THE UNIVERSITY OF CHICAGO

PROBLEMS IN MODERN INFERENCE: DISTRIBUTION FREE PREDICTION AND GOODNESS OF FIT TESTING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
AABESH BHATTACHARYYA

CHICAGO, ILLINOIS
AUGUST 2026

To my parents and my brother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I must begin by expressing my deepest gratitude to my advisor, Professor Rina Foygel Barber, without whose support and guidance this thesis would never have come to fruition. When I first joined the PhD program at UChicago, I was, unlike many of my friends and peers, quite unsure about the kind of research I wanted to pursue or how to choose an advisor. It all felt like a game I didn't know how to play. As someone with barely any experience in research, the few research papers I had read until that point had seemed either uninteresting or too difficult for me to comprehend.

Towards the end of my first year, I attended a mini-seminar in which the then-second-year students presented their research. It was there, through presentations by Rohan Hore and Yu Gui, who were both working with Rina, that I first learned about the field of conformal prediction. For the first time, I felt that I had encountered an area of research that genuinely interested me, perhaps because of the broad applicability of the problems despite the remarkable simplicity of their formulations and goals. I began working with Rina soon afterwards, and I have been immensely fortunate to have had her as my advisor for the remainder of my PhD.

As someone who often takes time to absorb new ideas, I have sometimes felt that my progress was slower and less impressive than I would have liked, but she has always been extraordinarily patient with me. I have learned an enormous amount from every meeting with her, often spending days afterward only gradually understanding what she had been trying to hint at about the problem we were working on. Her sheer brilliance in solving problems, her ability to devise entirely new techniques, and, above all, her research originality have deeply inspired me. Whenever I became stuck in my research, she would somehow find a way forward. This gave me the courage to pursue research without fearing the unknown or worrying that my efforts might ultimately lead nowhere, because I knew that, whenever I found myself facing a locked door, she would have a key to open it. She is undoubtedly one

of the most intelligent people I have ever encountered, and I feel incredibly fortunate to call her my advisor.

I am grateful to my committee members, Professor Nikos Ignatiadis and Professor Cong Ma, for their guidance and for the thoughtful questions they asked during my proposal and defense, which encouraged me to think more deeply about my research. I thank Professor Mei Wang, who has always been exceptionally kind and generous with her help, as well as Keisha Prowoznick, whose timely reminders about deadlines and administrative requirements helped me stay on track and allowed me to focus more fully on my research. I would also like to thank the members of Rina's research group for the many engaging group meetings, from which I learned a great deal. I am especially grateful to Rohan Hore, who was immensely helpful when I was just beginning my research. I often discussed my work with him, and I benefited greatly from his thoughtful advice.

I was fortunate to collaborate with Professor Lucas Janson on a project, and his insights were invaluable. My other collaborators included Ritwik Bhaduri, a friend from my undergraduate years, Tiffany Ding, and Boxuan Zhang, and I learned a great deal from each of them. My collaboration with Ritwik was especially enjoyable. The many hurdles posed by the research problem often felt far more manageable because of our discussions about its different aspects. Even spending long hours on the project rarely felt tiring; instead, it felt like learning something new with a friend, making mistakes together, and then finding ways to correct them. As someone who is often hesitant to voice a thought for fear that it might sound foolish, I particularly appreciated the open and judgment-free space we shared, where we could freely explore any possible research idea.

Life in Chicago over the past five years has been rewarding in countless ways, and I look back on this time with immense gratitude. Life is incomplete without friends, and I have been fortunate to form many meaningful relationships along the way. First and foremost, I would like to thank my partner, Tannistha, who is also my best friend, for being a constant source

of strength and for standing by me, especially during the difficult periods when I was not the best version of myself. I still remember the day in February 2023 when she received her offer letter from UChicago. I received a call from her, it was four or five in the morning in India, and she happened to be awake, as she often stays up very late. We were both overjoyed that we would finally be able to live in the same city and even be part of the same department for at least the next three years. We were incredibly fortunate in that regard, and the years that followed were nothing short of exhilarating, filled with countless adventures and cherished memories.

I would like to thank my roommates and two of my closest friends in Chicago, Tamsuk and Soumik. When I first moved to Chicago in 2021, thousands of miles away from home, I barely knew anyone and had little hope of finding a close-knit group of friends. Tamsuk was the first person who made me feel the warmth of friendship in this unfamiliar place. We often went for walks by Lake Michigan, sometimes late at night despite the many warnings we had received about Chicago being unsafe. I still remember Halloween in 2021, when we were walking around campus and she opened up to me about some personal aspects of her life. As we began sharing more about ourselves, our bond gradually grew stronger. I will always cherish our spontaneous evening and late-night hangouts at Jimmy's. She is truly one of the most loyal friends anyone could have.

I deeply admire Soumik for his many wonderful qualities, several of which have inspired me. I especially value the fact that we can talk just as easily about completely nonsensical and silly things as we can about deeply personal matters. Soumik also has a remarkable ability to make any group gathering entertaining, whether through hilariously outrageous antics such as prank calls or stories about his seemingly endless list of crushes. A fun fact is that he is the only person who has been my roommate throughout my entire time in Chicago.

Tamsuk, Soumik, and Tannistha have been my partners in crime through everything. We share the same spontaneity and enthusiasm for life: we are crazy enough to rent a car and

drive for an hour at midnight simply to have tea at Chai Ho Jai or to wander around the UChicago campus late at night. I will dearly miss these impromptu plans and the many memories we created together.

I am also immensely grateful to the many other friends I have made in Chicago—Sulagna Di, Pratyasha, Atreyie Di, Supriyo Da, Debaleen Da, Mandira Di, Shreya, Sunandita Di, Anchita Di, Anish Da, Gauri and Scarlett to name just a few—with whom I have created countless wonderful memories through game nights, potlucks, movie nights, and spontaneous addas that often continued for hours. I have also cherished the many trips I have taken with Tannistha, Tamsuk, Soumik, Pratyasha, Shreya and Sulagna Di. During our time here, some of my friends and I founded Kheyal, the first Bengali students' association at the university, to celebrate our culture and remain connected to our roots through food, music, events, and gatherings. Kheyal has brought many of us closer together and given us a small piece of home while living far away from it.

Apart from them, several of my friends in the department have been an integral part of my PhD journey. I have probably spent the most time with Amber and Annie, with whom I share a wonderful bond. I have always enjoyed our time together, whether over lunch or dinner on campus or elsewhere, at the movies, or during our many spontaneous hangouts. I would also like to thank Cheuk, Sean, Kulunu, Raphael, and many others. I have also shared many memorable moments with Soham, Anirban Da, Prasun, and Chandramouli, especially during our late-afternoon ping-pong breaks in the Physics building or at the John Crerar Library. Soham was the only other person from ISI in my cohort and was also my roommate when I first arrived. Although both of us were novice cooks, we shared a deep passion for preparing elaborate dishes and spent a great deal of time experimenting in the kitchen. I only wish I had met Anirban Da sooner; we began hanging out towards the very end of my PhD, but even within that short period, we made many wonderful memories.

Outside Chicago, many of my friends have always had my back. Sayan, Souvik, and

Anuraag have been my friends since high school, and we proudly boast that ours is one of the very few high school friend groups whose members still talk and debate almost every day about a wide range of topics, including politics, social issues, and sports. Although the four of us now live in different cities across three continents, we make it a point to take a trip together every year—an occasion we all eagerly anticipate. Whenever I visited home, another highlight was spending time with my neighbor and childhood friend Rahul, going on our daily walks, and stopping for street food and tea.

I am also fortunate to have several other close high school friends—Souradipto, Pritha, Soumyarup, and Anwesha—who live in the United States. I have always cherished my bond with Souradipto and Pritha, and they have been a constant source of emotional support. It is deeply reassuring to know that I can always count on them. I must also mention Shaswata, who has been my friend since around third grade in junior school. I feel extremely lucky that he now lives in Champaign, just a two-hour drive from Chicago. We have met several times in Chicago, and our bond remains just as strong as it was before. Spending time with him always brings back fond memories of the wonderful years we shared in school. I am also grateful to Ritoban, who has always been a kind and helpful friend. We speak every few days, either to gossip about random things, to discuss academics or our recent trips. We also share a love for trying different sports, and I hope that now that we will be living closer to each other we will get more opportunities to play together.

I would also like to thank Sumit Kaku, my only close relative living in North America, who warmly showed me around Toronto when I visited the city for JSM in 2023.

Finally, I would like to thank Chandrima Bandyopadhay, my mother; Barun Bhattacharyya, my father; and Aneesh Bhattacharyya, my brother. No words can adequately express the love and support that they have given me. I am the person I am today because of them, and I owe every bit of my success to their sacrifices and encouragement. I feel especially fortunate to have an elder brother who made my own journey smoother by experiencing many

of life's stages before me and allowing me to learn from his example. Over the years, my mother has also become my best friend. Sharing so many aspects of our lives has brought us even closer and given me immense strength and courage. My father ensured that we always had everything we needed, allowing us to focus wholeheartedly on our studies and pursue our goals. I am deeply grateful to my family for everything they have done for me.

ABSTRACT

This dissertation studies two important directions in modern statistical inference: distribution-free prediction and hypothesis testing. These directions are increasingly important in an era where complex machine-learning algorithms are widely deployed, but their statistical behavior is often difficult to characterize. In such settings, classical guarantees based on correctly specified parametric models can be unreliable or too restrictive. Distribution-free prediction aims to construct prediction sets that are valid without assumptions on the fitted model or the underlying distribution, typically under exchangeability, while modern goodness-of-fit testing seeks flexible procedures that can detect structured deviations from complex null models while maintaining rigorous type-I error control. The central theme of this dissertation is to develop methods that preserve finite-sample or approximate frequentist guarantees while remaining useful in realistic settings.

The first part of the dissertation contributes to conformal prediction. Conformal prediction constructs prediction sets around the output of a fitted model with coverage guarantees that do not rely on the model being correct. Standard full and split conformal methods, however, assume exchangeability between training and test data. We study two challenges that arise in practice. First, we consider deviations from exchangeability caused by covariate shift between the training and test distributions, focusing on the setting where the shift is explained by a discrete covariate. In this group-weighted setting, we show that existing guarantees for weighted conformal prediction can be substantially sharpened, especially when the number of groups is large. We then extend this perspective to label-weighted conformal prediction for settings with rare classes, showing how to obtain statistically efficient guarantees for macro-coverage. Second, we address the tradeoff between statistical efficiency and computation faced by full conformal prediction or other conformal prediction approaches. We develop a tournament-correction framework that converts a broad class of approximations to full conformal prediction into procedures with rigorous distribution-free coverage guarantees.

The second part of the dissertation contributes to flexible goodness-of-fit testing. We develop approximate co-sufficient sampling via Bayes (aCSS-B), a method for generating artificial data sets that are approximately exchangeable with the observed data under a composite null hypothesis. This enables valid p-values using essentially arbitrary test statistics, allowing the test statistic to be tailored to the scientific alternative of interest rather than restricted to generic likelihood-based summaries. We prove approximate exchangeability for the proposed samples and derive approximate type-I error control for the resulting tests.

CHAPTER 1

INTRODUCTION

This chapter provides background for the two inferential themes studied in this dissertation: distribution-free prediction and goodness-of-fit testing. We begin with conformal prediction, reviewing the split conformal and full conformal procedures for constructing distribution-free prediction sets with marginal coverage guarantees. We also review weighted conformal prediction, which extends conformal methods to certain forms of covariate shift.

We then turn to goodness-of-fit testing, with a focus on resampling-based approaches. These methods are useful because they avoid the need to derive the null distribution of a test statistic analytically and allow the statistic to be chosen flexibly. This background sets up the conformal prediction contributions in Chapters 2–4 and the goodness-of-fit testing contribution in Chapter 5.

1.1 Distribution-free predictive inference

Prediction is one of the central tasks in statistics and machine learning. Given observed data, the goal is not only to produce a point prediction for a future response, but also to quantify the uncertainty in that prediction. In many applications, a point prediction alone is insufficient: for example, a prediction system used in medicine, finance, or scientific discovery should also communicate the uncertainty in its prediction. Prediction intervals or prediction sets provide a natural way to express this uncertainty.

A major difficulty is that uncertainty quantification is often model-dependent. Classical prediction intervals are typically derived under assumptions such as linearity, Gaussian noise, or a correctly specified likelihood. In modern applications, however, the fitted predictor may be very complicated, and in some cases we may not even know the underlying architecture of the fitted model. This motivates treating the prediction rule as a black box. Conformal prediction

is model agnostic and provides prediction sets in a distribution-free regime, requiring no assumptions on the fitted model and only requiring exchangeability of the relevant data points.

1.1.1 Conformal prediction

Conformal prediction is a general framework for constructing prediction sets with finite-sample validity under exchangeability (Vovk et al., 2005; Papadopoulos et al., 2002; Papadopoulos, 2008; Lei et al., 2018; Angelopoulos and Bates, 2023). Let the observed data be

$$Z_i = (X_i, Y_i), \qquad i = 1, \ldots, N,$$

where $X_i \in \mathcal{X}$ is a feature vector and $Y_i \in \mathcal{Y}$ is a response. The test point is denoted

$$Z_{N+1} = (X_{N+1}, Y_{N+1}),$$

where the test feature X_{N+1} is observed, while the response Y_{N+1} is unobserved. Given a target miscoverage level $\alpha \in (0, 1)$, the goal is to construct a prediction set $\widehat{C}(X_{N+1}) \subseteq \mathcal{Y}$ satisfying the marginal coverage guarantee

$$\mathbb{P}\{Y_{N+1} \in \widehat{C}(X_{N+1})\} \geq 1 - \alpha.$$

The key feature of conformal prediction is that this guarantee does not require the fitted prediction model to be correct.

A conformal method is built from a score function

$$s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R},$$

where larger values of $s(x, y)$ indicate that the candidate pair (x, y) is more unusual. For

instance, in regression, a common choice is the absolute residual score

$$s(x, y) = |y - \widehat{f}(x)|,$$

where $\widehat{f}$ is a fitted regression function. In classification, the score may be based on the probability assigned by a fitted classifier to the candidate label.

The simplest computationally efficient version is split conformal prediction, also called inductive conformal prediction. The observed data are split into a training set and a calibration set. Let

$$\mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{cal}} = \{1, \ldots, N\}, \qquad \mathcal{I}_{\text{train}} \cap \mathcal{I}_{\text{cal}} = \varnothing,$$

and let $n_{\text{cal}} = |\mathcal{I}_{\text{cal}}|$. The training set is used to fit the predictive model and hence the score function s. The calibration set is then used to compute the calibration scores

$$S_i = s(X_i, Y_i), \qquad i \in \mathcal{I}_{\text{cal}}.$$

Let

$$\widehat{q} = \text{Quantile}_{(1-\alpha)(1+1/n_{\text{cal}})} \left(\{S_i : i \in \mathcal{I}_{\text{cal}}\} \right),$$

where the factor $(1 + 1/n_{\text{cal}})$ is the usual finite-sample conformal correction. The split conformal prediction set is

$$\widehat{C}_{\text{split}}(x) = \{y \in \mathcal{Y} : s(x, y) \leq \widehat{q}\}.$$

After conditioning on the training data, so that s can be treated as fixed, if

$$\{(X_i, Y_i) : i \in \mathcal{I}_{\text{cal}}\} \quad \text{and} \quad (X_{N+1}, Y_{N+1})$$

are exchangeable, then split conformal prediction satisfies

$$\mathbb{P}\{Y_{N+1} \in \widehat{C}_{\text{split}}(X_{N+1})\} \geq 1 - \alpha.$$

Thus, when the calibration and test data are drawn from the same distribution, split conformal prediction provides a finite-sample marginal coverage guarantee. Its main advantage is computational simplicity: the model is fitted only once.

