5.2. In a regression setting, there are no guarantees that $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ will be an interval; in fact, there are particular cases where it can be a union of intervals. This property is shared by other "hybrid" methods mentioned in Section 4.1.1. One can avoid having a union of intervals by taking the convex hull of the set (the interval formed by the furthest endpoints) as explained in Gupta *et al.* (2022). In addition, when the residual score is chosen as score function, the prediction set $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ is a subset of the CV+ set that is guaranteed to be an interval (see Appendix C.3).

5.1.2 Modified cross-conformal prediction

It is clear from the previous section that we can obtain a set with coverage at least equal to $1-2\alpha$ using a modification of the cross-conformal prediction set defined in (5.2). We define the modified cross-conformal prediction interval (the same name is used in Barber *et al.* (2021)) as

$$\mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \frac{1}{K}\sum_{k=1}^{K} P_k(y) > \alpha \right\}. \tag{5.7}$$

Using the result stated in (5.6), we have

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})\right) \geq 1 - 2\alpha.$$

The intervals defined in (5.2) and (5.7) usually have inflated coverage. In other words, with typically employed levels of α, the coverage obtained using these methods often fluctuates between the levels $1-\alpha$ and 1. This is due to the fact that the rule in (5.6) is valid under arbitrary dependence and it has to take into account the "worst-case" scenario of dependence, which typically differs from the scenario observed in the data. However, in some situations where the regression algorithm is unstable or with some particular distribution Q, the coverage can oscillate between the guaranteed level $1-2\alpha$ and $1-\alpha$. Linusson *et al.* (2017) offers some empirical observations regarding the miscalibration of the average of p-values obtained from different folds. In particular, since the p-values are dependent, the distribution of the averaged p-values is in between the Bates distribution and the uniform distribution, and this strictly depends on the stability of the underlying algorithm.

Since p-values take discrete values, in order to avoid having noninformative sets identical to $\mathcal{Y}$, the inequality $1 \leq \alpha(m+1)$ must hold. A slightly improvement can be obtained using randomized p-values $P_1(Y_{n+1};\tau), \ldots, P_K(Y_{n+1};\tau)$ defined by

$$\begin{aligned} &P_k(y;\tau) = \\ &= \frac{\tau + \sum_{i \in I_k} \tau \mathbb{1}\left\{s\left((X_{n+1},y);\mathcal{D}_{[n]\setminus I_k}\right) = S_i^{\text{CV}}\right\} + \sum_{i \in I_k} \mathbb{1}\left\{s\left((X_{n+1},y);\mathcal{D}_{[n]\setminus I_k}\right) < S_i^{\text{CV}}\right\}}{m+1} \end{aligned} \tag{5.8}$$

where τ is a uniform random variable in the interval $(0,1)$ drawn independently from the data. In this case, the p-values (for $y = Y_{n+1}$) are uniformly distributed in the interval $(0,1)$, rather than taking discrete values. However, the dependence among the p-values obtained from different folds is not broken.

5.2 New variants of cross-conformal prediction

In this section, we improve the prediction sets in (5.7) using the results regarding the combination of p-values introduced in Chapter 2. The results are obtained in a completely general manner and do not require the use of expensive computational procedures (Carlsson *et al.*, 2014) or the use of specific models (Boström *et al.*, 2017).

5.2.1 Exchangeable modified cross-conformal prediction

The interval in (5.7) can be improved using the results on the combination of exchangeable p-values. Before proceeding, we state a useful result.

Proposition 5.3. *Let $P_1(Y_{n+1}), \ldots, P_K(Y_{n+1})$ be the (cross-conformal) p-values obtained using data $Z_i = (X_i, Y_i)$, $i = [n+1]$, then $P_1(Y_{n+1}), \ldots, P_K(Y_{n+1})$ are exchangeable, meaning that $\mathbf{P} \overset{d}{=} \mathbf{P}^\pi$, where $\overset{d}{=}$ represents equality in distribution, $\mathbf{P} = (P_1(Y_{n+1}), \ldots, P_K(Y_{n+1}))$, $\mathbf{P}^\pi = (P_{\pi(1)}(Y_{n+1}), \ldots, P_{\pi(K)}(Y_{n+1}))$ and $\pi : [K] \to [K]$ is any permutation of the indices.*

A proof of the result is based on the following lemma.

Lemma 5.4 (Dean and Verducci (1990); Kuchibhotla (2020))**.** *Suppose $W = (W_1, \ldots, W_n) \in \mathcal{W}^n$ is a vector of exchangeable random variables. Fix a transformation $G : \mathcal{W}^n \to (\mathcal{W}')^m$. If for each permutation $\pi_1 : [m] \to [m]$ there exists a permutation $\pi_2 : [n] \to [n]$ such that*

$$\pi_1 G(w) = G(\pi_2 w), \quad \text{for all } w \in \mathcal{W}^n,$$

then $G(\cdot)$ preserves exchangeability.

Proof of Proposition 5.3. Let $G : (\mathcal{X} \times \mathcal{Y})^{n+1} = \mathcal{Z}^{n+1} \to [0,1]^K$ be the transformation that takes as input the $n+1$ exchangeable data points $Z_1, \ldots, Z_{n+1}$ and returns as output the p-values $P_1(Y_{n+1}), \ldots, P_K(Y_{n+1})$. In other words, the i-th element of G is computed by training the algorithm $\mathcal{A}$ using the dataset $\mathcal{D}_{[n]\setminus I_i}$ and then computing the scores and the corresponding p-value defined in (5.4) using data points in $\mathcal{D}_{I_i \cup \{n+1\}}$. It is important to note that the score function s satisfies the condition in (4.1), and so the scores do not depend on the order of the data points in $\mathcal{D}_{[n]\setminus I_i}$. Let $\sigma_1 : [n] \to [K]$ be the function that assigns the training data points to the K different folds and $\sigma_2 : [n] \to [m]$ be the function that assigns the positions of the training data points within the assigned folds. In words, each point $i \in [n]$ is assigned a unique pair $\{\sigma_1(i), \sigma_2(i)\}$ that identifies its fold and its position inside the fold. For example, if $\sigma_1(1) = 2$ and $\sigma_2(1) = 3$ then the first data point in the original dataset is the third data point in the second fold. Let $\pi_1 : [K] \to [K]$ be a permutation of the indices, then for all $z \in \mathcal{Z}^{n+1}$,

$$\pi_1 G(z_1, \ldots, z_n, z_{n+1}) = G(\pi_2(z_1, \ldots, z_n, z_{n+1})),$$

where $\pi_2 : [n+1] \to [n+1]$ is such that

$$\pi_2(i) = \begin{cases} [\pi_1(\sigma_1(i)) - 1] \cdot m + \sigma_2(i), & i \neq n+1, \\ n+1, & i = n+1. \end{cases}$$

In words, π_2 permutes the training data points into their respective permuted folds (i.e., $i \in I_{\pi_1(\sigma_1(i))}$), while the test point remains in the $(n+1)$-th position. It holds that $G(\cdot)$ preserves exchangeability and this concludes the proof. □

Remark 5.5. The assumption that $n/K = m$ is crucial to prove the result in Proposition 5.3. In fact, if the subsets $I_1, \ldots, I_K$ have different sample sizes, then the result in Proposition 5.3 does not hold. Notice that the p-values in (5.4) take discrete values $\{1/m, 2/m, \ldots, 1\}$. If the sample sizes differ, then the p-values assume values in different grids of values, and therefore the marginal distributions of $P_1(Y_{n+1}), \ldots, P_K(Y_{n+1})$ are different. This implies that p-values cannot be exchangeable. In addition, with different sample sizes the proof of the result breaks down and a permutation π_2 that satisfies the condition in Lemma 5.4 does not exist. We will see how to extend the result to the case where the folds have different sizes using a simple trick.

An improved version of the set in (5.7) can be defined as:

$$\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \min_{\ell \in [K]} \frac{1}{\ell} \sum_{k=1}^{\ell} P_k(y) > \alpha \right\}, \tag{5.9}$$

where, for a given y, the combination of the different $P_k(y)$ is asymmetric and depends on the order of the p-values.

Theorem 5.6. *It holds that* $\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})$. *In addition, if data are exchangeable,*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1})\right) \geq 1 - 2\alpha. \tag{5.10}$$

Proof. By definition

$$\min_{\ell \in [K]} \frac{1}{\ell} \sum_{k=1}^{\ell} P_k(y) \leq \frac{1}{K} \sum_{k=1}^{K} P_k(y),$$

so less points will be included in the set. From Proposition 5.3 we have that the conformal p-values are exchangeable. The coverage property in (5.10) is a direct consequence of the result stated in 2.27, which states that $\min_{\ell \in [K]}(1/\ell) \sum_{k=1}^{\ell} P_k(Y_{n+1})$ is a valid p-value up to a factor of 2 if p-values $P_1(Y_{n+1}), \ldots, P_K(Y_{n+1})$ are exchangeable. □

The theorem indicates that one can derive a set smaller than the modified cross-conformal prediction set while maintaining the same coverage guarantee. The same results hold if the p-values in (5.8) are used. Specifically, the randomized p-values are still exchangeable if τ is common across the folds. Indeed, conditional on τ the p-values are exchangeable due to Proposition 5.3. In particular, using the p-values in (5.8) we obtain a smaller set since $P_k(y;\tau) \leq P_k(y)$ almost surely.

Remark 5.7. Once K exchangeable (or more generally dependent) p-values are obtained, there are several methods to combine them. The proposed solution is to use the minimum (over ℓ) of the mean obtained using the first ℓ p-values, which is related to the valid combination rule "twice the average" used by Vovk *et al.* (2018, 2022a) to prove the coverage guarantee of cross-conformal prediction. However, similar results apply to other merging functions like the ones described in Chapters 1 and 2. However, it is important to emphasize that, as explained in Section 6 of Vovk and Wang (2020), the mean is an effective method of combining strongly dependent p-values, which is often the case for rank-based p-values obtained by cross-conformal prediction. Other merging rules, such as the Bonferroni method, are more appropriate near independence. For instance, Lei *et al.* (2018, Section 2.3) shows, under some assumptions, that the Bonferroni rule is overly conservative in the context of multisplit conformal prediction.

5.2.2 Randomized modified cross-conformal prediction

In the previous paragraph, we leveraged the exchangeability of p-values to obtain a smaller set. In this section, we move in a different direction and improve the set (5.7) using a simple "randomization trick" (introducing a uniform random variable). Indeed, in this case, the exchangeability of the p-values is not necessary. As before, the improvement does not alter the marginal validity of the set, but the new result is obtained in a different way. Although randomization is avoided in some statistical applications due to the extra randomness it introduces, in this case, it does not pose a major issue. Indeed, cross-conformal prediction is, by definition, a randomized method. More precisely, data are randomly divided into K different subsets in the first step, which means that the procedure inherently includes randomness (see Remark 5.10 for further discussion).

We can define a "randomized" improvement of the interval in (5.7) as follows:

$$\mathcal{C}^{\text{u-mod-cross}}_{n,K,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \frac{1}{2-U}\frac{1}{K}\sum_{k=1}^{K} P_k(y) > \alpha \right\}, \tag{5.11}$$

where U is a uniform random variable in the interval $(0, 1)$ independent of all the data.

Theorem 5.8. *It holds that* $\mathcal{C}^{\text{u-mod-cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})$. *In addition, if data are exchangeable,*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{u-mod-cross}}_{n,K,\alpha}(X_{n+1})\right) \geq 1 - 2\alpha. \tag{5.12}$$

Proof. By definition

$$\frac{1}{2-U}\frac{1}{K}\sum_{k=1}^{K} P_k(y) \leq \frac{1}{K}\sum_{k=1}^{K} P_k(y),$$

since $1/(2-U) \leq 1$ almost surely. The result implies that less points will be included in the set. The coverage property in (5.12) is a consequence of Theorem 2.28, which states that

$$\frac{2}{2-U}\frac{1}{K}\sum_{k=1}^{K} P_k(Y_{n+1})$$

is a valid p-value. In particular, the result holds under arbitrary dependence of the starting p-values $P_1(Y_{n+1}), \ldots, P_K(Y_{n+1})$. □

Even in this case, the guaranteed marginal coverage remains at least $1-2\alpha$, but the set size is enhanced using a simple result based on randomization.

5.2.3 Exchangeable and randomized modified cross-conformal prediction

The results in Section 5.2.1 and in Section 5.2.2 can be "combined" in order to obtain a prediction set that improves the one defined in (5.9). In this case as well, the exchangeability property outlined in Proposition 5.3 is crucial.

We define a randomized improvement of the conformal prediction set defined in (5.9):

$$\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \min\left\{ \frac{1}{2-U}P_1(y), \min_{\ell \in [K]} \frac{1}{\ell}\sum_{k=1}^{\ell} P_k(y) \right\} > \alpha \right\}, \quad (5.13)$$

where U is a uniform random variable in the interval $(0,1)$ independent of all the data.

Theorem 5.9. *It holds that* $\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})$. *In addition, if data are exchangeable,*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha}(X_{n+1})\right) \geq 1-2\alpha. \quad (5.14)$$

Proof. By definition

$$\min\left\{ \frac{1}{2-U}P_1(y), \min_{\ell \in [K]} \frac{1}{\ell}\sum_{k=1}^{\ell} P_k(y) \right\} \leq \min_{\ell \in [K]} \frac{1}{\ell}\sum_{k=1}^{\ell} P_k(y).$$

The result implies that less points will be included in the set and so $\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1})$. The fact that $\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})$ is outlined in Theorem 5.6.

From Proposition 5.3 we have that the conformal p-values are exchangeable. The coverage property in (5.14) is a consequence of the fact that

$$\min\left\{\frac{1}{2-U}P_1(Y_{n+1}), \min_{\ell\in[K]}\frac{1}{\ell}\sum_{k=1}^{\ell}P_k(Y_{n+1})\right\}$$

is a p-value up to a factor of two if p-values $P_1(Y_{n+1}),\ldots,P_K(Y_{n+1})$ are exchangeable due to Theorem A.1. □

The set in (5.13) can be considered an improvement of the set described in (5.9) but not of the (randomized) set in (5.11), since only the first p-value of the sequence is randomized.

Remark 5.10 (Randomization and "interval-hacking"). A direct use of external randomization is present in both procedures described in Section 5.2.2 and Section 5.2.3. The use of randomization is often avoided in statistical methods, as it can pose challenges to the reproducibility of results. Clearly, randomization becomes problematic when a human is in the loop and runs the procedure multiple times until the desired result is achieved (for example, in the described cases, one can sample U many times until it reaches a value close to zero). Some recommendations aimed at solving this problem are proposed, for example, in Ramdas and Manole (2024, Section 10). Actually, in the data pipeline of split and cross-conformal prediction methods, randomization comes into play in different parts: by default in the division of data into their respective folds; to smoothen p-values as described in (5.8); and potentially to improve the conditional coverage as described in Hore and Barber (2025). In particular, there exists a trade-off between reproducibility and statistical efficiency, and it is not always evident which should be prioritized. In other words, randomized procedures tend to be more efficient than standard procedures but may lack in terms of reproducibility, and vice versa. For instance, our methods may be particularly well-suited in industrial settings, where hundreds or thousands of predictions are made daily, and efficiency may be more important.

5.2.4 Improving cross-conformal prediction

The improvements proposed in the previous subsections are valid for modified cross-conformal prediction; in particular, the new variants are able to produce smaller prediction sets while preserving the same marginal coverage. Specifically, the marginal coverage does not depend on the number of folds K and the number of observations

n. When the folds have the same size, the techniques can be used to enhance cross-conformal prediction (Vovk, 2015): in particular, by examining (5.5), one can observe that it is possible to improve cross-conformal prediction simply by replacing the threshold α with $\alpha + (1-\alpha)(K-1)/(K+n)$ in the prediction sets defined in (5.9), (5.11), and (5.13).

Theorem 5.11. *It holds that*

$$\begin{aligned}
\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha'}(X_{n+1}) &\subseteq \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1}),\\
\mathcal{C}^{\text{u-mod-cross}}_{n,K,\alpha'}(X_{n+1}) &\subseteq \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1}),\\
\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha'}(X_{n+1}) \subseteq \mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha'}(X_{n+1}) &\subseteq \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1}),
\end{aligned}$$

where $\alpha' = \alpha+(1-\alpha)(K-1)/(K+n)$. *If data are exchangeable, the marginal coverage of the conformal prediction sets* $\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha'}(X_{n+1})$, $\mathcal{C}^{\text{u-mod-cross}}_{n,K,\alpha'}(X_{n+1})$ *and* $\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha'}(X_{n+1})$ *is at least* $1-2\alpha'$.

In practice, when $n \gg K$, the prediction sets $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ and $\mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})$ are similar. However, for moderate values of n, we will see that the sets defined in (5.9), (5.11), and (5.13) are typically narrower than $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$, even though $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ assures theoretically a lower coverage guarantee. Clearly, the proposed improvements are valid as long as the marginal coverage level $1-2\alpha'$ is meaningful, which in practical applications is the most common case. It follows that the improvements are not valid, for example, in the extreme case of leave-one-out conformal prediction (the case $K = n$).

5.3 Empirical results

We study the effectiveness of the proposed methods through a simulation study and real data examples. In all experiments, the score function used is the residual score, defined as:

$$s\left((x,y);\mathcal{D}\right) = |y - \hat{\mu}_{\mathcal{D}}(x)|, \tag{5.15}$$

where $\hat{\mu}_{\mathcal{D}}$ is the regression function obtained by applying the regression algorithm $\mathcal{A}$ on $\mathcal{D}$.

5.3.1 Simulation study

OLS. We examine the performance of the proposed methods on simulated data using least squares as our regression algorithm. Data are simulated as in Barber *et al.* (2021,

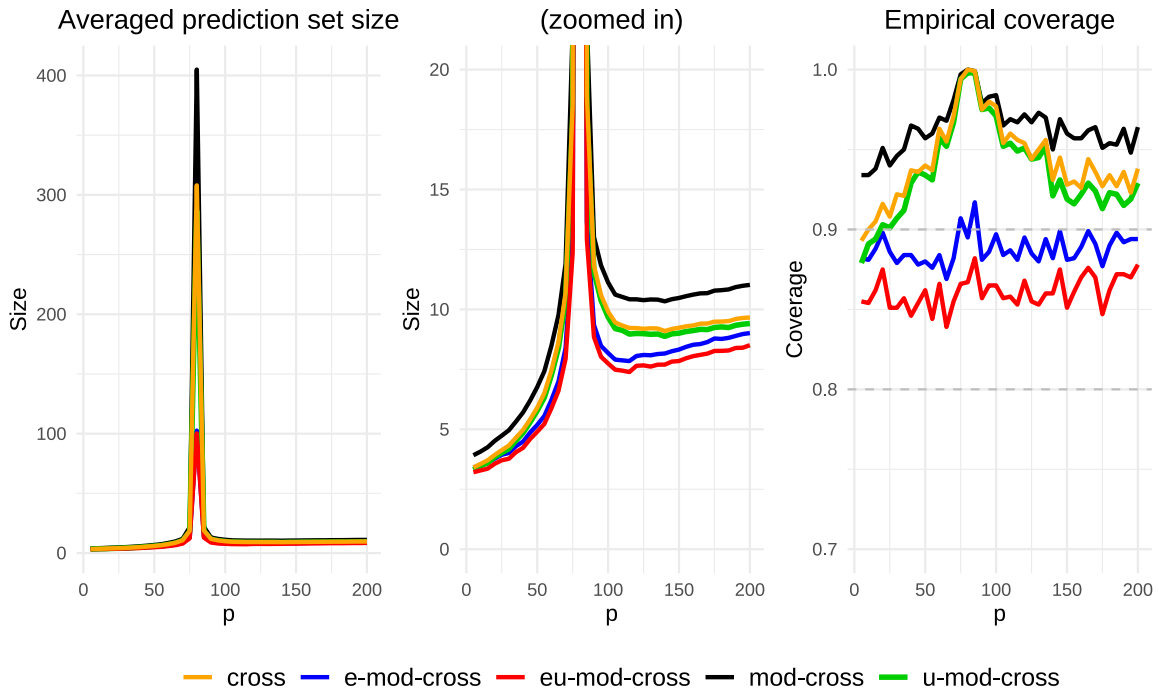

FIGURE 5.1: Simulation results, showing the size and coverage of the predictive sets for cross-conformal prediction and its variants. In the left plot, peaks are observed at 404, 102, 286, 100 and 307 for `mod-cross`, `e-mod-cross`, `u-mod-cross`, `eu-mod-cross` and `cross`, respectively. The parameter α is set to 0.1. The smaller sets are often obtained using `eu-mod-cross` that has coverage between $1 - 2\alpha$ and $1 - \alpha$. The randomized method (`u-mod-cross`) performs similarly to cross-conformal prediction.

Section 6); in particular, the number of observations is $n = 100$ and we let the number of regressors vary $p = \{5, 10, \ldots, 200\}$. The training data points are iid from

$$X_i \sim \mathcal{N}_p(0, I_p) \quad \text{and} \quad Y_i \mid X_i \sim \mathcal{N}(X_i^\top \beta, 1), \tag{5.16}$$

where the vector of coefficients is drawn as $\beta = \sqrt{10}\, v$ for a uniform random unit vector $v \in \mathbb{R}^p$. Ordinary least squares is employed as regression method (if the linear system is underdetermined, then we take the solution that minimizes the ℓ_2-norm). Formally, given the training data $(X_i, Y_i), i \in [n]$, we estimate the regression function $\hat{\mu}(x) = x^\top \hat{\beta}$, where $\hat{\beta} = X_{\text{mat}}^\dagger Y_{\text{vec}}$, Y_{vec} is the response vector, X_{mat} is the matrix of covariates of dimension $n \times p$ and $\dagger$ denotes the Moore-Penrose inverse. The nominal miscoverage rate equals $\alpha = 0.1$, the number of replications (for each p) is 1000 and for each replication, we generate a single test point (X_{n+1}, Y_{n+1}). The number of folds for cross-conformal prediction and its extensions is $K = 5$.

From Figure 5.1, we can see a spike in the size observed at $p = 80$. This is due to the fact that the prediction algorithm is unstable when the number of training points

is equal (or almost equal) to the number of covariates (Hastie *et al.*, 2022). Since the number of folds equals 5, the peak is observed at $p = 80$.

The smaller size is often observed by the exchangeable and randomized variant of cross-conformal prediction. Cross-conformal prediction (Vovk, 2015), is usually over-conservative, and in some cases, its coverage is closer to one rather than to the guaranteed level. This behavior is not shared by the proposed `e-mod-cross` and `eu-mod-cross`. The coverage of these methods lies between the levels $1 - 2\alpha$ and $1 - \alpha$, and remains essentially constant with respect to the number of covariates p. The coverage of the randomized variant `u-mod-cross` depends on p and exhibits a behavior similar to that of standard cross-conformal prediction. In general, our proposals outperform standard cross-conformal prediction in terms of set size.

For additional comparisons, we evaluate our proposed `eu-mod-cross` method with split conformal prediction trained at levels α and 2α. In particular, we note that the marginal coverage of the exchangeable and randomized variant is at least $1 - 2\alpha$. From Figure 5.2, we can observe that for some values of $p \in [25, 60]$, when the prediction algorithm is not stable for the split conformal method, the average length of the `eu-mod-cross` sets is smaller than that of split conformal prediction trained at level 2α. Described differently, both techniques ensure the same coverage level. However, there is no single method that performs best for all values of p. When p is sufficiently small compared to n and the algorithm is stable, split conformal prediction trained at level α performs well, although it uses half the points to train the model.

We compare the results with full conformal prediction and jackknife+. In particular, `eu-mod-cross` conformal prediction is added for comparison. Full conformal prediction is fitted using the package R **conformalInference**. The simulation scenario considered is the same as that described in Section 2.8 and all methods are trained at level $\alpha = 0.1$. The theoretical and empirical guarantees and the computational cost of the methods are reported in Table 5.2 (Section 5.4). The different methods will be compared in terms of coverage and interval size.

Also in Figure 5.3, we can see some spikes in the width of the sets at different levels of p. However, the peak for the jackknife+ is smaller than that observed for split conformal prediction and the `eu-mod-cross` method. The smaller sets are usually obtained using `eu-mod-cross` conformal prediction (or jackknife+). It is important to note that the jackknife+ has the same coverage guarantee as `eu-mod-cross` conformal prediction; however, the empirical coverage for the jackknife+ is around the level $1 - \alpha$ while for our method it lies between levels $1 - 2\alpha$ and $1 - \alpha$. As reported in Table 5.2, the computational cost of the jackknife+ is higher than that of cross-conformal prediction

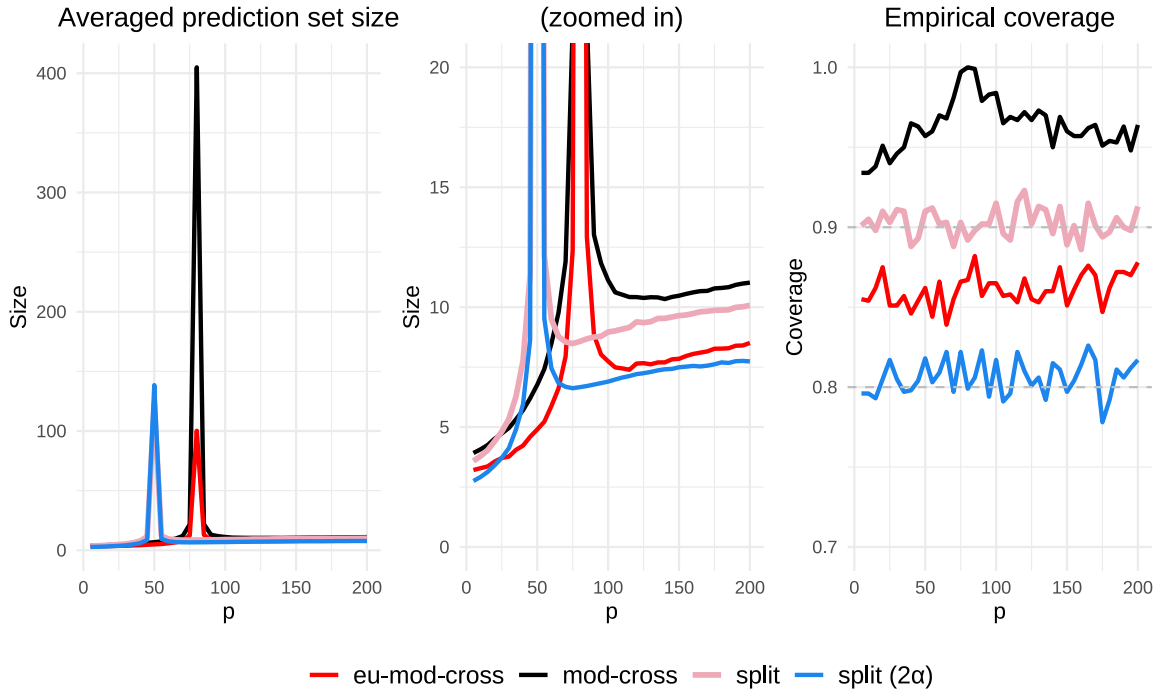

FIGURE 5.2: Simulation results, showing the size and coverage of the predictive intervals obtained using 4 methods. The parameter α is set to 0.1. Split conformal prediction is trained at levels α and 2α and it is compared with `mod-cross` and `eu-mod-cross`. The modified cross-conformal prediction method always overcovers and tends to produce large prediction sets. Its exchangeable and randomized variant gives good results in terms of size. When $p \in [25, 60]$, the average size of the `eu-mod-cross` method is smaller than that of the split conformal prediction method trained at level 2α.

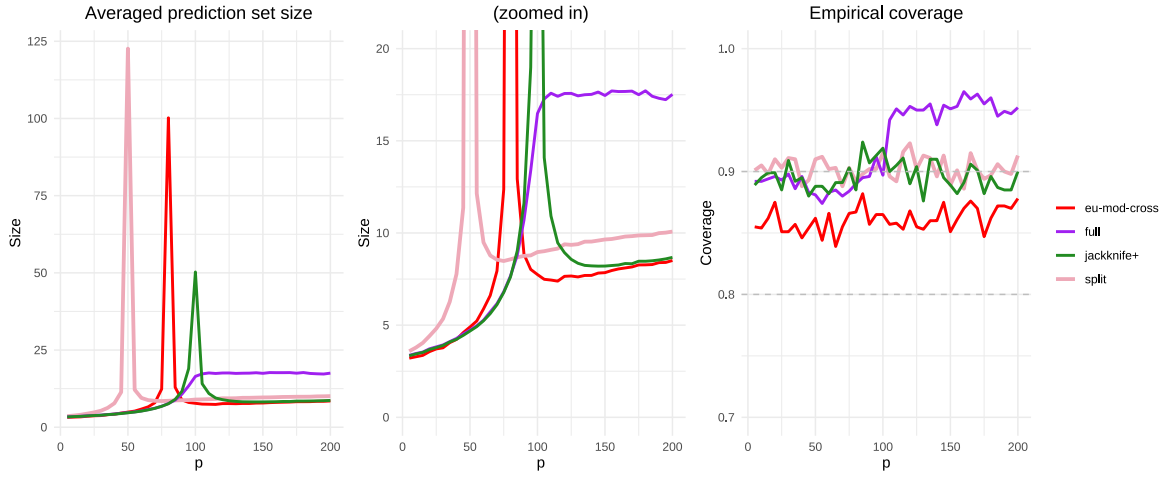

FIGURE 5.3: Simulation results, showing the size and the coverage of the predictive intervals for jackknife+, split conformal prediction and full conformal prediction. The `eu-mod-cross` method is added for comparison and the α-level is set to 0.1. The smaller sets are usually observed by `eu-mod-cross` conformal prediction. Split conformal prediction and jackknife+ have empirical coverage $\approx 1 - \alpha$.

methods: it requires n calls to the prediction algorithm (versus the K required by cross-conformal methods). However, the method is non-randomized, since it can be seen as an extension of cross-conformal prediction to the extreme case $K = n$.

As observed in Barber *et al.* (2021), when $p > n$, full conformal prediction results in intervals of infinite length because for each possible value of the response, all residuals are equal to zero. In practice, the interval is truncated to a finite range, which has a minimal effect on the marginal coverage (Chen *et al.*, 2018). Split conformal prediction and jackknife+ have similar coverage, but the intervals obtained from jackknife+ are usually smaller (except when the algorithm proves to be unstable).

Ridge regression. We repeat the simulation study using a different regression method. Data are generated as in (5.16), with $p \in \{10, 20, \ldots, 200\}$. In this case, we employ ridge regression with a penalty parameter set to $0.001\|X\|_2$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix. The number of replications is 200 and we set $\alpha = 0.1$.

The results are reported in Figure 5.4, and we can observe that coverage is always $\geq 1 - 2\alpha$ for all the considered methods. As in the previous experiment, there is a peak at $p = 80$, although it is much smaller than the one observed in Figure 5.1. This confirms that ridge regression is more stable than ordinary least squares in this setting. The more erratic behavior of the curves is due to the lower number of replications.