Full conformal prediction avoids the sample-splitting step. For each candidate value $y \in \mathcal{Y}$, define the augmented data set

$$\mathcal{D}^y_{N+1} = \{(X_1, Y_1), \ldots, (X_N, Y_N), (X_{N+1}, y)\}.$$

The model is refitted on $\mathcal{D}^y_{N+1}$, and conformity scores are computed for the observed data points and for the candidate test point:

$$S^y_i = s(Z_i; \mathcal{D}^y_{N+1}), \qquad i = 1, \ldots, N,$$

and

$$S^y_{N+1} = s((X_{N+1}, y); \mathcal{D}^y_{N+1}).$$

The candidate value y is included in the prediction set if its score is not larger than the corrected empirical quantile of the scores of the observed data points. Thus, the full conformal prediction set is

$$\widehat{C}_{\text{full}}(X_{N+1}) = \left\{ y \in \mathcal{Y} : S^y_{N+1} \leq \text{Quantile}_{(1-\alpha)(1+1/N)} \left(S^y_1, \ldots, S^y_N \right) \right\}.$$

Because the augmented data set treats the candidate test point symmetrically with the observed data points, full conformal prediction gives a finite-sample, distribution-free marginal

coverage guarantee under exchangeability:

$$\mathbb{P}\{Y_{N+1} \in \widehat{C}_{\text{full}}(X_{N+1})\} \geq 1 - \alpha.$$

Full conformal prediction is often statistically efficient because it uses all observed data points symmetrically. However, it is computationally expensive: for each candidate value y, one must refit or recompute the model on the augmented data set.

1.1.2 Weighted conformal prediction

Standard conformal prediction assumes that the calibration and test data are exchangeable. Weighted conformal prediction extends the framework to certain forms of distribution shift, especially covariate shift (Tibshirani et al., 2019). Suppose the calibration data are drawn from a distribution P, while the test point is drawn from a distribution Q, with

$$P_{Y|X} = Q_{Y|X},$$

but possibly $P_X \neq Q_X$. If the likelihood ratio

$$w(x) \propto \frac{\mathsf{d}Q_X}{\mathsf{d}P_X}(x)$$

is known, then the calibration scores can be reweighted to represent the test distribution.

Weighted conformal prediction is usually described in the split conformal framework. As above, let $\mathcal{I}_{\text{cal}}$ denote the calibration indices and let $m = |\mathcal{I}_{\text{cal}}|$. Assume the score function s has been fit on data independent of the calibration set and the test point. Given calibration scores

$$S_i = s(X_i, Y_i), \qquad i \in \mathcal{I}_{\text{cal}},$$

weighted conformal prediction computes, for a test feature value x,

$$\widehat{q}_w(x) = \text{Quantile}_{1-\alpha}\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \frac{w(X_i)}{\sum_{j\in\mathcal{I}_{\text{cal}}} w(X_j) + w(x)}\delta_{S_i} + \frac{w(x)}{\sum_{j\in\mathcal{I}_{\text{cal}}} w(X_j) + w(x)}\delta_{+\infty}\right).$$

The weighted conformal prediction set is then

$$\widehat{C}_{\text{WCP}}(x) = \{y \in \mathcal{Y} : s(x,y) \le \widehat{q}_w(x)\}.$$

When the weight function is known exactly, weighted conformal prediction restores marginal coverage under covariate shift:

$$\mathbb{P}_{P^m\times Q}\{Y_{N+1} \in \widehat{C}_{\text{WCP}}(X_{N+1})\} \ge 1-\alpha.$$

Here the probability is taken over calibration data drawn i.i.d. from P and an independent test point drawn from Q, after conditioning on the data used to fit the score function. Thus, WCP preserves the distribution-free spirit of conformal prediction while allowing an important deviation from exchangeability.

Our contribution to conformal prediction. The conformal prediction part of this dissertation studies three ways in which distribution-free predictive inference can be made more useful in modern applications. Chapter 2 considers conformal prediction under covariate shift, focusing on the setting where the shift between the training and test distributions is explained by a discrete group label. Weighted conformal prediction can handle covariate shift when the likelihood ratio is known, but in practice this weight function must be estimated. We show that, in the group-weighted setting, the effect of this estimation error can be controlled much more sharply than what follows from existing generic bounds for weighted conformal prediction. This yields stronger finite-sample coverage guarantees in settings where the population is divided into many groups.

Chapter 3 studies conformal prediction for classification problems with rare labels. Standard split conforma only targets marginal coverage that can hide poor performance on rare classes, while exact class-conditional coverage can be statistically inefficient or impossible when some classes have very few calibration examples. We develop label-weighted conformal prediction methods that guarantee macro-coverage and more general weighted averages of group-conditional coverages. This provides a meaningful middle ground between marginal coverage and class-conditional coverage, especially in long-tailed classification problems.

Chapter 4 addresses the tradeoff between statistical efficiency and computation. Full conformal prediction is often statistically efficient but computationally expensive because it requires recomputing the fitted model for many candidate test responses. Split conformal prediction is computationally cheap but can be less statistically efficient because it uses separate data for fitting and calibration. We develop a tournament-correction framework that can be used to correct many approximations to full conformal prediction that are used in practice, so that they enjoy the marginal coverage guarantee of $1-2\alpha$. These methods require more computation than split conformal prediction but much less than exact full conformal prediction, and the correction restores enough symmetry to obtain rigorous distribution-free coverage guarantees.

1.2 Goodness-of-fit testing

The second part of this dissertation turns from prediction to hypothesis testing. Goodness-of-fit testing asks whether the observed data are consistent with a proposed statistical model. In a parametric setting, the null hypothesis has the form

$$H_0 : X \sim f_\theta \quad \text{for some } \theta \in \Theta,$$

where $\{f_\theta : \theta \in \Theta\}$ is a family of distributions. Such tests are used throughout statistics and the sciences to assess whether a model is adequate before it is used for estimation, prediction, or scientific interpretation.

We mainly focus on testing via resampling where we do not have to rely on finding out the distribution of the test statistic; instead can approximate the distribution by using resamples of the data. When the null hypothesis is simple, meaning that θ is known, goodness-of-fit testing is conceptually straightforward. One can generate artificial data sets from the null distribution and compare the observed data to these null samples using any test statistic T. For example, if

$$\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$$

are independent draws from the known null distribution, then a valid Monte Carlo p-value can be constructed by comparing $T(X)$ to

$$T(\widetilde{X}^{(1)}), \ldots, T(\widetilde{X}^{(M)}).$$

The validity of this procedure follows from exchangeability: under the null, the observed data and the artificial data sets have the same distribution.

The composite null case is more difficult. If θ is unknown, then it is not immediately clear which distribution should be used to generate the artificial data. A plug-in approach, such as sampling from $f_{\widehat{\theta}}$, is often convenient but does not generally provide finite-sample validity. Classical likelihood ratio, Wald, or score tests address this issue in particular asymptotic regimes, but they also impose structure on the test statistic and may not be well suited to complex alternatives.

Also, in many settings we do not want to restrict ourselves with a fixed choice of a test statistic like in the generalized likelihood ratio test, instead we want to use a test statistic based on domain knowlegde or so that we have more power. A testing framework that allows

arbitrary test statistics can therefore be much more useful in practice, provided that it still controls type-I error.

1.2.1 Co-sufficient and approximate co-sufficient sampling

Co-sufficient sampling (Stephens, 2012) provides one route to such flexible testing. The idea is to condition on a sufficient statistic for the unknown parameter θ. After conditioning, the distribution of the data no longer depends on θ, so one can sample artificial data sets from the conditional null distribution and compare them to the observed data. This restores exchangeability and permits valid testing with arbitrary test statistics. However, exact sufficient statistics are available and useful only in limited settings. In many modern models, either no low-dimensional sufficient statistic exists, or conditioning on it removes too much information from the data and leads to low power. Approximate co-sufficient sampling methods (Barber and Janson, 2022; Zhu and Barber, 2023) or aCSS address this issue by conditioning on statistics that are nearly sufficient while retaining enough randomness to produce useful null samples. They show that such tests can be shown to control the type-I error level approximately at the nominal level.

In Chapter 5, we develop approximate co-sufficient sampling via Bayes, abbreviated aCSS-B which uses samples from a Bayesian posterior distribution as an approximate sufficient statistic. This significantly extends the scope of application of the aCSS framework making them applicable across a much broader class of problems. The theoretical contribution is to prove approximate exchangeability for the samples generated by aCSS-B and to translate this into approximate type-I error control for the resulting tests. This gives a flexible approach to goodness-of-fit testing for complex composite null models, while preserving the ability to use scientifically meaningful, problem-specific test statistics.

CHAPTER 2

GROUP-WEIGHTED CONFORMAL PREDICTION

This chapter is based on Bhattacharyya and Barber (2024) (joint work with Rina Foygel Barber). As discussed in Chapter 1 weighted conformal prediction (WCP) is used to tackle the problem of covariate shift between the training and test distributions. However, WCP requires knowledge of the nature of the covariate shift—specifically, the likelihood ratio between the test and training covariate distributions. In practice, since this likelihood ratio is estimated rather than known exactly, the coverage guarantee may degrade due to the estimation error. In this work, we consider a special scenario where observations belong to a finite number of groups, and these groups determine the covariate shift between the training and test distributions—for instance, this may arise if the training set is collected via stratified sampling. Our results demonstrate that in this special case, the predictive coverage guarantees of WCP can be drastically improved beyond the bounds given by existing estimation error bounds.

2.1 Introduction

In its standard form, conformal prediction deals with the case of i.i.d. data, or more generally exchangeable data. Namely, the training data used for constructing the prediction intervals and the test data on which the prediction intervals are deployed are assumed to be drawn from the same distribution.

To relax this assumption, Tibshirani et al. (2019) extend conformal prediction to the setting of covariate shift, where the test data distribution Q may differ from the training data distribution P in terms of the marginal distribution of the covariates—that is, Q_X may differ from P_X. However, the conditional distribution of the response given the features is assumed to be the same across training and test data, i.e., $P_{Y|X} = Q_{Y|X}$. The resulting

method, weighted conformal prediction, requires knowledge of the covariate shift, specifically the ratio between the marginal distributions P_X and Q_X, which is used to reweight the training data to better resemble the test data distribution. For example, if P_X and Q_X have densities f and g, respectively, implementing WCP requires access to a weight function $w(x) \propto g(x)/f(x)$. In a real data setting, however, it is not realistic to assume that $w(x)$ is known exactly. In practice the weight function would instead be approximated by some estimate $\widehat{w}(x)$, for instance by estimating the unknown densities from the training and test data. The error in this estimation process can then lead to a failure of coverage in the resulting prediction intervals.

In this work, we consider a special case of the covariate shift problem: we are interested in the setting where the unknown weight function $w(x)$ depends only on the subpopulation to which this data point belongs. Specifically, suppose that the feature vector X can be written as $X = (X^0, X^1)$, where $X^0 \in [K] = \{1, \ldots, K\}$ indicates the group or subpopulation to which the data point belongs, and X^1 captures any remaining features for this data point. This type of structure can arise in settings where we collect data from different groups, and can include sampling strategies such as stratified sampling. We will consider a setting where the weight function $w(x)$ depends only on the first part of the feature, i.e., if two data points X, X' belong to the same group, then $w(X) = w(X')$.

As a motivating example, suppose that our data points consist of students within a particular city's school system. It may be the case that the sampled data, drawn from P, are distributed differently across schools $k = 1, \ldots, K$ relative to the general population, which has distribution Q. This can occur, for example, if the K schools differ in terms of their budget constraints or their willingness to participate in the study. However, if within each school the study samples students uniformly at random, then the difference between the training sample distribution P and the general population distribution Q can be captured by a weight function $w(X)$ that depends only on the school to which the individual belongs, i.e.,

only on the value of $X^0 \in [K]$.

The contribution of our work lies in examining the role of the weight function w, and its estimate $\widehat{w}$, in quantifying the coverage guarantees for prediction intervals constructed via WCP. In general, an inaccurate estimate $\widehat{w}$ can lead to a substantial loss of coverage for the WCP prediction intervals. In the special case of group-weighted conformal prediction, where $w(X)$ depends only on the group X^0 to which the data point belongs, we derive bounds on the loss of coverage that are far tighter than what is currently known in the existing literature. Methodologically, our approach coincides with conventional WCP applied to discrete covariate shift. The novelty lies in our analysis, which provides significantly stronger coverage guarantees than those reported in previous works.

The remainder of the chapter is organized as follows. In Section 2.2, we review the weighted conformal procedure and some other related work. In Section 2.3, we introduce the setting of group-wise covariate shift and explain why prior guarantees for WCP give weak results in this setting. In Sections 2.4 and 2.5, we present our main results under two different formulations of the problem. Finally, we conclude with a short discussion in Section 2.6.

2.2 Background: conformal prediction under covariate shift

Chapter 1 described the split conformal, full conformal, and weighted conformal prediction procedures. In this chapter, we focus only on the split conformal setting, since group-weighted conformal prediction is built by modifying the calibration step of split conformal prediction. We assume the training data consists of $n^* + n$ many points and has been partitioned into a pretraining set and a calibration set — $\{(X_i^*, Y_i^*)\}_{i\in[n^*]}$ (used for model fitting) and $\{(X_i, Y_i)\}_{i\in[n]}$ (used for calibration) The pretraining data set is used to fit the score function . Throughout the theoretical results in this chapter, we condition on the pretraining data set and therefore treat s as fixed. The calibration scores are denoted as $s_i = s(X_i, Y_i), i \in [n]$.

2.2.1 Weighted conformal prediction and covariate shift

We consider the covariate shift setting, where the calibration data are drawn from a distribution P, while the test point is drawn from a potentially different distribution Q:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P, \qquad (X_{n+1}, Y_{n+1}) \sim Q.$$

The covariate shift assumption is that

$$P_{Y|X} = Q_{Y|X},$$

but possibly $P_X \neq Q_X$. Thus, the conditional distribution of the response given the features is the same in the training and test populations, but the marginal distribution of the features may differ.

As reviewed in Chapter 1, WCP corrects for this shift by reweighting the calibration scores using a weight function

$$w(x) \propto \frac{\mathsf{d}Q_X}{\mathsf{d}P_X}(x).$$

For reference, the WCP threshold for the test feature X_{n+1} is

$$\widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n} \frac{w(X_i)}{\sum_{i'=1}^{n+1} w(X_{i'})} \cdot \delta_{s_i} + \frac{w(X_{n+1})}{\sum_{i'=1}^{n+1} w(X_{i'})} \cdot \delta_{+\infty}\right), \tag{2.1}$$

where δ_s is the point mass at s. When w is known exactly, WCP satisfies

$$\mathbb{P}_{P^n \times Q}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha.$$

Here probability is taken with respect to the joint distribution $P^n \times Q$, i.e., the calibration data points are drawn i.i.d. from P, while the test point is drawn from Q. The usual split conformal method is recovered as the special case $w(x) \equiv 1$, corresponding to $P_X = Q_X$.

The WCP method has subsequently been applied to a range of statistical problems that can be reformulated as instances of covariate shift, including applications to survival analysis (Gui et al., 2022; Candès et al., 2023), causal inference (Lei and Candès, 2021; Yin et al., 2022; Jin et al., 2023), and adaptive learning (Fannjiang et al., 2022).