5.3.2 Real data application

The methods are applied to the "Online News Popularity" dataset (Fernandes *et al.*, 2015). The dataset contains information on $n = 39\,797$ articles published by the online news blog Mashable. After some preprocessing operations, the number of covariates is $p = 55$ and the covariates contain information about the text of the article. The goal is to predict the number of times the article was shared on a logarithmic scale. Three different regression algorithms are used, specifically: linear regression (as described in Section 5.3.1), lasso regression with penalty parameter set to 0.2 and random forest with 200 trees grown for each forest.

Conformal prediction methods are applied to 10 000 data points randomly sampled without replacement; while other 2500 observations chosen at random from those not part of the training set are used as the test set. The miscoverage rate is set to $\alpha = 0.1$ and the procedure is repeated 100 times to remove the randomness of the split. The method used are cross-conformal prediction and its variants (with $K = 10$) and split conformal prediction. In particular, split conformal prediction is trained both at levels α and 2α. The averages over 100 trials are reported as results in Figure 5.5 and Table 5.1.

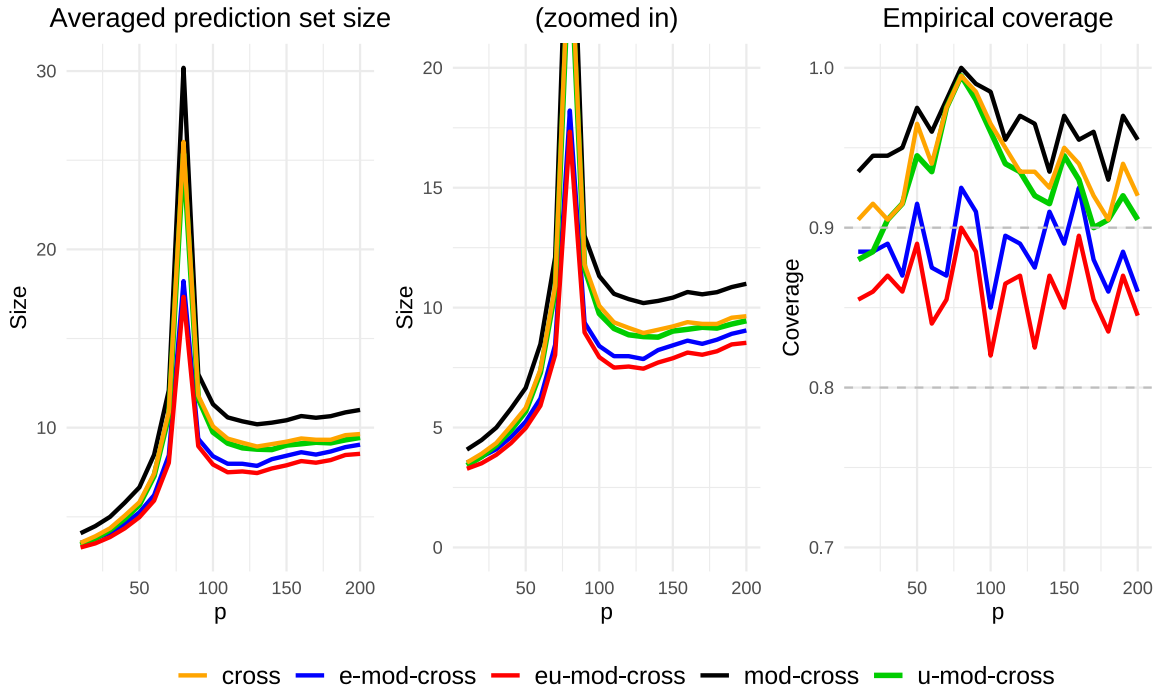

FIGURE 5.4: Simulation results, showing the size and coverage of the predictive sets for cross-conformal prediction and its variants. The parameter α is set to 0.1 and the algorithm used is Ridge regression.

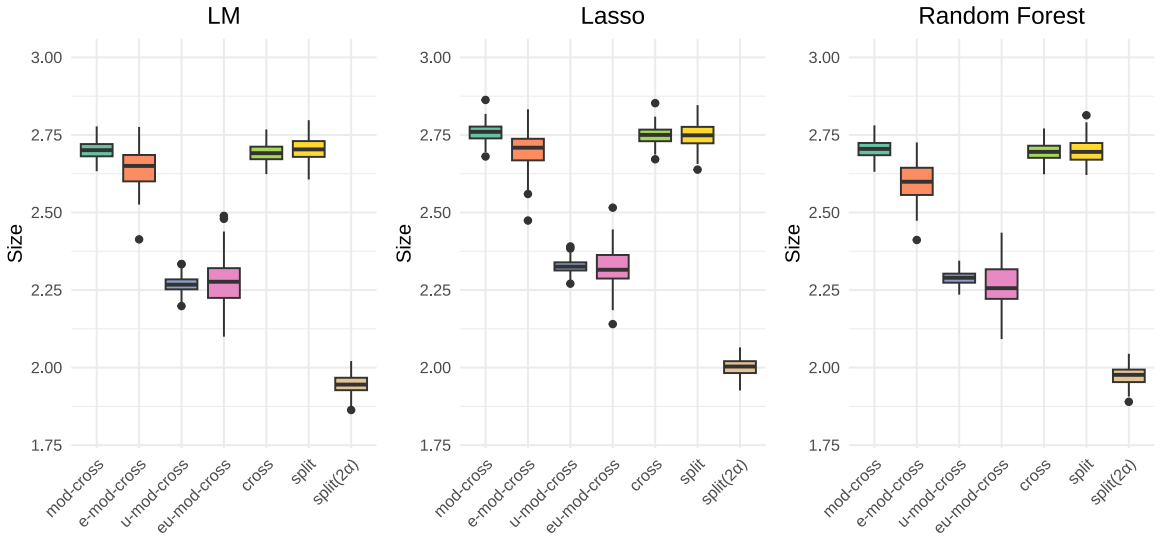

FIGURE 5.5: Empirical size obtained using different regression algorithms and different conformal prediction methods. The methods `mod-cross` and `cross` give similar results. The variants that use randomization (`u-mod-cross` and `eu-mod-cross`) have a smaller size with respect to the other methods trained at level α. The smaller sets are obtained using split conformal prediction trained at level 2α.

Method	Metric	mod-cross	e-mod-cross	u-mod-cross	eu-mod-cross	cross	split	split (2α)
LM	Mean	0.902	0.896	0.851	0.852	0.901	0.900	0.801
	Min	0.884	0.876	0.834	0.826	0.884	0.885	0.777
	Max	0.917	0.915	0.874	0.881	0.916	0.916	0.831
	Range	0.032	0.039	0.040	0.055	0.032	0.030	0.054
Lasso	Mean	0.901	0.896	0.851	0.845	0.900	0.900	0.800
	Min	0.888	0.872	0.831	0.826	0.887	0.884	0.772
	Max	0.916	0.914	0.870	0.876	0.915	0.915	0.822
	Range	0.028	0.042	0.039	0.050	0.028	0.030	0.050
RF	Mean	0.903	0.892	0.853	0.847	0.903	0.900	0.801
	Min	0.885	0.870	0.832	0.815	0.885	0.882	0.778
	Max	0.918	0.915	0.875	0.881	0.917	0.918	0.819
	Range	0.032	0.045	0.042	0.066	0.032	0.036	0.042

TABLE 5.1: Empirical coverage for the News Popularity Dataset using different regression algorithms and different conformal prediction methods. The α-level is set to 0.1. `Mod-cross` and `cross` have empirical coverage around $1-\alpha$ (while guaranteeing a coverage level of at least $1-2\alpha$). The coverage for the randomized methods lies between $1-2\alpha$ and $1-\alpha$. The `e-mod-cross` variant has coverage $\approx 1-\alpha$. Coverage variability is reported through the minimum and maximum values across the 100 replications, together with their difference.

From Figure 5.5, it is possible to note that cross-conformal prediction and its modified version give very similar results in terms of size and are usually slightly better than split conformal prediction trained at level α. The randomized methods `u-mod-cross` and `eu-mod-cross` show a significant improvement in terms of size. The improvement is not as evident for the `e-mod-cross` method, which turns out to be slightly better than the modified method. The smaller sets are obtained using split conformal prediction trained at level 2α. The level of coverage of cross-conformal prediction (and `mod-cross`) is around $1-\alpha$ and the two methods tend to overcover (indeed, they guarantee a miscoverage rate smaller than 2α). The `e-mod-cross` method exhibits similar performance to cross-conformal prediction in terms of coverage; while the coverage of `u-mod-cross` and `eu-mod-cross` is between the levels $1-2\alpha$ and $1-\alpha$. Clearly, there exists a direct relationship between set size and coverage, with smaller prediction sets attaining coverage values closer to $1-2\alpha$. The coverage of split conformal prediction is essentially equal to the target levels $1-\alpha$ or $1-2\alpha$ (see Appendix 4.3 for further details on the coverage properties of the method). As a final remark, the `eu-mod-cross` method demonstrates a higher degree of variability in coverage, while the remaining methods exhibit comparable and more stable variability levels.

Additional experiments on real-world datasets are reported in Appendix C.5.

Method	Theoretical guarantee	Typical empirical coverage	Model training cost
Split	$\geq 1-\alpha$	$\approx 1-\alpha$	1
Full	$\geq 1-\alpha$	$\approx 1-\alpha$ or $> 1-\alpha$ if $\hat{\mu}$ overfits	n_{grid}
Jackknife+	$\geq 1-2\alpha$	$\approx 1-\alpha$	n
Cross	$\geq 1-2\alpha-2/\sqrt{n}$	$\geq 1-\alpha$	K
Mod-cross	$\geq 1-2\alpha$	$> 1-\alpha$	K
e/u/eu-mod-cross	$\geq 1-2\alpha$	$\in [1-2\alpha, 1-\alpha]$	K

TABLE 5.2: Comparison of the properties of different conformal prediction methods. The first two columns regards the marginal coverage. The theoretical guarantees are valid in finite sample. The last column counts the number of times that the algorithm $\mathcal{A}$ is run on a data set containing n training points, for obtaining a prediction set for a new test point. The parameter n_{grid} represents the number of different possible y values.

5.4 Discussion

The chapter describes new variants of cross-conformal prediction that can achieve smaller prediction sets while maintaining valid coverage guarantees. The achievements are based on results on the combination of dependent and exchangeable p-values. In particular, starting from a target miscoverage rate equal to α, the new methods guarantee a marginal coverage of at least $1-2\alpha$. The same coverage is guaranteed by other methods, such as the jackknife+ introduced by Barber *et al.* (2021) or multi-split conformal prediction (with threshold set to a half) by Solari and Djordjilović (2022).

Specifically, similar to cross-validation, the proposed approaches require training the models only K times, unlike n times for the jackknife+ or even potentially an infinite number of times for full conformal prediction (see Table 5.2). The empirical coverage of the proposed methods usually oscillates between levels $1-2\alpha$ and $1-\alpha$, while for standard cross-conformal prediction the empirical coverage is usually around the nominal $1-\alpha$ level. As reported in the experimental results, the size of the sets is smaller than that obtained by split conformal prediction and cross-conformal prediction. Since the results depend on randomized or asymmetric combinations of p-values, the size is generally, though not consistently, more variable compared to cross-conformal prediction. In particular, while randomization and asymmetric combination improve efficiency of the prediction sets, they add an extra-layer of randomness to the procedure. The results presented in Sections 5.2.1, 5.2.2 and 5.2.3 can also be extended to the case where the number of observations in the folds are different. Cross-conformal prediction can be improved using the same techniques, as shown in Section 5.2.4.

The question of which variant of cross-conformal prediction to use in practice is a

subtle one. If it is crucial to obtain a $1 - \alpha$ guarantee against worst-case distributions and unstable algorithms (for example, when downstream decisions critically depend on the provided theoretical guarantees), then it makes sense to run our methods at level $\alpha/2$. On the other hand, if conformal prediction is used merely as "weak guidance" for downstream decisions, meaning the user is willing to tolerate violations of the $1 - \alpha$ target, then we recommend running the proposed cross-conformal variants at level α (where the worst-case coverage is $1 - 2\alpha$, but actual coverage lies between $1 - 2\alpha$ and $1 - \alpha$). Finally, if the situation is intermediate, where conformal prediction is neither applied extremely rigorously nor used as loose guidance, and the goal is to achieve $1 - \alpha$ coverage for "typical" datasets and algorithms, while accepting both overcoverage and undercoverage for unusual distributions or algorithms, then (modified) cross-conformal prediction may be the most appropriate choice.

In general, the sets presented in Section 5.2 show good properties in terms of size and coverage, both in simulations and in applications. The methods can be especially useful in settings where full conformal prediction or jackknife+ are computationally prohibitive.

Chapter 6

Conformal online model aggregation

In Chapter 4 and in Chapter 5, we introduced conformal prediction and described different variants of the method. All the described approaches involve fixing the underlying machine learning (or prediction) model in advance and then obtaining an interval with a prespecified level of marginal coverage. A significant challenge in the application of conformal prediction arises in the context of model selection or aggregation. Since different models yield sets with the same theoretical coverage guarantees (in iid settings), the natural goal is to select the best of a collection of models. In online settings with distribution drift, it appears imprudent to commit to one model (model selection), and hence a better idea is to aggregate the models in an adaptive manner (model aggregation).

In the following, we propose a novel model aggregation strategy based on the combination of prediction sets from multiple algorithms through the voting mechanism described in Chapter 3. Importantly, the weights assigned to each model in the aggregation process are dynamically adjusted over time, reflecting the past performance of the models. The performance of the various models will be assessed based on the size of the generated sets, such as cardinality for a discrete set or Lebesgue measure for a continuous set. Specifically, models consistently producing smaller intervals will be favored over others. Casella *et al.* (1993) outlines the fact that two quantities are important for evaluating a set: the level of coverage and the size. In the realm of conformal prediction, these two concepts are translated as *validity* and *efficiency* (Shafer and Vovk, 2008). This implies that if the starting intervals are valid, then they should naturally be evaluated for their efficiency.

6.1 Problem setup

We focus the attention on an online setting in which data arrive sequentially over time. A plausible scenario arises when dealing with a continuous flow of data; in symbols, for each iteration $t = 1, 2, ..., T$, a covariate-response pair $Z^{(t)} = (X^{(t)}, Y^{(t)}) \in \mathcal{X} \times \mathcal{Y}$[1] is observed. Technically, we first observe $X^{(t)}$, then output a prediction set $\mathcal{C}^{(t)} := \mathcal{C}^{(t)}(X^{(t)})$ that aims to contain $Y^{(t)}$, and we then observe $Y^{(t)}$.

We assume access to K different prediction algorithms (or *models* or *experts*) used to predict $Y^{(t)}$ based on covariates $X^{(t)}$. As a result, in each iteration, for any prespecified error tolerance $\alpha \in (0, 1)$, we can derive K distinct conformal prediction sets (one for each model) $\mathcal{C}_1^{(t)}, \ldots, \mathcal{C}_K^{(t)}$ based on the information from $(\{Z^{(i)}\}_{i=1}^{t-1}, X^{(t)})$. Actually, one can choose to use only a subset of $\{Z^{(i)}\}_{i=1}^{t-1}$ to produce the sets due to computational issues or distribution shifts. As outlined in Chapter 4, if the data are iid (or exchangeable), conformal prediction sets are known to have valid marginal coverage:

$$\mathbb{P}\left(Y^{(t)} \in \mathcal{C}_k^{(t)}\right) \geq 1 - \alpha, \quad k = 1, \ldots, K, \tag{6.1}$$

where the probability is taken with respect to all sources of randomness (the new data point, the training points, possibly the algorithm).

The main question that arises is how to efficiently merge or select among the different conformal prediction sets. This issue is particularly relevant in distributed settings, where the aggregator only has access to the K sets, but not to the underlying data or the individual model predictions. A key challenge is that these sets are not independent, since they are all constructed from the same underlying data.

6.1.1 Proposed solution

We build on the approach proposed in Chapter 3, which involves a weighted majority vote. Let $w^{(t)} = (w_1^{(t)}, \ldots, w_K^{(t)})$ be a set of weights such that

$$w_k^{(t)} \in [0, 1], \quad k = 1, \ldots, K, \text{ and } \sum_{k=1}^{K} w_k^{(t)} = 1,$$

[1]The superscript (t) replaces the subscript n from Chapter 4, to emphasize that we are now in an online setting.

where the weight $w_k^{(t)}$ represents the importance of the prediction algorithm k at time t in the voting procedure. We can define the *weighted majority vote set* as

$$\mathcal{C}_M^{(t)} := \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\left\{ y \in \mathcal{C}_k^{(t)} \right\} > \frac{1 + u^{(t)}}{2} \right\}, \tag{6.2}$$

where $u^{(t)}$ is a realization from $U^{(t)}$ that is a random variable in $\mathcal{U}$ independent of all the data. If the weights are chosen a priori (independent of the data), Corollary 4.2 shows that if the starting intervals are marginally valid (they respect (6.1)) then the coverage of $\mathcal{C}_M^{(t)}$ is at least $1 - 2\alpha$. The factor 2 is provably unavoidable and arises from the arbitrary nature of the dependence between sets in this scenario. Moreover, as pointed out in Appendix B.2, calculating the majority vote set can be computed in $O(K \log K)$ time for real-valued sets. This is due to the need of sorting the interval endpoints and making a single pass over them. In the discrete scenario, the process is straightforward and requires only the counting of intervals that include a specific label.

It is worth noting that the method requires only the K input sets to construct $\mathcal{C}_M^{(t)}$. This feature is particularly advantageous in decentralized settings, where the aggregator, at each time step, has access solely to the sets themselves rather than to the conformal p-values, the scores, or the raw outputs of the individual models.

There are two hurdles to applying these ideas directly. First, when weights are based on past data, the weights themselves become random variables, so the theory does not immediately apply. The second question is how one should update the weights in a theoretically sound way that performs well practically. The rest of this chapter will present ways to overcome these hurdles, resulting in a principled and practical conformal online model aggregation (COMA) algorithm. In particular, in Section 6.2, we provide the theoretical guarantees of the proposed method. In Section 6.3, we introduce a method based on (6.2), in which the weights are updated based on the performance obtained by the different models. The method relies on a modification of the EW algorithm, introduced in de Rooij *et al.* (2014), and enables the establishment of an upper bound on the observed loss in relation to the best-performing algorithm. In Section 6.4, we demonstrate the practical effectiveness of the method in iid settings. In Section 6.5, our approach is applied in scenarios of potential distribution shifts, in conjunction with the methods by Gibbs and Candès (2021), Angelopoulos *et al.* (2024a) and Angelopoulos *et al.* (2024a).

6.2 Weighted majority with data-dependent weights

Suppose that the weight vector $w^{(t)}$ is a realization of a random vector $W^{(t)}$ that maps the first $t-1$ observations of $\{Z^{(i)}\}_{i=1}^{t-1}$ into the $(K-1)$-dimensional simplex $\Delta_{K-1} := \{w \in [0,1]^K : \mathbf{1}^\top w = 1\}$, or more succinctly, $W^{(t)} : (\mathcal{X} \times \mathcal{Y})^{t-1} \to \Delta_{K-1}$. In the following, we define the miscoverage indicator for the k-th set at time t as $\phi_k^{(t)} := \mathbb{1}\{y^{(t)} \notin \mathcal{C}_k^{(t)}\}$ and $[K] := \{1, \ldots, K\}$. In the following, we will make the assumptions:

Assumption 1. Marginal coverage and negative total elementwise correlation between $\phi^{(t)}$, $W^{(t)}$:

$$\mathbb{E}\left[\phi_k^{(t)}\right] \leq \alpha \text{ for } k \in [K], \text{ and } \mathbb{E}\left[\sum_{k=1}^{K}\left(\phi_k^{(t)} - \mathbb{E}[\phi_k^{(t)}]\right)\left(W_k^{(t)} - \mathbb{E}[W_k^{(t)}]\right)\right] \leq 0. \quad (6.3)$$

$\phi^{(t)}$, $W^{(t)}$ could be negatively correlated if, for example, predictors that were assigned high weights (based on previous data) are less likely to make mistakes.

Theorem 6.1. *Let $\mathcal{C}_1^{(t)}, \ldots, \mathcal{C}_K^{(t)}$ be $K \geq 2$ different conformal prediction sets for $Y^{(t)}$, let $W^{(t)}$ be a random vector in Δ_{K-1} depending only on $\{Z^{(i)}\}_{i=1}^{t-1}$, and consider the weighted majority vote set $\mathcal{C}_M^{(t)}$ in (6.2). It always holds that $\mathbb{P}(Y^{(t)} \notin \mathcal{C}_M^{(t)}) \leq 2\mathbb{E}[\max_{k\in[K]}(\phi_1^{(t)}, \ldots, \phi_K^{(t)})]$. If Assumption 1 holds, then we further have $\mathbb{P}(Y^{(t)} \notin \mathcal{C}_M^{(t)}) \leq 2\alpha$. This implies that with probability at least $1-2\alpha$ the set $\mathcal{C}_M^{(t)}$ is non-empty.*

Proof. By direct calculation, similarly to Theorem 3.14, we can write

$$\mathbb{P}\left(Y^{(t)} \notin \mathcal{C}_M^{(t)}\right) = \mathbb{P}\left(\sum_{k=1}^{K} W_k^{(t)}\phi_k^{(t)} \geq \frac{1}{2} - U^{(t)}/2\right) \overset{(i)}{\leq} 2\mathbb{E}\left[\sum_{k=1}^{K} W_k^{(t)}\phi_k^{(t)}\right]$$

where (i) is valid due to the Uniformly-randomized Markov Inequality in Theorem (1.7).

The first part of the theorem is a direct consequence of Hölder inequality. Indeed, due to the fact that $||W||_1 = 1$, we notice that

$$\mathbb{E}\left[\sum_{k=1}^{K} \phi_k^{(t)} W_k^{(t)}\right] \leq \mathbb{E}\left[||\phi^{(t)}||_\infty ||W^{(t)}||_1\right] = \mathbb{E}\left[||\phi^{(t)}||_\infty\right],$$

where $\phi^{(t)} = (\phi_1^{(t)}, \ldots, \phi_K^{(t)})$.

In the case of negative total elementwise correlation it holds that $\sum_{k=1}^{K} \mathbb{E}[W_k^{(t)}\phi_k^{(t)}] \leq \sum_{k=1}^{K} \mathbb{E}[W_k^{(t)}]\mathbb{E}[\phi_k^{(t)}]$ and under marginal coverage, we have that

$$\mathbb{P}\left(Y^{(t)} \notin \mathcal{C}_M^{(t)}\right) \leq 2\sum_{k=1}^{K} \mathbb{E}\left[W_k^{(t)}\phi_k^{(t)}\right] \leq 2\sum_{k=1}^{K} \mathbb{E}[W_k^{(t)}]\,\mathbb{E}[\phi_k^{(t)}] \leq 2\alpha \sum_{k=1}^{K} \mathbb{E}[W_k^{(t)}] = 2\alpha.$$

Since the target is contained with probability at least $1 - 2\alpha$, it follows that the set is non-empty with at least the same probability. □

Remark 6.2. The first part of the theorem states, without any assumption, that the probability of miscoverage is bounded by $2\mathbb{E}[\max_{k\in[K]}(\phi_1^{(t)}, \ldots, \phi_K^{(t)})]$. This can translate into a trivial bound (like 1) if there is at least one algorithm that always makes an error, while it can be approximately 2α in the case where miscoverage events are strongly correlated and $\mathbb{E}[\phi_k^{(t)}] \leq \alpha$, for each $k \in [K]$.

Remark 6.3. The use of randomization enhances the efficiency of the prediction sets while preserving the theoretical validity of the procedure. If one opts to avoid randomization, for instance, to mitigate any potential for p-hacking, then setting $u^{(t)} = 0$ maintains the coverage guarantee, which is ensured by Markov's inequality.

Remark 6.4. On examining the proof, it is easy to see that the degradation of the coverage degrades smoothly with the degree of violation of Assumption 1. In particular, if we relax the assumption in (6.3) to $\sum_{k=1}^{K} \mathbb{E}[\phi_k^{(t)} W_k^{(t)}] \leq (1+\nu) \sum_{k=1}^{K} \mathbb{E}[\phi_k^{(t)}]\mathbb{E}[W_k^{(t)}]$, for some $\nu \geq 0$ which measures the degree to which assumption (6.3) is violated, we see that the corresponding bound on the miscoverage becomes $2\alpha(1+\nu)$, recovering the above theorem when $\nu = 0$.

A simple instance in which the negative total elementwise correlation stated in (6.3) holds is described here. Consider the scenario involving only two models, resulting in a weight vector denoted by $(W_1, 1-W_1)$, where the superscript (t) is omitted for simplicity. Moreover, assume $\phi_1 = \phi_2 = \phi$, which might occur when two algorithms deliver nearly identical results. If $\mathrm{cov}(W_1, \phi) = \rho$, then it follows that $\mathrm{cov}(1 - W_1, \phi) = -\rho$; and this implies $\sum_{k=1}^{2} \mathbb{E}[\phi W_k] = \sum_{k=1}^{2} \mathbb{E}[\phi]\,\mathbb{E}[W_k]$, so (6.3) holds with equality. This is just a simple illustration; the assumption is investigated later.

In addition, if the weight vector takes the form $w_k^{(t)} \approx 1$, and $w_j^{(t)} \approx 0$ for all $j \neq k$, then the procedure reduces to the selection of the k-th model, which is the desired result in the presence of a clearly best model. A potential drawback of the procedure is that it may produce a union of intervals, even when the input sets are intervals. Luckily, the combined set is an interval as long as the intersection of the K intervals is non-empty (see Lemma 3.6), which we expect to be the case more often than not. In practice, the probability of observing either an empty set or a union of sets is close to zero in our experiments. Moreover, when the weights are concentrated on a single algorithm, the majority vote reduces to the set of that algorithm.

As a final remark, we note that the expectation in the marginal coverage condition in (6.3) is taken over both the training data and the test point. An alternative condition

is the following:

$$\mathbb{E}\left[\phi_k^{(t)} \mid \{Z^{(i)}\}_{i=1}^{t-1}\right] \leq \alpha, \text{ for all } k \in [K], \tag{6.4}$$

where, in this case, the miscoverage error is bounded conditionally on the "training set" $\{Z^{(i)}\}_{i=1}^{t-1}$. In practice, the model may not have been updated after each additional point, but we still call this condition as *training conditional coverage*. This condition can be used to extend the results in Theorem 6.1.

Lemma 6.5. *Let $\mathcal{C}_1^{(t)}, \ldots, \mathcal{C}_K^{(t)}$ be $K \geq 2$ different prediction sets for $Y^{(t)}$, let $W^{(t)}$ be a random vector in Δ_{K-1} depending only on $\{Z^{(i)}\}_{i=1}^{t-1}$, and consider the weighted majority vote set $\mathcal{C}_M^{(t)}$ in* (6.2)*. If* (6.4) *holds, then $\mathbb{P}(Y^{(t)} \notin \mathcal{C}_M^{(t)}) \leq 2\alpha$.*

Proof. Following the same steps as in Theorem 6.1, if the assumptions holds then

$$\mathbb{P}\left(Y^{(t)} \notin \mathcal{C}_M^{(t)}\right) \leq 2\sum_{k=1}^{K} \mathbb{E}\left[W_k^{(t)}\phi_k^{(t)}\right] = 2\sum_{k=1}^{K} \mathbb{E}\left[W_k^{(t)}\mathbb{E}\left[\phi_k^{(t)} \mid \{Z^{(i)}\}_{i=1}^{t-1}\right]\right] \leq 2\alpha.$$

□

However, this condition does not hold in general for conformal prediction methods. A weaker type of guarantee is addressed by PAC-type results. In particular, it involves two parameters: the miscoverage rate α and the parameter δ. The training conditional coverage for the split conformal method is explored in Vovk (2012), whereas the training conditional coverage for the jackknife+ and the K-fold CV+ (Barber *et al.*, 2021) is examined in Bian and Barber (2023). In particular, under certain conditions, the training conditional coverage of these methods is asymptotically (in the size of the calibration set) $1-\alpha$ with probability at least $1-\delta$.

A different condition holds for full conformal prediction in the online case when data is iid (see Chapter 2.2 in Vovk *et al.* (2005)), specifically

$$\mathbb{E}\left[\phi_k^{(t)} \mid \{\phi_k^{(i)}\}_{i=1}^{t-1}\right] \leq \alpha, \text{ for all } k \in [K]. \tag{6.5}$$

The left hand side above conditions on less information than (6.4). This condition states that the errors are independent in the online setting when data are independent and identically distributed. However, this condition does not suffice to guarantee a miscoverage rate of at most 2α.

6.3 Dynamic merging via exponential weighted majority

So far, we have studied the theoretical properties of the method in the case where the weights themselves are random variables; what remains is to find a sensible and practical way to update these weights in order to improve the efficiency of our weighted majority vote set. Building on previous discussions on the *efficiency* of the sets, we begin by defining the loss functions relevant to the problem under consideration: for regression tasks, the loss is quantified by the interval's size, and for classification tasks, by the count of elements within the set. In particular, we define the loss function $\ell(\cdot)$ as

$$\ell := \ell(\mathcal{C}) = \begin{cases} \text{Lebesgue measure}(\mathcal{C}), & \text{if } \mathcal{Y} \subseteq \mathbb{R}, \\ |\mathcal{C}|, & \text{if } \mathcal{Y} = \{y_1, \ldots, y_D\}, \end{cases} \tag{6.6}$$

where $\mathcal{C}$ denotes any set, $|\mathcal{C}|$ represents the cardinality of $\mathcal{C}$, and $\{y_1, \ldots, y_D\}$ describes a discrete set with D elements (or labels). In general, one could choose some nondecreasing function of the Lebesgue measure (or of the cardinality), but the Lebesgue measure and the cardinality seem to be a sensible canonical choice.

At each iteration, we observe losses of experts $\ell^{(t)} = (\ell_1^{(t)}, \ldots, \ell_K^{(t)}) \in \{\mathbb{R}^+\}^K$ and the cumulative loss for a given model after t rounds is given by $L_k^{(t)} := \ell_k^{(1)} + \cdots + \ell_k^{(t)}$. A notable property of the majority vote is that it generates sets that are never larger than twice the weighted average size of the initial sets.

Lemma 6.6. *Let $\mathcal{C}_M^{(t)}$ be the set defined in* (6.2)*, $\ell_M^{(t)} = \ell(\mathcal{C}_M^{(t)})$ be the loss function defined in* (6.6) *associated with $\mathcal{C}_M^{(t)}$, and $w^{(t)}$ be any realization of $W^{(t)}$. Then,*

$$\ell_M^{(t)} = \ell(\mathcal{C}_M^{(t)}) \leq 2 \sum_{k=1}^{K} w_k^{(t)} \ell(\mathcal{C}_k^{(t)}). \tag{6.7}$$

In addition,

$$\mathbb{E}\left[\ell_M^{(t)}\right] \leq 2 \log 2 \, \mathbb{E}\left[\sum_{k=1}^{K} W_k^{(t)} \ell_k^{(t)}\right]. \tag{6.8}$$

Proof. The first part is a corollary of Theorem 3.14. For the second part, we have

$$\begin{aligned}
\mathbb{E}\left[\ell_M^{(t)}\right] &= \mathbb{E}\left[\sum_{d=1}^{D} \mathbb{1}\left\{\sum_{k=1}^{K} W_k^{(t)} \mathbb{1}\left\{y_d \in \mathcal{C}_k^{(t)}\right\} > \frac{1}{2} + \frac{U^{(t)}}{2}\right\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{d=1}^{D} \mathbb{1}\left\{\sum_{k=1}^{K} W_k^{(t)} \mathbb{1}\left\{y_d \in \mathcal{C}_k^{(t)}\right\} > \frac{1}{2} + \frac{U^{(t)}}{2}\right\} \mid U^{(t)}\right]\right] \\
&\leq \mathbb{E}\left[\sum_{d=1}^{D} \mathbb{E}\left[\frac{2}{1+U^{(t)}} \sum_{k=1}^{K} W_k^{(t)} \mathbb{1}\left\{y_d \in \mathcal{C}_k^{(t)}\right\} \mid U^{(t)}\right]\right] \\
&= \mathbb{E}\left[\frac{2}{1+U^{(t)}} \mathbb{E}\left[\sum_{k=1}^{K} W_k^{(t)} \ell_k^{(t)}\right]\right] = 2\log 2\, \mathbb{E}\left[\sum_{k=1}^{K} W_k^{(t)} \ell_k^{(t)}\right],
\end{aligned}$$

The proof in the case $\mathcal{Y} \subseteq \mathbb{R}$ is analogous, due to the Fubini-Tonelli Theorem. (Note that $\log 2 \approx 0.7 < 1$) □

From (6.7) and (6.8), it is now clear that our goal is to find a good weighting system in order to narrow down the size of $\mathcal{C}_M^{(t)}$, in particular, we aim to assign a higher weight to experts that consistently yield smaller intervals compared to others. In other words, we want to prioritize methods that yield smaller sets. Furthermore, we want the cumulative loss of our proposed method $L_M^{(t)} := \ell_M^{(1)} + \cdots + \ell_M^{(t)}$ to be not much larger than (and possibly smaller than) the cumulative loss incurred by the best expert in hindsight $L_*^{(t)} := \min_{k\in[K]} L_k^{(t)}$. As a remark, the factor $\log 2$ in (6.8), can be seen as the improvement due to randomization.

6.3.1 Exponential weighted majority algorithm

A possible way to update our weights is the EW (or Hedge) Algorithm (Littlestone and Warmuth, 1994; Freund and Schapire, 1997), which corresponds to Algorithm 2 with $\eta^{(t)} = \eta$, for all t. A key quantity is the *hedge* loss $H^{(t)} = h^{(1)} + \cdots + h^{(t)}$, where $h^{(t)}$ is the dot product $h^{(t)} = w^{(t)} \cdot \ell^{(t)} = \sum_k w_k^{(t)} \ell_k^{(t)}$. This can be interpreted as the expected loss achieved by randomly selecting the model k with probability $w_k^{(t)}$.

The algorithm adjusts the weights based on the performance of the methods in various rounds, with adjustments made in inverse proportion to the exponential of their total loss scaled by a parameter η. This "learning rate" η plays a crucial role; if η approaches zero, the weights approach a uniform weighting, and if $\eta \to \infty$ the algorithm reduces to the Follow-the-Leader strategy, which puts all the weight on the method with the smallest loss so far.

In certain scenarios, determining the appropriate value for this parameter may be difficult; indeed, we do not know if there is a method outperforming the others or any

Data: $\mathcal{C}_1^{(t)}, \ldots, \mathcal{C}_K^{(t)}$ at each round t, initial learning rate $\eta^{(0)} \geq 0$
$w_k^{(1)} \leftarrow 1/K, L_k^{(0)} \leftarrow 0$, $k = [K]$;
for *rounds* $t = 1, \ldots, T$ **do**
  $\mathcal{C}_M^{(t)} \leftarrow \left\{ y \in \mathcal{Y} : \sum_{k=1}^K w_k^{(t)} \mathbb{1}\{y \in \mathcal{C}_k^{(t)}\} > \frac{1+u^{(t)}}{2} \right\}$;
  Receive loss $\ell_k^{(t)}$, update $L_k^{(t)} := L_k^{(t-1)} + \ell_k^{(t)}$;
  Update learning rate $\eta^{(t)}$;
  $w_k^{(t+1)} \leftarrow \exp\{-\eta^{(t)} L_k^{(t)}\} / \sum_{j=1}^K \exp\{-\eta^{(t)} L_j^{(t)}\}$;
end

Algorithm 2: Exponentially Weighted Majority Vote

constant works well in some situations but not in others. Our solution is to use the AdaHedge algorithm, introduced in de Rooij *et al.* (2014), which dynamically adjusts the learning parameter over time. At each round, weights are assigned according to $w_k^{(t+1)} \propto e^{-\eta^{(t)} L_k^{(t)}}, k \in [K]$, with

$$\eta^{(t)} := \frac{\log K}{\delta^{(1)} + \cdots + \delta^{(t-1)}},$$

where $\delta^{(t)} = h^{(t)} - m^{(t)}$ is the difference between the hedge loss $h^{(t)}$ and the *mix* loss,

$$m^{(t)} := -\frac{1}{\eta^{(t)}} \log \left(w^{(t)} \cdot \exp\{-\eta^{(t)} \ell^{(t)}\} \right).$$

The AdaHedge algorithm provides an upper bound on regret, defined as the difference between the hedge loss and the loss of the best method up to time t, specifically: $H^{(t)} - L_*^{(t)}$.

Consider $\ell_+^{(t)} = \max_{k \in [K]} \ell_k^{(t)}$ and $\ell_-^{(t)} = \min_{k \in [K]} \ell_k^{(t)}$ to represent the smallest and largest loss in round t, and let $L_+^{(t)}$ and $L_-^{(t)}$ represent their cumulative sum. Additionally, define

$$V^{(t)} := \max\{v^{(1)}, \ldots, v^{(t)}\}$$

denote the largest loss range, where $v^{(t)} = \ell_+^{(t)} - \ell_-^{(t)}$. Suppose that our objective is to find an upper bound for regret after an arbitrary number of rounds T. According to Theorem 8 in de Rooij *et al.* (2014), we have

$$H^{(T)} \leq L_*^{(T)} + 2 \left(V^{(T)} \log K \frac{(L_+^{(T)} - L_*^{(T)})(L_*^{(T)} - L_-^{(T)})}{L_+^{(T)} - L_-^{(T)}} \right)^{1/2} + V^{(T)} \left(\frac{16}{3} \log K + 2 \right). \tag{6.9}$$

It is now possible to bound the loss of the sets produced by the majority vote method.

Theorem 6.7. *If Assumption 1 holds, then the exponential weighted majority vote algorithm satisfies* $\mathbb{P}(Y^{(t)} \in \mathcal{C}_M^{(t)}) \geq 1 - 2\alpha$. *The loss of the weighted majority vote set over* T *rounds can be bounded as*

$$L_M^{(T)} \leq 2L_*^{(T)} + 4\left(V^{(T)} \log K \frac{(L_+^{(T)} - L_*^{(T)})(L_*^{(T)} - L_-^{(T)})}{L_+^{(T)} - L_-^{(T)}}\right)^{1/2} + 2V^{(T)}\left(\frac{16}{3}\log K + 2\right). \tag{6.10}$$

In addition, if the loss range is bounded, i.e. $\sup_t v^{(t)} \leq D$, *then*

$$\frac{L_M^{(T)}}{T} \leq \frac{2L_*^{(T)}}{T} + o(1) \tag{6.11}$$

Proof. The first part of the theorem is just a consequence of Theorem 6.1 applied to Algorithm 2. From Lemma 6.6 we have $\ell_M^{(t)} \leq 2h^{(t)}$. This implies that $L_M^{(T)} = \ell_M^{(1)} + \cdots + \ell_M^{(T)} \leq 2H^{(T)}$; combining this result with (6.9) gives the desired bound. If the loss range is bounded, then $v^{(t)} \leq D$, and using Corollary 9 in de Rooij *et al.* (2014) we have

$$\begin{aligned} L_M^{(T)} &\leq 2L_*^{(T)} + 4\left(\sum_{t=1}^{T}(v^{(t)})^2 \log K\right)^{1/2} + 2V^{(T)}\left(\frac{4}{3}\log K + 2\right) \\ &\leq 2L_*^{(T)} + 4(TD^2 \log K)^{1/2} + 2D\left(\frac{4}{3}\log K + 2\right) \end{aligned}$$

Dividing both sides by T gives the result. □

The result in (6.10) is identical to the regret bound (6.9), except for a multiplicative factor of 2. In addition, under the assumption of bounded loss range (which is automatically satisfied in the classification setting, for example), the empirical average loss of the majority vote set is at most twice the empirical average loss of the best method in hindsight, up to an asymptotically vanishing term. Although initially developed for aggregating point forecasts, the AdaHedge algorithm is adapted to a distinctly different context: the aggregation of sets. It is somewhat surprising that the regret bound can also be applied to this scenario. Clearly, other methods are available to update the weights in Algorithm 2. However, we recommend the proposed solution because it does not require any tuning parameters (it is completely data-driven), ensures a regret bound even in the case of unbounded loss functions, and demonstrates strong empirical performance.

As outlined in Theorem 6.7, Assumption 1 guarantees a miscoverage rate of at most 2α. However, since the weights are updated based on the losses of the different algorithms, a sufficient condition which lies between (6.4) and (6.5) and guarantees a

miscoverage rate of at most 2α for Algorithm 2 is

$$\mathbb{E}\left[\phi_k^{(t)} \mid \{\phi_k^{(i)}\}_{i=1}^{t-1}, \{\ell_k^{(i)}\}_{i=1,k=1}^{t-1,K}\right] \leq \alpha, \text{ for all } k \in [K].$$

However, it remains unclear whether this condition is likely to hold either in theory or in practice. For this reason, we regard Assumption 1 as the most reasonable choice in practical applications. From a theoretical point of view, this condition is difficult to verify, since the sets are based on the same underlying data and both the errors and the losses may be arbitrarily correlated. Nevertheless, it can be empirically assessed and appears to hold reasonably well in the experiments presented in the following sections. In addition, if the assumption is not met, Remark 6.4 shows that the coverage guarantee weakens gradually as the violation grows.

In certain instances, the Lebesgue measure of the set cannot be directly employed because of situations where the interval corresponds to entire real line. For example, we will see that in adaptive conformal prediction (Gibbs and Candès, 2021) it may happen that the predictive set of an agent coincides with $\mathcal{Y}$. A possible solution is to define the loss function as an increasing, nonnegative, bounded function of the measure of the set (so that the weight of a predictor does not get permanently set to zero if its loss happens to be unbounded in some round). We will discuss this example in the following sections.

6.4 Experiments in iid setting

We now test the solution proposed in the previous section in different scenarios. When the described dynamic merging algorithm is used (with constant, adaptive or another stepsize rule) in this conformal context, we call the method as "conformal online model aggregation", or COMA for short. In the iid setting, it is often the case that our adaptive stepsize update rule results in the weight vector rapidly putting almost full (unit) weight on the best predictor, resulting effectively in online model "selection". In contrast, in non-iid settings with possible distribution shifts, we will see that the weights can meaningfully fluctuate across the different experts, favoring those that perform best at different times.

We start by examining our proposed methods, in a scenario where data are iid both in a classification and regression task.

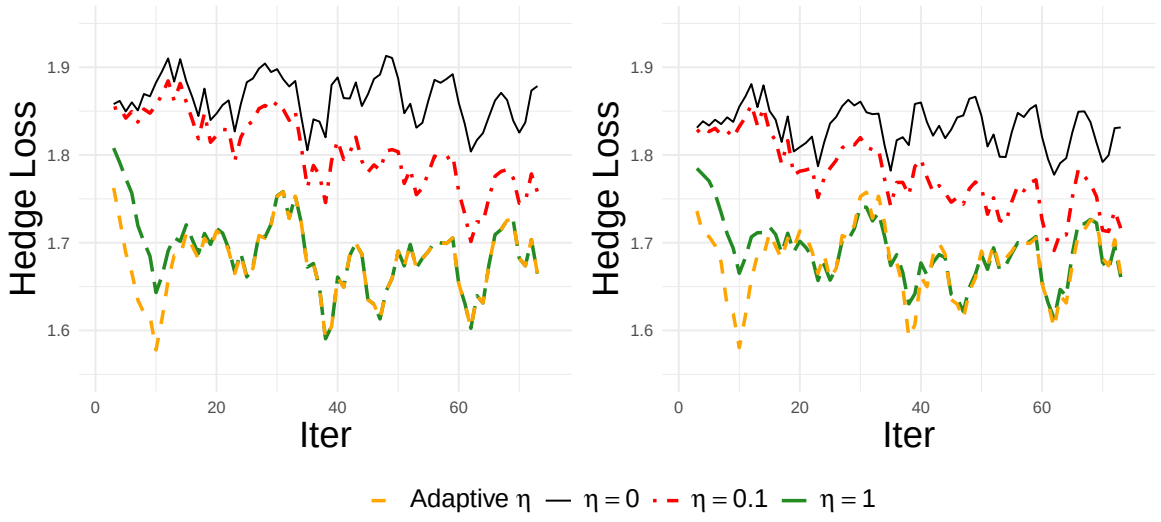

FIGURE 6.1: Hedge loss (h_t) obtained during various iterations with either a constant or adaptive learning rate scheme. The case $\eta = 0$ coincides with the standard (non-weighted) majority vote. The case with $K = 4$ algorithms is shown in the left plot, while the case with $K = 5$ algorithms is shown in the right plot. The series have been smoothed using a moving average $(5, \frac{1}{5})$. In both cases, COMA with adaptive η quickly achieves the smallest loss.

6.4.1 Regression in iid setting

To examine the practical behavior of COMA, we apply it to a real-world dataset. Specifically, the dataset comprises 75 000 observations pertaining to Airbnb apartments in New York City. The response variable is represented by the logarithm of the nightly price, while the covariates include information about the apartment and its geographical location. The data has been partitioned into 75 rounds, each consisting of 1000 observations and a single test point. At each iteration, K split conformal prediction intervals are constructed based on the data observed at round t using various models. In the first simulation, $K = 4$ regression algorithms are employed: linear model, lasso regression with penalty parameter set to 0.1, ridge regression with penalty parameter set to 0.1, and random forest with 250 trees. In the second scenario, a neural net with 8 nodes is added, and so $K = 5$. In both cases, the random forest emerges as the best model, producing narrower intervals. However, whereas the differences with other models are substantial in the first scenario, in the second scenario the neural net also delivers satisfactory results.

The weights are adjusted using the COMA method with either adaptive or fixed values of η. For Algorithm 2, when $\eta^{(t)} = \eta$ is kept constant across iterations, two values are considered: $\eta = 0.1$ and $\eta = 1$. The case $\eta = 0$ is included as a benchmark; it

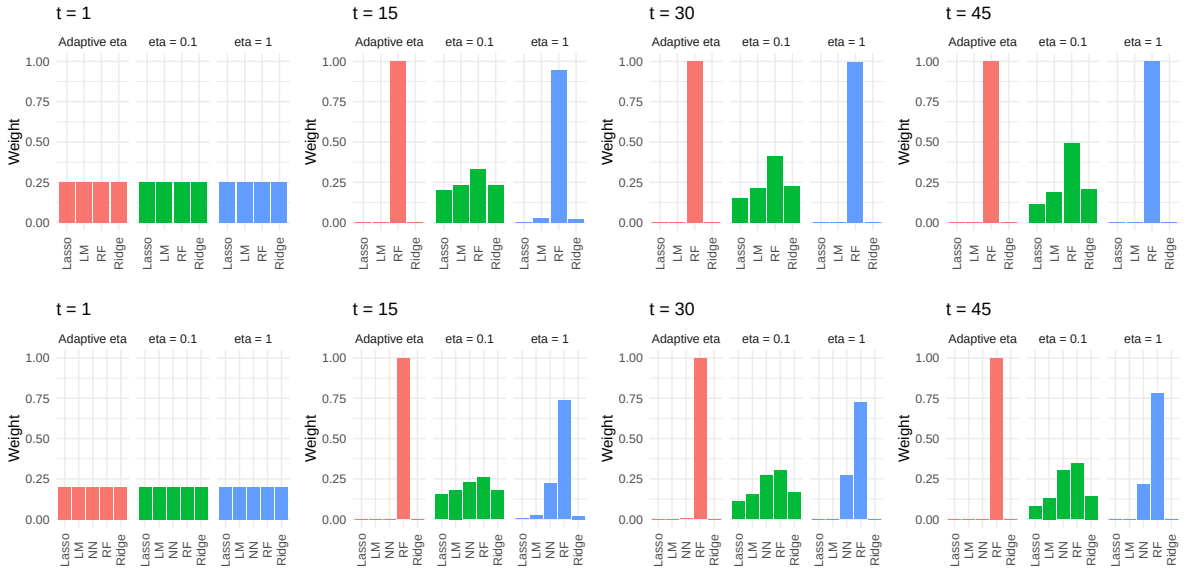

FIGURE 6.2: Weights assumed by the regression algorithms during different iterations. The case with $K = 4$ algorithms is shown in the top row, while the case with $K = 5$ algorithms is shown in the bottom row. After few iterations, the strategy with adaptive η assigns full weight to the random forest. The COMA method with fixed η is more conservative in assigning weights.

essentially corresponds to the solution proposed by Cherubin (2019), where the weights remain uniform and fixed over time. In both cases, the loss of the COMA algorithm with adaptive η decreases rapidly and then stabilizes. This behavior is explained by the fact that the algorithm quickly assigns unit mass to the most efficient model that is the random forest (Figure 6.2). In contrast, when η is set at 0.1, the decrease in loss is slower, and the algorithm behaves more conservatively in assigning weights. Finally, when the weights remain constant throughout the iterations, the loss stays stable and no improvements are observed.

We also compare our approach with several alternative methods. As a first benchmark, we apply conformal prediction to an ensemble model constructed from the $K = 5$ algorithms used in our analysis, where predictions are aggregated using a simple average. Although this approach guarantees valid coverage, it requires access to the estimated outputs of all K models to construct the prediction set. In contrast, our method relies only on the K sets themselves and can thus be applied in decentralized scenarios where the different experts cannot communicate. In addition, when conformal sets are constructed from density estimation rather than from discrepancy measures relative to a point predictor, implementing an ensemble method becomes challenging. A second competitor is the naive union of the K sets. This strategy also guarantees valid coverage, but typically results in excessively wide prediction sets (in terms of size, this

Method	Ensemble	Linear Regression	Bonf.	Union	Adapt. COMA ($K = 4$)	Adapt. COMA ($K = 5$)
Size	1.738	1.874	1.863	2.229	1.665	1.670
Coverage	0.973	0.987	0.987	0.987	0.973	0.973

TABLE 6.1: Comparison of empirical size and coverage with $\alpha = 0.05$. The best method in terms of efficiency is COMA with adaptive η.

can be interpreted as a worst-case benchmark). On the other hand, by the Bonferroni inequality, the intersection of the K sets ensures a coverage of at least $1 - K\alpha$, which may become meaningless as K grows. To provide a fair comparison, we obtain the conformal prediction sets at level $2\alpha/K$ using the $K = 5$ different algorithm. Under this choice, the intersection achieves a coverage of at least $1 - 2\alpha$ (as our method). As a final benchmark, we consider standard linear regression to construct a prediction set at level $1 - \alpha$. Although this method does not guarantee validity due to the potential misspecification of the underlying assumptions, it is often considered robust in practice. The miscoverage level α is set to 0.05.

As reported in Table 6.1, the union method yields excessively large sets, whereas the ensemble provides better results. Overall, COMA with adaptive stepsize outperforms the competing methods while maintaining comparable coverage. In general, COMA exhibits strong empirical performance, while still ensuring adequate empirical coverage. In particular, the adaptive–η version is able to perform model selection in scenarios where one model clearly outperforms the others. For the COMA method, the observed output is always a non-empty interval, and the maximum value of the empirical elementwise correlation defined in (6.3) is 0.002.

6.4.2 Classification in iid setting

For the classification experiments, we generate a synthetic dataset with 2000 observations as follows. We consider $p = 8$ covariates sampled from a multivariate Gaussian distribution with mean zero and equicorrelation structure, with $\sigma_{ij} = 0.5$ for $i \neq j$. The response $y \in \{1, \dots, 8\}$ is then drawn from a multinomial distribution with probabilities proportional to $w_j(x) \propto \exp\{x^\top \beta_j\}, j = 1, \dots, 8$, where each coefficient vector $\beta_j \in \mathbb{R}^p$ is independently sampled from a standard normal distribution. Specifically, following a burn-in phase of 1000 iterations, we re-trained four different classifiers at each step, and the weights for these algorithms are adjusted using the COMA method. In particular, the four classifier chosen from the R package `mlr` are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest and Neural Net (we refer to Azzalini and Scarpa (2012) for an introduction to the methods). At each time

Method	NN	RF	QDA	LDA	COMA (Adap. η)	COMA ($\eta = 0.1$)	Union	Bonf.	Ensemble
Size	3.056	3.106	2.897	2.488	2.489	2.476	4.159	2.647	2.619
Coverage	0.902	0.903	0.889	0.899	0.896	0.890	0.965	0.897	0.897

TABLE 6.2: Comparison of empirical size and coverage at significance level $\alpha = 0.1$ between the underlying algorithms and COMA (with both adaptive and fixed η). For reference, we also include the union, the intersection, and the sets obtained via conformal prediction using an ensemble of the base algorithms as the predictive model.

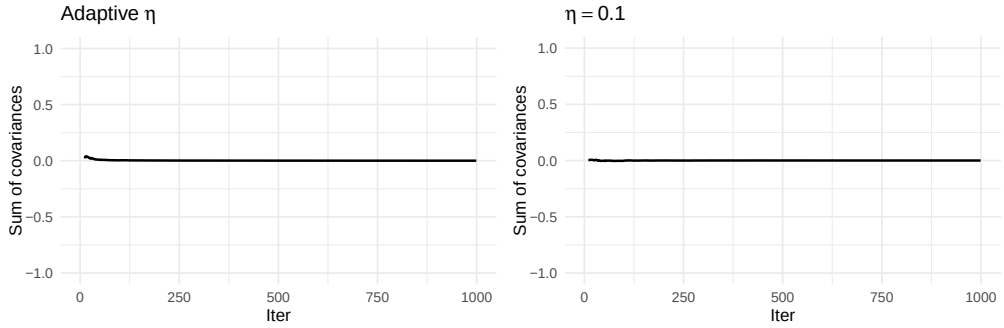

FIGURE 6.3: Empirical sum of the covariances between $w_k^{(t)}$ and $\phi_k^{(t)}$ over time. Both series remain close to zero across all iterations.

step, split conformal inference is used to obtain a prediction set for $y^{(t)}$ and the score used are $s(x, y) = 1 - \hat{\mu}(x; y_d)$ where $\hat{\mu}(x; y_d)$ is the predicted probability for the label $y_d \in \mathcal{Y} = \{1, \ldots, 8\}$.

We apply our COMA procedure with both adaptive and fixed η (set to 0.1). As a comparison, we also include, as before, the union of the label sets and the intersection obtained with each method trained at level $2\alpha/K$. In addition, we apply conformal prediction using an ensemble of base models. From Table 6.2, our methods achieve the smallest set sizes while maintaining empirical coverage close to $1 - \alpha$. The factor of 2 does not appear to be present in the miscoverage, as the miscoverage rate remains close to α. The same behavior is observed when applying the intersection of the sets (Bonferroni). This can be explained by the fact that Bonferroni's inequality is tight near-independence, whereas in this case there is strong dependence, since all methods are trained on the same data. Moreover, the empirical counterpart of the quantity defined in (6.3) remains close to zero throughout all iterations, as illustrated in Figure 6.3. Figure 6.4 shows that the weights quickly concentrate on LDA, which is the best-performing model.

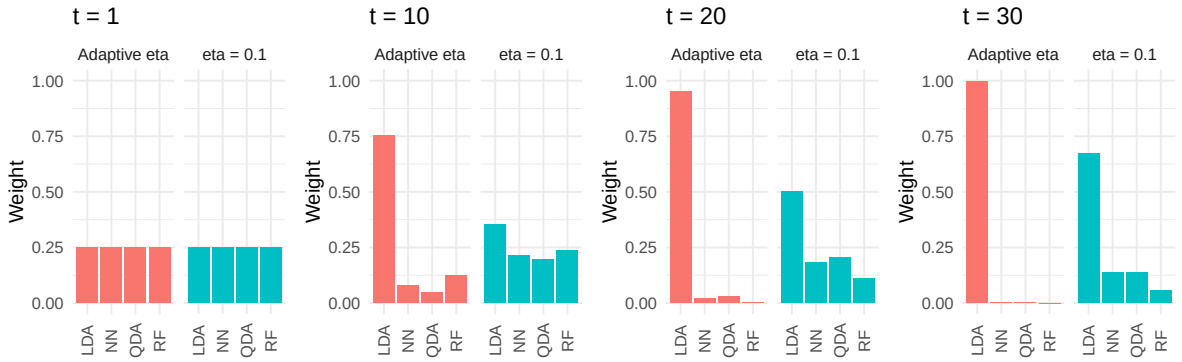

FIGURE 6.4: Weights assumed by the classification algorithms during different iterations. After few iteration the strategy with adaptive η puts unit mass in the LDA, which is the best method.

6.5 COMA under distribution shift

Up to this point, we have focused on the iid setting and observed that COMA performs effectively as a wrapper in such scenario, often concentrating the weights on the best model(s). However, what occurs if distributional shifts arise and different models exhibit varying levels of performance at different points in time? First of all, as discussed in Chapter 4, conformal prediction provides a general framework to transform the output of any predictive algorithm into a prediction set; however, its validity is based on the assumption of independence and identical distribution (or exchangeability). This assumption is violated when the underlying data distribution shifts over time and standard conformal inference can no longer be expected to yield reliable guarantees. To address this challenge, several methods have been proposed in recent years, including Gibbs and Candès (2021) and Angelopoulos *et al.* (2024a). We now briefly outline their approaches before extending our COMA procedure to this setting. In particular, we will see that COMA can be used in combination with these approaches, serving as a flexible tool in the considered scenario. For clarity of exposition, we first introduce the methods in the case of a single algorithm, deferring the multi-expert extensions to Section 6.5.1 and Section 6.5.3.

In Gibbs and Candès (2021), *adaptive conformal inference* (ACI) is introduced as a novel method for constructing prediction sets in an online manner, providing robustness to changes in the marginal distribution of the data. Their work is based on the update of the error level α to achieve the desired level of confidence. Indeed, given the possible non-stationary nature of the data-generating distribution, conventional results do not guarantee $1-\alpha$ coverage. However, at each time t, an alternative value $\alpha^{(t)}$ might exist, allowing the attainment of the desired coverage.

The procedure described in Gibbs and Candès (2021) recursively updates the error rate α as follows:

$$\alpha^{(1)} = \alpha, \tag{6.12}$$

$$\alpha^{(t)} = \alpha^{(t-1)} + \gamma(\alpha - \phi^{(t-1)}), \quad t \geq 2, \tag{6.13}$$

where $\phi^{(t)} = \mathbb{1}\{y^{(t)} \notin \mathcal{C}^{(t)}(\alpha^{(t)})\}$ is the sequence of miscoverage events setting the miscoverage rate to $\alpha^{(t)}$, while $\gamma > 0$ represents a step size parameter. Possible improvements are proposed in Gibbs and Candès (2024) and Zaffran *et al.* (2022); in particular, their proposals tune the parameter γ over time using an online learning procedure. A drawback arises from the fact that in certain situations, ACI can return either $\mathcal{Y}$ or $\varnothing$ as a set, attributed to the circumstance that in some iterations, the parameter $\alpha^{(t)}$ can be smaller than 0 or greater than 1.

Another method is presented in Angelopoulos *et al.* (2024a), where instead of updating the error level α, there is an update of the quantile of the distribution of the *scores*. In this setting, we define $s^{(t)} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as the *conformal score* function, which measures the accuracy of the prediction at time t. If different models are involved one can use different scores (e.g. residual or density based scores). The procedure, starting from an initial threshold $q^{(1)} = q$, updates the quantiles as follows:

$$q^{(t)} = q^{(t-1)} + \gamma(\phi^{(t-1)} - \alpha), \quad t \geq 2,$$

where, in this case, $\phi^{(t)} = \mathbb{1}\{y^{(t)} \notin \mathcal{C}^{(t)}(q^{(t)})\}$ and the sets at each time are simply defined by

$$\mathcal{C}^{(t)}(q^{(t)}) = \left\{y \in \mathcal{Y} : \, s^{(t)}(x^{(t)}, y) \leq q^{(t)}\right\}. \tag{6.14}$$

As noted in Angelopoulos *et al.* (2024a), both ACI and quantile tracking, can be rephrased to an online descent algorithm with respect to the pinball loss (Koenker, 2005). In addition, both methods achieve a *long-run coverage* in time, defined as

$$\frac{1}{T}\sum_{t=1}^{T} \phi^{(t)} = \alpha + o(1), \tag{6.15}$$

under few or no assumptions on $\{z^{(t)}\}_{t=1}^{T}$ and $\{s^{(t)}\}_{t=1}^{T}$, where $o(1)$ is a quantity that tends to zero as $T \to \infty$. In particular, (6.15) holds deterministically and is different from the probabilistic assumption in (3.1) valid in an iid setting. Improvements are also suggested for the quantile tracking method, such as an adaptive learning rate or the introduction of an *error integrator*. The proposal in Angelopoulos *et al.* (2024a) is to

set a decaying step size $\gamma^{(t)}$. They show that if scores are bounded and $\gamma^{(t)} \propto t^{-1/2-\epsilon}$ for some $\epsilon \in (0, 1/2)$, then the long-run coverage is bounded as $O(\frac{1}{T^{1/2-\epsilon}})$ in the adversarial setting, while in the iid case the algorithm achieves a valid coverage guarantee.

In light of the dynamics shift observed in the marginal distribution of the data, it is conceivable that the weights assigned to different algorithms may also undergo variations throughout iterations, since different models may be preferable at different time points. This feature makes our method particularly appealing. We now describe two different ways to combine our dynamic ensembling algorithm with ACI: either as a wrapper that does not alter the inner workings of ACI (Subsection 6.5.1), or by altering the feedback provided to ACI itself (Subsection 6.5.3).