WCP with estimated weights. In practice, a major challenge in implementing WCP is that the weight function

$$w(x) \propto \frac{\mathsf{d}Q_X}{\mathsf{d}P_X}(x)$$

is not known exactly. Indeed, Tibshirani et al. (2019)'s initial results for WCP offer no theoretical guarantees for an estimated $w(x)$, although their empirical results demonstrate that estimating w on a large sample performs well in practice. More recent results in the literature offer guarantees that quantify the extent to which errors in estimating w can lead to loss of coverage. For example, Lei and Candès (2021) prove the following guarantee. Assume we are given a fixed function $\widehat{w}(x)$, or more generally a function $\widehat{w}(x)$ that is estimated independently of the calibration data and test point and can thus be treated as fixed. If the true and estimated weights are normalized to satisfy

$$\mathbb{E}_{P_X}[w(X)] = \mathbb{E}_{P_X}[\widehat{w}(X)] = 1,$$

then the WCP predictive interval satisfies

$$\mathbb{P}_{P^n \times Q}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - \mathbb{E}_{P_X}\left[\frac{|\widehat{w}(X) - w(X)|}{2}\right]. \tag{2.2}$$

Related results, also establishing bounds on the loss of coverage in terms of the difference $\widehat{w} - w$, can be found throughout the literature; see, for instance, Candès et al. (2023); Gui et al. (2022); Yang et al. (2022); Jin et al. (2023); Yin et al. (2022). Another work similar in flavor, though in the context of label shift rather than covariate shift, is the work of Plassier

et al. (2023), which considers the problem of label shift in federated learning and establishes coverage guarantees in this setup.

An alternative approach: within-group coverage. Finally, we mention another line of existing literature, which seeks to provide group-conditional coverage type guarantees, i.e., guarantees of the type

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \mid X^0_{n+1} = k\}.$$

For instance, works such as Dunn et al. (2022); Ding et al. (2023) study problems related to this framework. Any guarantee of group-conditional coverage will automatically yield coverage with respect to a group-wise covariate shift. That is, this type of guarantee is strictly stronger than coverage with respect to Q, which we study here. However, a group-conditional guarantee is meaningfully achievable only when the within-group sample size n_k is fairly large, while our work allows for small n_k, as long as K, the number of groups, is large. Thus, without additional assumptions, when some of the n_k's are very small, any method that achieves within-group coverage would be highly conservative and hence not meaningful. A more useful prediction interval may be obtained if we instead want coverage in an average sense across groups, which allows us to share information across the different groups. Additional related approaches in the literature include the work of Gupta et al. (2020) on bin-wise calibration and the works of Gibbs et al. (2023); Jung et al. (2022), which focus on group-conditional coverage guarantees.

2.3 Problem setting: group-wise covariate shift

As described in Section 2.1, we consider the setting where the covariate shift arises from the presence of different groups or subpopulations within the data, which have different probabilities under the training versus test distributions. Specifically, we assume that the

distribution P of a training data point $(X, Y) = (X^0, X^1, Y)$ has the form

$$\begin{cases} X^0 \sim \text{Multinomial}(p_1, \ldots, p_K), \\ (X^1, Y) \mid (X^0 = k) \sim \Pi_k, \end{cases} \tag{2.3}$$

i.e., the training data point is drawn from group k with probability p_k, and then conditional on being drawn from group k, the remaining features X^1 and response Y are drawn from some joint distribution Π_k. The test data distribution Q is instead given by

$$\begin{cases} X^0 \sim \text{Multinomial}(q_1, \ldots, q_K), \\ (X^1, Y) \mid (X^0 = k) \sim \Pi_k, \end{cases} \tag{2.4}$$

so that the test point is sampled from group k with probability q_k rather than p_k, but then conditional on being drawn from group k, the remaining features and response (X^1, Y) follow the *same* joint distribution Π_k.

In this setting, the weight function w characterizing the covariate shift is simple to compute: at a data point $X = (X^0, X^1)$, the weight function is given by

$$w(X) \propto q_k / p_k$$

if $X^0 = k$, for each $k \in [K]$. In particular, implementing WCP and thus ensuring a predictive coverage guarantee relies only on knowing these weights; coverage does not rely on any knowledge or assumptions for the distributions Π_k, which determine the feature and response distributions within each group k, since these are assumed to be the same across the training and test data.

For most of this chapter, we will assume that the test proportions q_k are known. Effectively, we are assuming that there is ample unlabeled data i.e., test points, for which we observe X

but not Y, so that we can estimate the proportion of subgroups within the target population with extremely high accuracy or we simply want coverage with respect to the empirical assignment of test points to the different groups. In Section 2.5.3, we will extend to the setting where these proportions are unknown, and the q_k's are instead estimated empirically. Alternatively we can interpret the assumption that the q_k's are known in a different way: we are simply requiring that coverage holds relative to some prespecified distribution over the K groups. For example, if we set $q_k \equiv 1/K$, this indicates that we are interested in the average group-wise coverage. The problem of estimating a weight function $w(X)$ then reduces to the question of estimating the proportions $p_1, \ldots, p_K$, to characterize the sampling bias of the training data, drawn from P, as compared to the target distribution Q.

2.3.1 Group-weighted conformal prediction (GWCP)

We now define the group-weighted conformal prediction (GWCP) method that will be our main focus. This method is essentially a variant of the standard WCP method proposed in literature applied to the context of a discrete covariate shift. This equivalence is made clearer in this section. As before, we work within a split conformal framework, where we are given a pretrained score function $s = s(x, y)$ (e.g., the residual score, $s(x, y) = |y - \widehat{f}(x)|$, where $\widehat{f}$ is a pretrained regression model), and where we evaluate the scores $s_i = s(X_i, Y_i)$ for the calibration data points, $i \in [n]$.

We define the group-weighted split conformal prediction interval as follows. Let $n_k = \sum_{i \in [n]} \mathbf{1}_{X_i^0 = k}$ denote the number of observations in the calibration set that belong to the kth group, and define

$$\widehat{P}_{\text{score}}^{(k)} = \frac{1}{n_k} \sum_{i \in [n], X_i^0 = k} \delta_{s_i},$$

which is the empirical distribution of scores for training data points in group k, for any k with $n_k > 0$. If instead $n_k = 0$ (i.e., group k does not appear at all in the calibration data

set), then we simply take $\widehat{P}^{(k)}_{\text{score}} = \delta_{+\infty}$, the point mass at infinity. [1]Finally, define

$$\widehat{C}_n(x) = \{y \in \mathcal{Y} : s(x,y) \le \widehat{q}\} \text{ where } \widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{k=1}^{K} q_k \widehat{P}^{(k)}_{\text{score}}\right). \tag{2.5}$$

In other words, the prediction interval is defined via a score threshold given by the $(1-\alpha)$-quantile of the following distribution: compute the empirical distribution $\widehat{P}^{(k)}_{\text{score}}$ of scores from data points in group k, and then take the mixture of these empirical distributions by placing weight q_k on the kth component, for each group $k \in [K]$. A useful feature of this construction is that we do not require the group assignments of the points in the test set to be known.

Dunn et al. (2022) used this kind of CDF pooling technique in the context of a two-layer hierarchical model, but only asymptotic guarantees are shown; this type of approach is also studied by Lee et al. (2023), who show finite sample guarantees, but again in a different setting where new groups are sampled from a hierarchical model.

Comparing to WCP. To better compare to our earlier notation for the WCP method, we can rewrite the threshold $\widehat{q}$ in (2.5) as follows:

$$\widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n} \frac{q_{X_i^0}}{n_{X_i^0}} \cdot \delta_{s_i}\right).$$

In other words, each data point $i \in [n]$ is given weight $q_{X_i^0}/n_{X_i^0}$ in this weighted quantile calculation.

We will now see that we can view this as a *slightly less conservative* version of the WCP method. To see why, suppose we let $\widehat{p}_k = n_k/n$—that is, if the calibration data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled from the K groups with probabilities $p_1, \ldots, p_K$,

1. In contrast if any group does not appear in the test data or the test distribution Q is known to have lesser than K groups then we can simply set $q_k = 0$ for those groups and run our method. This is essentially equivalent to not using the calibration data in those groups to find the prediction intervals.

as in (2.3), then $\widehat{p}_k$ is an estimate of p_k for each k. Then, for any X with $X^0 = k$, the weight for the covariate shift under this model is given by $w(X) = q_k/p_k$; we can estimate this weight with $\widehat{w}(X) = q_k/\widehat{p}_k$. In this comparison, we will assume for simplicity that $n_k > 0$ for all k, i.e., all groups are observed in the training set—otherwise the function $\widehat{w}$ would not be well-defined. Note that if $\hat{p}_k = n_k/n$ and $\hat{w}(X) = q_k/\hat{p}_k = q_k n/n_k$, then

$$\sum_{i\in[n]} \hat{w}(X) = \sum_k \sum_{i\in[n], X_i^0=k} q_k n/n_k = \sum_k q_k n = n,$$

and so

$$\frac{\hat{w}\left(X_i\right)}{\sum_{i'=1}^n \hat{w}\left(X_{i'}\right)} = \frac{q_{X_i^0}}{n\hat{p}_{X_i^0}} = \frac{q_{X_i^0} n/n_{X_i^0}}{n} = \frac{q_{X_i^0}}{n_{X_i^0}}.$$

Then rewriting the definition of $\widehat{q}$ one more time,

$$\widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^n \frac{\widehat{w}(X_i)}{\sum_{i'=1}^n \widehat{w}(X_{i'})} \cdot \delta_{s_i}\right). \tag{2.6}$$

In contrast, if we were to run WCP with the weight function $\widehat{w}$,[2] the prediction set would be given by $\widehat{C}_n(x) = \{y \in \mathcal{Y} : s(x,y) \leq \widehat{q}_+\}$, where

$$\widehat{q}_+ = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^n \frac{\widehat{w}(X_i)}{\sum_{i'=1}^{n+1} \widehat{w}(X_{i'})} \cdot \delta_{s_i} + \frac{\widehat{w}(X_{n+1})}{\sum_{i'=1}^{n+1} \widehat{w}(X_{i'})} \cdot \delta_{+\infty}\right). \tag{2.7}$$

That is, the definition of the threshold $\widehat{q}$ for GWCP is strictly less conservative since it is exactly the same as the WCP threshold $\widehat{q}_+$ except that the weight on $\delta_{+\infty}$ has been removed.

As we will see below, the prediction interval defined here in (2.5) leads to a coverage guarantee that is slightly smaller than the target level $1-\alpha$. Interestingly, however, this cannot be fixed by returning to the slightly more conservative WCP method given in (2.1).

2. While the theory for WCP, described above in Section 2.2.1, requires the estimated weight function $\widehat{w}$ to be prefitted—i.e., independent of the calibration and test data—the WCP algorithm itself can nonetheless be defined with a data-dependent $\widehat{w}$.

We choose to work with this present formulation, (2.5), since avoiding the additional weight on $\delta_{+\infty}$ leads to a simpler construction and a (slightly) less conservative method without any additional loss of coverage in the theoretical guarantee. Of course, any coverage guarantee that can be established for the prediction interval defined in (2.5) will also hold for the version of WCP that adds a weight to $\delta_{+\infty}$, since this leads to a strictly more conservative prediction interval.

2.3.2 Estimation error in the group-wise setting: applying prior results

As mentioned above, in this special case, where the covariate shift is solely determined by the K groups, the problem of estimating w is easier: in particular, we only need to estimate a K-dimensional parameter, i.e., $(p_1, \ldots, p_K)$ (since we have assumed the q_k's are known), in contrast to the general case, where estimating an arbitrary function $w(x)$ is an infinite-dimensional problem.

However, even in this setting of a finite-dimensional unknown parameter, existing coverage guarantees provide a fairly pessimistic view and imply that the loss of coverage may be severe. In particular, if our estimate of the probabilities $p_1, \ldots, p_K$ (which determine the relative proportions of the K groups in the training distribution P) is based on a sample of size $\mathcal{O}(n)$, then we would expect that each p_k is estimated with error scaling as $\mathcal{O}(\sqrt{p_k/n})$ (due to the variance of a Binomial distribution); equivalently, the *relative* error in estimating p_k scales as $\mathcal{O}(1/\sqrt{np_k})$. For example, if $p_k \equiv 1/K$ (i.e., the true distribution places equal weight on each group), then the relative error in estimating each p_k scales as $\mathcal{O}(\sqrt{K/n})$, and so the coverage loss term in Lei and Candès (2021)'s guarantee (2.2) can be expected to scale as

$$\mathbb{E}\left[\frac{|\widehat{w}(X) - w(X)|}{2}\right] \asymp \sqrt{\frac{K}{n}}.$$

To demonstrate this intuition more concretely, we compute Lei and Candès (2021)'s

coverage guarantee (2.2) empirically, across a range of sample sizes $n = 100, 110, 120, \ldots, 1000$, in three regimes: a constant number of groups, $K = 10$; a slowly increasing number of groups, $K = \lfloor\sqrt{n}\rfloor$; and a proportional number of groups, $K = n/10$. In each case, the probabilities are given by $p_k = q_k \equiv 1/K$ for all groups $k \in [K]$, for both the training and test distributions. The estimated weight function $\widehat{w}$ is then given by $\widehat{w}(X) = q_k/\widehat{p}_k$ for any data point with $X^0 = k$, where $q_k \equiv 1/K$ is known, while $\widehat{p}_k$ is given by the empirical proportion of group k in a pretraining sample of size n. To avoid dividing by zero, we add a $+1$ adjustment to each group's count—that is, $\widehat{p}_k = \frac{n_k+1}{n+K}$, where n_k is the number of samples observed in group k within the training set of size n (since the true proportions p_k are uniform, this adjustment can only make the estimates more accurate).

Figure 2.1 illustrates the guaranteed coverage level (2.2) for each choice of n and K, where the term $\mathbb{E}\left[\frac{|\widehat{w}(X)-w(X)|}{2}\right]$ is estimated empirically over 100 trials.[3] We can see that, even though estimating w is a finite-dimensional parameter estimation problem, the resulting guarantee suggests that there may be substantial loss of coverage; in the proportional regime, $K = n/10$, the guaranteed coverage level does not even converge to $1 - \alpha$ as $n \to \infty$.

2.4 GWCP with fixed group sizes

We now turn to the theoretical properties of the GWCP prediction interval. We begin by considering a different sampling framework: suppose that the within-group sample sizes $n_1, \ldots, n_k \geq 0$ are *fixed* rather than random.

The problem setting can be formalized as follows: independently for each group $k \in [K]$,

3. Code for reproducing all figures in the chapter is available at `https://github.com/aabeshb/Group-Weighted-Conformal-Prediction`.

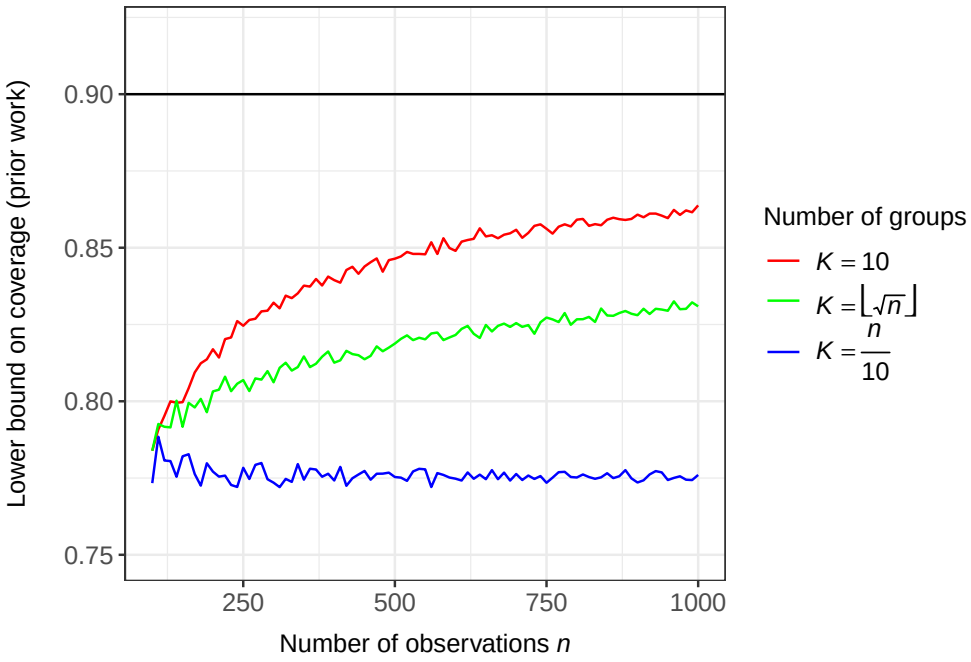

Figure 2.1: (Lei and Candès, 2021)'s WCP coverage guarantee (2.2) for varying numbers of groups. The horizontal line marks the target coverage level 90% (i.e., $\alpha = 0.1$). See Section 2.3.2 for details.

we sample n_k many calibration points from the kth group,

$$X_i^0 = k,\ (X_i^1, Y_i) \overset{\text{iid}}{\sim} \Pi_k, \quad \text{for each } i \in \{n_1 + \cdots + n_{k-1} + 1, \ldots, n_1 + \cdots + n_{k-1} + n_k\}. \tag{2.8}$$

We then want to ensure predictive coverage with respect to a test point $(X, Y) = (X^0, X^1, Y)$ drawn from the target distribution Q, defined as in (2.4), as before. At a high level, we can think of this framework as strictly harder than the random setting, because here we will require coverage to hold for any fixed $n_1, \ldots, n_K$, while in the random setting described above in Section 2.3, a marginal coverage guarantee only requires coverage to hold *on average* over the draw of $n_1, \ldots, n_K$ (i.e., when the calibration data points are sampled i.i.d. from P, as in (2.3)).