6.5.1 Decentralized COMA under distribution shift

In the scenario described above, suppose that each agent operates independently by providing an interval at each iteration and, subsequently, the *aggregator* merges these intervals according to the vector $w^{(t)}$. This framework can be referred to as a decentralized dynamic merging. In this setting, the advantage of the quantile tracking method is that it never returns sets of infinite measure. This implies that the loss function used can remain the size of the different sets, ensuring that the bound described in (6.10) is valid. In the case of using the described ACI approach, it becomes necessary to employ a function $g(\cdot)$ that is capable of returning a finite number when $\alpha_k^{(t)} \leq 0$. Before proceeding with the rest of the section, we introduce a proposition which holds under the following assumptions:

1. Negative total elementwise empirical correlation between $\{\phi^{(t)}\}_{t=1}^T$ and $\{w^{(t)}\}_{t=1}^T$:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}(\phi_k^{(t)} - \bar{\phi}_k)(w_k^{(t)} - \bar{w}_k) \leq 0, \tag{6.16}$$

   where $\bar{\phi}_k := \frac{1}{T}\sum_{t=1}^T \phi_k^{(t)}$ and $\bar{w}_k := \frac{1}{T}\sum_{t=1}^T w_k^{(t)}$.

2. No randomization: the value $u^{(t)}$ in (6.2) equals 0:

$$u^{(t)} = 0, \quad t = 1, \ldots, T. \tag{6.17}$$

Proposition 6.8. *Let $\{z^{(t)} = (x^{(t)}, y^{(t)})\}_{t=1}^T$ be an arbitrary sequence of data points, $\{w^{(t)} = (w_1^{(t)}, \ldots, w_k^{(t)})\}_{t=1}^T$ be a sequence of weight vectors, and $\{\phi_k^{(t)}\}_{t=1}^T$ be the sequence of miscoverage events for the k-th algorithm, $k \in [K]$. Under assumptions* (6.16) *and*

(6.17), *for all* $\alpha \in (0, 1/2]$, *decentralized dynamic merging in conjunction with adaptive conformal inference* (6.13) *satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{y^{(t)} \notin \mathcal{C}_M^{(t)}\} \leq 2\alpha + 2\frac{1-\alpha+\gamma}{T\gamma}. \tag{6.18}$$

Proof. We remark that, if $u^{(t)} = 0$, then $y^{(t)}$ is not contained in $\mathcal{C}^{(t)}$ if and only if $\sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\{y^{(t)} \notin \mathcal{C}_k^{(t)}(\alpha_k^{(t)})\} \geq 1/2$. Then, since $\phi_k^{(t)} = \mathbb{1}\{y^{(t)} \notin \mathcal{C}_k^{(t)}(\alpha_k^{(t)})\}$, we can write

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}\{y^{(t)} \notin \mathcal{C}_M^{(t)}\} &= \sum_{t=1}^{T} \mathbb{1}\left\{\sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\{y^{(t)} \notin \mathcal{C}_k^{(t)}(\alpha_k^{(t)})\} \geq \frac{1}{2}\right\} \\
&\overset{(i)}{\leq} \sum_{t=1}^{T} 2\sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\{y^{(t)} \notin \mathcal{C}_k^{(t)}(\alpha_k^{(t)})\}, \\
&= 2\sum_{k=1}^{K}\sum_{t=1}^{T} w_k^{(t)} \phi_k^{(t)} \\
&\overset{(ii)}{\leq} 2\sum_{k=1}^{K} \frac{1}{T}\left(\sum_{t=1}^{T} w_k^{(t)}\right)\left(\sum_{t=1}^{T} \phi_k^{(t)}\right) \\
&\overset{(iii)}{\leq} 2\sum_{k=1}^{K} \frac{1}{T}\left(\sum_{t=1}^{T} w_k^{(t)}\right)\left(T\alpha + \frac{1-\alpha+\gamma}{\gamma}\right) \\
&= 2\left(\alpha + \frac{1-\alpha+\gamma}{T\gamma}\right)\sum_{k=1}^{K}\sum_{t=1}^{T} w_k^{(t)} \\
&= 2T\alpha + 2\frac{1-\alpha+\gamma}{\gamma},
\end{aligned}$$

where (i) holds due to the fact that $\mathbb{1}\{x \geq 1\} \leq x$, if $x \geq 0$ and (ii) is due to assumption (6.16), in fact if the assumption holds, then $T\sum_{k=1}^{K}\sum_{t=1}^{T} w_k^{(t)}\phi_k^{(t)} - \sum_{k=1}^{K}(\sum_{t=1}^{T}\phi_k^{(t)})(\sum_{t=1}^{T} w_k^{(t)}) \leq 0$. Proposition 4.1 in Gibbs and Candès (2021) along with $\alpha \in (0, 1/2]$ implies (iii). Dividing both sides by T gives the result. □

Similar results can be obtained using the quantile tracking method under the assumption of bounded scores (see Proposition 1 in Angelopoulos *et al.* (2024a) and Theorem 1 in Angelopoulos *et al.* (2024a)). Randomization is not included because the above is a deterministic result; however, the empirical results using randomization appear to be reasonably good also in this setting. The result (6.18) guarantees long-run coverage at least of level $1-2\alpha$ for any distribution, provided specific conditions about the observed weights and miscoverage errors are met. In our case the weights will be obtained using the method described in Section 6.3.

The algorithm is presented in Algorithm 3, and it requires an additional step to update the quantiles after observing the true response value at each iteration. This update can be performed independently by each expert in a decentralized setting, where the experts work separately and the aggregator only receives the K sets at each iteration. The algorithm can be modified using the miscoverage rate α instead of the quantiles.

Data: Initial learning rate $\eta^{(0)} \geq 0$ and starting quantiles $q_1, \ldots, q_K$
$w_k^{(1)} \leftarrow 1/K, L_k^{(0)} \leftarrow 0$, $k = [K]$;
$q_k^{(1)} = q_k$, $k = [K]$; **for** *rounds* $t = 1, \ldots, T$ **do**
  Require $\mathcal{C}_k^{(t)}(q_k^{(t)})$, $k \in [K]$; $\mathcal{C}_M^{(t)} \leftarrow \left\{ y \in \mathcal{Y} : \sum_{k=1}^K w_k^{(t)} \mathbb{1}\{y \in \mathcal{C}_k^{(t)}\} > \frac{1}{2} \right\}$;
  Receive loss $\ell_k^{(t)}$, update $L_k^{(t)} := L_k^{(t-1)} + \ell_k^{(t)}$;
  Update learning rate $\eta^{(t)}$;
  $w_k^{(t+1)} \leftarrow \exp\{-\eta^{(t)} L_k^{(t)}\} / \sum_{j=1}^K \exp\{-\eta^{(t)} L_j^{(t)}\}$;
  Observe $y^{(t)}$ and miscoverage errors for every experts $\phi_k^{(t)}, k \in [K]$;
  Update quantiles $q_k^{(t+1)}, k \in [K]$;
end
Algorithm 3: Decentralized Exponentially Weighted Majority Vote under distribution shift

6.5.2 Experiments in non-iid setting

We now evaluate our COMA procedure as a wrapper for quantile tracking and ACI through some experiments on real datasets.

6.5.2.1 ELEC2 dataset

We now use the methods described in Section 6.3 and in Section 6.5 in a real-world application. We define $K = 3$ different regression algorithms, and we employ quantile tracking with a decaying step size $\gamma^{(t)} \propto t^{-1/2-\epsilon}$. The score function corresponds to absolute residuals. In particular, we use the `ELEC2` (Harries, 2003) data set that monitors electricity consumption and pricing in the states of New South Wales and Victoria in Australia, with data recorded every 30 minutes over a 2.5-year period from 1996 to 1999. For our experiment, we utilize four covariates: `nswprice` and `vicprice`, representing the electricity prices in each respective state, and `nswdemand` and `vicdemand`, denoting the usage demand in each state. The response variable is `transfer`, indicating the quantity of electricity transferred between the two states. We narrow our focus to a subset of the data, retaining only observations within the time range of 9 am to 12 pm to mitigate daily fluctuation effects. The models used correspond to three distinct linear models based on different covariates. The first model relies on the entire set of covariates at

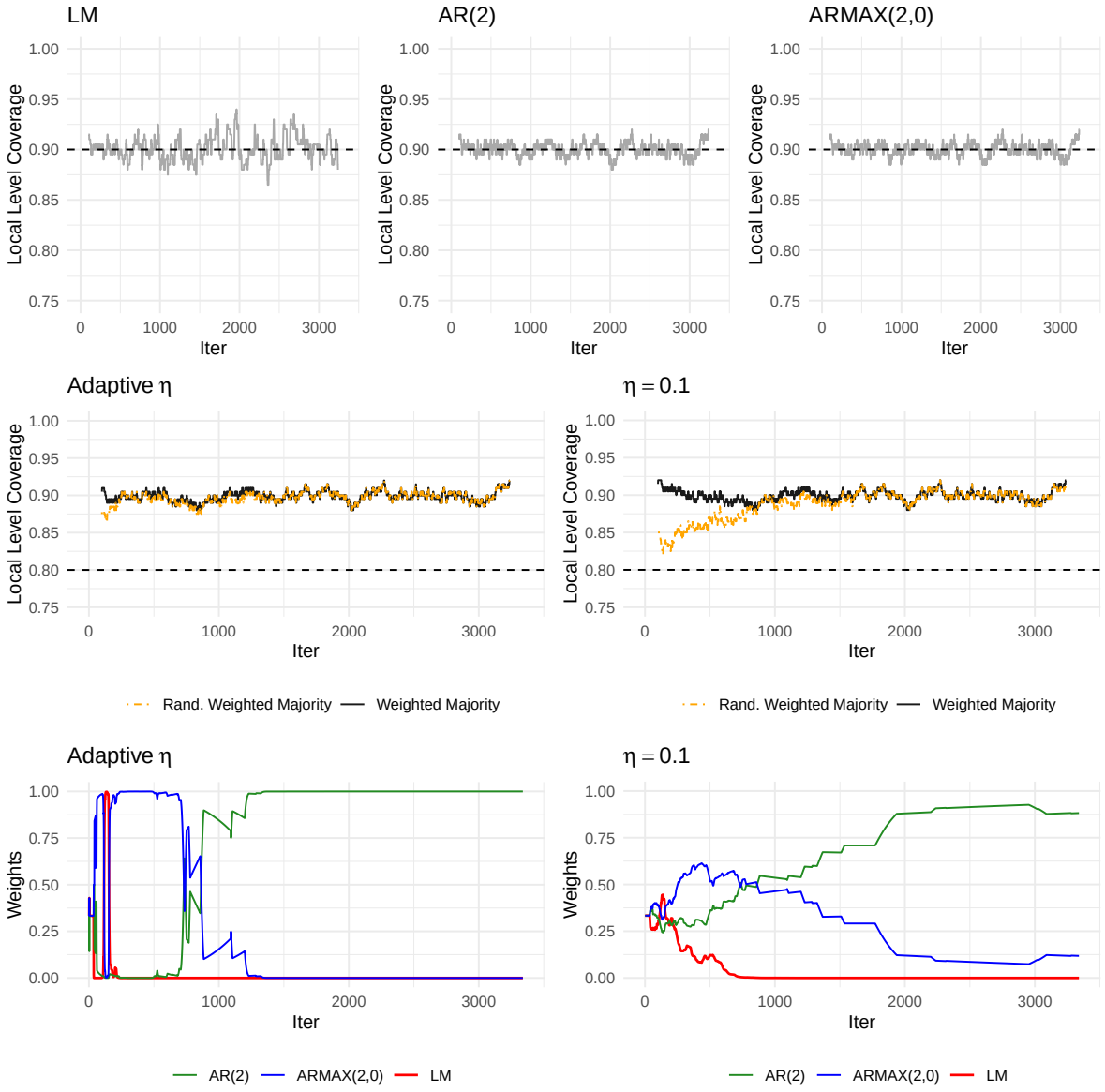

FIGURE 6.5: First row: Local level coverage for the quantile tracking using the 3 regression models described in Section 6.5.1. The dashed line represents the level $1-\alpha$. Second row: Local level coverage for the COMA method with adaptive η or $\eta = 0.1$. The dashed line represents the level $1-2\alpha$. Third row: Weights obtained by the various model for the decentralized COMA algorithm applied in `ELEC2` dataset. The strategy with adaptive $\eta^{(t)}$ is more aggressive than the one with $\eta^{(t)} = \eta$ in assigning weights and it performs model selection in most of the iterations.

time t, the second model employs $y^{(t-1)}$ and $y^{(t-2)}$ as covariates (it corresponds to an AR(2)), while the third model supplements the latter with the covariate `nswprice`.

The performance of the methods is compared using the local level coverage on 200 data points:

$$\text{localCov}^{(t)} := \frac{1}{200} \sum_{i=t-100+1}^{t+100} \mathbb{1}\left\{ y^{(i)} \in \mathcal{C}_k^{(i)}\left(q_k^{(i)}\right) \right\}, \tag{6.19}$$

	LM	AR	ARMAX	Adapt. η	Adapt. η + Rand.	$\eta = 0.01$	$\eta = 0.01$ + Rand.	$\eta = 0$
$L^{(T)}$	1288.76	787.04	807.16	788.20	780.34	789.51	745.53	808.24
Coverage	0.900	0.902	0.902	0.900	0.897	0.901	0.889	0.905

TABLE 6.3: Cumulative loss ($L^{(T)}$) and empirical coverage obtained by the various methods. The losses of the proposed methods are better (or similar) than the cumulative loss obtained by the best algorithm. The column $\eta = 0$ represents a baseline where weights are not updated.

where $\mathcal{C}_k^{(t)}(q_k^{(t)})$ is the set obtained by the k-th regression algorithms with the quantile fixed at $q_k^{(t)}$. The local coverage is also used to measure the performance of our merging procedure. In particular, we compare the use of an adaptive learning parameter and a small fixed $\eta = 0.1$. As can be seen in Figure 6.5, the COMA procedure with adaptive $\eta^{(t)}$ obtains a coverage that is similar to the prespecified level $1 - \alpha$. This is because the weights concentrate on one model over time; in particular, the chosen model is not always the same, as can be seen in the last row in Figure 6.5. The procedure proves appealing in this setting: when a clear winner emerges, it concentrates the weights on the best model while, in the absence of dominance, it allocates them across the most appropriate models while maintaining valid coverage.

In Table 6.3, we compare the cumulative loss and the empirical coverage obtained by the various methods proposed with and without the use of randomization. In addition, the case $\eta = 0$ corresponds to the majority vote with uniform weights. As can be seen, the cumulative loss of our proposed methods (with fixed and adaptive η) is better than the standard majority vote approach in all cases. The best result in terms of cumulative loss, in this case, is obtained by fixing $\eta = 0.01$ and adding randomization. This is due to the fact that during some iterations the two best models have a similar weight and their sum is close to one. In this situation, the majority vote reduces to the intersection of the two sets most of the time. As an example, suppose that $w^{(t)} \approx (0.6, 0.4, 0)$ and $u^{(t)} \approx 1/2$, then the threshold is $3/4$ and in this case the majority vote set must contain the points of the first two sets (the intersection of $\mathcal{C}_1^{(t)}$ and $\mathcal{C}_2^{(t)}$). Indeed, during the first iterations, the local level coverage for COMA with fixed η and randomization is observed to be closer to $1-2\alpha$. The cumulative loss of the methods using randomization is smaller than the cumulative loss incurred by the best method, which in this case is AR(2). In particular, the factor of 2 in (6.9) does not appear. The empirical coverage is approximately 0.9 in all scenarios and the maximum correlation between the weight and error vectors is 0.02. The proportion of empty sets produced by both methods is below 1%.

6.5.2.2 Google stock prices

Inspired by Angelopoulos *et al.* (2024a), we analyze the historical series comprising the single stock open prices of Google from January 1, 2006, to December 31, 2014 (Nguyen, 2018). The compared models consist of six different autoregressive models incorporating temporal lags from 1 to 6, trained on the logarithmic scale of the opening price, while the forecasts are obtained on the original scale. The series exhibits nonstationary behavior, as can be observed in the last row in Figure 6.6. The method used to obtain the sets is the quantile tracking method described in Section 6.5 with $\alpha = 0.1$, a similar analysis is performed in Angelopoulos *et al.* (2024a) with an autoregressive model with 3 lags. The parameter γ is updated as $\gamma^{(t)} = 0.1B^{(t)}$, where $B^{(t)}$ is the maximum observed score in a trailing window of 100 iterations, and the models are re-trained at every time step. The score function corresponds to absolute residuals $s(x^{(t)}, y^{(t)}) = |y^{(t)} - \hat{\mu}_k^{(t)}(x^{(t)})|$, $k \in [K]$, and prediction sets are obtained as in (6.14). The weights are updated using an adaptive η or fixing η to the level 0.01, this small value of η allows the mixing of the different intervals during iterations.

It appears that autoregressive models with two and four lags perform better than the others. Particularly, in the case of adaptive η, after the first iterations, they alternate during different times (Figure 6.6). As expected, the strategy with a small fixed η is more conservative in assigning weights and continues to assemble the different models during all iterations. In both methods, with adaptive η and fixed η, local coverage remains around the level of $1 - \alpha$. The quantity $\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}(\phi_k^{(t)} - \bar{\phi}_k)(w_k^{(t)} - \bar{w}_k)$ is around zero during all iterations.

To summarize the results, it seems that our proposed method is able to adapt to the changes of the distributions observed in the data. The total negative empirical correlation assumption seems to be reasonable in both cases.

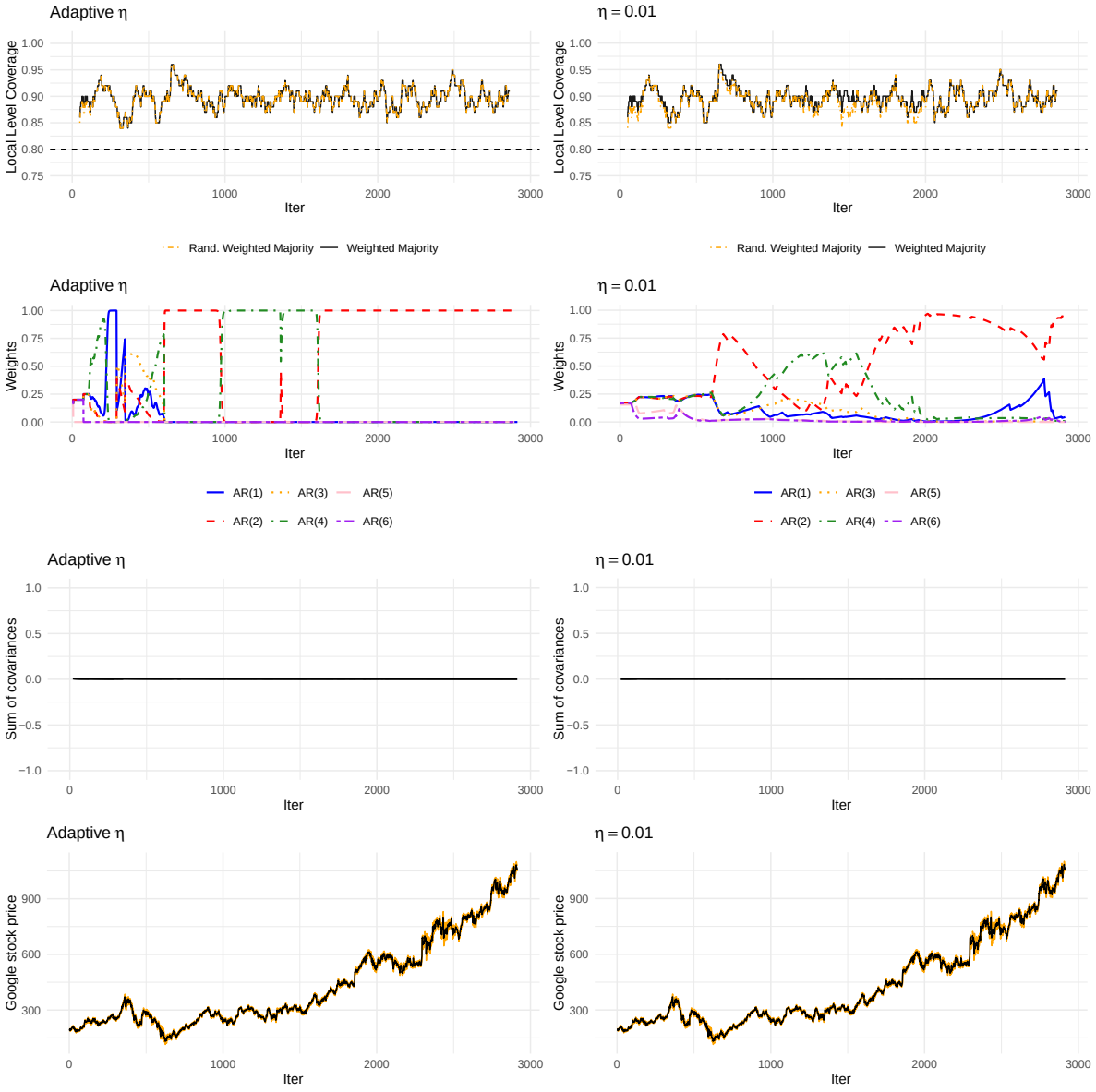

FIGURE 6.6: Local level coverage obtained in a window containing 100 data points (first row), weights assumed during the different iterations by the various models (second row), empirical sum of the quantity $\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}(\phi_k^{(t)}-\bar{\phi}_k)(w_k^{(t)}-\bar{w}_k)$ obtained using an increasing number of observations (third row) and series of the original stock prices with corresponding prediction intervals (fourth row). The black dashed lines in the first row represent the level $1-2\alpha$. The results corresponding to the strategy with adaptive η are presented in the first column, while the results corresponding to the strategy with $\eta = 0.01$ are presented in the second column.

6.5.2.3 Additional simulations on quantile tracking

We investigate the quantile tracking method in conjunction with the COMA with adaptive η, we conduct a case study in which changes in marginal distribution also impact the weights of the two algorithms. In this case, we use simulated data and we

let the distribution vary during the iterations as described below. We suppose to have two covariates, x_1 and x_2, and the data are generated in blocks, with only one of the two covariates influencing the response at a time. Specifically, at each time t we have $y^{(t)} = 2x^{(t)} + \varepsilon^{(t)}$ where

$$x^{(t)} = \begin{cases} x_1^{(t)}, & t \in \{[0,50) \cup [100,150) \cup [250,350) \cup [450,550) \cup \cdots \cup [1250,1350)\} \\ x_2^{(t)}, & t \in \{[50,100) \cup [150,250) \cup [350,450) \cup \cdots \cup [1350,1450]\} \end{cases}$$

and $x_1^{(t)}, x_2^{(t)}, \varepsilon^{(t)}$ are generated independently from a standard normal distribution. Two standard linear models are used, one using x_1 as a regressor and the other using x_2. Prediction intervals are generated using the quantile tracking method with the absolute residual as the score function. The models are re-trained at every time step using the most recent 100 observations. The parameter γ is set to 1 and the entire procedure is repeated 1000 times.

The weight assignment to each method is dynamically adjusted based on the performance of the two regression algorithms. In the case where only two models are utilized, the majority vote (without randomization) is reduced to selecting the model with the higher weight (it performs model selection by definition). From Figure 6.7, the COMA method alternates the weights based on changes in the marginal distribution. In other words, the method can adapt to the changes and perform model selection. In particular, within each time window, the selected model switches between the two models around the midpoint of the window. The less aggressive behavior in assigning weights is due to the fact that $\eta^{(t)}$ decreases over time. The empirical miscoverage over the 1000 replications of our COMA procedure is 0.113; while the miscoverage rates for the two linear models are respectively 0.101 and 0.102. The maximum level of empirical correlation between the weights and the errors across replications is 0.167.

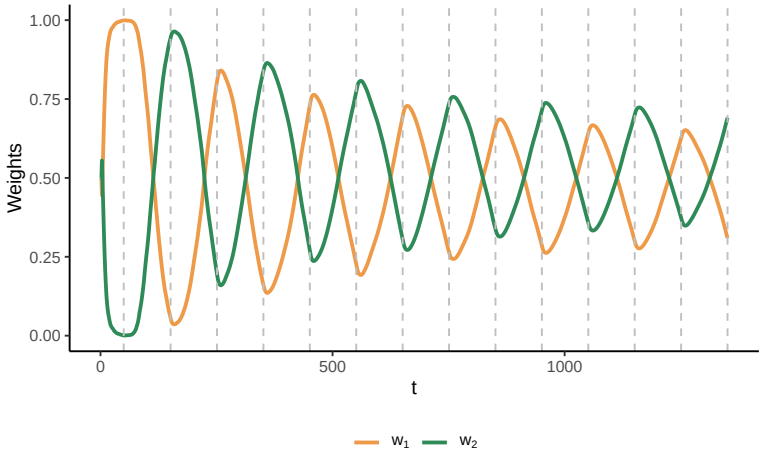

FIGURE 6.7: Average weights across 1000 iterations of the linear model, using either x_1 or x_2 as the regressor. The dotted lines indicate the time points at which the distribution changes.

6.5.2.4 Classification in non-iid setting

We conclude the experimental section with a classification example conducted under a non-iid setting. The dataset consists of hourly air quality measurements collected in the city of Chanping. The response variable corresponds to the levels of PM10 concentration, discretized into five classes of equal cardinality, while the covariates comprise meteorological indicators and the concentrations of other atmospheric pollutants. The models used include a Random Forest with 150 trees, a neural network with 4 nodes, a neural network with 6 nodes, and Linear Discriminant Analysis (LDA). For this experiment, we employed the ACI method with the parameter γ set to 0.01. Since we are dealing with a classification problem, the loss is bounded and therefore cannot take infinite values (also if $\alpha^{(t)} < 0$). The significance level α is set to 0.1.

Also in this case, where the loss takes only integer values, the sets obtained via our COMA procedure attain a loss that is comparable to, and in some instances even better than, that of the best model. Moreover, the empirical coverage remains close to the nominal level $1 - \alpha$ (see Table 6.4). The maximum observed value of the sum of empirical covariances is 0.012 for COMA with adaptive η, and 0.006 for COMA with fixed η. These results indicate that the assumption of total negative correlation seems reasonable in this case.

	RF	NN(4)	NN(6)	LDA	Adapt. η	Adapt. η + Rand.	$\eta = 0.01$	$\eta = 0.01$ + Rand.	$\eta = 0$
$L^{(T)}$	1962	2355	2267	2414	1964	1959	1967	1945	2016
Coverage	0.895	0.894	0.895	0.894	0.896	0.895	0.896	0.890	0.891

TABLE 6.4: Cumulative loss ($L^{(T)}$) and empirical coverage obtained by the various methods. The column $\eta = 0$ represents a baseline where weights are not updated.

6.5.3 Adaptive conformal inference directly applied on dynamic merging

The approach described in the preceding section had agents updating the quantiles of their score function based on *their own* past errors. This makes sense when all agents can observe the ground truth and calculate their own errors, and they do not care about the aggregator's goals. In other words, the dynamic merging algorithm was just an outer wrapper that did not interfere with the inner functioning of the K adaptive conformal inference algorithms.

Below, we show that our dynamic ensembling method can provide direct feedback to the adaptive conformal inference framework, and this can have some advantages in settings where the end goal is good aggregator performance only, and individual agents do no have their own goals (of maintaining coverage, say). First, we observe that if the absolute residuals are employed as the score function and if the quantile ($q^{(t)}$) is common across all algorithms, then the size of the sets at time t is equal for all intervals. This implies that the weights are not updated during the various iterations. In this case, we propose using a value $\alpha^{(t)}$ that is common for all K agents, and it is updated by the aggregator based on the performance of the set obtained by the exponential weighted majority vote. In other words, the aggregator, at each iteration, tries to learn the coverage level required to obtain a $(1-\alpha)$-prediction set. The procedure updates the α level according to the miscoverage event $\phi^{(t)} = \mathbb{1}\{y^{(t)} \notin \mathcal{C}_M^{(t)}(\alpha^{(t)})\}$, where

$$\mathcal{C}_M^{(t)}(\alpha^{(t)}) := \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\left\{ y \in \mathcal{C}_k^{(t)}\left(\alpha^{(t)}\right)\right\} > 1/2 \right\},$$

and $w_k^{(t)}$ are the weights learned by the COMA procedure. The error level (common among agents) is updated as in (6.12) and (6.13). It is important to note that the interval lengths can be different between the various agents even if $\alpha^{(t)}$ is common. In such a scenario, ACI operates directly on the set produced by the aggregator, and so we directly recover the properties of the ACI method. This is due to the fact that the sets are nested with respect to the parameter $\alpha^{(t)}$ and the set is $\mathcal{Y}$ if $\alpha^{(t)} \leq 0$ and $\varnothing$ if

$\alpha^{(t)} \geq 1$ (Szabadváry and Löfström, 2026). Further, there is no need for randomization to enhance the length of the intervals. This scenario is plausible when the different agents (models) can collaborate with each other, and the aggregator is able to update both the weights and the miscoverage rate.

The procedure is outlined in Algorithm 4, where the α-level is shared among the K models and updated by the aggregator itself, based on its own miscoverage events.

Data: Initial learning rate $\eta^{(0)} \geq 0$ and miscoverage rate α
$w_k^{(1)} \leftarrow 1/K, L_k^{(0)} \leftarrow 0$, $k = [K]$;
$\alpha^{(1)} = \alpha$; **for** *rounds* $t = 1, \ldots, T$ **do**
  Require $\mathcal{C}_k^{(t)}(\alpha^{(t)})$, $k \in [K]$; $\mathcal{C}_M^{(t)} \leftarrow \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\{y \in \mathcal{C}_k^{(t)}\} > \frac{1}{2} \right\}$;
  Receive loss $\ell_k^{(t)}$, update $L_k^{(t)} := L_k^{(t-1)} + \ell_k^{(t)}$;
  Update learning rate $\eta^{(t)}$;
  $w_k^{(t+1)} \leftarrow \exp\{-\eta^{(t)} L_k^{(t)}\} / \sum_{j=1}^{K} \exp\{-\eta^{(t)} L_j^{(t)}\}$;
  Observe $y^{(t)}$ and miscoverage event $\phi^{(t)} = \mathbb{1}\{y^{(t)} \notin \mathcal{C}_M^{(t)}\}$;
  Update miscoverage rate $\alpha^{(t+1)}$;
end

Algorithm 4: Centralized Exponentially Weighted Majority Vote under distribution shift

6.5.4 Real data example: ELEC2 dataset

We applied the procedure to the ELEC2 dataset, fixing the target coverage level at 0.9 and adopting the loss function $\arctan(\ell^{(t)})$. The use of a bounded loss function is necessary, since ACI can output $\mathbb{R}$ as an interval. We consider both the adaptive η strategy and a fixed $\eta = 0.1$, using the three regression algorithms described in Section 6.5.2.1. In addition, we apply ACI with a regression model obtained by assembling the three algorithms by a simple average. The learning rate of the ACI algorithm is set to 0.05 for all methods, and we use (6.19) to compare their coverage performance.

The results, reported in Figure 6.8, show that the local coverage oscillates around the target level $1 - \alpha$ for all three methods. The methods yield comparable results in terms of empirical size and in the proportion of cases where the sets coincide with $\mathbb{R}$ (i.e., when $\alpha \leq 0$). However, our strategy with adaptive η achieves the best performance in terms of set size, whereas the strategy with fixed η has the smallest proportion of cases where the sets coincide with $\mathbb{R}$.

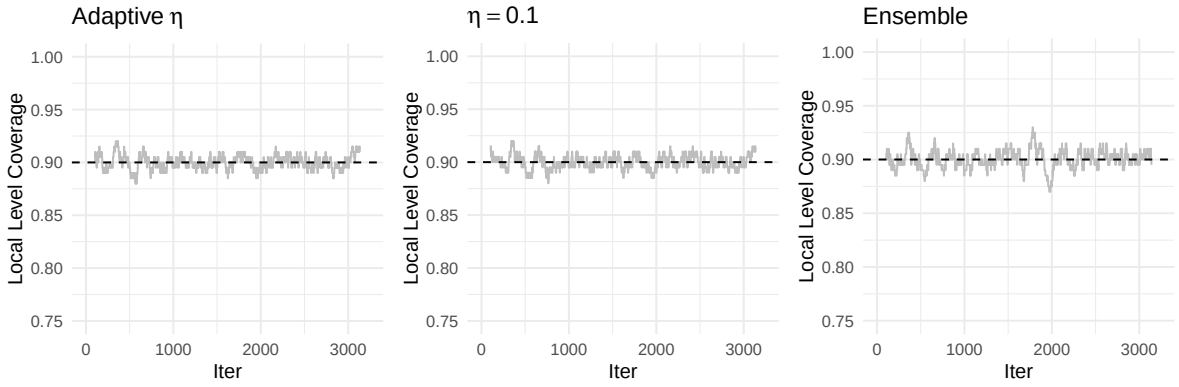

FIGURE 6.8: Local level coverage for *direct* ACI for COMA with adaptive η and $\eta = 0.1$ applied to `ELEC2` dataset (Section 6.5.3). The third column represents ACI directly applied to an ensemble of the three base models. The dashed lines represent the target level $1 - \alpha$.

	Adaptive η	$\eta = 0.1$	Ensemble
Empirical size	0.230	0.234	0.238
% set= $\mathcal{Y}$	0.009	0.006	0.016

TABLE 6.5: Empirical size and percentage of times the sets equal $\mathcal{Y}$. The strategy with adaptive η achieves the best performance in terms of empirical size, while the proportion of cases where the methods produce sets of infinite size is negligible.

6.6 Discussion

The chapter presents a method for aggregating conformal prediction sets derived from different prediction algorithms. Specifically, a method based on majority vote is proposed in which the weight of each individual prediction model is updated based on its past performance. The proposed method requires access only to the prediction sets, which can be particularly advantageous in decentralized settings and is computationally efficient. The approach can be seen as a principled wrapper around conformal prediction, specifically designed for streaming and online settings.

In some cases, when there is an algorithm that outperforms the others, all the mass is assigned to the better-performing model, effectively resulting in model selection. In other cases it moves between models, depending on which is. Particularly, in the case of iid data, it is very common for there to exist a *best* model that outperforms the others in terms of size of the outputted sets.

It is feasible to establish an upper bound for the loss incurred by our method relative to the loss of the best-performing algorithm, translating into a form of regret for our method. We also tested our method in cases where the data underwent a distribution shift, on top of the quantile tracking and ACI methods. Even when the assumptions

may not be fully satisfied, empirical results appear to be promising.

One limitation is that in practice we often see $1 - \alpha$ coverage, even when the current theory only predicts it to be at least $1 - 2\alpha$. Another is that we assumed that we get to observe the true label in every round; extending it to intermittent or delayed feedback would be interesting future work.

Conclusions

Summary

Statistical inference provides the theoretical framework for drawing valid conclusions about the phenomenon of interest based on the data. This manuscript addresses the problem of deriving valid inference when combining dependent sources of evidence, such as varying p-values for the same hypothesis or different uncertainty sets for the same parameter of interest.

Chapter 2 focuses on the case where p-values are exchangeable, a situation that can naturally arise when p-values are obtained through sample-splitting methods. In this setting, new and more powerful combination strategies are developed, improving existing approaches. In addition, a simple randomization technique is shown to enhance procedures that are valid under arbitrary dependence.

Chapter 3 studies the problem of combining dependent uncertainty sets into a single set with desirable properties in terms of coverage and size. A majority-vote approach is examined, and several extensions are introduced to yield smaller sets without compromising coverage guarantees. The main results of this first part rely on e-values (Ramdas and Wang, 2025) and extensions of Markov's inequality (Ramdas and Manole, 2024).

The second part of the thesis applies the results in the framework of conformal prediction, a general methodology for constructing distribution-free prediction sets. In particular, Chapter 5 is devoted to improving the efficiency of cross-conformal prediction, an extension of split conformal prediction proposed by Vovk (2015). In the final chapter, a new method for combining conformal prediction sets in an online setting is proposed. The COMA procedure guarantees a regret bound for the loss of the combined set while maintaining valid coverage under additional assumptions. The method can be used in conjunction with the recently proposed Adaptive Conformal Inference (ACI) by Gibbs and Candès (2021) and quantile tracking by Angelopoulos *et al.* (2024a).

Discussion

The thesis aims to bridge the gap in existing methodologies for combining p-values or dependent uncertainty sets. Particular attention is devoted to the exchangeable case. Under this assumption, it follows that the only way to achieve an improvement, in terms of power or statistical efficiency, is through the use of asymmetric combinations. The proposed rules process the p-values (or sets) sequentially, allowing the procedure to stop once stabilization is achieved. The rationale behind this result lies in the fact that, under exchangeability, the order of the random variables is not relevant (i.e., each ordering is equally probable). Consequently, no specific order can be favored, and asymmetry can thus be interpreted as a form of randomization. The same intuition leads to an improvement in combination rules for arbitrarily dependent sources of uncertainty through the use of external randomization. Such randomization can be represented either by an independent uniform random variable or by a random permutation, which enables the exploitation of the principle described above.

It is clear that the use of randomization may be undesirable in certain contexts due to potential issues such as p-hacking. However, there exists a clear trade-off between replicability and statistical efficiency, and this work provides a clear illustration of the "no free lunch" principle. It follows that the use of randomization may be preferable in settings where statistical efficiency is prioritized. It is important to emphasize that the use of randomization is deeply rooted in both statistics and machine learning. The motivations are diverse. Randomization is used to improve statistical efficiency, examples include the median-of-means estimator and randomized tests (Stevens, 1950). It is also used to enhance computational efficiency, as in stochastic gradient descent or permutation-based tests. Moreover, randomization can simplify theoretical analysis and ensure the validity of results, as exemplified by methods based on sample splitting (e.g., Cox (1975)) or by the recent proposal of Leiner *et al.* (2025). Other applications of randomization can be found in differential privacy and Monte-Carlo methods.

It is also worth highlighting an observation reported by Spjøtvoll (Cox *et al.*, 1977) regarding the use of multiple tests for the same hypothesis. In particular, the author notes something seemingly paradoxical yet realistic:

> *"The imaginative statistician can think of many relevant test statistics, while the one with less imagination may produce one or two. Seemingly, the imaginative one is penalized since he has to consider a given significant probability as less significant than his colleague just because started out with a large number of statistics."*

This fact is evident when using the Bonferroni correction, where the significance level is set to α/K. However, this phenomenon generally holds under arbitrary dependence, where the multiplicative correction factor can indeed be substantial (for instance, a factor of two when using the mean and the median). This observation also implies that the iterative application of the methods presented in Chapters 2 and 3 (e.g., repeatedly sampling different values of U and subsequently recombining the corresponding p-values) yields no practical benefit, as the multiplicative cost is incurred at each iteration. This point is also noted in Section 10 of Ramdas and Manole (2024).

As a final remark, the results obtained generally do not entail any additional computational cost. This aspect is particularly relevant in Chapter 5, where the improvements are especially significant, allowing for the derivation of smaller sets.

Future directions of research

While this work provides results on the combination of evidence under dependence, it also leaves open several interesting questions, which are highlighted in the following.

In Chapter 2, the central objective is to identify combination rules that control the type-I error while maintaining an adequate level of power. Several methods based on quantiles and generalized averages are proposed; however, it remains unclear which combination should be preferred under different configurations of the alternative. Moreover, the admissibility of the proposed ex-p-merging function is still an open question.

In Chapter 5, we address the problem of improving the efficiency of cross-conformal prediction. Although the method guarantees coverage of at least $1-2\alpha$, empirical results typically show coverage closer to $1-\alpha$. Our proposal is to employ more efficient p-value combination strategies in order to obtain smaller prediction sets while preserving the same coverage guarantees. Empirically, our variants achieve coverage within the interval $(1-2\alpha,\, 1-\alpha)$. An alternative line of research would be to investigate the stability assumptions of the underlying algorithm that ensure $1-\alpha$ coverage for the standard approach.

Finally, it would be of interest to study under which assumptions the weights in the COMA method converge to a single algorithm. In addition, further empirical investigations using different algorithms could help assess the plausibility of the assumptions.

Appendix A

A.1 Exchangeable and randomized p-merging function

In this part, we will integrate the results in Section 2.2, using both exchangeability and randomization. In fact, starting from exchangeable p-values, it is possible to prove that if randomization is allowed, then it is possible to improve some of the results obtained in Section 2.2. We start by defining a "randomized" version of Theorem 2.5.

Theorem A.1. *Let f be a calibrator, and $(\mathbf{P}, U) = (P_1, \dots, P_K, U) \in \mathcal{U}^K \otimes \mathcal{U}$ such that $\mathbf{P}$ is exchangeable. For each $\alpha \in (0,1)$, we have*

$$\mathbb{P}\left(f\left(\frac{P_1}{\alpha}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \geq 1 \right) \leq \alpha.$$

Proof. The proof invokes the exchangeable and uniformly-randomized Markov inequality (EUMI) introduced in Ramdas and Manole (2024); see Theorem 1.8. In particular,

$$\begin{aligned}\mathbb{P}\left(f\left(\frac{P_1}{\alpha}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\alpha}\right) \geq 1 \right) &\overset{(i)}{\leq} \mathbb{E}\left[f\left(\frac{P_1}{\alpha}\right) \right] \\ &= \alpha \mathbb{E}\left[\frac{1}{\alpha} f\left(\frac{P_1}{\alpha}\right) \right] \overset{(ii)}{\leq} \alpha,\end{aligned}$$

where (i) is due to EUMI and (ii) is due to Lemma 2.3. □

Similarly to how it was done in the Chapter 2, let us now define a randomized ex-p-merging function.

Definition A.2. A randomized ex-p-merging function is an increasing Borel function $F : [0,1]^{K+1} \to [0,1]$ such that $\mathbb{P}(F(\mathbf{P}, U) \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ and $(\mathbf{P}, U) \in \mathcal{U}^K \otimes \mathcal{U}$ with $\mathbf{P}$ exchangeable. It is homogeneous if $F(\gamma \mathbf{p}, u) = \gamma F(\mathbf{p}, u)$ for all $\gamma \in (0,1]$

and $(\mathbf{p}, u) \in [0,1]^{K+1}$. A randomized ex-p-merging function is admissible if for any randomized ex-p-merging function G, $G \leq F$ implies $G = F$.

Let f be a calibrator; then for $\alpha \in (0,1)$, we define the exchangeable and randomized rejection region by

$$R_\alpha = \left\{ (\mathbf{p}, u) \in [0,1]^{K+1} : f\left(\frac{p_1}{\alpha}\right) \geq u \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \right\},$$

where we set $f(p_i/u) = 0$ if $u = 0$. Starting from R_α, we can define the function $F : [0,1]^{K+1} \to [0,1]$ by

$$\begin{aligned} F(\mathbf{p}, u) &= \inf \{\alpha \in (0,1) : (\mathbf{p}, u) \in R_\alpha\} \\ &= \inf \left\{ \alpha \in (0,1) : f\left(\frac{p_1}{\alpha}\right) \geq u \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \geq 1 \right\}, \end{aligned} \tag{A.1}$$

with the convention $\inf \varnothing = 1$ and $0 \times \infty = \infty$.

Theorem A.3. *If f is a calibrator and $(\mathbf{P}, U) \in \mathcal{U}^K \otimes \mathcal{U}$ with $\mathbf{P}$ exchangeable, then F in* (A.1) *is a homogeneous randomized ex-p-merging function.*

Proof. It is clear that F is increasing and Borel since R_α is a lower set. For an exchangeable $\mathbf{P} \in \mathcal{U}^K$ and $\alpha \in (0,1)$, using Theorem A.1 and the fact that $(R_\beta)_\beta \in (0,1)$ is nested, we have

$$\begin{aligned} \mathbb{P}(F(\mathbf{P}, U) \leq \alpha) &= \mathbb{P}\left(\inf \left\{ \beta \in (0,1) : f\left(\frac{P_1}{\beta}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1 \right\} \leq \alpha \right) \\ &= \mathbb{P}\left(\bigcap_{\beta > \alpha} \left\{ f\left(\frac{P_1}{\beta}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1 \right\} \right) \\ &= \inf_{\beta > \alpha} \mathbb{P}\left(f\left(\frac{P_1}{\beta}\right) \geq U \text{ or } \exists k \leq K : \frac{1}{k}\sum_{i=1}^{k} f\left(\frac{P_i}{\beta}\right) \geq 1 \right) \\ &\leq \inf_{\beta > \alpha} \beta = \alpha. \end{aligned}$$

therefore F is a valid randomized ex-p-merging function. Homogeneity comes directly from the definition of (A.1). □

A.2 Generalized Hommel combination rule

We can generalize the Hommel combination by allowing the selection of certain quantiles from the possible K different quantiles of $\mathbf{p} = (p_1, \ldots, p_K)$, for example, one can

select the minimum between K times the minimum, 2 times the median and the maximum. See, for example, Falk (1989). This can be obtained by noting that the Hommel combination rule can be rewritten in terms of quantiles. In particular, the following holds:

$$F'_{\text{Hom}}(\mathbf{p}) = h_K \bigwedge_{k=1}^{K} \frac{1}{\lambda_k} p_{(\lceil \lambda_k K \rceil)}, \quad \text{with } h_K = \sum_{j=1}^{K} \frac{\lambda_j - \lambda_{j-1}}{\lambda_j},$$

where $(\lambda_0, \lambda_1, \ldots, \lambda_K)$ is the vector of quantiles such that $\lambda_j = j/K$, $j = 0, 1, \ldots, K$. This gives the intuition to define a generalization of the aforementioned rule.

Let us define the vector of quantiles $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_M)$, such that $\lambda_0 = 0$, $\lambda_j \in (0, 1]$, if $j = 1, \ldots, M$ and $\lambda_j < \lambda_{j+1}$. Then, we can define

$$F'_{\text{GHom}}(\mathbf{p}) := h_M \bigwedge_{k=1}^{M} \frac{1}{\lambda_k} p_{(\lceil \lambda_k K \rceil)}, \quad \text{with } h_M = \sum_{j=1}^{M} \frac{\lambda_j - \lambda_{j-1}}{\lambda_j}. \tag{A.2}$$

Lemma A.4. *Let f be a function defined by*

$$f(p) = \sum_{j=1}^{M} \frac{1}{\lambda_j} \mathbb{1}\left\{ p \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M} \right] \right\} + \infty \mathbb{1}\{p = 0\}.$$

Then f is an admissible calibrator. Moreover, the p-merging function induced by f is

$$F_{\text{GHom}}(\mathbf{p}) = \inf \left\{ \alpha \in (0,1) : \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{M} \frac{1}{\lambda_j} \mathbb{1}\left\{ \frac{p_k}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M} \right] \right\} \geq 1 \right\},$$

that is valid and dominates (A.2).

Proof. It is simple to see that f is decreasing and upper semicontinuous. In addition,

$$\int_0^1 f(p) \mathrm{d}p = \sum_{j=1}^{M} \frac{1}{\lambda_j} \left(\frac{\lambda_j - \lambda_{j-1}}{h_M} \right) = \frac{1}{h_M} \sum_{j=1}^{M} \frac{\lambda_j - \lambda_{j-1}}{\lambda_j} = 1.$$

This implies that F_{GHom} is admissible. To prove the last part, we see that for any $\mathbf{p} \in (0, 1]^K$ and $\alpha \in (0, 1)$ we have that $F_{\text{GHom}}(\mathbf{p}) \leq \alpha$, if and only if,

$$\bigwedge_{k=1}^{M} \frac{1}{\lambda_k} p_{(\lceil \lambda_k K \rceil)} \leq \frac{\alpha}{h_M}.$$

This implies that, for some $m \in \{1, \ldots, K\}$, we have that

$$\sum_{i=1}^{K} \frac{1}{K} \mathbb{1}\left\{ p_i \leq \alpha \frac{\lambda_m}{h_M} \right\} \geq \lambda_m.$$

We now define this chain of inequalities,

$$1 \leq \sum_{i=1}^{K} \frac{1}{K\lambda_m} \mathbb{1}\left\{p_i \leq \alpha \frac{\lambda_m}{h_M}\right\} = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{\lambda_m} \sum_{j=1}^{m} \mathbb{1}\left\{\frac{p_i}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\}$$
$$\leq \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{m} \frac{1}{\lambda_j} \mathbb{1}\left\{\frac{p_i}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\} \leq \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{M} \frac{1}{\lambda_j} \mathbb{1}\left\{\frac{p_i}{\alpha} \in \left(\frac{\lambda_{j-1}}{h_M}, \frac{\lambda_j}{h_M}\right]\right\}.$$

□

It is possible to see that (A.2) is a special case of the Rüger combination when $M = 1$, while it coincides with the Hommel combination rule when $\lambda = (0, 1/K, \ldots, (K-1)/K, 1)$.

A.3 Improving generalized mean

In this section, we discuss the generalized mean combination rule, for $r \in \mathbb{R} \setminus \{0\}$. This combination rule, introduced in Vovk and Wang (2020), is quite broad and contains some important cases well known in the literature. In particular, if $r = 1$ it reduces to the sample average (Section 2.5), while if $r = -1$ it coincides with the harmonic mean described in Section 2.6. We introduce a lemma characterizing the calibrator used in the context of the generalized mean combination rule.

Lemma A.5. *Let $r \in \mathbb{R} \setminus \{0\}$ and $f : [0, \infty) \to [0, \infty]$ be given by*

$$f(p) = \min\left\{\frac{r(1-p^r)}{T_{r,K}}, K\right\} \mathbb{1}\{p \in [0,1]\}, \tag{A.3}$$

where $T_{r,K} > 0$ is any constant, possibly dependent on K, such that $\int_0^1 f(p)\mathrm{d}p \leq 1$. Then f is a calibrator.

Proof. Since $f(p)$ is continuously decreasing in $T_{r,K}$, it can be verified that there exists $T_{r,K} > 0$ such that $\int_0^1 f(p)\mathrm{d}p = 1$. Moreover, f is decreasing and non-negative which completes the claim. □

It is simple to see that for $r > 0$, we have that $T_{r,K} = r^2/(r+1)$ satisfies $\int_0^1 f(p)\mathrm{d}p \leq 1$. From previous sections, we obtain $T_{1,K} = 1/2$ while a more complex result appears for $T_{-1,K}$. We now see how the calibrator defined in (A.3) is related to the rule defined in Section 2.2. First, we define the function $F'_{\mathrm{M_r}}$ as

$$F'_{\mathrm{M_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K} \frac{r(1-(p_k/\alpha)^r)}{T_{r,K}} \geq 1\right\} = \frac{M_{\mathrm{r}}(\mathbf{p})}{(1 - T_{r,K}/r)^{1/r}}, \tag{A.4}$$

where $M_{\mathrm{r}}(\mathbf{p})$ is the r-generalized mean of $\mathbf{p}$, defined by $M_{\mathrm{r}}(\mathbf{p}) = (\sum_{k=1}^{K} p_k^r/K)^{1/r}$. In particular, $F'_{\mathrm{M_r}}(\mathbf{p})$ coincides with the generalized mean combination rule where $a_{r,K} = (1 - T_{r,K}/r)^{-1/r}$. In addition, if $r > 0$ then $F'_{\mathrm{M_r}}(\mathbf{p}) = (r+1)^{1/r} M_r(\mathbf{p})$ which coincides with the asymptotically precise merging function studied by Vovk and Wang (2020). It is possible to prove that $F'_{\mathrm{M_r}}(\mathbf{p}) \leq F_{\mathrm{M_r}}(\mathbf{p})$, where

$$F_{\mathrm{M_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right)_+ \geq 1\right\}, \tag{A.5}$$

which is the p-merging function induced by the calibrator defined in (A.3). In particular, according to Theorem 2.2 we have that $F_{\mathrm{M_r}}(\mathbf{p})$ is a p-merging function.

A.4 Exchangeable generalized mean

We now define the function $F_{\mathrm{EM_r}}$ in the following way:

$$F_{\mathrm{EM_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{r(1-(p_i/\alpha)^r)}{T_{r,K}}\right)_+ \geq 1\right\}. \tag{A.6}$$

If we define $F'_{\mathrm{EM_r}}$ by

$$F'_{\mathrm{EM_r}}(\mathbf{p}) = \inf\left\{\alpha \in (0,1) : \bigvee_{\ell \leq K} \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{r(1-(p_i/\alpha)^r)}{T_{r,K}}\right) \geq 1\right\}, \tag{A.7}$$

then it holds that $F_{\mathrm{EM_r}} \leq F'_{\mathrm{EM_r}}$.

Theorem A.6. *Let $r \in \mathbb{R} \setminus \{0\}$, then the function $F'_{\mathrm{EM_r}}$ defined in* (A.7) *equals*

$$\left(1 - \frac{T_{r,K}}{r}\right)^{-1/r}\left(\bigwedge_{m=1}^{K} M_r(\mathbf{p}_m)\right),$$

is an ex-p-merging function, and it strictly dominates the function $F_{\mathrm{M_r}}$ in (A.5)*. However, it is strictly dominated by the function $F_{\mathrm{EM_r}}$ defined in* (A.6) *that is also an ex-p-merging function.*

Proof. According to Theorem 2.7 and Lemma A.5, it follows that the function $F_{\mathrm{EM_r}}$ is an ex-p-merging function. In addition, fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$, then $F'_{\mathrm{EM_r}} \leq \alpha$

if and only if

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{r(1-(p_i/\alpha)^r)}{T_{r,K}} \geq 1 \text{ for some } \ell \leq K \implies \left(1-\frac{T_{r,K}}{r}\right)^{-1/r}\left(\frac{1}{\ell}\sum_{i=1}^{\ell}p_i^r\right)^{1/r} \leq \alpha$$

for some $\ell \leq K$.
Taking a minimum over ℓ yields the desired formula. □

A.5 Randomized generalized mean

According to the the previous sections, we define the randomized p-merging function $F_{\mathrm{UM_r}}$ as follows

$$F_{\mathrm{UM_r}}(\mathbf{p},u) = \inf\left\{\alpha\in(0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right)_+ \geq u\right\}. \tag{A.8}$$

The function $F_{\mathrm{UM_r}} \leq F'_{\mathrm{UM_r}}$ where the function $F'_{\mathrm{UM_r}}$ is defined by

$$F'_{\mathrm{UM_r}}(\mathbf{p},u) = \inf\left\{\alpha\in(0,1) : \frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right) \geq u\right\}, \tag{A.9}$$

and can be considered as the randomized version of (A.4).

Theorem A.7. *Let $r \in \mathbb{R}\setminus\{0\}$, then the function $F'_{\mathrm{EM_r}}$ defined in* (A.9) *equals*

$$\frac{M_r(\mathbf{p})}{(1-uT_{r,K}/r)^{1/r}},$$

is a randomized p-merging function, and it strictly dominates the function $F_{\mathrm{M_r}}$ in (A.5)*. However, it is strictly dominated by the function $F_{\mathrm{UM_r}}$ defined in* (A.8) *that is also a randomized p-merging function.*

Proof. According to Corollary 2.18 and Lemma A.5, it follows that $F_{\mathrm{UM_r}}$ is a valid randomized p-merging function. In addition, fix any $\alpha \in (0,1)$ and $\mathbf{p} \in (0,1]^K$, then $F'_{\mathrm{UM_r}} \leq \alpha$ if and only if

$$\frac{1}{K}\sum_{k=1}^{K}\left(\frac{r(1-(p_k/\alpha)^r)}{T_{r,K}}\right) \geq u \implies \frac{M_r(\mathbf{p})}{(1-uT_{r,K}/r)^{1/r}} \leq \alpha.$$

□

A.6 General algorithm

One general algorithm to compute the ex-p-merging function defined in (2.3) induced by a calibrator f is proposed in Algorithm 5. The algorithm employs the bisection method and consistently yields a p-value that exceeds that of the induced ex-p-merging function by at most 2^{-B}.

Data: A calibrator f, $B \in \mathbb{N}$, and a sequence of p-values $(p_1, \dots, p_K)$.
$L := 0$ and $U := 1$;
for $m = 1, \dots, B$ **do**
  $\alpha := (L + U)/2$;
  if $\bigvee_{k=1}^{K} \left(\frac{1}{k} \sum_{i=1}^{k} f\left(\frac{p_i}{\alpha}\right) \right) \geq 1$ **then**
    $U := \alpha$
  end
  else
    $L := \alpha$
  end
end
Result: U

Algorithm 5: Ex-p-merging algorithm

A.7 Additional simulation results

In this section we report some additional simulation results. We first compare the merging rules introduced in Section 2.3 and in Section 2.5, specifically the rules: "twice the median" and "twice the average". In particular, $K = 100$ p-values are generated as in (2.19), and two different values of ρ are chosen, 0.9 and 0.1, respectively. The parameter k for the randomized Rüger combination rule is set to $K/2$ and μ varies in the interval $[0, 3]$. The results are computed by averaging the outcomes obtained in $10,000$ replications.

In Figure A.1, we can see that the type-I error is controlled at the nominal level 0.05 for all proposed methods. In the case of the Rüger combination, in both dependence scenarios, the power of the combinations obtained by exploiting exchangeability or employing external randomization is highly similar. In the case of the combination based on the arithmetic mean, it appears that the rules obtained using randomization exhibit greater power than those based on exchangeability (and, naturally, than the original rules). In general, across all observed scenarios, the new rules demonstrate a quite significant improvement in terms of power.

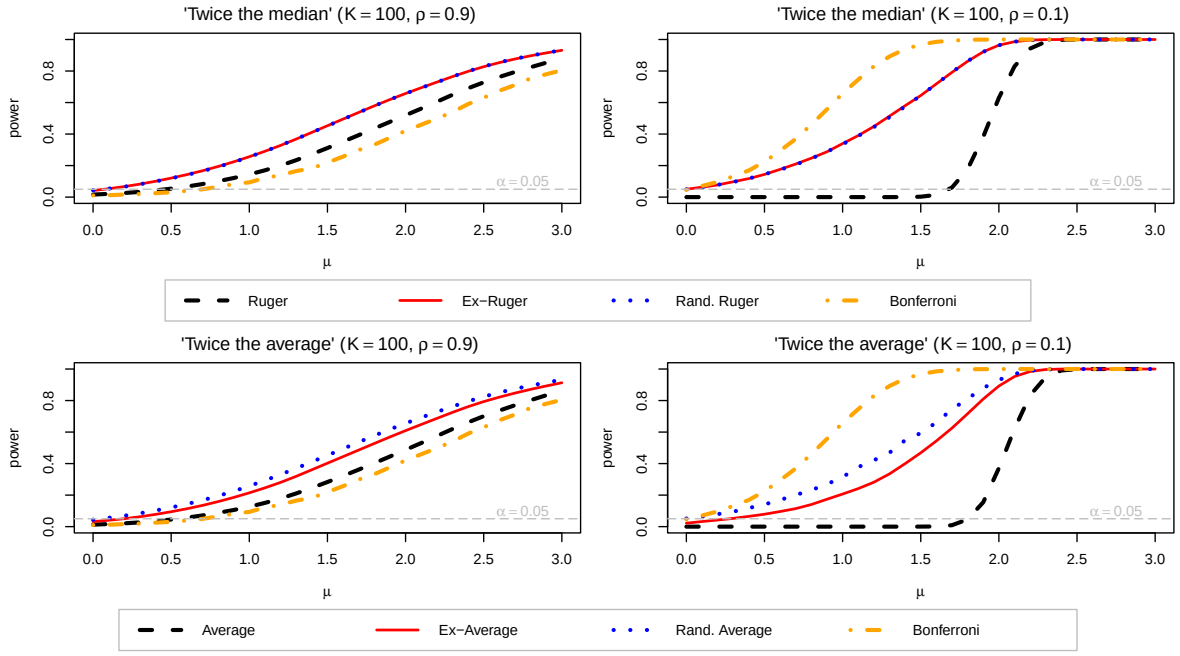

FIGURE A.1: Combination of p-values using different rules. Every subplot illustrates power against μ. The left endpoint of $\mu = 0$ actually represents the empirical type-I error, which is controlled at the nominal level $\alpha = 0.05$ for all methods proposed. The first column has $\rho = 0.9$, while the second column has $\rho = 0.1$ — as expected, the Bonferroni correction is more powerful near independence, but is less powerful under strong dependence. Further, our exchangeable and randomized improvements offer sizeable increases in power over the original variants in all settings.

In addition, we compare all the randomized combination rules reported in the main text. We omit the randomized functions that are dominated by other randomized p-merging functions. As before, $K = 100$ p-values are generated as in (2.19), and $\rho = \{0.1, 0.9\}$. The parameter k for the randomized Rüger combination rule is set to $K/2$ and μ varies in the interval $[0, 3]$. The results are obtained by repeating the procedure $10,000$ times and reporting the average.

The results of the randomized versions of "twice" the median and (improved) "twice" the average are similar in both scenario. In addition, F_{UHom} and F_{UH} are quite similar in terms of power (Figure A.2). From Figure A.2, we can observe an opposite behavior of the functions in the case where $\rho = 0.9$ or $\rho = 0.1$. In general, the most powerful functions in the left graph tend to be the least powerful in the right graph, and vice versa.

At the end, we examine the 3 different randomized combination rules defined in Section 2.5, respectively, $F_{\text{UA}}, F'_{\text{UA}}, F^*_{\text{UA}}$. In particular, we recall that, F_{UA} dominates F'_{UA}, while the combination F^*_{UA} (introduced in Wang (2024, Appendix B.2)) neither dominates nor is dominated by either of the two.