2.4.1 Theoretical guarantee

Our first main result establishes a marginal coverage guarantee for the fixed group size setting.

Theorem 2.4.1 (Coverage under fixed group sizes.)**.** *Suppose the training data points* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are sampled from the fixed-group-size model, as in* (2.8)*, while the test point* (X_{n+1}, Y_{n+1}) *is drawn independently from the distribution* Q *as defined in* (2.4)*. Then, for any fixed (i.e., pretrained) score function* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$*, the GWCP prediction interval* $\widehat{C}_n$ *defined in* (2.5) *satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - \max_{k:n_k>0}\{q_k/n_k\}.$$

In the special case where $q_k \equiv 1/K$, i.e., the target distribution for the test data places equal weight on each of the K groups, we can rewrite this guarantee as

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - \frac{1}{K \min_{k:n_k>0} n_k}. \tag{2.9}$$

We emphasize that this theorem shows the potential loss of coverage is low as long as *either* the number of groups K or the minimum group size n_k is fairly large. We will examine the strength of this result more later on, when we compare to previous guarantees. Moreover, in Appendix A.1.2, we will verify that this lower bound on coverage is tight, while in Appendix A.1.4, we will define a slightly more conservative version of the method that can be used if we wish to ensure a coverage guarantee of at least $1 - \alpha$, without an error term.

We now turn to the proof of the theorem.

Proof of Theorem 2.4.1. First we prove the result for the case that $n_k > 0$ for all k. We begin by defining some notation. For convenience of the proof, we will redefine the indexing

of the training set: for each k, and for $i = 1, \ldots, n_k$, let

$$(X_{k,i}, Y_{k,i}) = (X_{n_1+\cdots+n_{k-1}+i}, Y_{n_1+\cdots+n_{k-1}+i}).$$

That is, the data points are now indexed within each group, rather than across the entire data set of size n. Next let

$$\widehat{P}^{(k)}_{\text{score}} = \frac{1}{n_k} \sum_{i=1}^{n_k} \delta_{s(X_{k,i}, Y_{k,i})}$$

be the empirical distribution of the training data set scores within the kth group, as in the definition of the GWCP method, and let

$$\widehat{P}_{\text{score}} = \sum_{k=1}^{K} q_k \cdot \widehat{P}^{(k)}_{\text{score}}$$

be the weighted mixture across groups. Then the threshold $\widehat{q}$ for GWCP, defined in (2.5), can equivalently be written as $\widehat{q} = \text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}})$, and thus the GWCP prediction set is given by

$$\widehat{C}_n(x) = \left\{ y \in \mathcal{Y} : s(x, y) \leq \text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}}) \right\}.$$

Next, independently for each $k \in [K]$, sample a new data point in the kth group,

$$(X_{k,0}, Y_{k,0}) = (k, X^1_{k,0}, Y_{k,0}) \text{ where } (X^1_{k,0}, Y_{k,0}) \sim \Pi_k,$$

independently from the training data. Then, since the test point (X_{n+1}, Y_{n+1}) is sampled from the mixture distribution Q which is defined by placing weight q_k on each of the k groups, the coverage probability can equivalently be expressed as

$$\begin{aligned}\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right\} &= \sum_{k=1}^{K} q_k \mathbb{P}\left\{ Y_{k,0} \in \widehat{C}_n(X_{k,0}) \right\} \\ &= \sum_{k=1}^{K} q_k \mathbb{P}\left\{ s(X_{k,0}, Y_{k,0}) \leq \text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}}) \right\}.\end{aligned}$$

Next, for each k and each $i = 1, \ldots, n_k$, we define a modified version of the weighted empirical distribution $\widehat{P}_{\text{score}}$:

$$\widehat{P}^{k,i}_{\text{score}} = \widehat{P}_{\text{score}} + \frac{q_k}{n_k}\left(\delta_{s(X_{k,0},Y_{k,0})} - \delta_{s(X_{k,i},Y_{k,i})}\right).$$

In particular, for $i = 0$, this is the same distribution as before, i.e., $\widehat{P}^{k,0}_{\text{score}} = \widehat{P}_{\text{score}}$, for each k. On the other hand, for $i \geq 1$, $\widehat{P}^{k,i}_{\text{score}}$ modifies the original weighted empirical distribution by replacing data point $(X_{k,i}, Y_{k,i})$ with data point $(X_{k,0}, Y_{k,0})$.

With this new notation in place, we can then rewrite the coverage probability as

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} = \sum_{k=1}^{K} q_k \mathbb{P}\left\{s(X_{k,0}, Y_{k,0}) \leq \text{Quantile}_{1-\alpha}(\widehat{P}^{k,0}_{\text{score}})\right\}. \tag{2.10}$$

For each k, since $(X_{k,0}, Y_{k,0}), \ldots, (X_{k,n_k}, Y_{k,n_k})$ are i.i.d. and are therefore exchangeable, by symmetry we must have

$$\mathbb{P}\left\{s(X_{k,0}, Y_{k,0}) \leq \text{Quantile}_{1-\alpha}(\widehat{P}^{k,0}_{\text{score}})\right\} = \mathbb{P}\left\{s(X_{k,i}, Y_{k,i}) \leq \text{Quantile}_{1-\alpha}(\widehat{P}^{k,i}_{\text{score}})\right\}$$

for every $i = 1, \ldots, n_k$, and therefore,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{q_k}{n_k} \mathbb{P}\left\{s(X_{k,i}, Y_{k,i}) \leq \text{Quantile}_{1-\alpha}(\widehat{P}^{k,i}_{\text{score}})\right\}.$$

Moreover, defining $\alpha' = \alpha + \max_k\{q_k/n_k\}$, we must have

$$\text{Quantile}_{1-\alpha}(\widehat{P}^{k,i}_{\text{score}}) \geq \text{Quantile}_{1-\alpha'}(\widehat{P}_{\text{score}})$$

for every k, i almost surely. This is because the total variation distance between these two weighted empirical distributions is bounded as $\text{d}_{\text{TV}}(\widehat{P}_{\text{score}}, \widehat{P}^{k,i}_{\text{score}}) \leq q_k/n_k$ (recall that we

have assumed $n_k > 0$ for all k). Therefore,

$$\begin{aligned}\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} &\geq \sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{q_k}{n_k}\mathbb{P}\left\{s(X_{k,i},Y_{k,i}) \leq \text{Quantile}_{1-\alpha'}(\widehat{P}_{\text{score}})\right\} \\ &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{q_k}{n_k}\mathbf{1}\left\{s(X_{k,i},Y_{k,i}) \leq \text{Quantile}_{1-\alpha'}(\widehat{P}_{\text{score}})\right\}\right].\end{aligned}$$

Finally, the quantity inside the last expected value must be $\geq 1-\alpha'$, almost surely, by definition of the weighted empirical distribution $\widehat{P}_{\text{score}}$ (see, e.g., Harrison (2012, Lemma A.1)).

Now we turn to the case where we may have $n_k = 0$ for some k. Define $q_+ = \sum_{k:n_k>0} q_k$. If $1-\alpha > q_+$, then deterministically we must have $\widehat{q} = +\infty$ in the definition of the method (2.5), which implies coverage holds with probability 1. from this point on, then, we will consider the case $1-\alpha \leq q_+$.

Let Q' be the distribution of $(X,Y) \sim Q$ conditional on the event that $X^0 = k$ for some k with $n_k > 0$, and let Q'' be the same conditional on the event that $X^0 = k$ for some k with $n_k = 0$, such that we can express the target distribution Q as the mixture

$$Q = q_+ \cdot Q' + (1-q_+) \cdot Q''.$$

Then

$$\mathbb{P}_Q\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq q_+ \cdot \mathbb{P}_{Q'}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\},$$

where the subscript on $\mathbb{P}$ indicates the distribution from which the test point (X_{n+1}, Y_{n+1}) is drawn. Next observe that by construction, we have

$$\widehat{P}_{\text{score}} = q_+ \cdot \widehat{P}'_{\text{score}} + (1-q_+) \cdot \delta_{+\infty}$$

where $\widehat{P}'_{\text{score}}$ is the average of empirical distributions for all *observed* groups,

$$\widehat{P}'_{\text{score}} = \sum_{k:n_k>0} q'_k \cdot \widehat{P}^{(k)} \text{ where } q'_k = q_k/q_+.$$

We therefore have

$$\text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}}) = \text{Quantile}_{1-\alpha'}(\widehat{P}'_{\text{score}})$$

where $1-\alpha' = (1-\alpha)/q_+$. Therefore,

$$\widehat{C}_n(x) = \left\{y \in \mathcal{Y} : s(x,y) \leq \text{Quantile}_{1-\alpha'}(\widehat{P}'_{\text{score}})\right\}.$$

We then have

$$\mathbb{P}_{Q'}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha' - \max_{k:n_k>0}\{q_k/n_k\},$$

which holds by applying our proof above (for the case where $n_k > 0$ for all k) to the target distribution Q' in place of Q. Finally, returning to the original distribution q, we therefore have

$$\mathbb{P}_{Q}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq q_+ \cdot \left[1-\alpha' - \min_{k:n_k>0}\{q'_k/n_k\}\right] = 1-\alpha - \max_{k:n_k>0}\{q_k/n_k\},$$

which completes the proof. □

2.4.2 Simulation with fixed group sizes

The theoretical guarantee given in Theorem 2.4.1 above for the fixed group size setting suggests that the potential coverage loss of the GWCP method depends on the *smallest* nonzero group size—specifically, in the case where the test distribution is uniform over the K groups, the loss of coverage in the guarantee (2.9) depends on $\min_{k:n_k>0} n_k$. It is intuitive that if many groups are small, then we should expect to see a loss of coverage due to issues of

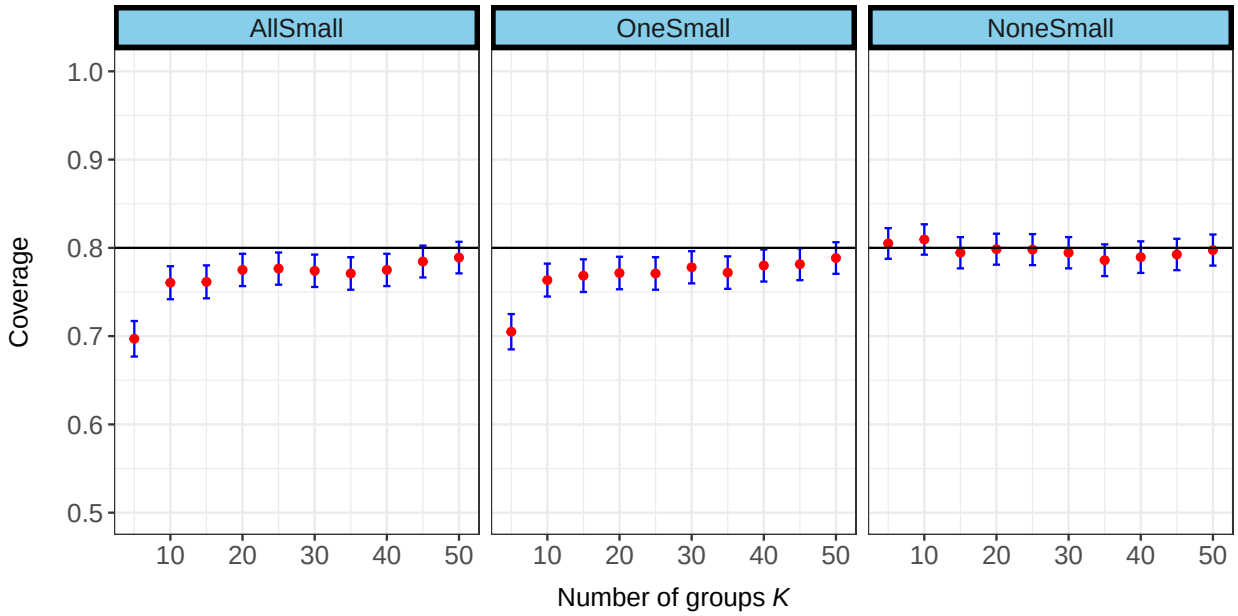

Figure 2.2: Simulation results for the setting of fixed group sizes, in three regimes for the group sizes $n_1, \ldots, n_K$. The horizontal line marks the target coverage level $1 - \alpha = 80\%$. Results are averaged over 2000 independent trials, and the figure displays the mean coverage rate along with a 95% confidence interval. See Section 2.4.2 for details.

low effective sample size, but we might suspect that the guarantee for GWCP should depend on some kind of *average* measure of group size, rather than the minimum. In fact, however, the lower bound in Theorem 2.4.1 is essentially tight: in a worst case scenario, the loss of coverage can indeed be governed by the smallest group size, even if most groups are large.

To illustrate this, we simulate data with fixed group sizes, in three regimes: `AllSmall`, where all the group sizes n_k are small; `OneSmall`, where one of the n_k's is small but the remaining $K - 1$ values are large; and `NoneSmall`, where all n_k's are large. If we believe that coverage loss is controlled by average group size, then we would expect to see the results of the `OneSmall` setting to resemble the `NoneSmall` setting. However, the simulation shows that the loss of coverage in the `OneSmall` setting behaves very similarly to the `AllSmall` setting, confirming that the guarantee of Theorem 2.4.1 is tight. (In Appendix A.1.2, we will discuss the question of tightness of the lower bound in more theoretical detail.)

The simulation is designed as follows. The target coverage level is $1 - \alpha = 0.8$. The

simulation is repeated for each value $K \in \{5, 10, 15, \ldots, 50\}$ as the number of groups. The distribution of (X, Y) in the kth group is given by $\delta_k \times \text{Uniform}[\frac{k-1}{K}, \frac{k}{K}]$ (i.e., $X = X^0$ is simply equal to the group label, k, and Y is sampled from the Uniform$[\frac{k-1}{K}, \frac{k}{K}]$ distribution). The pretrained score function is given by $s(x, y) = y$. The group sizes are defined for each of the three regimes as:

- `AllSmall` regime: $n_k = 1$ for all k.
- `OneSmall` regime: $n_{0.8K+1} = 1$, and $n_k = 100$ for all $k \neq 0.8K + 1$.
- `NoneSmall` regime: $n_k = 100$ for all k.

The empirical coverage of the resulting prediction set is then computed based on 2000 many independent trials where for each trial a calibration data set is drawn from the training data distribution P and a test point is drawn from Q. Figure 2.2 shows the mean coverage along with the 95% confidence interval for each setting.