The set of $K = 100$ p-values is generated as in (2.19), and two different values of ρ are

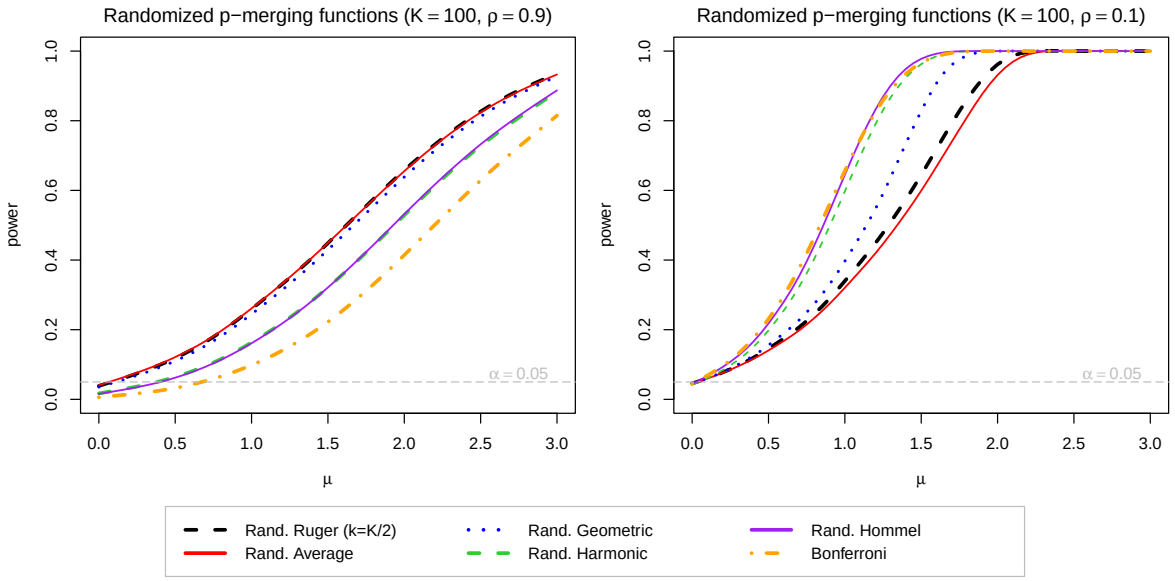

FIGURE A.2: Combination of p-values using different randomized p-merging functions. The order of the performance of the different ex-p-merging functions is almost the opposite in the two situations.

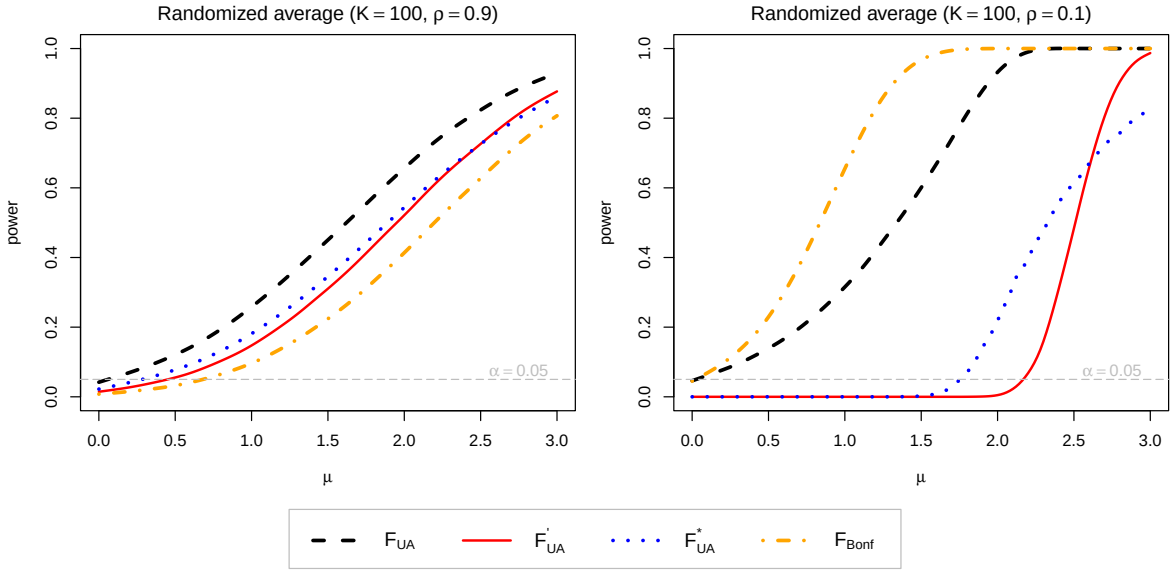

FIGURE A.3: Combination of p-values using different randomized combination rules based on the average of p-values (and the Bonferroni method, for comparison). F^{*}_{UA} is more powerful than F'_{UA} only when $\mu \lesssim 2.5$.

chosen, 0.9 and 0.1, respectively. The results are obtained by repeating the procedure $10,000$ times and reporting the average. From Figure A.3, we can see that the power F_{UA} is always higher than the power of the other two combination rules. The function F^*_{UA} is more powerful than F'_{UA} only in the first part of the y-axis in both cases (for $\mu \leq 2.5$, indicatively), so there is no clear preference between the two. As expected, the Bonferroni correction is more powerful near independence ($\rho = 0.1$).

Before concluding, let us see what our procedures give for testing a global null. The goal of this last part is to explore a case where the order of the p-values in our proposed ex-p-merging functions is chosen in a data-driven manner. In the considered scenario, we will see how a particular statistics can be chosen to order the p-values with the aim of improving the statistical power under the alternative. In particular, this data-driven ordering does not alter the exchangeability under the null that is required in our ex-p-merging functions. Specifically, we investigate the issue of performing simultaneous one-sample t-tests using n observations for each hypothesis. Let us suppose to have K samples (one for each hypothesis) from a normal distribution,

$$X_{ki} \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad i = 1, \dots, n, \quad \mu_k \in \mathbb{R}, \quad \sigma_k > 0,$$

where the observations X_{ki}, $k \in \{1, \dots, K\}$, $i \in \{1, \dots, n\}$, are mutually independent. Our goal is to test the hypothesis $H_k : \mu_k = 0$ and to do so we use t-test. In addition, we define the global null as $H_0 : \bigcap_{k=1}^{K} H_k$, that can be tested by merging the different p-values into a single p-value. The same problem has been studied, for example, in Westfall *et al.* (2004) and Ignatiadis *et al.* (2024). Let us define

$$\bar{X}_k = \frac{1}{n} \sum_{i=1}^{n} X_{ki}, \quad \hat{\sigma}_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ki} - \bar{X}_k)^2, \quad T_k = \frac{\sqrt{n} \bar{X}_k}{\hat{\sigma}_k^2},$$

and p-values are $P_k := 2G_{n-1}(-|T_k|)$, where G_{n-1} is the t-distribution with $n-1$ degrees of freedom.

Our p-values will be ordered using the statistics $S_k^2 = \sum_{i=1}^{n} X_{ki}^2$, and this can be done since the proposed statistics are independent from T_k (and so P_k). The key observation is that, under H_0 (i.e., when $\mu_k = 0$ for all $k = 1, \dots, K$), then the statistic S_k^2 is sufficient and complete for the inference on the parameter σ_k^2. On the other hand, the test statistic T_k is constant in distribution with respect to σ_k^2. By Basu's Theorem (Basu, 1955), S_k^2 and T_k are independent and, in addition, the test T_k and the statistics S_k^2 are independent among themselves since they are functions of independent random variables.

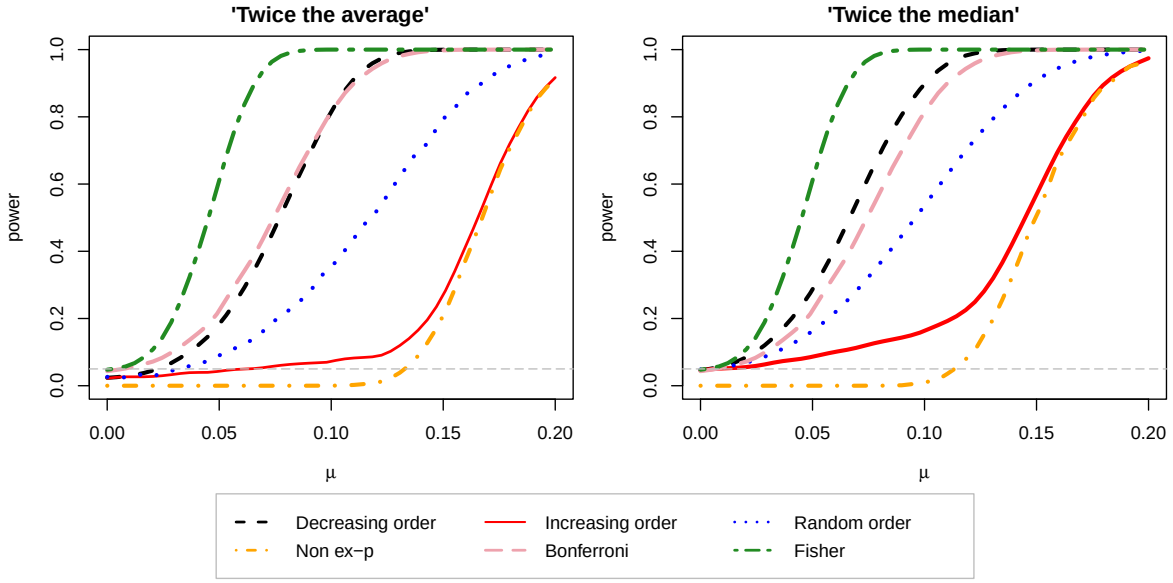

FIGURE A.4: Combination of p-values for testing a global null. If p-values are ordered in decreasing order with respect to S_k^2 we have a power that is comparable with the Bonferroni rule. However, in all situations Fisher's combination is the most powerful.

We assume $\sigma_k = \sigma = 1$ for all $k = 1, \ldots, K$, and $\mu_k = k \cdot \mu$, where μ is a parameter that varies in $[0, 0.2]$. The parameter α is set to 0.05, while $K = 20$ and $n = 10$. As can be seen in Figure A.4, in all situations the type-I error is controlled at the level α under H_0 (i.e., when $\mu = 0$). If p-values are ordered in decreasing order with respect to the statistics S_k^2 then we have a rule that has a power comparable (or higher) than the Bonferroni rule. It is expected that the descending order is more effective, as a large S_k^2 indicates that H_k is likely false. The Fisher combination, which requires the strong assumption of independence, has the largest power.

Appendix B

B.1 Merging confidence distributions

As outlined in the Chapter 3, if the aggregator knows the *confidence distribution* for each agent, then it could be straightforward to combine them in a single confidence distribution. In particular, the *confidence distribution* can be conceptualized as the distribution derived from the p-values corresponding to each point in the parameter space. In particular, for each agent and each point s in the parameter space, we have the corresponding p-value for the hypothesis $H_0 : c = s$. The confidence distribution can be obtained when c is the functional or parameter of interest of a given distribution, but also in the conformal prediction setting where it is common to work with so called conformal (or rank-based) p-values. This suggests that, in order to derive the distribution of the aggregator, we can combine the p-values obtained by the K different agents for each point s using a valid p-merging function.

The simple majority vote rule can be viewed as an inversion of the fact that for K dependent p-values, $2 \cdot \text{median}(p_1, \dots, p_K)$ yields a valid p-value (Rüger, 1978). To see this, let $p_k = p_k(z; s)$ be the observed p-value by the k-th agent for the null hypothesis $H_0 : c = s$; then, using the duality between hypothesis testing and confidence sets, we have that $\mathcal{C}_k = \{s \in \mathcal{S} : p_k(z; s) > \alpha\}$. Suppose (for the sake of contradiction) that $p_{(\lceil K/2 \rceil)}(z; s) \leq \alpha$ and $s \in \mathcal{C}^M$. This implies that

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{s \in \mathcal{C}_k\} > \frac{1}{2} \implies \sum_{k=1}^{K} \mathbb{1}\{p_k(z; s) > \alpha\} > \left\lfloor \frac{K}{2} \right\rfloor \implies p_{(\lceil K/2 \rceil)}(z; s) > \alpha,$$

which contradicts the supposition, establishing the claim. More generally, Rüger (1978) showed that $(K/k)p_{(k)}$ is a valid p-value, where $p_{(k)}$ is the k-th order statistic, recovering the Bonferroni correction at $k = 1$, the union at $k = K$, and the median rule for $k = K/2$. The result in Theorem 2.22, states that $(K/k)p_{(\lceil Uk \rceil)}$ is also valid, and this rule is always smaller or equal than the Rüger combination since $U \in \mathcal{U}$. In Figure B.1 we report an

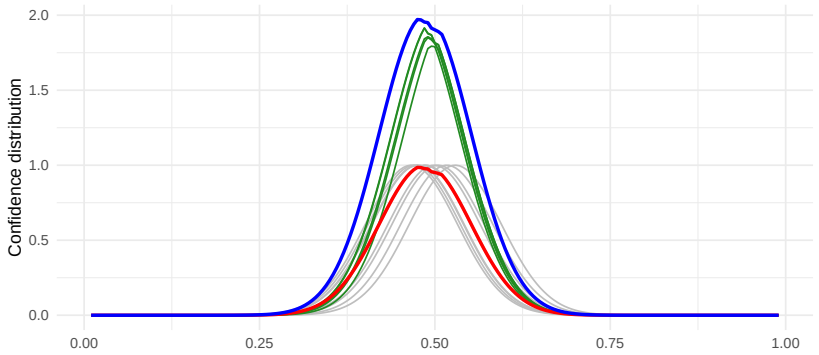

FIGURE B.1: Example of *confidence distributions* obtained in a simulation scenario (private multi-agent setting). Gray: confidence distributions of the single agents, Red: confidence distribution of the median, Blue: confidence distribution of the median multiplied by 2, Green: four possible distributions obtained using the randomized version of Rüger's combination; we fix $U = \{0.2, 0.4, 0.6, 0.8\}$ in the latter method, though it would be random in practice. The red curve is not a valid combination of the grey ones, but the blue and green curves are valid.

example of the *confidence distribution* obtained using two times the median of p-values as a merging function and its randomized extension. In particular, the example refers to an iteration of the simulation scenario described in Section 3.2.

B.2 Algorithm for interval construction

We propose an algorithm that returns the resulting set from the majority voting procedure. The starting point, for simplicity, is a collection of closed intervals, but it can be easily adapted to cases where intervals are open, or only some of them are open. In particular, the algorithm returns two vectors: one containing the lower bounds and the other containing the upper bounds.

Clearly, a naive solution involves dividing the space of interest $\mathcal{S}$ into a grid of points and evaluating how many intervals each point belongs to. However, this approach can become computationally burdensome, especially when the number of points is significantly high. Therefore, an alternative algorithm presented in Algorithm 6 is recommended. In particular, the algorithm is based solely on the endpoints of the various intervals, and it does not require any discretization of the space.

Data: K different intervals with lower bounds a_k and upper bounds b_k, $k = 1, \ldots, K$; $w = (w_1, \ldots, w_K)$ vector of weights; $\tau \in (0, 1)$ threshold ($\tau = 0.5$ for the majority vote procedure)
Result: lower; upper
$q \leftarrow (q_1, \ldots, q_{2K})$ vector containing the endpoints of the intervals in ascending order;
$i \leftarrow 1$;
lower $\leftarrow \varnothing$;
upper $\leftarrow \varnothing$;
while $i \leq 2K$ **do**
  if $\sum_{k=1}^{K} w_k \mathbb{1}\{a_k \leq \frac{q_i + q_{i+1}}{2} \leq b_k\} > \tau$ **then**
    lower $\leftarrow$ lower $\cup\, q_i$;
    $j \leftarrow i$;
    while $(j < 2K)$ *and* $\left(\sum_{k=1}^{K} w_k \mathbb{1}\{a_k \leq \frac{q_j + q_{j+1}}{2} \leq b_k\} > \tau\right)$ **do**
      $j \leftarrow j + 1$;
    end
    $i \leftarrow j$;
    upper $\leftarrow$ upper $\cup\, q_i$;
  end
  else
    $i \leftarrow i + 1$;
  end
end

Algorithm 6: Majority Vote Algorithm

B.3 Combining an infinite number of sets

There are instances where the cardinality of the set K can be uncountably infinite. Consider a sequence of sets identified by some parameters — such as in lasso regression, where each value of the penalty parameter corresponds to a distinct set. However, since the parameter can assume values on the positive semi-axis, the resulting number of sets is uncountable.

Specifically, we assume the existence of a mapping from $\lambda \in \Lambda \subseteq \mathbb{R}^d$ to $2^{\mathcal{S}}$, signifying that for each fixed λ value, there exists a corresponding $1 - \alpha$ uncertainty set $\mathcal{C}_\lambda$. In addition, let us define a nonnegative *weight function*, denoted as $w(\cdot)$, such that:

$$\int_{\Lambda} w(\lambda) d\lambda = 1, \tag{A.1}$$

and $w(\cdot)$ can be interpreted as a prior distribution on λ. In this case, we can define

$$\mathcal{C}^W = \left\{ s \in \mathcal{S} : \int_{\Lambda} w(\lambda) \mathbb{1}\{s \in \mathcal{C}_\lambda\} d\lambda > \frac{1}{2} + U/2 \right\}. \tag{A.2}$$

Proposition A.1. *Let $\mathcal{C}_\lambda$ be a sequence of $1 - \alpha$ uncertainty sets indexed by $\lambda \in \Lambda$. Then the set $\mathcal{C}^W$ defined in* (A.2) *is a level $1 - 2\alpha$ uncertainty set. Furthermore, the*

measure associated with the set $\mathcal{C}^W$ *satisfies*

$$\nu(\mathcal{C}^W) \leq 2 \int_\Lambda w(\lambda)\nu(\mathcal{C}_\lambda)d\lambda.$$

In this scenario, in order to make the computation of $\mathcal{C}^W$ computationally feasible, it is necessary to have a finite number of possible λ values. A potential solution is to sample N independent instances of λ from the *prior distribution*, calculate their respective $1-\alpha$ intervals, and then construct the set $\mathcal{C}^R$ described in (3.11).

B.4 Derandomizing cross-fitting in causal inference

In many problems of semiparametric inference, it is common to be interested in making inference about a parameter θ in the presence of nuisance parameters ξ, possibly of high dimension. Often, the estimation of ξ is carried out using machine learning methods that tend to perform adequately well in high-dimensional settings. However, issues such as overfitting and regularization bias can pose challenges for the inference of the parameter of interest. A solution is proposed in Chernozhukov *et al.* (2018a), where cross-fitting is employed to circumvent such problems. Subsequently, we will discuss this technique in the context of estimating the Average Treatment Effect (ATE).

We now define the problem setup and the notation, suppose that we observe a sample of n iid random variables $Z_1, \ldots, Z_n$ from a distribution P. In particular, Z_i consists on the triplet $Z_i := (X_i, A_i, Y_i)$; where $X_i \in \mathbb{R}^d$ are the baseline covariates of the i-th observation, A_i is the treatment that they receive while $Y_i \in \mathbb{R}$ is the outcome observed after the treatment. As an example, consider observations that represent a cohort of patients undergoing two different types of treatment (each patient receives only one of the two treatments), where their covariates include control variables such as age and sex. Our target parameter is represented by the ATE, $\theta := \mathbb{E}[Y^1 - Y^0]$, where Y^a represents the counterfactual outcome for a randomly selected subject that received the treatment; see VanderWeele (2015, Chapter 2) and Robins and Greenland (1992) for an introduction. Subject to conventional causal identification assumptions, commonly known as consistency, positivity, and no unmeasured confounding (refer, for instance, to Kennedy (2016)), it follows that θ can be expressed as a functional of the distribution P, without invoking counterfactual considerations. In particular, we have

$$\theta = \theta(P) = \mathbb{E}\left\{\mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x]\right\}.$$

Once the data have been observed, we have to find an efficient estimator of ATE. First of all, let us define $h(z)$ as

$$h(z) := \left(\mu^1(x) - \mu^0(x)\right) + \left(\frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)}\right)(y - \mu^a(x)),$$

where $\mu^a(x) := \mathbb{E}[Y \mid X = x, A = a]$ is the regression function for observations whose treatment level is equal to $a = \{0, 1\}$, while $\pi(x) := \mathbb{P}(A = 1 \mid X = x)$ is the so-called propensity score. The *nuisance* functions $\xi = (\mu^1(x), \mu^0(x), \pi(x))$ are unknown, and need to be estimated even if not of direct interest (actually in RCTs $\pi(x)$ is known and only $\mu^a(x), a = \{0, 1\}$, need to be estimated). An efficient estimator for the ATE is given by $n^{-1} \sum_{i=1}^{n} h(z_i)$; see Kennedy (2016).

Adopting a parametric structure for nuisance functions can be restrictive, especially in high-dimensional contexts. Frequently, these functions are estimated through machine learning methods, which, however, encounter challenges such as overfitting and selection bias and render the convergence complex. As described in Chernozhukov *et al.* (2018a), a potential solution to mitigate these issues is the use of the cross-fit estimator, using a random data splitting approach. In particular, the functions ξ are estimated on a first (random) part of the dataset, called $\mathcal{Z}^{trn}$, in order to obtain $\hat{\xi}$ and an estimator of θ is given by $\hat{\theta}_1 = |\mathcal{Z}^{eval}|^{-1} \sum_{i \in \mathcal{Z}^{eval}} h(z_i)$ where $\mathcal{Z}^{eval}$ is $\{Z_i\}_{i=1}^{n} \setminus \mathcal{Z}^{trn}$ and ξ is substituted by its estimated counterpart. Another estimator, $\hat{\theta}_2$, is simply obtained by switching the roles of the set $\mathcal{Z}^{trn}$ and $\mathcal{Z}^{eval}$. The cross fit estimator is simply $\hat{\theta} = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$. In general, if observations are partitioned into B non-overlapping samples, then one can obtain B different estimators and, in this case, the cross-fit estimator is given by their average. Under suitable conditions, the cross-fit estimator is asymptotically normal, so the computation of a confidence interval is possible. Also in this case, the estimator and the intervals depend on the random split of the data. One can repeat the procedure K times in order to construct K different intervals and subsequently merge them using the majority vote procedure.

Real data application. We investigate the impact of K in an example using real data on the effectiveness of a treatment for type 2 diabetes. Data are used in Berchialla *et al.* (2022) and are publicly available. Specifically, the study contains $n = 92$ patients, and for each patient a series of baseline covariates is measured. Treatment refers to the drug administered to the patient. Specifically, `glimepiride` serves as the baseline drug, while `sitagliptin` represents the novel drug under evaluation. It is noteworthy that the study does not adhere strictly to a RCT design, as the allocation of medications to participants is not random, but based on specific characteristics of the participants. The

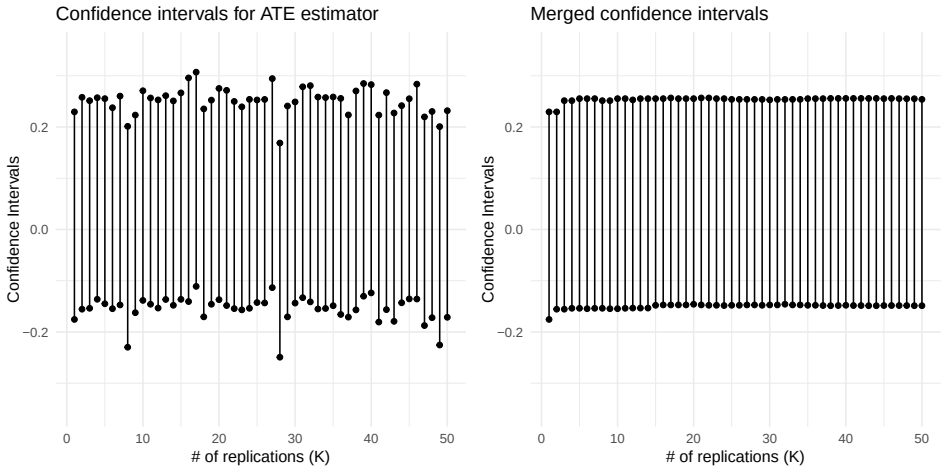

FIGURE B.2: The left plot shows 50 estimates of the ATE via cross-fitting using different sample splits on the same dataset. The right plot shows their combination via majority vote. Both plots can be produced one-by-one from left to right, and stopped when the intersection of the sets in the right plot is deemed stable enough.

outcome is represented by the difference of the level of `HbA1c` before and after treatment, and takes a value of one if the difference is less than -0.5, and zero otherwise. Due to the difficulty in specifying a parametric model for $\mu^a(x)$ and $\pi(x)$, they were estimated using machine learning algorithms. In this case, a random forest is employed to estimate $\mu^a(x)$ while a penalized logistic regression is used to estimate the propensity score. The number of non-overlapping subsets B is fixed at 4, as suggested in Chernozhukov *et al.* (2018a); and the cross-fit estimator is computed using the R package `DoubleML` (Bach *et al.*, 2024). As it is possible to see from Figure B.2 the intervals obtained for ATE can differ between them, while intervals obtained using the majority vote remain essentially the same for $K > 15$. Given that the sets are obtained with the same generating mechanism but using different random data splits, they are exchangeable and the set $\mathcal{C}^E$ is given by the running intersection of the majority vote sets.

Appendix C

C.1 Merging sets with conformal risk control

In Section 4.4, we have used conformal prediction to obtain prediction intervals that allow the derivation of an upper bound for the probability of miscoverage. However, as explained in Angelopoulos and Bates (2023), miscoverage is not the primary and natural error metric in many machine learning problems. Consequently, a more general metric may be necessary to assess the loss between the target of interest and an arbitrary set $\mathcal{C}$. To achieve this, one may proceed by choosing a loss function

$$\mathcal{L} : 2^{\mathcal{Y}} \times \mathcal{Y} \to [0, B], \quad B \in (0, \infty), \tag{A.1}$$

where $\mathcal{Y}$ is the space of the target being predicted, and $2^{\mathcal{Y}}$ is the power set of $\mathcal{Y}$. In addition, we require that the loss function satisfies the following properties:

$$\mathcal{C} \subset \mathcal{C}' \implies \mathcal{L}(\mathcal{C}, c) \geq \mathcal{L}(\mathcal{C}', c),$$
$$\mathcal{L}(\mathcal{C}, c) = 0 \text{ if } c \in \mathcal{C}.$$

By definition, the loss function in (A.1) is bounded and shrinks if $\mathcal{C}$ grows (eventually shrinking to zero when the set contains the target). Similar to the conformal prediction framework described in Section 4.4, we consider the target of interest as $Y_{n+1} \in \mathcal{Y}$, while $\mathcal{C} = \mathcal{C}(x), x \in \mathcal{X}$, is a set based on an observed collection of feature-response instances $z_i = (x_i, y_i)$, $i = 1, \ldots, n$.

Angelopoulos *et al.* (2024b) generalize (split) conformal prediction to prediction tasks where the natural notion of error is defined by a loss function that can be different from miscoverage. In particular, their extension of conformal prediction provides guarantees of the form

$$\mathbb{E}\Big[\mathcal{L}\left(\mathcal{C}(X_{n+1}), Y_{n+1}\right)\Big] \leq \alpha, \tag{A.2}$$

where $\alpha \in (0, B)$. It can be seen that *standard* conformal prediction intervals can be obtained simply by choosing $\mathcal{L}\left(\mathcal{C}(X_{n+1}), Y_{n+1}\right) = \mathbb{1}\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\}$.

C.1.1 Majority vote for conformal risk control

It appears possible to extend the majority vote procedure, described in Section 3.1.1 and in Section 3.1.8, to sets with a conformal risk control guarantee.

Proposition A.1. *Let $\mathcal{C}_1(x), \ldots, \mathcal{C}_K(x)$ be $K \geq 2$ different sets with the property in* (A.2)*, $x \in \mathcal{X}$ and $w = (w_1, \ldots, w_K)$ a vector of weights defined as in* (3.13)*. Define*

$$\mathcal{C}^M(x) = \left\{ y \in \mathcal{Y} : \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}\left(\mathcal{C}_k(x), y\right) < \frac{B}{2} \right\}, \tag{A.3}$$

$$\mathcal{C}^W(x) = \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} w_k \mathcal{L}\left(\mathcal{C}_k(x), y\right) < U\frac{B}{2} \right\}, \tag{A.4}$$

where $U \sim \text{Unif}(0, 1)$. Then, $\mathcal{C}^M(x)$ and $\mathcal{C}^W(x)$ control the conformal risk at level 2α.

The proof initially involves calculating the miscoverage of the set, followed by establishing an upper bound for the risk defined as the expected value of the loss function.

Proof. Using Uniformly-randomized Markov inequality (UMI) it is possible to obtain,

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \notin \mathcal{C}^W(X_{n+1})) &= \mathbb{P}\left(\sum_{k=1}^{K} w_k \mathcal{L}(C_k(X_{n+1}), Y_{n+1}) \geq U\frac{B}{2} \right) \\
&\leq \frac{2}{B}\mathbb{E}\left[\sum_{k=1}^{K} w_k \mathcal{L}(C_k(X_{n+1}), Y_{n+1}) \right] \leq \frac{2}{B}\alpha.
\end{aligned}$$

The same holds true using Markov's inequality and choosing $w_k = \frac{1}{K}, k = 1, \ldots, K$. The risk can be bounded as follows,

$$\begin{aligned}
\mathbb{E}\Big[\mathcal{L}(\mathcal{C}^W(X_{n+1}), Y_{n+1}) \Big] &= \int \mathcal{L}(\mathcal{C}^W(X_{n+1}), Y_{n+1}) dP_{XY}^{n+1} \\
&= \int \mathcal{L}(\mathcal{C}^W(X_{n+1}), Y_{n+1}) \mathbb{1}\{Y_{n+1} \notin \mathcal{C}^W(X_{n+1})\} dP_{XY}^{n+1} \\
&\leq B \int \mathbb{1}\{Y_{n+1} \notin \mathcal{C}^W(X_{n+1})\} dP_{XY}^{n+1} \\
&= B\, \mathbb{P}(Y_{n+1} \notin \mathcal{C}^W(X_{n+1})) \leq 2\alpha.
\end{aligned}$$

The same result can be obtained using $\mathcal{C}^M(x)$. □

The obtained bound may be excessively conservative, as it involves substituting the value of the loss function with its upper limit. Consequently, the resulting sets can be too large, particularly when the loss function is uniform over the interval $[0, B]$ or centered on an internal point, or exhibits skewness towards smaller values.

C.1.2 Experiment on simulated data

We now present a simulation study regarding majority vote for conformal risk control. In classification problems, it often occurs that misclassified labels may incur a different cost based on their importance. An example of a loss function used for this purpose is

$$\mathcal{L}(\mathcal{C}, y) = L_y \mathbb{1}\{y \notin \mathcal{C}\},$$

where L_y is the cost related to the misclasification of the label $y \in \mathcal{Y}$ and $\mathcal{Y}$, in this case, denotes the finite set of possible labels.

The methodology introduced by Angelopoulos *et al.* (2024b) uses predictions generated from a model $\hat{\mu}$ to formulate a function $\mathcal{C}_\gamma(\cdot)$ that assigns features $x \in \mathcal{X}$ to a set. The parameter γ denotes the degree of conservatism in the function, with smaller values of γ producing less conservative outputs. The primary objective of their approach is to infer the value of γ using a calibration set, with the aim of achieving the guarantee outlined in (A.2). Given an error threshold α, Angelopoulos *et al.* (2024b) define

$$\hat{\gamma} := \inf \left\{ \gamma : \frac{n}{n+1} \sum_{i=1}^{n} \mathcal{L}(\mathcal{C}_\gamma(x_i), y_i) + \frac{B}{n+1} \leq \alpha \right\}.$$

For classification problems, $\mathcal{C}_\gamma(x_i)$ is simply expressed as $\mathcal{C}_\gamma(x_i) = \{y \in \mathcal{Y} : \hat{\mu}(x_i)_y \geq 1 - \gamma\}$, where $\hat{\mu}(x_i)_y$ represents the probability assigned to the label y by the model.

The approach is used in a classification task with simulated data. We simulated data from 10 classes, each originating from a bivariate normal distribution with a mean vector (i, i) and a randomly generated covariance matrix, where $i = 1, \ldots, 10$. In addition, two covariates were incorporated to add noise and α is set to 0.05. For each class, we generated 600 data points, partitioning them into three equal subsets: one-third for the estimation set, one-third for the calibration set, and one-third for the test set. Loss values are represented by $L_y = \frac{8+y}{18}$, where $y \in \{1, \ldots, 10\}$. This indicates that the cost of misclassifying a label in the last class is twice that of a label in the first class. We used $K = 7$ different classification algorithms, and the parameters $\hat{\gamma}_k$ were estimated within the calibration set, for all $k = 1, \ldots, K$. An example of the majority vote procedure is shown in Figure C.1. The empirical losses computed in the test set of the methods are,

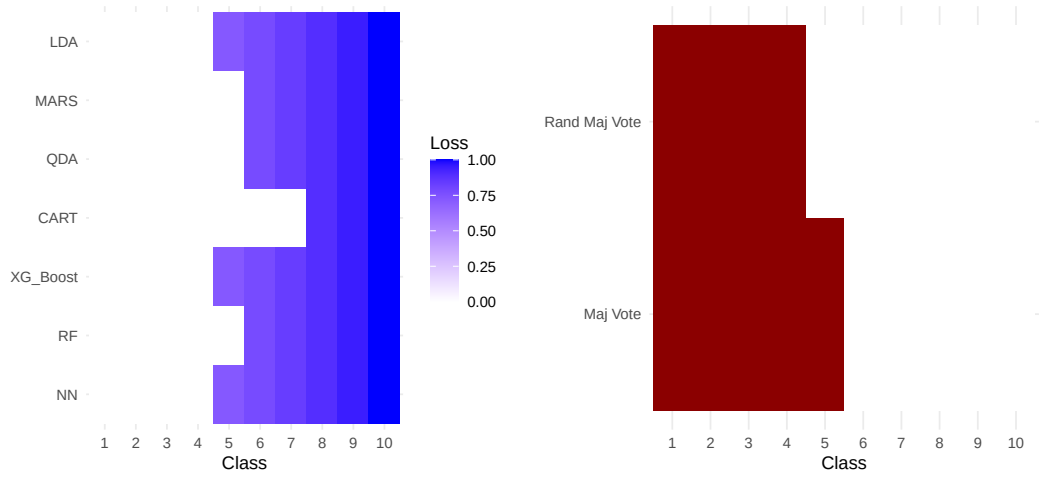

FIGURE C.1: Left: losses of various algorithms, where the loss is zero if the point is included in the interval. Right: labels included by majority vote and randomized majority vote.

respectively, 0.042 for the simple majority vote and 0.084 for the randomized version of the method. It is important to highlight that, in some situations, the majority vote procedure can produce too large sets. Suppose that the loss for a single point is less than half; then the procedure will include the point also if it is not included in any of the sets. The randomized method can present the same problem if the values of the loss function are close to zero. A possible solution to mitigate the problem is to tune the threshold parameter of the majority vote to a smaller value achieving different levels of guarantee.

C.2 Conformal prediction: Additional simulations with different weights

In this section we study the impact of different weight vectors in our proposed method through a simulation study. In particular, data are simulated as in Section 4.4.1, and five different vector of weights are used. The five different cases can be summarized as follows: (i) the weights are uniform and this case coincide with the standard majority vote; (ii) the weight assigned to the last ten sets is thrice the weight assigned to the first ten, i.e. $w \propto (1, \ldots, 1, 3, \ldots, 3)$; (iii) the scenario is opposite to the earlier one and $w \propto (3, \ldots, 3, 1, \ldots, 1)$; (iv) the weights are increasing and $w \propto (1, 2, \ldots, 20)$; (v) the weights are decreasing and $w \propto (20, 19, \ldots, 1)$. The process is repeated 5000 times and in Table C.1 we report the empirical coverage and the empirical size of the set $\mathcal{C}^W$ without randomization.

Case	Coverage	Width
(i)	0.985	8.719
(ii)	0.981	11.832
(iii)	0.987	7.681
(iv)	0.979	12.112
(v)	0.987	7.676

TABLE C.1: Empirical coverage and empirical size of the sets $\mathcal{C}^W$ using different weights. The size is smaller if the weights are more concentrated in the first positions (case (iii) and case (v)).

From Table C.1 it can be observed that the empirical coverage of the methods appears to be consistent across different cases; in contrast, there are substantial differences in terms of interval width. Specifically, if the weights are more concentrated in the first part of the vector, narrower intervals are obtained. This result is in line with Figure 4.1, as smaller penalty parameters seem to correspond to narrower intervals.

As expected, the weights play a key role, like that of a prior, and a "bad" prior can enlarge the size of the weighted majority vote sets. However, coverage guarantees remain unaltered, as pointed out in Section 3.1. If there is no prior knowledge regarding the different methods used to obtain the various intervals, it seems prudent to use a vector of uniform weights, but other (more complicated) options exist. In the considered example, one might set aside a portion of the dataset to select an appropriate penalty parameter and subsequently use this parameter to obtain the prediction interval on the portion not used for selection. In a similar spirit, rather than choosing a single parameter, one could opt for multiple parameters, assigning weights inversely proportional to the error metric observed on the selection set. This can be particularly convenient if there is no a clear winner. Outside the considered scenario, the prior weights can be elicited according to the information available to the user (like the sample sizes used by the different agents or the expected performance of the different algorithms).

C.3 Marginal coverage of cross-conformal prediction and connection with CV+

As stated in Remark 5.1, it is possible to establish an alternative bound for the marginal coverage of cross-conformal prediction, distinct from the one shown in (5.3). In particular, Barber *et al.* (2021) proves that

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})\right) \geq 1 - 2\alpha - 2(1-\alpha)\frac{1 - K/n}{K+1}. \tag{A.5}$$

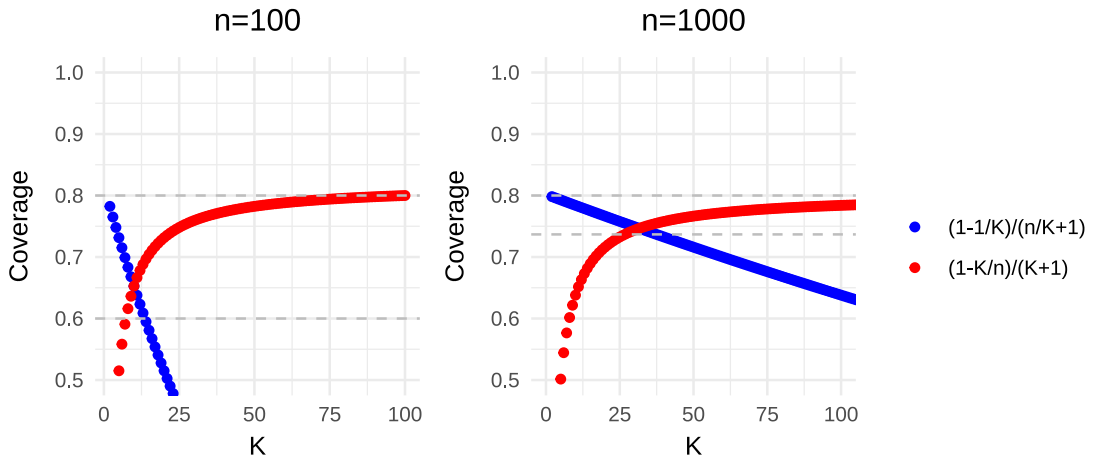

FIGURE C.2: Comparison of the bounds in (5.3) and (A.5) for different values of K with $\alpha = 0.1$. Dashed lines represent the levels $1 - 2\alpha - 2/\sqrt{n}$ and $1 - 2\alpha$.

The proof technique is completely different from the technique presented in Section 5.1 based on p-values and relies on counting arguments applied to tournament matrices (Barber *et al.*, 2021; Angelopoulos *et al.*, 2025). Combining the results in (5.3) and (A.5), we obtain

$$\begin{aligned} \mathbb{P}\left(Y_{n+1} \in \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})\right) &\geq 1 - 2\alpha - 2(1-\alpha)\min\left\{\frac{1-1/K}{n/K+1}, \frac{1-K/n}{K+1}\right\} \\ &\geq 1 - 2\alpha - 2/\sqrt{n}. \end{aligned} \tag{A.6}$$

The two bounds are compared in Figure C.2 and it is possible to see that the two bounds have opposite behaviors. As depicted in Figure C.2, even for small or moderate n, the bound in (5.3) is the one that applies to commonly employed values of K.

In addition, in a regression setting, cross-conformal prediction (Vovk, 2015) is closely related to K-fold CV+ introduced in Barber *et al.* (2021). In particular, both methods can be used to obtain prediction sets with finite sample coverage guarantees. Cross-conformal prediction is covered in Section 5.1; here, we introduce CV+ and explain its connection to cross-conformal prediction. In this case as well, the data points are divided into K disjoint folds $I_1, \ldots, I_K$ of size $m = n/K$, and $\hat{\mu}_{-I_k}$ refers to the regression function trained using data in $[n] \setminus I_k$, $k \in [K]$. The K-fold CV+ prediction set is defined

as

$$\mathcal{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1}) = \Bigg[-\text{quantile}\left(\left(-\left(\hat{\mu}_{-I_{k(i)}}(X_{n+1}) - S_i^{\text{CV+}}\right)\right)_{i\in[n]}; (1-\alpha)(1+1/n)\right), \\ \text{quantile}\left(\left(\hat{\mu}_{-I_{k(i)}}(X_{n+1}) + S_i^{\text{CV+}}\right)_{i\in[n]}; (1-\alpha)(1+1/n)\right)\Bigg], \tag{A.7}$$

where $S_i^{\text{CV+}} = |Y_i - \hat{\mu}_{-I_{k(i)}}(X_i)|$, $i \in [n]$, are the residual scores (or absolute residuals). An attractive property of the set in (A.7) is that it is interpretable, since it is always an interval (rather than, possibly, a union of intervals). In particular, when $n = K$, it corresponds to the jackknife+ interval by Barber *et al.* (2021).

At first glance, the sets $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ and $\mathcal{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1})$ can appear distinct; however, Barber *et al.* (2021, Appendix B.2) proves that, when the score function in (5.2) is the residual score, then

$$\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1}) \subseteq \mathcal{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1}).$$

As a corollary, it follows that the marginal coverage guarantee in (A.6) also holds for the CV+ method. This implies that the marginal coverage guarantee for the jackknife+, the case $K = n$, is at least $1 - 2\alpha$. The empirical coverage often exceeds the stated $1 - 2\alpha - 2/\sqrt{n}$, typically aligning closer to the level $1 - \alpha$, and sometimes approaching one. In fact, the value 2α can be considered as the worst-case scenario for the method. Under some assumptions about the stability of the prediction algorithm, a modified version of the jackknife+ is shown to have marginal coverage close to $1 - \alpha$. A similar problem is studied in Steinberger and Leeb (2023), where the authors prove a conditional coverage probability statement for K-fold cross-validation (a set similar to the one in (A.7)) valid under some assumptions on the algorithm and the distribution of the data. We refer to Angelopoulos *et al.* (2025, Ch. 6) for a detailed discussion of the properties of cross-conformal prediction and jackknife+.

C.4 Cross-conformal prediction with varying fold sizes

In this section we treat the case where the number of observation in each fold can differ. We consider the same setup as at the beginning of Section 5.1, and we allow different sizes among the subsets. Let m_k denote the number of observations in subset I_k, $k \in [K]$. By definition, the sum $m_1 + \cdots + m_K$ equals the number of observations n. In this case, the definition of conformal p-values in (5.4) change slightly, allowing for dependence on

m_k in the denominator:

$$P_k(y) = \frac{1 + \sum_{i \in I_k} \mathbb{1}\left\{s\left((X_{n+1}, y); \mathcal{D}_{[n]\setminus I_k}\right) \leq S_i^{\text{CV}}\right\}}{m_k + 1}, \tag{A.8}$$

where S_i^{CV} is defined in (5.1). However, $\mathbb{P}(P_k(Y_{n+1}) \leq \alpha) \leq \alpha$, still holds for any $\alpha \in (0, 1)$. In addition, we define the weights

$$w_k = \frac{m_k + 1}{n + K}, \tag{A.9}$$

where we note that the weights are positive, sum to one and it holds that $w_k = 1/K$ if $m_1 = \cdots = m_K$.

It is now possible to prove that the marginal coverage of cross-conformal prediction with varying fold sizes remains the same.

Lemma A.2. *Suppose that $m_k = |I_k|, k \in [K]$, then the set $\mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ in (5.2), is such that*

$$\mathbb{P}\big(Y_{n+1} \in \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})\big) \geq 1 - 2\alpha - 2(1-\alpha)\frac{1 - 1/K}{n/K + 1}.$$

Proof. According to the definition of the cross-conformal prediction set in (5.2), we can see that $y \in \mathcal{C}^{\text{cross}}_{n,K,\alpha}(X_{n+1})$ if and only if

$$\frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{s\left((X_{n+1}, y); \mathcal{D}_{[n]\setminus I_{k(i)}}\right) \leq S_i^{\text{CV}}\right\}}{n+1} > \alpha \Longleftrightarrow \sum_{k=1}^{K} w_k P_k(y) > \alpha + (1-\alpha)\frac{K-1}{n+K}, \tag{A.10}$$

where w_k and $P_k(y)$ are defined in (A.9) and (A.8), respectively. To complete the proof, we apply the fact that the weighted average of p-values provides a quantity that is a p-value up to a factor of 2 (Vovk and Wang, 2020). □

At this point, one may wonder whether the validity of the sets defined in Sections 5.1.2, 5.2.1, 5.2.2 and 5.2.3 can also be extended to the case where the fold sizes vary. Since twice the (simple) average of p-values is itself a p-value under arbitrary dependence of the starting p-values, it follows that the coverage guarantee of sets $\mathcal{C}^{\text{mod-cross}}_{n,K,\alpha}(X_{n+1})$ and $\mathcal{C}^{\text{u-mod-cross}}_{n,K,\alpha}(X_{n+1})$ is preserved even if $P_1(y), \ldots, P_K(y)$ are obtained using different m_k.

The coverage guarantee for sets $\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1})$ and $\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha}(X_{n+1})$ is valid when the underlying p-values are exchangeable, and this is related to the number of data points in each fold, as stated in Remark 5.5. However, p-values can be made exchangeable through a random permutation of the indices. For example, assume $n = 101$ and $K = 5$; in this case, the subset with 21 observations does not always have to be the

same, but should be randomly selected among the 5 folds. This implies that the coverage guarantees for the sets $\mathcal{C}^{\text{e-mod-cross}}_{n,K,\alpha}(X_{n+1})$ and $\mathcal{C}^{\text{eu-mod-cross}}_{n,K,\alpha}(X_{n+1})$ can still hold.

More attention should be paid to Section 5.2.4. In fact, as seen on the right side of (A.10), y belongs to the set only if the weighted average of the conformal p-values exceeds a certain threshold. However, the weighted average is an asymmetric function, and results on the combination of exchangeable p-values do not hold in this case. Only the result using randomization remains valid.

C.5 Additional experiments on cross-conformal prediction

Communities and Crime dataset. We apply the proposed methods to the Communities and Crime dataset. The dataset contains information on $n = 1994$ communities in the United States and the goal is to predict the per capita violent rate. After removing the columns containing missing values and categorical variables, the number of regressors is $p = 99$. Two regression algorithms are used, specifically lasso regression with penalty parameter set to 0.01 and random forest with 50 trees grown for each forest.

The α-level is set to 0.1 and the conformal prediction methods are applied on 1000 data points randomly sampled without replacement. The remaining part is used as a test set to compute the metrics. The procedure is repeated 100 times to remove the randomness of the split and we report the averages over these 100 trials. The methods used are cross-conformal prediction and its variants (with $K = 10$), and split conformal prediction is added for comparison.

The results are reported in Figure C.3, where it is possible to see that the smaller sets are obtained using the `eu-mod-cross` method. The modified variants using exchangeability and randomization exhibit higher variability in interval width, likely due to the use of randomization and the asymmetry of the combination rules. All proposed methods have an empirical coverage of at least $1 - 2\alpha$. We remark that cross-conformal prediction guarantees a coverage of at least $1 - 2\alpha$, but is usually conservative. The coverage of the new methods is closer to the level $1 - 2\alpha$ and the new variants outperform cross-conformal prediction in terms of set size.

Boston Housing dataset. We apply conformal prediction methods on a dataset of moderate dimensions, with $p = 13$ and $n = 506$. The aim is to predict the cost of a house in Boston given some information on the neighborhood. The algorithm used is standard linear regression. We apply conformal prediction methods using 200 training

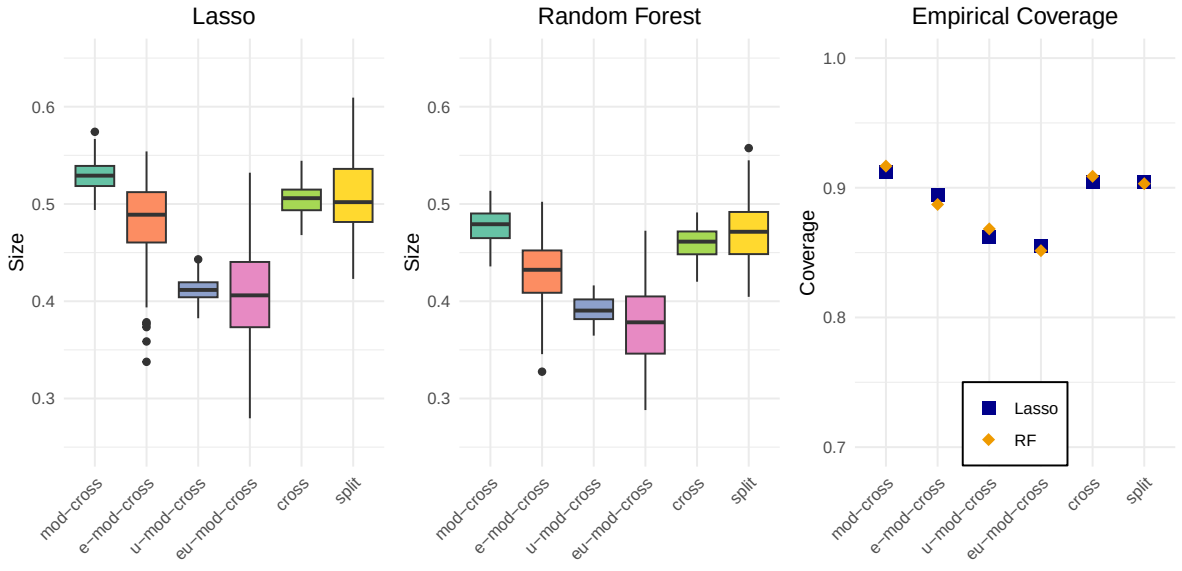

FIGURE C.3: Results for the "Communities and Crime" dataset. Empirical size and empirical coverage of different conformal prediction algorithms are reported. The α-level is set to 0.1. The smaller sets are obtained using `eu-mod-cross` whose empirical coverage is around 0.85. Cross-conformal prediction is conservative, but it tends to produce stable sets.

points, the remaining part is used as test set. The number of different subsets for cross-conformal prediction is set to $K = 5$ and the miscoverage rate is $\alpha = 0.1$. The procedure is repeated 100 times, and we report the averages over the 100 replications.

From Table C.2, we see that smaller sets are obtained using `eu-mod-cross` conformal prediction, while larger ones are produced by split conformal prediction, which exhibits high variability in set size. The methods `e-mod-cross` and `u-mod-cross` have an empirical coverage slightly lower than $1 - \alpha$, with an average size generally smaller than that obtained using cross-conformal prediction. Full conformal prediction exhibits low variability in terms of size, with the sets typically being smaller than those produced by split conformal prediction and cross-conformal prediction. However, as already seen, these advantages are counterbalanced by a high computational cost.

UPDRS dataset. We tested our methods on a dataset containing information on patients with early-stage Parkinson's disease (Tsanas and Little, 2009). The goal is to predict the total UPDRS (Unified Parkinson's Disease Rating Scale) using a range of biomedical voice measurements. In particular, after some preprocessing operations, the data set includes $n = 5875$ points and $p = 13$ covariates. The two regression algorithms used are lasso regression (with penalty parameter equal to 0.01) and random forest (with 25 trees grown for each forest). The α-level is set to 0.1, the number of

	mod-cross	e-mod-cross	u-mod-cross	eu-mod-cross	cross	split	full
Mean	17.303	14.920	14.143	13.370	15.753	16.357	14.516
Sd	1.440	1.998	1.152	1.899	1.307	2.751	1.244
Median	17.188	14.692	14.113	13.128	15.679	16.041	14.656
Min	13.883	10.068	11.551	8.822	12.766	11.730	11.998
Max	20.935	20.537	17.150	18.339	19.101	27.469	17.756
Coverage	0.927	0.888	0.878	0.855	0.908	0.897	0.888

TABLE C.2: Results for the "Boston Housing" dataset using OLS as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The α-level is set to 0.1. The smaller sets are obtained using `eu-mod-cross` conformal prediction. The variability is especially high when using split conformal prediction.

folds is $K = 10$ and the conformal prediction methods are applied on 3000 data points randomly sampled without replacement. The remaining part is used as a test set to compute the metrics. The procedure is repeated 100 times to remove the randomness of the split. The results reported are the averages over these 100 trials. We compare our proposals with cross-conformal prediction. In addition, split conformal prediction with miscoverage rate set to α and 2α is added for comparison.

The results are reported in Table C.3 and Table C.4. In Table C.3, we can see that for lasso regression the smaller sets are obtained using split conformal with miscoverage rate set to 2α. Overall, our approaches typically yield smaller sets compared to those obtained using standard cross-conformal prediction and split conformal prediction. However, we can observe a higher variability derived from the use of randomization (or sequential processing of the p-values). Interestingly, when the random forest is used as a regression algorithm (Table C.4), the smaller sets are obtained using the exchangeable and randomized cross-conformal prediction method. In particular, on average, the method also outperforms the split conformal prediction trained at level 2α. In both cases, the marginal coverage of the proposed methods fluctuates between levels $1 - 2\alpha$ and $1 - \alpha$. On the other hand, from Table C.4 it is possible to see that cross conformal is conservative with coverage higher than the nominal $1 - \alpha$.

Abalone dataset. The proposed methods are applied to the abalone dataset. The goal is to predict the age of abalones (the number of rings) using $p = 8$ physical measurements. The dataset contains $n = 4177$ observations where 4000 observations are used as training points, while the remaining part is used as a test set. In this experiment, we directly modify the cross-conformal prediction as described in Section 5.2.4 (indeed, we remove the word `mod` from the labels in Tables C.5, C.6 and C.7). The procedure is repeated 100 times to remove the randomness of the split and the results reported are the average over the 100 trials. The α-level is set to 0.1 and we use different number of

	mod-cross	e-mod-cross	u-mod-cross	eu-mod-cross	cross	split	split (2α)
Mean	29.965	28.956	26.427	26.252	29.686	29.901	23.773
Sd	0.404	1.019	1.744	1.942	0.383	0.878	0.402
Median	29.946	29.175	26.202	26.092	29.725	29.703	23.734
Min	11.450	10.019	8.477	9.248	11.340	28.539	22.913
Max	32.808	30.826	31.267	30.716	32.698	32.318	25.223
Coverage	0.904	0.892	0.854	0.850	0.901	0.901	0.800

TABLE C.3: Results for the UPDRS dataset using lasso as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The α-level is set to 0.1. On average the smaller sets are obtained using split conformal with miscoverage rate 2α. The marginal coverage of the proposed methods lies between $1 - 2\alpha$ and $1 - \alpha$.

	mod-cross	e-mod-cross	u-mod-cross	eu-mod-cross	cross	split	split (2α)
Mean	17.180	15.186	14.737	13.718	17.015	19.210	14.550
Sd	0.918	1.058	1.481	1.370	0.908	0.825	0.638
Median	17.175	15.303	14.643	13.762	16.954	19.232	14.541
Min	8.697	7.376	6.716	5.725	8.697	16.842	13.079
Max	23.890	17.945	22.129	17.725	23.560	20.954	16.551
Coverage	0.931	0.883	0.887	0.846	0.929	0.900	0.802

TABLE C.4: Results for the UPDRS dataset using random forest as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The α-level is set to 0.1. On average the smaller sets are obtained using the exchangeable and randomized version of cross-conformal prediction. The marginal coverage of the proposed methods lies between $1 - 2\alpha$ and $1 - \alpha$.

folds, in particular, $K = \{5, 10, 20\}$. The regression algorithm used is a random forest with 25 trees grown for each forest.

The results are reported in Tables C.5, C.6 and C.7. The coverage level for the proposed method oscillates between levels $1 - 2\alpha$ and $1 - \alpha$. The smaller sets are obtained on average by the split conformal prediction with a miscoverage rate equal to 2α. The suggested methods improve quite significantly the performance of cross-conformal prediction in terms of set size, although the variability is generally higher. There is a slight decrease in the size of sets, and a slight increase in variability, of the `e` and `eu-mod-cross` methods as K increases. The results for the split conformal conformal prediction are essentially the same for all tables as a different number of folds is applicable just for cross conformal prediction and its variants.

Electricity consumption dataset. We additionally analyze a real dataset on electricity consumption, where accurate uncertainty quantification is crucial as the supplier's revenue depends on customer energy use. The dataset contains 35 411 observations and 20 covariates; 30 000 observations are used for training, while the remaining ones are

	cross	e-cross	u-cross	eu-cross	split	split (2α)
Mean	6.864	6.372	5.708	5.477	6.730	4.617
Sd	0.074	0.255	0.040	0.316	0.155	0.103
Median	6.858	6.332	5.713	5.390	6.703	4.583
Min	6.798	6.049	5.659	5.195	6.512	4.516
Max	6.986	6.762	5.755	6.015	6.912	4.777
Coverage	0.911	0.890	0.873	0.854	0.903	0.803

TABLE C.5: Results for the "Abalone dataset" using random forest as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The number of folds is $K = 5$ and the α-level is set to 0.1. The marginal coverage of the proposed methods lies between $1 - 2\alpha$ and $1 - \alpha$.

	cross	e-cross	u-cross	eu-cross	split	split (2α)
Mean	6.869	6.323	5.695	5.444	6.816	4.741
Sd	0.060	0.225	0.069	0.258	0.171	0.110
Median	6.865	6.360	5.692	5.450	6.815	4.750
Min	6.725	5.559	5.545	4.667	6.351	4.419
Max	7.030	6.800	5.878	6.061	7.235	5.060
Coverage	0.910	0.886	0.863	0.843	0.899	0.798

TABLE C.6: Results for the "Abalone dataset" using random forest as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The number of folds is $K = 10$ and the α-level is set to 0.1. The marginal coverage of the proposed methods lies between $1 - 2\alpha$ and $1 - \alpha$.

	cross	e-cross	u-cross	eu-cross	split	split (2α)
Mean	6.855	6.210	5.649	5.379	6.817	4.717
Sd	0.062	0.340	0.063	0.381	0.190	0.089
Median	6.850	6.283	5.642	5.403	6.837	4.728
Min	6.648	4.992	5.520	4.321	6.396	4.503
Max	7.032	6.760	5.787	6.315	7.358	4.971
Coverage	0.909	0.879	0.862	0.839	0.899	0.795

TABLE C.7: Results for the "Abalone dataset" using random forest as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The number of folds is $K = 20$ and the α-level is set to 0.1. The marginal coverage of the proposed methods lies between $1 - 2\alpha$ and $1 - \alpha$.

used for testing. The splitting is repeated 100 times, and the algorithm used is a random forest (with `ntree = 100`). In this case, methods such as full conformal prediction or jackknife+ can be computationally expensive due to the relatively large number of observations, making sample-splitting-based methods preferable.

The results are presented in Table C.8. The column "uneven split" represents split conformal prediction, where the training set comprises a $(K-1)/K$ fraction of the data points. This corresponds to split conformal prediction with the same fraction of training data points used by a single round of cross conformal prediction. The last two columns represent cases where the p-values are combined using the Bonferroni rule at level α (i.e., $K \min(\mathbf{p})$) or at level 2α (i.e., $K/2 \min(\mathbf{p})$) rather than the simple average. This

implies that the corresponding prediction sets have coverage at least equal to $1-\alpha$ and $1-2\alpha$, respectively.

The smaller sets are obtained by the `eu-mod-cross` method, and in general there is an improvement in the set size if we use cross conformal prediction instead of split conformal prediction. Bonferroni method produces very large prediction sets and this is expected since the Bonferroni rule is not powerful when p-values (or sets) are highly dependent. Indeed, the Bonferroni correction is tightest when the p-values are nearly independent, while the conformal p-values across folds are highly dependent. On the other hand, the rule "twice the mean" used in cross conformal prediction is more powerful when p-values are dependent; see Section 6.1 Vovk and Wang (2020) for a discussion. This aligns with Theorem 4 in Lei *et al.* (2018), which states that the use of multisplit conformal prediction in conjunction with the Bonferroni rule produces wide sets (specifically, under some assumptions, sets are wider than "single split" conformal prediction). The average set size of "uneven" split conformal prediction is smaller than that of split conformal prediction but slightly larger than that of cross conformal prediction. Moreover, its variability (Sd) is higher.

	mod-cross	e-mod-cross	u-mod-cross	eu-mod-cross	cross	split	uneven split	split (2α)	Bonf	Bonf (2α)
Mean	50.85	47.10	33.78	32.27	50.71	59.59	52.25	26.52	196.68	149.63
Sd	0.49	1.66	0.31	1.54	0.49	1.64	2.97	0.54	6.51	2.92
Median	50.83	47.38	33.80	32.47	50.69	59.64	52.12	26.47	197.46	149.71
Min	49.73	41.97	32.99	27.74	49.60	56.54	44.45	25.25	174.87	141.06
Max	52.38	49.67	34.41	36.20	52.25	64.95	59.84	27.96	210.94	157.85
Coverage	0.91	0.89	0.86	0.85	0.90	0.90	0.90	0.80	0.98	0.97

TABLE C.8: Results for the "Electricity consumption dataset" using random forest as regression algorithm. The results refer to set size, except for the last row, which refers to the marginal coverage. The number of folds is $K = 10$ and the α-level is set to 0.1. The marginal coverage of the proposed methods lies between $1-2\alpha$ and $1-\alpha$. The Bonferroni method gives large prediction sets.