We observe that the empirical coverage levels for the `OneSmall` regime behave very similarly to those for the `AllSmall` regime, with substantial undercoverage when K is small. That is, for the `OneSmall` regime, the minimum group size, $\min_k n_k = 1$, indeed seems to determine its behavior in terms of a loss of coverage in the finite sample regime. Of course, the theoretical guarantee ensures that coverage will converge to (at least) $1 - \alpha$ as the number of groups K increases, even if $\min_k n_k = 1$; our simulation results confirm this, since all three regimes show coverage at approximately the target level $1 - \alpha$ once K is large.

However, while this result confirms that our theoretical guarantee is tight in the worst case, the specific construction of the `OneSmall` regime is designed specifically to be as challenging as possible. The specific design of the simulation ensures that the smallest group (i.e., the k for which $n_k = 1$) is likely to have its score distribution concentrated near the $(1-\alpha)$-quantile of the overall distribution of scores. This type of worst-case behavior may be unlikely to occur in practice; while the $\min_k n_k$ term in the theoretical guarantee cannot be improved

without further assumptions (see Appendix A.1.2 for more details), it is likely that for a typical problem, loss of coverage may depend more on average group size than on minimum group size.

2.5 GWCP under covariate shift

We now return to the original problem that motivates this work: the setting of a group-wise covariate shift. Returning to our earlier problem formulation, we assume that the n training points are sampled i.i.d. from P, and the target test distribution is Q, where these distributions are defined as in (2.3) and (2.4) above. We will next present our main results for finite-sample coverage guarantees in the case of a group-wise covariate shift, and then compare to existing coverage guarantees in the prior literature.

2.5.1 Theoretical guarantee

We now give the theoretical guarantee for the group-wise covariate shift setting. Our theoretical guarantee for the group-wise covariate shift setting is a straightforward consequence of the guarantee for the fixed group size setting: since Theorem 2.4.1 ensures that coverage holds for *any* group sizes n_k, this means that coverage also holds *on average* over a random draw of the group sizes n_k.

Theorem 2.5.1 (Coverage under group-wise covariate shift.)**.** *Suppose the calibration data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled i.i.d. from the distribution P as defined in* (2.3)*, while the test point (X_{n+1}, Y_{n+1}) is drawn independently from the distribution Q as defined in* (2.4)*. Then, for any fixed (i.e., pretrained) score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the GWCP prediction interval $\widehat{C}_n$ defined in* (2.5) *satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - \mathbb{E}\left[\max_{k:n_k>0}\{q_k/n_k\}\right].$$

In particular, if $\min_k p_k \geq \frac{8 \log n}{n}$ *then*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - 4n^{-1} \max_k \{q_k/p_k\}.$$

The proof is given in Appendix A.1.1; essentially it is achieved by simply applying the guarantee for fixed n_k's, given in Theorem 2.4.1, and then marginalizing over the random draw of the n_k's. The lower bound on p_k of $O\left(\frac{\log n}{n}\right)$ is a mild condition that ensures that all groups are observed with high probability.

2.5.2 Uniform groups and unobserved groups: a closer look

In the results above, we have assumed that the group probabilities $q_1, \ldots, q_K$ under the target distribution (i.e., the distribution of the test data) are known. For example, if we know ahead of time that we will need to predict the response Y for a particular collection of X values, $X_{n+1}, \ldots, X_{n+m}$, we can simply take the q_k's to be the *empirical* proportions of groups $1, \ldots, K$ within this list, to guarantee coverage on average over this test set—that is, to guarantee

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{P}\{Y_{n+i} \in \widehat{C}_n(X_{n+i})\} \geq 1 - \alpha.$$

On the other hand, we can also imagine settings in which the set of test feature vectors is not available in advance. In this case, perhaps we do not even know ahead of time the total number of groups in the population. Suppose, say, we want to ensure coverage relative to a uniform distribution over all groups (i.e., $q_k \equiv 1/K$), but we do not know the total number of groups K; instead, we simply run GWCP with the target distribution given by a uniform distribution over all *observed* groups. This leads to the following modification of the GWCP

method (2.5):

$$\widehat{C}_n(x) = \{y \in \mathcal{Y} : s(x,y) \le \widehat{q}\} \text{ where } \widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{k:n_k \ge 1} \frac{1}{\widehat{K}} \widehat{P}^{(k)}_{\text{score}}\right),\ \widehat{K} = \sum_{k=1}^{K} \mathbf{1}_{n_k>0}. \tag{2.11}$$

Corollary 2.5.2. *Suppose the training data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled i.i.d. from the distribution P as defined in (2.3), while the test point (X_{n+1}, Y_{n+1}) is drawn independently from the distribution Q as defined in (2.4) with $q_k \equiv 1/K$. Then, for any fixed (i.e., pretrained) score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the prediction interval $\widehat{C}_n$ defined in (2.11) satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \ge (1-\alpha) - \mathbb{E}\left[\frac{1}{K \min_{k:n_k>0} n_k} + (1-\alpha)\frac{\sum_k \mathbf{1}_{n_k=0}}{K}\right].$$

In particular, if $\min_k p_k \ge \frac{8 \log n}{n}$ then

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \ge 1 - \alpha - \frac{4}{Kn \min_k p_k}.$$

The proof is given in Appendix A.1.1.

2.5.3 Unknown test group probabilities

In this section, we consider the setting where the the group probabilities in the test distribution (i.e., $q_1, q_2, \ldots, q_K$) are not known. Instead we have m many draws $\tilde{X}^0_1, \ldots, \tilde{X}^0_m \in \{1, \ldots, K\}$ that offer an empirical approximation to this distribution over groups—for instance, we might have a sample of size m from the test population, and the $\tilde{X}^0_i$'s are the group assignments of these m data points.

Can we simply use the empirical proportions in each group $k = 1, \ldots, K$ in place of the q_k's when constructing the GWCP prediction set (2.5)? That is, our empirically weighted

prediction set is given by

$$\widehat{C}_n(x) = \{y \in \mathcal{Y} : s(x,y) \le \widehat{q}\} \text{ where } \widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{k=1}^{K} \tilde{q}_k \widehat{P}^{(k)}_{\text{score}}\right),\ \tilde{q}_k = \frac{\sum_{i=1}^{m} \mathbf{1}\{\tilde{X}^0_i = k\}}{m}. \tag{2.12}$$

The following result verifies that this approach offers nearly the same level of coverage as before—the additional loss of coverage in the guarantee is only $\frac{1}{m+1}$.

Corollary 2.5.3. *Suppose the training data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled i.i.d. from the distribution P as defined in* (2.3)*, and the test point is independently drawn as $(X_{n+1}, Y_{n+1}) \sim Q$ as defined in* (2.4)*. Moreover, suppose $\tilde{X}^0_1, \ldots, \tilde{X}^0_m$ are i.i.d. draws from the Multinomial$(q_1, \ldots, q_K)$ distribution, drawn independently of the data. Then, for any fixed (i.e., pretrained) score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the prediction interval $\widehat{C}_n$ defined in* (2.12) *satisfies*

$$\mathbb{P}\left\{Y_{n+m+1} \in \widehat{C}_n(X_{n+m+1})\right\} \ge 1 - \alpha - \frac{1}{m+1} - \mathbb{E}\left[\max_{k:n_k>0}\{\tilde{q}_k/n_k\}\right].$$

The proof is given in Appendix A.1.1.

2.5.4 Comparison to existing guarantees

We next provide an empirical comparison between the theoretical guarantee obtained in Corollary 2.5.2, which provides a lower bound for predictive coverage under group-wise covariate shift, against the existing guarantee obtained by Lei and Candès (2021), shown above in (2.2), where the loss of coverage is bounded as $\mathbb{E}\left[\frac{|\widehat{w}(X)-w(X)|}{2}\right]$. We use the same settings as in Section 2.3.2, with three regimes: $K = 10$ (a constant number of groups), $K = \lfloor\sqrt{n}\rfloor$ (a slowly growing number of groups), and $K = n/10$ (a proportional number of groups). The training and test distributions are again uniform, $p_k = q_k \equiv 1/K$. Here the coverage guarantee given by Lei and Candès (2021) is the same as that shown in Figure 2.1,

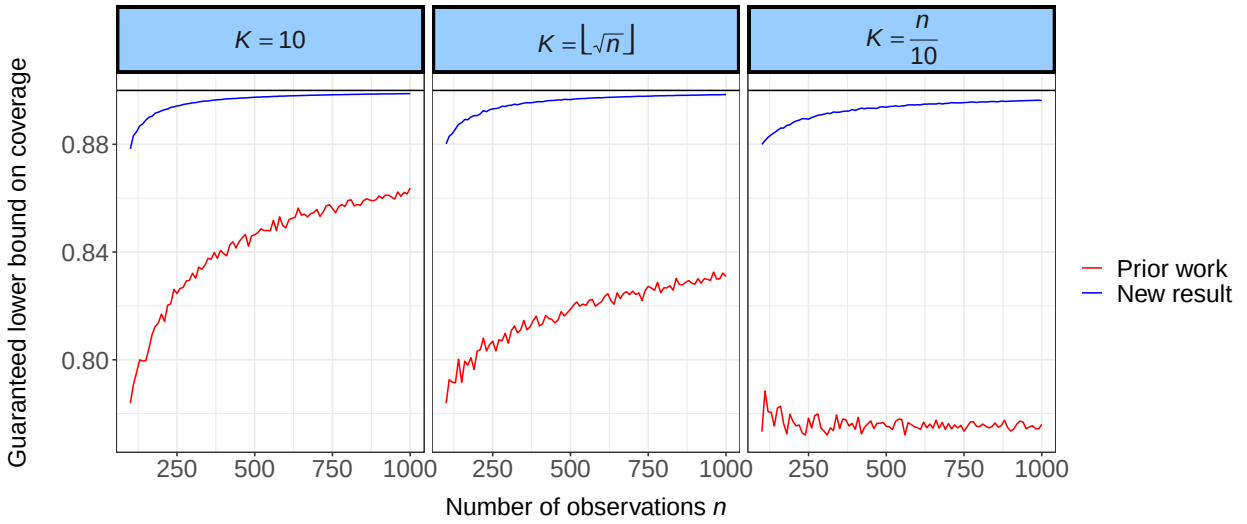

Figure 2.3: (Lei and Candès, 2021)'s WCP coverage guarantee (2.2), compared to the new guarantee given in Corollary 2.5.2, for varying numbers of groups. The horizontal line marks the target coverage level 90% (i.e., $\alpha = 0.1$). See Section 2.5.4 for details.

above; our new result, the coverage guarantee of Corollary 2.5.2, is now added to the plot for comparison. Note that this guarantee is for the version (2.11) of the group-weighted conformal prediction interval, which does not assume that we know the total number of groups.

As in Section 2.3.2, to ensure that $\widehat{w}$ is always well-defined even though there is a nonzero probability that not all groups are observed, we define $\widehat{w}(x)$ using estimated proportions $\widehat{p}_k = \frac{n_k+1}{n+K}$, i.e., augmenting each group count by 1. Therefore, we are giving two advantages to Lei and Candès (2021)'s existing bound relative to ours: (1) the definition of $\widehat{p}_k$, which (since the true p_k's are uniform) is more accurate than the actual empirical distribution; (2) Lei and Candès (2021)'s coverage guarantee applies to the more conservative form of the method, given in (2.7), relative to ours.

We can see in Figure 2.3 that the new coverage guarantee in Corollary 2.5.2 is substantially stronger, converging much more rapidly to the target coverage level of 90% relative to the existing lower bound. Indeed, as observed earlier, in the third setting, $K = n/10$, where the sample size within a group does not increase with n, the existing lower bound does not

approach the target coverage level $1-\alpha$ even as $n \to \infty$; in contrast, we can observe that the lower bound in Corollary 2.5.2 converges to $1-\alpha$ as long as *either* the number of groups K or the (minimum) sample size within a group n_k diverges to infinity.

Finally, we remark that another difference between our work and the existing literature is that, for our theoretical guarantee, it is important that the weight function $\widehat{w}$ is estimated using the calibration set—in contrast, in existing work, it is more common to assume $\widehat{w}$ is pretrained, e.g., fitted to the pretraining data. We discuss this distinction more in Appendix A.1.3.

2.6 Discussion

This chapter considers a supervised learning setting where the underlying population consists of various groups, with the relative proportions of the different groups being potentially different between the training data (i.e., the observed sample) and the test data (i.e., the target distribution or general population). We view this as a covariate shift problem, where the distribution shift depends only on a finitely-valued covariate, i.e., the group to which an observation belongs. Thus, the prediction problem can be addressed with weighted conformal prediction, which offers distribution-free coverage guarantees as long as the covariate shift can be accurately estimated. Our results provide sharp coverage guarantees that bound the influence of this estimation error, and are substantially closer to the target coverage level $1-\alpha$ than previously known bounds; in particular, our results can even achieve coverage converging to $1-\alpha$ when the group sizes remain constant as sample size n increases (i.e., the number of groups is proportional to n).

More broadly, this problem can be viewed as an instance of a more general class of settings, where covariate shift can be expressed via a finite dimensional parametrization. Whether these techniques can be extended to broader scenarios, moving beyond the setting of disjoint subpopulations, is an important direction for further exploration. Another future direction of

interest is the problem of allowing for noisy or corrupted group labels in the test or training data, to handle scenarios where group labels may not be known exactly—for instance, due to issues such as privacy constraints or inaccurate data collection.

CHAPTER 3

CONFORMAL PREDICTION WITH MACRO-COVERAGE GUARANTEES

This chapter is based on the paper Bhattacharyya et al. (2026a) (joint work with Tiffany Ding and Rina Foygel Barber). Prediction sets should have high coverage to be useful, but some coverage notions are more practically relevant than others. In the classification setting, class-conditional coverage requires that the prediction set (i.e., the set of candidate labels for a new test point) must achieve the target accuracy level within each class, which may be challenging to satisfy when many classes are rare and have few calibration points. At the other extreme, marginal coverage requires only that coverage holds on average over the distribution of all classes, which can lead to low-probability labels being essentially ignored. To find a middle ground, recent work has introduced macro-coverage, defined as the unweighted average of class-conditional coverages. Macro-coverage offers a compromise between marginal coverage and class-conditional coverage that is particularly appropriate for long-tailed settings. In this work, we show that label-weighted conformal prediction which is a direct extension of group-weighted conformal prediction described in Chapter 2, can be used to obtain a finite-sample macro-coverage guarantee, and more generally a guarantee on a family of generalized macro-coverage objectives that aggregate coverage at the level of arbitrary class groupings and take a weighted average. We further characterize the form of the smallest prediction sets satisfying a given generalized macro-coverage objective and propose a corresponding conformal score function. We validate our theoretical results on two large-scale image classification datasets.

3.1 Introduction

In this work, we focus on multiclass classification, where the goal is to use features $X \in \mathcal{X}$ to predict a response Y that takes values in a finite label space $\mathcal{Y}$. An especially important area of focus is long-tailed classification problems, where the label space is very large and the class frequencies vary by orders of magnitude. Such long-tailed distributions appear in problems such as species identification and medical diagnosis, and in order for prediction sets to be useful in such settings, they should be small enough to be inspected by a human while still reliably containing the true label, even when the true label belongs to a rare class.

Given a target miscoverage level $\alpha \in [0, 1]$, standard split conformal prediction (Vovk et al., 2005) provides a way to construct prediction sets $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ that satisfy

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

under the assumption that the observed data is exchangeable. This guarantee holds on average over the population, which can lead to undesirable behavior. Writing $p(y) = \mathbb{P}(Y = y)$, we can decompose marginal coverage as

$$\mathbb{P}(Y \in \mathcal{C}(X)) = \sum_{y \in \mathcal{Y}} p(y) \cdot \mathbb{P}(Y \in \mathcal{C}(X) \mid Y = y).$$

For a long-tailed distribution, where $p(y)$ is much larger for some y than others, the marginal guarantee places most of its weight on the frequent classes. A method may therefore attain marginal coverage while substantially undercovering tail classes. This phenomenon has been observed in recent work on conformal prediction for long-tailed classification, where standard conformal methods tend to overcover common classes and undercover rare classes (Ding et al., 2026; Liu et al., 2026).