C.6 Additional experiments on conformal online model aggregation

C.6.1 Additional simulations on the negative correlation assumption

In this section, we investigate the assumption of total negative correlation in Theorem 6.1 via a simulation study. We study the specific case introduced in Chernozhukov

et al. (2018b), where the data is simulated as follows:

$$y^{(t)} = \beta^{\mathrm{T}} x^{(t)} + \varepsilon^{(t)}, \quad t = 1, \dots, 200,$$

where $x^{(t)}$ is distributed as $\mathcal{N}(0, I_{100})$. The serial dependence is induced by generating the error as follows:

$$\varepsilon^{(t)} = \rho \varepsilon^{(t-1)} + \xi^{(t)}, \quad \xi^{(t)} \overset{iid}{\sim} \mathcal{N}(0, 1 - \rho^2),$$

where $\rho = 0.1$ and $\varepsilon^{(0)} = 0$. The coefficients are $\beta = (4/\sqrt{5}, 4/\sqrt{5}, 4/\sqrt{5}, 4/\sqrt{5}, 4/\sqrt{5}, 0, \dots, 0)$. At each time step $t = 100, \dots, 200$, the model is re-trained and a prediction interval for $y^{(t)}$ based on $x^{(t)}$ is constructed using Algorithm 1 in Chernozhukov *et al.* (2018b) with $\alpha = 0.1$ and absolute residuals as score function. In particular, the algorithm chosen in this case is Lasso with 4 different penalty parameters. Under suitable conditions, Chernozhukov *et al.* (2018b) proves that the coverage is approximately valid.

The procedure is repeated 1000 times, and we report the empirical average over the number of replications. As can be seen in Figure C.4, the empirical counterpart of $\mathbb{E}[\sum_{k=1}^{K}(\phi_k^{(t)} - \mathbb{E}[\phi_k^{(t)}])(W_k^{(t)} - \mathbb{E}[W_k^{(t)}])]$, obtained as the empirical mean of the quantities across the replications, is around 0 for all iterations $t = 100, \dots, 200$. Although it is positive in some cases, it is only slightly above zero. The weights for each replication are obtained using the COMA method with an adaptive η, and on average they concentrate their mass in the penalty parameters λ_2 and λ_3. The coverage level for the COMA method is between levels $1 - 2\alpha$ and $1 - \alpha$.

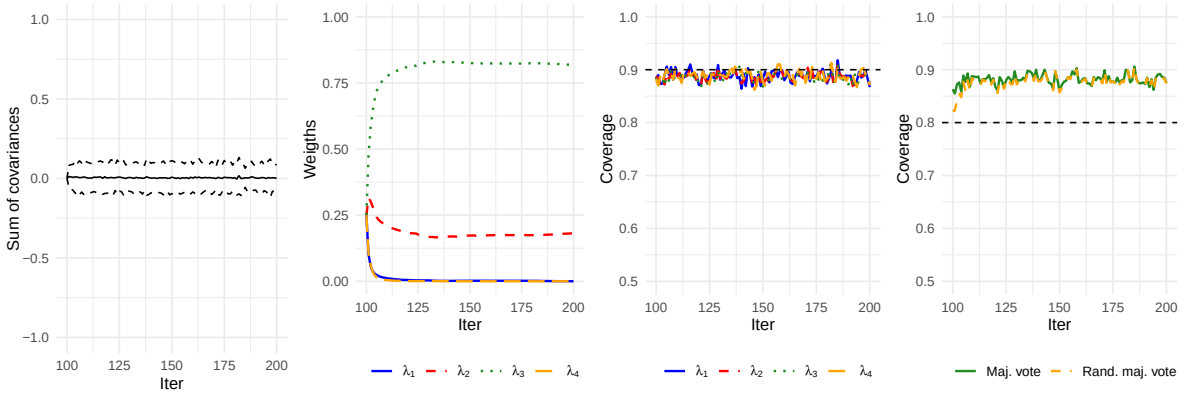

FIGURE C.4: In the first plot, the quantity $\mathbb{E}[\sum_{k=1}^{K}(\phi_k^{(t)} - \mathbb{E}[\phi_k^{(t)}])(W_k^{(t)} - \mathbb{E}[W_k^{(t)}])]$ is reported ($\pm$sd), for all $t = 100, \dots, 200$. In the second plot, the weights assumed by the various algorithms are reported. The last two plots represent the coverage of the different algorithms (third column) and the coverage of the COMA method with and without randomization (fourth column). The black dashed lines represent the levels $1 - \alpha$ and $1 - 2\alpha$.

C.6.2 Amazon stock prices

We replicate the analysis described in Section 6.5.2.2 using the time series of Amazon stock prices from January 1, 2006, to December 31, 2014.

The results, reported in Figure C.5, show that the empirical coverage remains close to the level $1 - \alpha$. In addition, the weights fluctuate across replications, illustrating that different methods yield varying results in different periods. The total covariance appears to be stable around zero across all iterations.

C.6.3 Additional experiments on real datasets

We evaluated our COMA method, described in Section 6.3, on multiple real datasets to assess its performance in terms of coverage and size. To construct the sets, we used split conformal prediction with absolute residuals as conformity scores. We considered two scenarios with $K = 6$ base models. In the first scenario, the models used are: a linear regression, a lasso regression with regularization parameter 0.1, a ridge regression with regularization parameter 0.1, a random forest with 250 trees, a neural network with 4 nodes, and a neural network with 8 nodes. In the second case, the models are three neural nets with $4, 6, 8$ nodes and three random forests with $50, 150, 250$ trees.

As benchmarks, we included prediction intervals from standard linear regression, split conformal prediction with an ensemble of the six models as base predictor (the ensemble is obtained by using a simple average), the union of the sets obtained from split conformal prediction, and the intersection of the sets obtained from split conformal prediction calibrated at level $\alpha' = 2\alpha/K$. For each dataset, we sampled 500 observations and applied the aforementioned methods using data from the 250-th observation onward. At each iteration, all models are retrained, and the results are reported in Table C.9 and Table C.10. In Table C.11 and in Table C.12, we report the empirical total correlation observed which is near zero for all experiments. Our COMA method never produces empty sets in the experiments and overall is the best method. The performance of COMA with fixed η and COMA with adaptive η can differ, and in some cases the fixed-η version outperforms the adaptive version. An explanation is that, after a few iterations, the adaptive-η version concentrates all the weight on the best model, whereas the fixed-η version distributes the weight across two or three algorithms. As a result, the randomized majority vote among two or three methods may align with the intersection, as discussed in the text.

Dataset	COMA Adapt. η	COMA $\eta = 1$	Ensemble	Linear Regression	Union	Bonf.
Crime	0.447	0.427	0.509	0.490	0.822	0.610
	0.900	0.888	0.928	0.972	0.984	0.940
Aquatic toxicity	0.990	0.956	0.974	1.040	1.290	1.266
	0.900	0.896	0.940	0.944	0.972	0.960
Diamonds	0.810	0.782	0.887	1.115	1.252	0.928
	0.892	0.884	0.908	0.900	0.968	0.936
Cement	23.104	23.665	28.814	33.864	41.944	28.042
	0.916	0.924	0.932	0.888	0.988	0.940
Productivity	0.487	0.477	0.524	0.484	0.672	0.633
	0.944	0.932	0.912	0.860	0.960	0.944

TABLE C.9: Empirical results of the different methods across multiple datasets. The base models are: lasso regression, ridge regression, linear model, random forest and two different neural nets. For each dataset, the first row reports the empirical set size, while the second row reports the empirical coverage.

Dataset	COMA Adapt. η	COMA $\eta = 1$	Ensemble	Linear Regression	Union	Bonf.
Crime	0.444	0.413	0.483	0.490	0.615	0.606
	0.916	0.896	0.932	0.972	0.980	0.980
Aquatic toxicity	0.910	0.876	0.980	1.040	1.260	1.238
	0.912	0.904	0.940	0.944	0.992	0.956
Diamonds	0.791	0.768	0.834	1.115	1.058	0.896
	0.892	0.888	0.932	0.900	0.988	0.940
Cement	21.535	23.999	24.968	33.864	35.670	26.208
	0.908	0.924	0.928	0.888	0.996	0.928
Productivity	0.431	0.416	0.501	0.524	0.633	0.569
	0.896	0.888	0.912	0.880	0.972	0.940

TABLE C.10: Empirical results of the different methods across multiple datasets. The base models are three random forests and three neural nets with different tuning parameters. For each dataset, the first row reports the empirical set size, while the second row reports the empirical coverage.

	Crime	Aquatic toxiticity	Diamonds	Cement	Productivity
COMA (Adapt. η)	-0.001	0.000	0.000	-0.003	0.001
COMA (η=1)	0.003	0.001	0.003	0.000	0.005

TABLE C.11: Empirical total correlation observed in the different datasets. The base models are: lasso regression, ridge regression, linear model, random forest and two different neural nets.

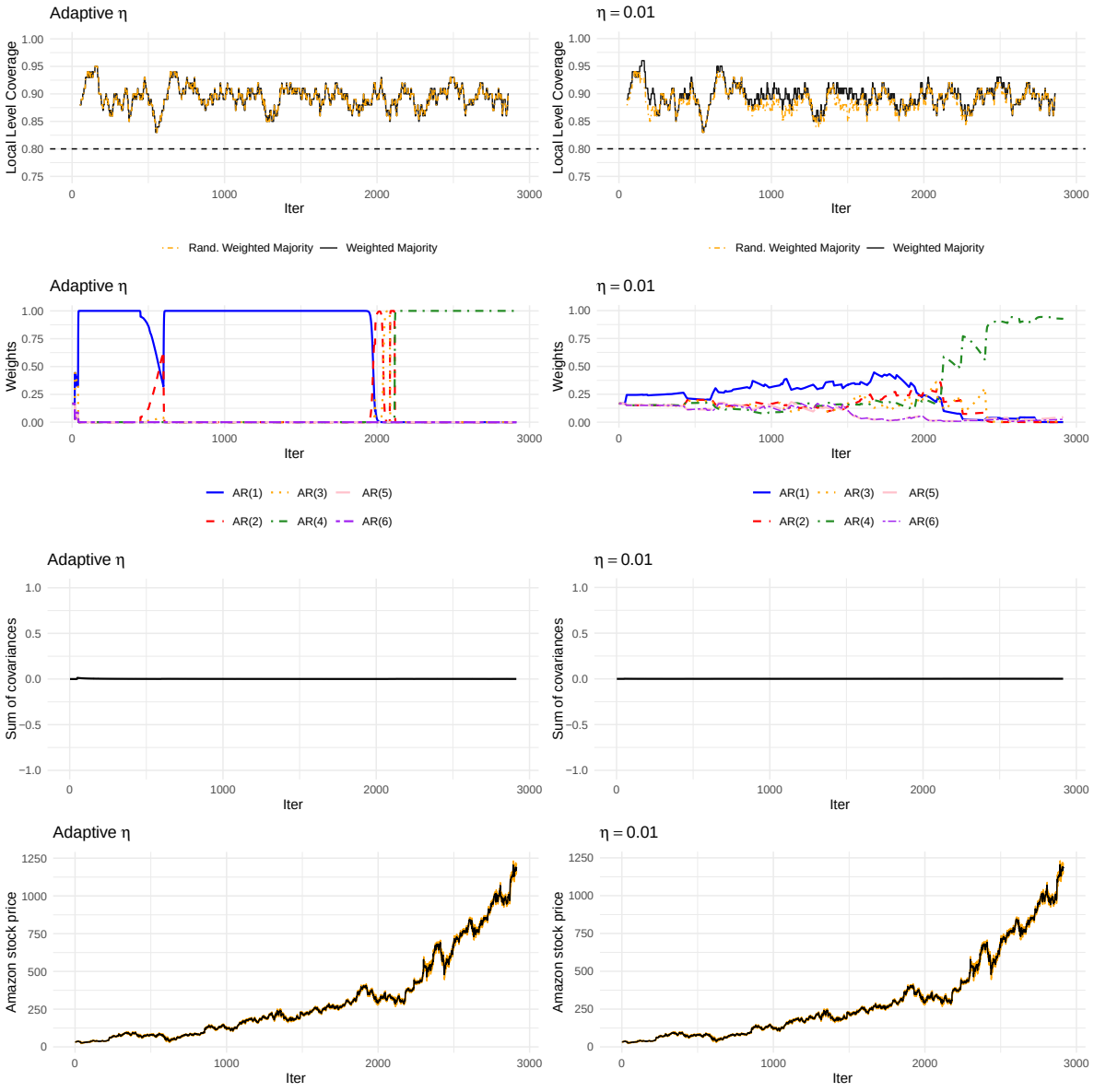

FIGURE C.5: Local level coverage obtained in a window containing 100 data points (first row), weights assumed during the different iterations by the various models (second row), empirical sum of the quantity $\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}(\phi_k^{(t)}-\bar{\phi}_k)(w_k^{(t)}-\bar{w}_k)$ obtained using an increasing number of observations (third row) and series of the original stock prices with corresponding prediction intervals (fourth row). The black dashed lines in the first row represent the level $1-2\alpha$. The results corresponding to the strategy with adaptive η are presented in the first column, while the results corresponding to the strategy with $\eta = 0.01$ are presented in the second column

	Crime	Aquatic toxicity	Diamonds	Cement	Productivity
COMA (Adapt. η)	0.006	-0.010	0.001	-0.006	-0.001
COMA (η=1)	0.002	-0.007	0.003	-0.012	-0.002

TABLE C.12: Empirical total correlation observed in the different datasets. The base models are three random forests and three neural nets with different tuning parameters.

Bibliography

Ajroldi, N., Diquigiovanni, J., Fontana, M. and Vantini, S. (2023) Conformal prediction bands for two-dimensional functional time series. *Computational Statistics & Data Analysis* **187**, 107821.

Angelopoulos, A., Candes, E. and Tibshirani, R. J. (2024a) Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23047–23074.

Angelopoulos, A. N., Barber, R. F. and Bates, S. (2024a) Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1616–1630. PMLR.

Angelopoulos, A. N., Barber, R. F. and Bates, S. (2025) Theoretical foundations of conformal prediction. arXiv preprint arXiv:2411.11824.

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

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L. and Schuster, T. (2024b) Conformal risk control. *The Twelfth International Conference on Learning Representations* .

Apple Inc. (2022) Differential Privacy Overview. `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`.

Azzalini, A. and Capitanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **65**(2), 367–389.

Azzalini, A. and Scarpa, B. (2012) *Data analysis and data mining: An introduction*. OUP USA.

Bach, P., Kurz, M. S., Chernozhukov, V., Spindler, M. and Klaassen, S. (2024) DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software* **108**(3), 1–56.

Banerjee, M., Durot, C. and Sen, B. (2019) Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics* **47**(2), 720 – 757.

Barber, R. F., Candès, E. J., Ramdas, A. and Tibshirani, R. J. (2021) Predictive inference with the jackknife+. *The Annals of Statistics* **49**(1), 486 – 507.

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

Basu, D. (1955) On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **15**(4), 377–380.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**(4), 1165–1188.

Berchialla, P., Lanera, C., Sciannameo, V., Gregori, D. and Baldi, I. (2022) Prediction of treatment outcome in clinical trials under a personalized medicine perspective. *Scientific Reports* **12**(1), 4115.

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

Boland, P. J., Singh, H. and Cukic, B. (2002) Stochastic orders in partition and random testing of software. *Journal of Applied Probability* **39**(3), 555–565.

Bonferroni, C. E. (1936) *Teoria statistica delle classi e calcolo delle probabilità*. Annuario dell'Istituto Superiore di Scienze Economiche e Commerciali.

Boström, H., Linusson, H., Löfström, T. and Johansson, U. (2017) Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence* **81**, 125–144.

Breiman, L. (1996) Stacked regressions. *Machine Learning* **24**, 49–64.

Carlsson, L., Eklund, M. and Norinder, U. (2014) Aggregated conformal prediction. In *Artificial Intelligence Applications and Innovations (AIAI)*, pp. 231–240.

Casella, G. and Berger, R. (2001) *Statistical inference*. Second edition edition. Chapman and Hall/CRC.

Casella, G., Hwang, J. T. G. and Robert, C. (1993) A paradox in decision-theoretic interval estimation. *Statistica Sinica* **3**(1), 141–155.

Chen, W., Chun, K.-J. and Barber, R. F. (2018) Discretized conformal prediction for efficient distribution-free inference. *Stat* **7**(1), e173.

Chen, Y., Liu, P., Tan, K. S. and Wang, R. (2023) Trade-off between validity and efficiency of merging p-values under arbitrary dependence. *Statistica Sinica* **33**(2), 851–872.

Chen, Y., Wang, R., Wang, Y. and Zhu, W. (2025) Subuniformity of harmonic mean p-values. *Canadian Journal of Statistics* p. e70017.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018a) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, 1–68.

Chernozhukov, V., Wüthrich, K. and Yinchu, Z. (2018b) Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pp. 732–749.

Cherubin, G. (2019) Majority vote ensembles of conformal predictors. *Machine Learning* **108**(3), 475–488.

Chi, Z., Ramdas, A. and Wang, R. (2025) Multiple testing under negative dependence. *Bernoulli* **31**(2), 1230–1255.

Choi, W. and Kim, I. (2023) Averaging p-values under exchangeability. *Statistics & Probability Letters* **194**, 109748.

Cousins, R. D. (2007) Annotated bibliography of some papers on combining significances or p-values. arXiv preprint arXiv:0705.2209.

Cox, D. R. (1975) A note on data-splitting for the evaluation of significance levels. *Biometrika* **62**(2), 441–444.

Cox, D. R., Spjøtvoll, E., Johansen, S., van Zwet, W. R., Bithell, J., Barndorff-Nielsen, O. and Keuls, M. (1977) The role of significance tests [with discussion and reply]. *Scandinavian Journal of Statistics* pp. 49–70.

Dean, A. and Verducci, J. (1990) Linear transformations that preserve majorization, schur concavity, and exchangeability. *Linear algebra and its applications* **127**, 121–138.

Decrouez, G. and Hall, P. (2014) Split sample methods for constructing confidence intervals for binomial and poisson parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**(5), 949–975.

Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016) Sub-Gaussian mean estimators. *The Annals of Statistics* **44**(6), 2695 – 2725.

DiCiccio, C. J., DiCiccio, T. J. and Romano, J. P. (2020) Exact tests via multiple data splitting. *Statistics & Probability Letters* **166**, 108865.

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

Efron, B. (2010) *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

Efron, B. and Tibshirani, R. J. (1994) *An introduction to the bootstrap*. Chapman and Hall/CRC.

Falk, R. W. (1989) Hommel's bonferroni-type inequality for unequally spaced levels. *Biometrika* **76**(1), 189–191.

Fernandes, K., Vinagre, P., Cortez, P. and Sernadela, P. (2015) Online News Popularity. UCI Machine Learning Repository.

Fisher, R. (1948) Combining independent tests of significance. *The American Statistician* pp. 2–30.

Fisher, R. A. (1934) *Statistical methods for research workers*. Number 5. Oliver and Boyd.

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

Freund, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**(1), 119–139.

Gasparin, M. and Ramdas, A. (2024a) Conformal online model aggregation. arXiv preprint arXiv:2403.15527.

Gasparin, M. and Ramdas, A. (2024b) Merging uncertainty sets via majority vote. arXiv preprint arXiv:2401.09379.

Gasparin, M. and Ramdas, A. (2025) Improving the statistical efficiency of cross-conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18848–18867. PMLR.

Gasparin, M., Wang, R. and Ramdas, A. (2025) Combining exchangeable p-values. *Proceedings of the National Academy of Sciences* **122**(11), e2410849122.

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

Gibbs, I. and Candès, E. J. (2024) Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research* **25**(162), 1–36.

Goeman, J. and Solari, A. (2011) Multiple testing for exploratory research. *Statistical Science* **26**(4), 584–597.

Grünwald, P., de Heide, R. and Koolen, W. (2024) Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology: Statistical Methodology* **86**(5), 1091–1128.

Gui, L., Jiang, Y. and Wang, J. (2025) Aggregating dependent signals with heavy-tailed combination tests. *Biometrika* .

Guo, F. R. and Shah, R. D. (2024) Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology* .

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

Harries, M. (2003) Splice-2 comparative evaluation: Electricity pricing. Technical report, The University of South Wales.

Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R. J. (2022) Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics* **50**(2), 949–986.

Hommel, G. (1983) Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal* **25**(5), 423–430.

Hore, R. and Barber, R. F. (2025) Conformal prediction with local weights: Randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology: Statistical Methodology* p. 549–578.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021) Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* **49**(2), 1055 – 1080.

Humbert, P., Le Bars, B., Bellet, A. and Arlot, S. (2023) One-shot federated conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 14153–14177. PMLR.

Ignatiadis, N., Wang, R. and Ramdas, A. (2024) E-values as unnormalized weights in multiple testing. *Biometrika* **111**(2), 417–439.

Kennedy, E. H. (2016) Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research* pp. 141–167.

Kim, B., Xu, C. and Barber, R. (2020) Predictive inference is free with the jackknife+-after-bootstrap. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4138–4149.

Kim, I. and Ramdas, A. (2024) Dimension-agnostic inference using cross U-statistics. *Bernoulli* **30**(1), 683 – 711.

Koenker, R. (2005) *Quantile regression*. Volume 38. Cambridge: Cambridge University press.

Kost, J. T. and McDermott, M. P. (2002) Combining dependent p-values. *Statistics & Probability Letters* **60**(2), 183–190.

Kuchibhotla, A. K. (2020) Exchangeability, conformal prediction, and rank tests. arXiv preprint arXiv:2005.06095.

Kuchibhotla, A. K., Balakrishnan, S. and Wasserman, L. (2023) Median regularity and honest inference. *Biometrika* **110**(3), 831–838.

Kuchibhotla, A. K., Balakrishnan, S. and Wasserman, L. (2024) The HulC: Confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology* p. 586–622.

Kuncheva, L. I. (2014) *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A. and Duin, R. P. (2003) Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications* **6**, 22–31.

Lehmann, E. L., Romano, J. P. and Casella, G. (1986) *Testing Statistical Hypotheses*. Volume 3. Springer.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018) Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113**(523), 1094–1111.

Lei, L. and Sudijono, T. (2024) Inference for synthetic controls via refined placebo tests. arXiv preprint arXiv:2401.07152.

Leiner, J., Duan, B., Wasserman, L. and Ramdas, A. (2025) Data fission: Splitting a single data point. *Journal of the American Statistical Association* **120**(549), 135–146.

Linusson, H., Johansson, U. and Boström, H. (2020) Efficient conformal predictor ensembles. *Neurocomputing* **397**, 266–278.

Linusson, H., Norinder, U., Boström, H., Johansson, U. and Löfström, T. (2017) On the calibration of aggregated conformal predictors. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60, pp. 154–173. PMLR.

Littlestone, N. and Warmuth, M. (1994) The weighted majority algorithm. *Information and Computation* **108**(2), 212–261.

Lugosi, G. and Mendelson, S. (2019) Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* **19**(5), 1145–1190.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**(3), 655–660.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009) P-values for high-dimensional regression. *Journal of the American Statistical Association* **104**(488), 1671–1681.

Minsker, S. (2023) U-statistics of growing order and sub-Gaussian mean estimators with sharp constants. *Mathematical Statistics and Learning* .

Morgenstern, D. (1980) Berechnung des maximalen signifikanzniveaus des testes "Lehne H_0 ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika* **27**, 285–286.

Nemirovskij, A. S. and Yudin, D. B. (1983) *Problem complexity and method efficiency in optimization*. Wiley-Interscience.

Nguyen, C. (2018) S&P 500 stock data. `https://www.kaggle.com/datasets/camnugent/sandp500`.

Owen, A. B. (2009) Karl Pearson's meta-analysis revisited. *The Annals of Statistics* **37**(6B), 3867 – 3892.

Pace, L. and Salvan, A. (1997) *Principles of statistical inference: from a Neo-Fisherian perspective*. Volume 4. World scientific.

Papadopoulos, H., Proedrou, K., Vovk, V. and Gammerman, A. (2002) Inductive confidence machines for regression. In *European conference on machine learning (ECML)*, pp. 345–356.

Pearson, K. (1934) On a new method of determining "Goodness of Fit". *Biometrika* **26**(4), 425–442.

Pesarin, F. and Salmaso, L. (2010) *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Politis, D. N., Romano, J. P. and Wolf, M. (1999) *Subsampling*. Springer.

Prinster, D., Liu, A. and Saria, S. (2022) Jaws: Auditing predictive uncertainty under covariate shift. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35907–35920.

Ramdas, A., Grünwald, P., Vovk, V. and Shafer, G. (2023) Game-theoretic statistics and safe anytime-valid inference. *Statistical Science* **38**(4), 576 – 601.

Ramdas, A. and Manole, T. (2024) Randomized and Exchangeable Improvements of Markov's, Chebyshev's and Chernoff's inequalities. *Statistical Science* **forthcoming**.

Ramdas, A. and Wang, R. (2025) *Hypothesis Testing with E-values*. Volume 1.

Ritzwoller, D. M. and Romano, J. P. (2023) Reproducible aggregation of sample-split statistics. arXiv preprint arXiv:2311.14204.

Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**(2), 143–155.

Romano, Y., Patterson, E. and Candes, E. (2019) Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553.

de Rooij, S., van Erven, T., Grünwald, P. D. and Koolen, W. M. (2014) Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research* **15**(37), 1281–1316.

Rüger, B. (1978) Das maximale signifikanzniveau des tests: "Lehne H_0 ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika* **25**, 171–178.

Rüschendorf, L. (1982) Random variables with maximum sums. *Advances in Applied Probability* **14**(3), 623–632.

Sarkar, S. K. (1998) Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *The Annals of Statistics* **26**(2), 494–504.

Sarkar, S. K. (2008) On the Simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1, pp. 231–243. Institute of Mathematical Statistics.

Saunders, C., Gammerman, A. and Vovk, V. (1999) Transduction with confidence and credibility. *Proceedings of the 16th international joint conference on Artificial intelligence (IJCAI)* pp. 722–726.

Sellke, T., Bayarri, M. J. and Berger, J. O. (2001) Calibration of p-values for testing precise null hypotheses. *The American Statistician* **55**(1), 62–71.

Severini, T. A. (2000) *Likelihood methods in statistics*. Oxford University Press.

Shafer, G. (2021) Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society* **184**(2), 407–431.

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

Shekhar, S., Kim, I. and Ramdas, A. (2022) A permutation-free kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18168–18180.

Shekhar, S., Kim, I. and Ramdas, A. (2023) A permutation-free kernel independence test. *Journal of Machine Learning Research* **24**(369), 1–68.

Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**(3), 751–754.

Solari, A. and Djordjilović, V. (2022) Multi-split conformal prediction. *Statistics & Probability Letters* **184**, 109395.

Steinberger, L. and Leeb, H. (2023) Conditional predictive inference for stable algorithms. *The Annals of Statistics* **51**(1), 290 – 311.

Stevens, W. L. (1950) Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37**(1/2), 117–129.

Stutz, D., Roy, A. G., Matejovicova, T., Strachan, P., Cemgil, A. T. and Doucet, A. (2023) Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research* .

Szabadváry, J. H. and Löfström, T. (2026) Beyond conformal predictors: Adaptive conformal inference with confidence predictors. *Pattern Recognition* **170**, 111999.

Tang, W. and Tang, F. (2023) The poisson binomial distribution — Old and new. *Statistical Science* **38**(1), 108 – 119.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288.

Tibshirani, R. J., Barber, R. F., Candes, E. and Ramdas, A. (2019) Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2530–2540.

Tippett, L. (1941) *The Methods of Statistics: An Introduction Mainly for Experimentalists*. Williams and Norgate.

Tsanas, A. and Little, M. (2009) Parkinsons Telemonitoring. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5ZS3N.

VanderWeele, T. (2015) *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Vovk, V. (1993) A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **55**(2), 317–351.

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

Vovk, V. (2015) Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* **74**, 9–28.

Vovk, V., Gammerman, A. and Shafer, G. (2005) *Algorithmic Learning in a Random World*. Volume 29. Springer.

Vovk, V., Gammerman, A. and Shafer, G. (2022a) *Algorithmic Learning in a Random World*. Second edition. Springer Nature.

Vovk, V., Nouretdinov, I., Manokhin, V. and Gammerman, A. (2018) Cross-conformal predictive distributions. In *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pp. 37–51. PMLR.

Vovk, V., Wang, B. and Wang, R. (2022b) Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics* **50**(1), 351–375.

Vovk, V. and Wang, R. (2020) Combining p-values via averaging. *Biometrika* **107**(4), 791–808.

Vovk, V. and Wang, R. (2021) E-values: Calibration, combination and applications. *The Annals of Statistics* **49**(3), 1736–1754.

Wang, B. and Wang, R. (2016) Joint mixability. *Mathematics of Operations Research* **41**(3), 808–826.

Wang, R. (2024) Testing with p*-values: Between p-values, mid p-values, and e-values. *Bernoulli* **30**(2), 1313–1346.

Wang, R. (2025) The only admissible way of merging arbitrary e-values. *Biometrika* **112**(2).

Wang, R. and Ramdas, A. (2022) False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(3), 822–852.

Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020) Universal inference. *Proceedings of the National Academy of Sciences* **117**(29), 16880–16890.

Wasserman, L. and Roeder, K. (2009) High-dimensional variable selection. *The Annals of Statistics* **37**(5A), 2178–2201.

Waudby-Smith, I., Wu, S. and Ramdas, A. (2023) Nonparametric extensions of randomized response for private confidence sets. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 36748–36789. PMLR.

Westfall, P. H., Kropf, S. and Finos, L. (2004) Weighted FWE-controlling methods in high-dimensional situations. *Lecture Notes-Monograph Series* pp. 143–154.

Wilson, D. J. (2019) The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* **116**(4), 1195–1200.

Xu, Z. and Ramdas, A. (2023) More powerful multiple testing under dependence via randomization. arXiv preprint arXiv:2305.11126.

Xu, Z., Solari, A., Fischer, L., de Heide, R., Ramdas, A. and Goeman, J. (2025) Bringing closure to false discovery rate control: A general principle for multiple testing. arXiv preprint arXiv:2509.02517.

Xu, Z., Wang, R. and Ramdas, A. (2024) Post-selection inference for e-value based confidence intervals. *Electronic Journal of Statistics* **18**(1), 2292–2338.

Zaffran, M., Féron, O., Goude, Y., Josse, J. and Dieuleveut, A. (2022) Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866.