A natural remedy is to seek *class-conditional coverage*:

$$\mathbb{P}(Y \in \mathcal{C}(X) \mid Y = y) \geq 1 - \alpha \qquad \text{for all } y \in \mathcal{Y}.$$

Class-conditional coverage treats every class separately and is therefore attractive in class-imbalanced settings. However, it can be too stringent in the long-tailed regime; when the calibration set contains only a few examples from a rare class, or none at all, class-specific quantile estimates become overly conservative. Consequently, classwise conformal methods may produce prediction sets that are very large and hence uninformative (Vovk, 2012; Ding et al., 2023). This creates a tension: standard (marginal) conformal prediction is efficient but can fail on rare classes, whereas class-conditional conformal prediction ensures coverage for every class but can be impractical when many classes have limited calibration data.

To balance this tension, Ding et al. (2026) introduced *macro-coverage*, which can be thought of as a relaxation of class-conditional coverage. This concept is inspired by macro-accuracy in multiclass classification (Lewis, 1991), where "macro" refers to the idea of zooming out to the class level before aggregating. Macro-coverage averages the class-conditional coverages uniformly over labels:

$$\text{MacroCov}(\mathcal{C}) := \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \mathbb{P}(Y \in \mathcal{C}(X) \mid Y = y). \tag{3.1}$$

Unlike marginal coverage, macro-coverage gives each class equal weight in the coverage criterion, regardless of its prevalence. By aiming to achieve macro-coverage, we can hope to do better on rare classes, without going to the extreme of asking for high class-conditional coverage for each class individually.

Our contribution. In this chapter, we propose to use *label-weighted conformal prediction* to produce prediction sets that directly achieve a macro-coverage guarantee (rather than a

stronger, but excessively stringent, class-conditional coverage guarantee), which no existing method is able to do. We prove that label-weighted conformal prediction can be used to not only achieve a macro-coverage guarantee but also to guarantee generalized notions of macro-coverage, where aggregation occurs at the level of an arbitrary grouping of classes and the weights of each group do not have to be equal. Notably, these weights can even be chosen after seeing partial information about the calibration data, allowing us to downweight coverage of groups that do not appear in the calibration data to avoid producing uninformative sets. Furthermore, given a desired generalized macro-coverage objective, we derive the form of the smallest sets that satisfy this objective and propose a corresponding conformal score function for approximating these theoretically optimal sets.

3.1.1 Related work

Class-conditional and group-conditional conformal prediction provide stronger guarantees than marginal conformal prediction by requiring coverage within each class or group (Vovk, 2012). In multiclass problems with many classes, however, classwise calibration can lead to large sets because rare classes may have very few calibration examples, necessitating large finite-sample adjustments. Clustered conformal prediction addresses this by targeting the data scarcity issue and groups data from classes with similar score distributions (Ding et al., 2023). On the other hand, rank-calibrated class-conditional conformal prediction directly targets set sizes by introducing a rank-based threshold to the set construction, while maintaining class-conditional coverage (Shi et al., 2024). Whereas these methods aim for class-conditional coverage, our work targets weighted averages of group-conditional coverages.

Our work is most directly motivated by conformal prediction for long-tailed classification (Ding et al., 2026), where the authors introduced macro-coverage and derived the prevalence-adjusted softmax score as an oracle-efficient score for this objective. Concurrently, Liu et al. (2026) proposed tail-aware conformal methods to reduce empirical disparities

between head and tail classes. These works highlight the limitations of marginal coverage in long-tailed settings, but the methods they propose still provide marginal rather than macro-coverage guarantees. We instead use label-weighted conformal calibration to obtain finite-sample guarantees for macro-coverage and generalized macro-coverage objectives.

Technically, our label-weighted conformal prediction approach is related to weighted conformal prediction, which has been used to address covariate shift, label shift, and other departures from exchangeability (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021; Barber et al., 2023). We build most directly on group-weighted conformal prediction (Bhattacharyya and Barber, 2024), adapting its finite-sample arguments to label-defined groups. Lastly, our optimal score construction is connected to least ambiguous set-valued classification (Sadinle et al., 2019), but specialized to the generalized macro-coverage objective studied here.

3.2 Method

We first generalize the concept of macro-coverage, as defined in (3.1), then define an algorithm for producing prediction sets guaranteed to have generalized macro-coverage of at least $1 - \alpha$ for any user-chosen $\alpha \in [0, 1]$.

Setting. Let $\alpha \in [0, 1]$ be a desired miscoverage level. Consider independent and identically distributed samples $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$, where $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ form the calibration dataset and (X_{n+1}, Y_{n+1}) is the test point, with Y_{n+1} unobserved. We operate in the split conformal prediction setting and assume access to a fixed conformal score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ trained on data that is separate from the calibration samples, where a larger value of $s(x, y)$ indicates that y is less likely to be the label corresponding to x.

3.2.1 Generalized macro-coverage

Let $g : \mathcal{Y} \to \mathcal{K}$ be a grouping function that assigns each class to a group and let $w : \mathcal{K} \to \mathbb{R}_{\geq 0}$ be a fixed function assigning weights to each group such that $\sum_{k \in \mathcal{K}} w(k) = 1$. For any

prediction set $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$, we define the (g, w) *macro-coverage* as

$$\text{MacroCov}_{g,w}(\mathcal{C}) := \sum_{k \in \mathcal{K}} w(k) \cdot \mathbb{P}\Big(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid g(Y_{n+1}) = k\Big),$$

the weighted average of group-conditional coverages. More generally, we can also choose the weights assigned to each group *after observing partial information about the calibration data*. Formally, denote the group membership of each calibration example as $G_i := g(Y_i)$ and record the realized group memberships of the calibration examples in

$$\mathcal{H} := \sigma(G_1, \ldots, G_n).$$

The weight function w can be any $\mathcal{H}$-measurable function. For such weight functions, we express the (g, w) macro-coverage as

$$\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) := \sum_{k \in \mathcal{K}} w(k) \cdot \mathbb{P}\Big(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid g(Y_{n+1}) = k, \mathcal{H}\Big).$$

By choosing g and w appropriately, practitioners can prioritize coverage across head and tail classes, clinically meaningful subpopulations, demographic groups, or other pre-specified partitions of interest.

Examples. The following are instantiations of generalized macro-coverage.

1. *Vanilla macro-coverage.* Using the identity group function $g(y) = y$ and uniform weights $w(y) = 1/|\mathcal{Y}|$ yields MacroCov, as defined in (3.1). We will refer to this as simply "macro-coverage."

2. *Tail-focused macro-coverage.* Let $\mathcal{Y}_{\text{tail}} \subset \mathcal{Y}$ denote the classes in the tail of the class distribution (e.g., $\mathcal{Y}_{\text{tail}}$ could be the 10% of classes with the fewest examples). In some settings, it may be especially important to correctly identify these rare classes. For some

upweighting factor $\lambda > 1$, the tail-focused macro-coverage (denoted $\text{MacroCov}_{\text{tail}}$) is defined to be the (g, w) macro-coverage with the identity grouping function $g(y) = y$ and weight function

$$w(y) = \begin{cases} \frac{\lambda}{W} & \text{if } y \in \mathcal{Y}_{\text{tail}} \\ \frac{1}{W} & \text{otherwise,} \end{cases}$$

where $W = \lambda|\mathcal{Y}_{\text{tail}}| + |\mathcal{Y} \setminus \mathcal{Y}_{\text{tail}}|$ is a normalization constant.

3. *Genus-level macro-coverage.* Rather than aggregating at the class level, in settings where classes have a hierarchical structure, we may instead wish to aggregate at a higher level. For example, in plant identification, each class is a plant species that is further grouped into a genus. We get GenusMacroCov by using the grouping function $g : \mathcal{Y} \to \{1, 2, \ldots, M\}$ that groups plant species by genus, where M denotes the total number of genera, and the weight function $w(k) = 1/M$ for $k \in [M]$ that assigns uniform weights to each genus. By targeting GenusMacroCov rather than MacroCov, we upweight the importance of species that are abundant *relative to other species in the same genus* (and downweight species that are relatively rare within a genus).

4. *Marginal coverage.* The standard conformal prediction objective of marginal coverage is also an instance of generalized macro-coverage, where g groups all classes into a single group (e.g., $g(y) = 1$ for all y), so the only valid weight function is $w(1) = 1$.

3.2.2 Label-weighted conformal prediction

For each group $k \in \mathcal{K}$, define $\mathcal{I}_k = \{i \in [n] : g(Y_i) = k\}$ to be the indices of calibration examples in group k and let $N_k = |\mathcal{I}_k|$ be the number of calibration examples in group k. We index the calibration scores corresponding to group k as $S_i^{(k)}$ for $i = 1, 2, \ldots, |\mathcal{I}_k|$.

We then define the (g, w) *label-weighted conformal prediction set* as

$$\mathcal{C}(x) := \left\{ y \in \mathcal{Y} : s(x, y) \leq \mathsf{Q}_{1-\alpha}\left(\sum_{k \in \mathcal{K}} w(k) P_k \right) \right\}, \tag{3.2}$$

where $\mathsf{Q}_{1-\alpha}(\mathcal{P}) := \min\{t \in \mathbb{R} : \mathbb{P}_{V \sim \mathcal{P}}(V \leq t) \geq 1 - \alpha)\}$ denotes the $1 - \alpha$ quantile of distribution P and

$$P_k = \begin{cases} \frac{1}{N_k} \sum_{j=1}^{N_k} \delta_{S_j^{(k)}} & \text{if } N_k > 0 \\ \delta_\infty & \text{if } N_k = 0. \end{cases}$$

In words, we compute the $1-\alpha$ quantile of the weighted average of empirical score distributions (where we substitute in a point mass at infinity for groups with no calibration examples), then use this quantile to threshold the score to determine which y values to include in the set.

Theorem 3.2.1. *Let $\mathcal{K}_+ = \{k \in \mathcal{K} : N_k > 0\}$ be the set of groups with non-zero calibration examples. Then the prediction set given in* (3.2) *satisfies*

$$\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) \geq 1 - \alpha - \max_{k \in \mathcal{K}_+} \frac{w(k)}{N_k}.$$

Note that conditional on $\mathcal{H}$, $\max_{k \in \mathcal{K}_+} \frac{w(k)}{N_k}$ is a constant. The proof of this result closely follows the proof of validity for the group-weighted conformal prediction approach in Chapter 3, but adapted to label-based groupings. The detailed proof is given in Appendix B.1. The fact that the guarantee holds conditional on $\mathcal{H}$ makes it stronger than the unconditional macro-coverage guarantee.

This theorem directly implies a procedure for producing prediction sets with a (g, w) macro-coverage guarantee of $1 - \alpha$, which we describe in Algorithm 1. In brief, we must simply compute the correction factor $\Delta := \max_{k \in \mathcal{K}_+} \frac{w(k)}{N_k}$, then construct the label-weighted conformal prediction set at a corrected miscoverage level $\alpha' = \alpha - \Delta$.

Algorithm 1 Label-weighted conformal prediction for guaranteed (g, w) macro-coverage

Input: Calibration data $\{(X_i, Y_i)\}_{i=1}^n$, score function s, grouping function $g : \mathcal{Y} \to \mathcal{K}$, group weights $w : \mathcal{K} \to \mathbb{R}_{\geq 0}$ with $\sum_{k \in \mathcal{K}} w(k) = 1$, target miscoverage $\alpha \in [0, 1]$, test point X_{n+1}

Output: Prediction set $\mathcal{C}(X_{n+1})$ with (g, w) macro-coverage of at least $1 - \alpha$

Compute calibration scores $S_i \leftarrow s(X_i, Y_i)$ and groups $G_i \leftarrow g(Y_i)$ for $i \in [n]$
for each group $k \in \mathcal{K}$ **do**
$\mathcal{I}_k \leftarrow \{i \in [n] : G_i = k\}$
$N_k \leftarrow |\mathcal{I}_k|$
$P_k \leftarrow \frac{1}{N_k} \sum_{i \in \mathcal{I}_k} \delta_{S_i}$ if $N_k > 0$, else $P_k \leftarrow \delta_\infty$
end for
$\mathcal{K}_+ \leftarrow \{k \in \mathcal{K} : N_k > 0\}$
$\alpha' \leftarrow \alpha - \max_{k \in \mathcal{K}_+} \frac{w(k)}{N_k}$
$\mathcal{C}(X_{n+1}) \leftarrow \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \mathsf{Q}_{1-\alpha'}(\sum_{k \in \mathcal{K}} w(k)\, P_k)\}$ if $\alpha' \geq 0$, else $\mathcal{C}(X_{n+1}) = \mathcal{Y}$

When implementing label-weighted conformal prediction, it can be useful to recast it as vanilla weighted conformal prediction, in which each calibration point is assigned a weight. Let $\mathcal{K}_0 := \{k \in \mathcal{K} : N_k = 0\}$ be the groups that have no calibration examples. Then (3.2) is equivalent to

$$\mathcal{C}(x) = \left\{ y \in \mathcal{Y} : s(x, y) \leq \mathsf{Q}_{1-\alpha} \left(\sum_{i=1}^{n} w_i \delta_{S_i} + \sum_{k \in \mathcal{K}_0} w(k) \delta_\infty \right) \right\}, \tag{3.3}$$

where the weight of calibration point (X_i, Y_i) is $w_i := w(g(Y_i))/N_{g(Y_i)}$. Note that the calibration point weights do not sum to one ($\sum_{i=1}^n w_i \neq 1$) if there are groups $k \in \mathcal{K}$ that have zero calibration examples. Thus, the $\sum_{k \in \mathcal{K}_0} w(k)\delta_\infty$ term ensures that we are taking the quantile of a valid probability distribution, with total mass of one. From this, it is also easy to see that if the total weight of the missing groups exceeds α, the quantile will be infinite, resulting in infinite sets.

3.2.3 Optimal score function for generalized macro-coverage

When we construct prediction sets, we generally have two objectives. First, we want the sets to satisfy our coverage criterion. Second, we want the sets to be as informative as possible, subject to the coverage constraint. In this section, we focus on the case where the weights $w(k)$ are fixed and do not depend on the calibration data. We propose a way to approximately achieve the theoretically optimal sets, where we use "optimal" to mean minimizing expected set size subject to the chosen generalized macro-coverage objective. Let $p(y) = \mathbb{P}(Y = y)$ be the probability of class y and define $\rho(k) = \sum_{y:g(y)=k} p(y)$ for $k \in \mathcal{K}$ to be the probability that a label belongs to group k. The following proposition characterizes the optimal prediction set that satisfies a chosen generalized macro-coverage constraint.

Proposition 3.2.2 (informal)**.** *The solution to*

$$\min_{\mathcal{C}:\mathcal{X}\to 2^{\mathcal{Y}}} \mathbb{E}|\mathcal{C}(X)| \quad \textit{subject to } \mathrm{MacroCov}_{g,w}(\mathcal{C}) \geq 1-\alpha$$

is of the form

$$\mathcal{C}_t^*(x) = \left\{ y \in \mathcal{Y} : \frac{w(g(y))}{\rho(g(y))} \cdot p(y|x) \geq t \right\} \quad \textit{for some threshold } t. \tag{3.4}$$

This result is obtained by showing that the (g, w) macro-coverage constraint can be expressed as $(g_I, \tilde{w})$ macro-coverage for the identity mapping $g_I(y) = y$ and weight function

$$\tilde{w}(y) := \frac{w(g(y)) \cdot p(y)}{\rho(g(y))}, \tag{3.5}$$

then applying Proposition 6 of Ding et al. (2026). The formal statement of the result and its proof is given in Appendix B.2.

Observe that we can rewrite (3.4) as

$$\mathcal{C}_t^*(x) = \left\{y \in \mathcal{Y} : s_{g,w}^*(x,y) \leq t\right\} \quad \text{for } s_{g,w}^*(x,y) = -\frac{w(g(y))}{\rho(g(y))} \cdot p(y|x).$$

Thus, combining Proposition 3.2.2 and Theorem 3.2.1, we know that if we were able to use $s_{g,w}^*$ as our score function and construct our prediction set according to Algorithm 1, this would yield the smallest set with (g, w) macro-coverage of at least $1-\alpha$. However, in practice, $s_{g,w}^*$ cannot be used since it depends on unknown population-level probabilities, so we instead substitute in plug-in estimates, yielding the score function

$$\hat{s}_{g,w}(x,y) = -\frac{w(g(y))}{\widehat{\rho}(g(y))} \cdot \widehat{p}(y|x). \tag{3.6}$$

We can get $\widehat{p}(y|x)$ by training any probabilistic classifier. $\widehat{\rho}(g(y))$ can be obtained using the label distribution in the classifier training dataset. Since Theorem 3.2.1 holds for any score function, it holds for $\hat{s}_{g,w}(x,y)$, so applying Algorithm 1 with this score function yields the desired generalized macro-coverage guarantee while approximating the smallest set achieving this guarantee.

3.2.4 Simultaneous coverage guarantees

In some applications, we want prediction sets that satisfy multiple coverage guarantees simultaneously. For example, we might want to guarantee that our sets have high macro-coverage *and* high marginal coverage. Formally, consider $m \geq 1$ constraints: for $j = 1, 2, \ldots, m$, g_j is a grouping function, $\mathcal{H}_j$ records the group membership of the calibration examples (as determined by g_j), w_j is an $\mathcal{H}_j$-measurable weighting function, and α_j is a miscoverage level. Our goal is to generate prediction sets that simultaneously satisfy

$$\text{MacroCov}_{g_j,w_j}(\mathcal{C} \mid \mathcal{H}_j) \geq 1-\alpha_j \quad \text{for } j = 1, 2, \ldots, m \tag{3.7}$$

Proposition 3.2.3. *Fix a score function s and run Algorithm 1 for each condition $j = 1, 2, \ldots, m$ and extract $\widehat{q}_j := \mathsf{Q}_{1-\alpha'_j}\left(\sum_{k \in \mathcal{K}_j} w_j(k)\, P_k^j\right)$ from line 9, where $\mathcal{K}_j$ are the groups given by g_j and P_k^j and α'_j are as computed by Algorithm 1. Then the prediction set*

$$\bar{\mathcal{C}}(x) := \{y \in \mathcal{Y} : s(x, y) \leq \max(\widehat{q}_1, \widehat{q}_2, \ldots, \widehat{q}_m)\} \tag{3.8}$$

satisfies (3.7).

Proofs for results in this section are given in Appendix B.3. Proposition 3.2.3 says that we can produce sets that satisfy multiple generalized macro-coverage conditions by simply taking the max of their score thresholds. As with the single coverage constraint problem, we can motivate a choice of score function for the multiple constraint setting by deriving the form of the optimal prediction set.

Proposition 3.2.4 (Informal)**.** *The solution to*

$$\min_{\mathcal{C}:\mathcal{X} \to 2^{\mathcal{Y}}} \mathbb{E}|\mathcal{C}(X)| \quad \text{s.t.} \quad \text{MacroCov}_{g_j, w_j}(\mathcal{C}) \geq 1 - \alpha_j \quad \textit{for all } j = 1, 2, \ldots, m$$

is of the form

$$\mathcal{C}_t^*(x) = \left\{ y \in \mathcal{Y} : \frac{p(y|x)}{p(y)} \cdot \sum_{j=1}^{m} \lambda_j \tilde{w}_j(y) \geq t \right\} \quad \textit{for some } \lambda_1, \lambda_2, \ldots, \lambda_m \textit{ and } t, \tag{3.9}$$

where $\tilde{w}_j : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is the class-level weighting function induced by (g_j, w_j), as defined in (3.5).

We state a formal version of this proposition in Appendix B.3 and also provide the proof. In practice, to approximate the optimal set given by (3.9) we use plug-in estimates of $p(y|x)$ and $p(y)$, as we did in Section 3.2.3. The additional difficulty in the simultaneous-coverage setting is that the score also depends on the vector $\lambda := (\lambda_1, \ldots, \lambda_m)$. However, in many applications, only a subset of the coverage constraints is active. For example, if we target

both marginal coverage and macro-coverage at the same miscoverage level α in long-tailed classification problems, it is often the case that the macro-coverage constraint is the more demanding one. This because it places more emphasis on covering rare classes, which classifiers struggle more with. In such cases, a procedure that attains the desired macro-coverage level may also attain the marginal coverage target, so that the marginal coverage constraint is inactive in the corresponding oracle problem. Note that for any inactive constraints, the corresponding λ_j in Proposition 3.2.4 will be zero. However, in general, we can set λ using data separate from the calibration data (e.g., held-out data or the data used for classifier training) to do grid search over potential values and selecting the value that gives the smallest estimated average set size.

3.3 Experiments

We perform experiments using real data to demonstrate that label-weighted conformal prediction achieves the macro-coverage objectives guaranteed by our theoretical results. We use the softmax score function, given by $s_{\mathsf{softmax}}(x, y) = -\hat{p}(y|x)$, which is also called the Least Ambiguous Classifier (LAC) score (Sadinle et al., 2019). We obtain $\hat{p}(y|x)$ from a fine-tuned ResNet-50 classifier initialized to ImageNet-pretrained weights and use the resulting softmax scores, following the procedure described in Ding et al. (2026). We also use the approximately optimal score function $\hat{s}_{g,w}$ defined in (3.6). Our primary evaluation metrics are macro-coverage (MacroCov) and average set size (AvgSize). We also compute marginal coverage (MarginalCov).

Methods. We compare three methods. The first baseline method is STANDARD conformal prediction, which constructs prediction sets as

$$\mathcal{C}(x) = \left\{ y : s(x, y) \leq \mathsf{Q}_{1-\alpha} \left(\frac{1}{n+1} \sum_{i=1}^{n} \delta_{S_i} + \frac{1}{n+1} \delta_{\infty} \right) \right\}$$

and guarantees $1-\alpha$ marginal coverage. The second baseline is CLASSWISE conformal prediction, which constructs prediction sets as

$$\mathcal{C}(x) = \left\{ y : s(x,y) \leq \mathsf{Q}_{1-\alpha} \left(\frac{1}{N_y+1} \sum_{i=1}^{N_y} \delta_{S_i^{(y)}} + \frac{1}{N_y+1} \delta_\infty \right) \right\},$$

where N_y and $S_i^{(y)}$ are defined analogously to N_k and $S_i^{(k)}$ in Section 3.2, for the special case where $\mathcal{K} = \mathcal{Y}$ (each class is its own group). CLASSWISE guarantees $1-\alpha$ class-conditional coverage for all classes. We compare these two baselines against LABEL-WEIGHTED conformal prediction, as described in Algorithm 1, for group and weight functions (g, w) described below.

Data. We use two large-scale image classification datasets: the plant species dataset Pl@ntNet-300K (Garcin et al., 2021) and the 2018 version of iNaturalist (Van Horn et al., 2015), which contains images of plants, animals, insects, and more. We choose these datasets as a test bed due to their high class imbalance, which is the setting in which macro-coverage is most relevant. To ensure reliable evaluation of macro-coverage, which requires computing class-conditional coverage, we use the truncated versions of these datasets created in Ding et al. (2026) by filtering out classes with too few examples for testing. This leaves 330 Pl@ntNet classes and 857 iNaturalist classes. Pl@ntNet is highly skewed, with the most common class being 100 times more common than the rarest class. iNaturalist is less skewed, with a ratio of 10. To compute standard errors, we combine the provided calibration and test splits (resulting in 98,061 total examples for Pl@ntNet and 50,906 for iNaturalist) and do our own random splitting into calibration and test datasets, where each example is assigned to the calibration dataset with probability 0.1. All metrics are computed using 20 random seeds.

3.3.1 Main results: Achieving macro-coverage

The goal of this experiment is to construct prediction sets with MacroCov $\geq 1-\alpha$ for a user-chosen $\alpha \in [0,1]$. To target this goal, we run LABEL-WEIGHTED with the identity grouping function $g(y) = y$ and uniform weights $w(y) = 1/|\mathcal{Y}|$ for all y. Note that for this g and w, the score $\hat{s}_{g,w}$ is equivalent to the prevalence-adjusted softmax score $s(x,y) = \hat{p}(y|x)/\hat{p}(y)$ proposed in Ding et al. (2026) for approximating the optimal score function for targeting macro-coverage.

As a high-level summary, this experiment empirically validates our two main theoretical results: First, that label-weighted conformal prediction can be used to achieve $1-\alpha$ macro-coverage for a user-chosen α, and second, that $\hat{s}_{g,w}$ achieves the smallest set size at a given macro-coverage level. Tables 3.1–3.2 show the result of running the various conformal prediction methods on Pl@ntNet and iNaturalist at levels $\alpha = 0.05$ and $\alpha = 0.1$. In the MacroCov column, we bold entries with macro-coverage at least $1-\alpha$, or within one standard error of it. In the AvgSize column, we bold the minimum average size among rows achieving the desired macro-coverage. We report the standard errors as subscripts. We use this presentation style throughout the rest of the chapter.

We observe that STANDARD, which guarantees $1-\alpha$ marginal coverage, does not in general achieve $1-\alpha$ macro-coverage. CLASSWISE results are only reported for s_{softmax} because CLASSWISE is invariant to class-dependent adjustments to score functions, so s_{softmax} and $\hat{s}_{g,w}$ produce identical results (for any g and w). CLASSWISE does achieve $1-\alpha$ macro-coverage, but it overshoots. Although overshooting the coverage target is not itself undesirable, here it comes at the cost of extremely large prediction sets. The cause of this overshooting is that CLASSWISE guarantees $1-\alpha$ class-conditional coverage, which is stronger than the macro-coverage condition we ask for. By contrast, LABEL-WEIGHTED is able to achieve the desired macro-coverage without overshooting. Furthermore, comparing the rows with bolded MacroCov entries (denoting that $1-\alpha$ macro-coverage is achieved), we see that

LABEL-WEIGHTED with $\hat{s}_{g,w}$ has the smallest average set size, supporting the theoretical results in Section 3.2.3.

Table 3.1: Targeting macro-coverage on Pl@ntNet

		Goal : MacroCov ≥ 0.9			Goal : MacroCov ≥ 0.95		
Method	Score	MarginalCov	MacroCov	AvgSize	MarginalCov	MacroCov	AvgSize
STANDARD	s_{softmax}	$0.901_{\pm 0.001}$	$0.861_{\pm 0.001}$	$2.8_{\pm 0.0}$	$0.951_{\pm 0.001}$	$0.929_{\pm 0.001}$	$5.9_{\pm 0.1}$
	$\hat{s}_{g,w}$	$0.900_{\pm 0.001}$	$0.886_{\pm 0.001}$	$2.2_{\pm 0.0}$	$0.950_{\pm 0.001}$	$0.939_{\pm 0.001}$	$3.7_{\pm 0.0}$
CLASSWISE	s_{softmax}	$0.940_{\pm 0.001}$	$\mathbf{0.940}_{\pm 0.001}$	$58.1_{\pm 1.4}$	$0.987_{\pm 0.000}$	$\mathbf{0.992}_{\pm 0.000}$	$263.4_{\pm 0.9}$
LABEL-WEIGHTED	s_{softmax}	$0.930_{\pm 0.001}$	$\mathbf{0.901}_{\pm 0.001}$	$4.0_{\pm 0.0}$	$0.966_{\pm 0.001}$	$\mathbf{0.951}_{\pm 0.001}$	$9.2_{\pm 0.1}$
	$\hat{s}_{g,w}$	$0.915_{\pm 0.001}$	$\mathbf{0.901}_{\pm 0.001}$	$\mathbf{2.4}_{\pm 0.0}$	$0.959_{\pm 0.000}$	$\mathbf{0.949}_{\pm 0.001}$	$\mathbf{4.4}_{\pm 0.0}$

Table 3.2: Targeting macro-coverage on iNaturalist

		Goal : MacroCov ≥ 0.9			Goal : MacroCov ≥ 0.95		
Method	Score	MarginalCov	MacroCov	AvgSize	MarginalCov	MacroCov	AvgSize
STANDARD	s_{softmax}	$0.901_{\pm 0.001}$	$0.894_{\pm 0.001}$	$10.1_{\pm 0.1}$	$0.950_{\pm 0.001}$	$0.946_{\pm 0.001}$	$26.9_{\pm 0.3}$
	$\hat{s}_{g,w}$	$0.901_{\pm 0.001}$	$0.898_{\pm 0.001}$	$7.2_{\pm 0.1}$	$0.950_{\pm 0.001}$	$\mathbf{0.949}_{\pm 0.001}$	$\mathbf{18.6}_{\pm 0.3}$
CLASSWISE	s_{softmax}	$0.941_{\pm 0.001}$	$\mathbf{0.940}_{\pm 0.001}$	$206.4_{\pm 3.0}$	$0.998_{\pm 0.000}$	$\mathbf{0.998}_{\pm 0.000}$	$826.9_{\pm 1.4}$
LABEL-WEIGHTED	s_{softmax}	$0.908_{\pm 0.001}$	$\mathbf{0.901}_{\pm 0.001}$	$11.3_{\pm 0.1}$	$0.955_{\pm 0.001}$	$\mathbf{0.951}_{\pm 0.001}$	$30.2_{\pm 0.4}$
	$\hat{s}_{g,w}$	$0.903_{\pm 0.001}$	$\mathbf{0.901}_{\pm 0.001}$	$\mathbf{7.5}_{\pm 0.1}$	$0.953_{\pm 0.000}$	$\mathbf{0.951}_{\pm 0.001}$	$19.6_{\pm 0.2}$

STANDARD achieves closer to the desired macro-coverage on iNaturalist compared to Pl@ntNet. This is likely because the class distribution of iNaturalist, although imbalanced, is closer to uniform than Pl@ntNet. Recall that for a perfectly uniform class distribution, MarginalCov equals MacroCov, so STANDARD, by targeting MarginalCov, can also indirectly do well on MacroCov when the class distribution is not too imbalanced.

3.3.2 Results beyond macro-coverage

We now demonstrate that label-weighted conformal prediction can be used to produce prediction sets that satisfy generalized macro-coverage conditions. We consider two notions

of generalized macro-coverage motivated by plant identification.

(a) Tail-focused macro-coverage. For biodiversity monitoring purposes, it is particularly important to correctly identify rare plant species. Motivated by this application, in this experiment, we seek to produce Pl@ntNet prediction sets with high tail-focused macro-coverage, as defined in Section 3.2.1. That is, we seek $\text{MacroCov}_{\text{tail}} \geq 1 - \alpha$. To construct $\mathcal{Y}_{\text{tail}}$, we take the 33 classes (10% of the 330 total classes) with the fewest examples. We upweight coverage on these classes by $\lambda = 10$.

The left panel of Table 3.3 shows results for $\alpha = 0.1$. We observe that LABEL-WEIGHTED successfully achieves the desired $1 - \alpha$ tail-focused macro-coverage level and that applying the optimal $\hat{s}_{g,w}$ score results in the smallest average set size, by many factors, compared to the commonly used softmax score. Results for $\alpha = 0.05$ are qualitatively similar (see Appendix B.4).

(b) Genus-level macro-coverage. If we are more interested in equalizing coverage at the genus level rather than the species level, we can aim to construct Pl@ntNet prediction sets satisfying $\text{GenusMacroCov} \geq 1 - \alpha$ (see Section 3.2.1). The right panel of Table 3.3 shows the results for $\alpha = 0.1$, with $\alpha = 0.5$ results deferred to Appendix B.4. We observe that LABEL-WEIGHTED achieves the desired GenusMacroCov level. STANDARD with the $\hat{s}_{g,w}$ score also manages to achieve $1 - \alpha$ GenusMacroCov, but it overshoots slightly, resulting in a suboptimal average set size. Conversely, LABEL-WEIGHTED achieves $1 - \alpha$ GenusMacroCov almost exactly and yields the smallest average set size.

3.4 Discussion

We propose label-weighted conformal prediction as a way to produce prediction sets with finite-sample guarantees of (generalized) macro-coverage. Rather than defaulting to the marginal guarantee of standard conformal, which implicitly assumes that the importance of

Table 3.3: Targeting generalized macro-coverage on Pl@ntNet at $\alpha = 0.1$. The left panel targets tail-focused macro-coverage; the right panel targets genus-level macro-coverage.

		Goal : $\text{MacroCov}_{\text{tail}} \geq 0.9$			Goal : $\text{GenusMacroCov} \geq 0.9$		
Method	Score	MarginalCov	$\text{MacroCov}_{\text{tail}}$	AvgSize	MarginalCov	GenusMacroCov	AvgSize
STANDARD	s_{softmax}	$0.901_{\pm 0.001}$	$0.656_{\pm 0.002}$	$2.8_{\pm 0.0}$	$0.901_{\pm 0.001}$	$0.883_{\pm 0.001}$	$2.8_{\pm 0.0}$
	$\hat{s}_{g,w}$	$0.900_{\pm 0.001}$	$0.849_{\pm 0.001}$	$2.3_{\pm 0.0}$	$0.900_{\pm 0.001}$	$\mathbf{0.941}_{\pm 0.000}$	$2.8_{\pm 0.0}$
CLASSWISE	s_{softmax}	$0.940_{\pm 0.001}$	$\mathbf{0.945}_{\pm 0.001}$	$58.1_{\pm 1.4}$	$0.941_{\pm 0.001}$	$\mathbf{0.949}_{\pm 0.001}$	$10.7_{\pm 0.2}$
LABEL-WEIGHTED	s_{softmax}	$0.984_{\pm 0.000}$	$\mathbf{0.901}_{\pm 0.002}$	$24.3_{\pm 0.9}$	$0.919_{\pm 0.001}$	$\mathbf{0.902}_{\pm 0.001}$	$3.4_{\pm 0.0}$
	$\hat{s}_{g,w}$	$0.949_{\pm 0.001}$	$\mathbf{0.899}_{\pm 0.002}$	$\mathbf{3.8}_{\pm 0.1}$	$0.764_{\pm 0.002}$	$\mathbf{0.902}_{\pm 0.001}$	$\mathbf{1.8}_{\pm 0.0}$

covering a class y is given by its prevalence, our work expands the scope of distribution-free guarantees that are achievable in classification settings. Using our framework, practitioners can precisely specify how much coverage for each class matters, define an aggregated coverage notion that respects the relative importances, then produce prediction sets with a guarantee that their customized coverage is at least $1 - \alpha$. Furthermore, using our score function proposal means that these sets will be approximately optimal, in the sense of set size.

Limitations and future work. Conformal prediction methods, including ours, have an inherent tradeoff between coverage and set size, which worsens as data becomes more limited. As the number of calibration examples of the smallest class (or, more generally, group) shrinks, the size of the α correction in Algorithm 1 grows, causing prediction set size to increase with it. However, the amount of increase is controlled by the weight assigned to the smallest group. Since our theory allows us to adjust the weight assigned to a group after seeing its calibration count, this allows us to avoid producing uninformative sets. Furthermore, the adjustment our method applies is favorable compared to that required by class-conditional conformal. Whereas the latter applies an adjustment for *every* class, our method only applies an adjustment for the "worst class" (the class with the highest weight, normalized by the number of calibration examples). Directions for future work include exploring practical ways of simultaneously achieving guaranteed macro-coverage and feature-conditional coverage

and operationalizing our framework for real-world classification tasks, such as the plant identification app from which the Pl@ntNet dataset from our experiments is derived.

CHAPTER 4

APPROXIMATING FULL CONFORMAL PREDICTION: DISTRIBUTION FREE GUARANTEES VIA THE TOURNAMENT CORRECTION

This chapter is based on the paper Bhattacharyya et al. (2026b) (joint work with Boxuan Zhang and Rina Foygel Barber). The two main variants of conformal prediction approaches are full conformal prediction and split conformal prediction (also called transductive and inductive) as introduced in Chapter 1. Full conformal prediction is widely considered to be statistically more efficient (since split conformal prediction requires data splitting, and therefore can lead to wider prediction intervals due to the resulting loss in sample size), but its implementation is computationally prohibitive, as it requires the underlying model to be refit for every candidate value in the response space. Existing computational shortcuts, such as using a discrete grid of values to approximate the full conformal prediction construction, frequently lack theoretical guarantees on marginal coverage and can fail in practice.

To address this limitation, we introduce a novel class of approximations to the full conformal prediction method, based on the idea of *tournaments*, which enables the construction of prediction sets with a rigorous marginal coverage guarantee of $1 - 2\alpha$. Under stability conditions, the theoretical coverage guarantee tightens to approximately $1 - \alpha$. This new framework generalizes the existing method of leave-one-out cross-conformal prediction, while allowing for flexible use of various existing approximation strategies.

4.1 Introduction

Let $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n+1$, and suppose that $Z_1, \ldots, Z_n, Z_{n+1}$ are exchangeable, with Y_{n+1} unobserved at test time. Given a target miscoverage level $\alpha \in (0, 1)$,

the goal is to construct a prediction set $\widehat{C}_{n,\alpha}(X_{n+1})$ such that

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\Big) \geq 1 - \alpha. \tag{4.1}$$

As described in Chapter 1, in *full conformal prediction* (Vovk et al., 2005), for each candidate label $y \in \mathcal{Y}$, one forms the augmented dataset $\mathcal{D}^y_{n+1} = \{Z_1, \ldots, Z_n, Z^y_{n+1}\}$ where $Z^y_{n+1} = (X_{n+1}, y)$ and the prediction set is defined as

$$\widehat{C}^{\text{full}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(Z^y_{n+1}; \mathcal{D}^y_{n+1}) \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n})}\left(\left\{s(Z_i, \mathcal{D}^y_{n+1})\right\}_{i=1}^n\right)\right\}. \tag{4.2}$$

Here $s(z; \mathcal{D})$ is a *score function*, with large values indicating that z appears unusual (does not 'conform') to the trends observed in $\mathcal{D}$. The function $s(z; \mathcal{D})$ is required to be *symmetric*, meaning that it is invariant to reordering the data points in $\mathcal{D}$. For intuition, we can consider a canonical example: the *residual score*, given by

$$s(z; \mathcal{D}) = |y - \hat{f}_{\mathcal{D}}(x)|, \tag{4.3}$$

where $\hat{f}_{\mathcal{D}}$ denotes a fitted model produced by a (symmetric) regression algorithm trained on the dataset $\mathcal{D}$. Full conformal prediction offers a distribution-free guarantee of predictive coverage: if the training and test data points $Z_1, \ldots, Z_{n+1}$ are exchangeable, then for any symmetric score function s, the marginal coverage guarantee (4.1) is guaranteed to hold (Vovk et al., 2005).

An important aspect of conformal prediction is the trade-off between statistical efficiency and computational cost. In particular, full conformal suffers from a high computational cost: one must refit or reevaluate the underlying model on the augmented dataset $\mathcal{D}^y_{n+1}$ for every possible value y, which is often infeasible (e.g., for a real-valued response, when $\mathcal{Y} = \mathbb{R}$). A popular alternative is *split conformal prediction* (Papadopoulos, 2008). Here the

sample is divided into a training set $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$, each of size $n/2$. Then the prediction interval is defined as

$$\widehat{C}^{\text{split}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s((X_{n+1}, y); \mathcal{D}_{\text{tr}}) \leq \text{Quantile}_{(1-\alpha)\left(1+\frac{1}{n/2}\right)} \left(\{ s(Z_i; \mathcal{D}_{\text{tr}}) \}_{i: Z_i \in \mathcal{D}_{\text{cal}}} \right) \right\}.$$

This procedure is computationally attractive since it requires only a single model fit, but it pays a statistical price for sample splitting—only half of the data is used to train the predictor, and the other half is used to calibrate the interval width. As a result, split conformal prediction often produces wider prediction sets than full conformal.

Motivated by this tension, many practical works consider approximations to full conformal that aim to retain its statistical benefits without its computational burden. Common examples include evaluating full conformal only on a grid of candidate values y, or using a naive plug-in score in which the model is fit once on the training data and then reused for all candidate responses. Although such methods can be effective in practice, they typically do not come with general finite-sample coverage guarantees and can fail badly in certain situations.

Our contribution. We introduce a framework for modifying approximations to the full conformal prediction set, using a *tournament correction*, that gives rise to a broad family of conformal prediction methods that are implementable in practice and enjoy a finite-sample coverage guarantee of $1 - 2\alpha$. In addition, under suitable stability assumptions, we show that a small inflation of the resulting interval can further improve the guarantee, yielding coverage close to $1 - \alpha$. This framework includes the existing leave-one-out cross-conformal procedure as a special case.

4.2 Preliminaries: approximations to full conformal

In this section, we summarize several common approaches towards approximating full conformal prediction, in the setting where $\mathcal{Y}$ is large (e.g., a real-valued response) so that computing

$\widehat{C}^{\text{full}}_{n,\alpha}(X_{n+1})$ is infeasible.

Example 0: Deletion. We first consider a baseline method, which neglects to use conformal prediction entirely: we train a model on the training dataset $\mathcal{D}_n$ only, and use the training residuals or scores to form a prediction set. To compare to full conformal (4.2), we can think of this approach as taking an approximation via deletion: we simply delete (or rather, fail to include) the test point into the model training process. That is, instead of computing the score function using the augmented dataset $\mathcal{D}^y_{n+1}$, we use only the training points $\mathcal{D}_n$, leading to the prediction set

$$\widehat{C}^{\text{delete}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(Z^y_{n+1}; \mathcal{D}_n) \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n})}(\{s(Z_i; \mathcal{D}_n)\}_{i=1}^n) \right\}.$$

Of course, this above approach will often lead to severe undercoverage, since we are ignoring the issue of overfitting. For example, if we use the residual score (4.3), we obtain the prediction set

$$\widehat{C}^{\text{delete}}_{n,\alpha}(X_{n+1}) = \hat{f}_{\mathcal{D}_n}(X_{n+1}) \pm \text{Quantile}_{(1-\alpha)(1+\frac{1}{n})}(\{|Y_i - \hat{f}_{\mathcal{D}_n}(X_i)|\}_{i=1}^n),$$

which is likely much too narrow since the training residuals $|Y_i - \hat{f}_{\mathcal{D}_n}(X_i)|$ are typically smaller than the test point's residual $|Y_{n+1} - \hat{f}_{\mathcal{D}_n}(X_{n+1})|$. Nonetheless, we can think of this as a (very inaccurate) 'approximation' to full conformal prediction.

Example 1: Rounding. In practice, the most commonly used approximation for full conformal prediction is to simply use rounding: that is, we consider a discrete grid of y values and only fit the model at those values (Lei et al., 2018). Concretely, instead of computing the score function using the augmented dataset $\mathcal{D}^y_{n+1}$, we replace y with its rounded value. Writing $[y]$ to denote the operation of rounding y to a finite grid of values $y^{(1)}, \ldots, y^{(M)} \in \mathcal{Y}$,

we then have the prediction set

$$\widehat{C}^{\text{round}}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(Z^y_{n+1}; \mathcal{D}^{[y]}_{n+1}) \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n})}(\{s(Z_i; \mathcal{D}^{[y]}_{n+1})\}_{i=1}^n) \right\}.$$

Note that, in this approach, we only need to fit a finite number of models: for instance, for the residual score (4.3), we only need to train M models, $\hat{f}_m = \hat{f}_{\mathcal{D}^{y^{(m)}}_{n+1}}$ for each $m = 1, \ldots, M$, and can then compute the prediction set as

$$\widehat{C}^{\text{round}}_{n,\alpha}(X_{n+1}) = \bigcup_{m=1,\ldots,M} \Big\{ y \in \mathcal{Y} : [y] = y^{(m)}, \\ |y - \hat{f}_m(X_{n+1})| \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n})}(\{|Y_i - \hat{f}_m(X_i)|\}_{i=1}^n) \Big\}.$$

Example 2: One-step update. In settings where the fitted model is obtained by solving an optimization problem, another common approach is the one-step update: for example, if $\hat{f}_{\mathcal{D}}$ is obtained by optimizing some penalized likelihood method, we might train the model on $\mathcal{D}_n$, and then add in the hypothesized test point (X_{n+1}, y) by taking one step of gradient descent. More generally, we approximate the full conformal set (4.2) by adding in the test point approximately through a one-step update (Martinez et al., 2023; Tailor et al., 2025): we write $\text{Update}(\hat{f}_{\mathcal{D}}, z)$ to denote that a trained model $\hat{f}_{\mathcal{D}}$ is then updated approximately with a new data point z.

For example, if we assume a parametric predictor $f_\theta : \mathcal{X} \to \mathbb{R}$ and $\widehat{f}_{\mathcal{D}} = f_{\widehat{\theta}(\mathcal{D})}$ where

$$\widehat{\theta}(\mathcal{D}) := \arg\min_\theta \; L_{\mathcal{D}}(\theta).$$

Then $\text{Update}(\hat{f}_{\mathcal{D}}, z) = f_{\widetilde{\theta}}$ where $\widetilde{\theta}$ is obtained by taking one gradient descent step:

$$\tilde{\theta} := \hat{\theta}(\mathcal{D}) - \eta \, \nabla L_{\mathcal{D} \cup \{z\}}(\hat{\theta}(\mathcal{D})),$$

assuming that the loss is differentiable (here $\eta > 0$ is some step size parameter). Let $\ell(z, \hat{f})$ denote any loss that compares a data point z against a fitted model $\hat{f}$, e.g., we might choose the absolute residual score $\ell(z, \hat{f}) = |y - \hat{f}(z)|$ where $z = (z, y)$. We then have the prediction set

$$\widehat{C}_{n,\alpha}^{\text{one-step}}(X_{n+1}) = \Bigg\{ y \in \mathcal{Y} : \ell\left(Z_{n+1}^y, \text{Update}(\hat{f}_{\mathcal{D}_n}, Z_{n+1}^y)\right) \leq Q_{(1-\alpha)(1+\frac{1}{n})}\left(\left\{\ell\left(Z_i, \text{Update}(\hat{f}_{\mathcal{D}_n}, Z_{n+1}^y)\right)\right\}_{i=1}^n\right) \Bigg\}.$$

Example 3: Approximating the Bayesian PPD. Finally, we consider a Bayesian setting, where our aim is to use the posterior predictive distribution (PPD) for the score. Given a prior π on parameter θ, and a likelihood $f_\theta(y \mid x)$ for the conditional distribution of $Y \mid X$, the PPD is given by

$$\mathbb{E}_{\theta \sim \pi(\cdot \mid \mathcal{D})}[f_\theta(y \mid x)],$$

i.e., the expected likelihood when θ is sampled from the posterior, given some dataset $\mathcal{D}$. Since we would like to retain values of Y that are likely under the PPD, we would ideally like to use the score $s((x, y); \mathcal{D}) = -\mathbb{E}_{\theta \sim \pi(\cdot \mid \mathcal{D})}[f_\theta(y \mid x)]$. In practice, the PPD is generally estimated with a Monte Carlo approximation, and

$$s((x, y); \mathcal{D}) = -\frac{1}{K} \sum_{k=1}^{K} f_{\theta_k}(y \mid x) \text{ where } \theta_1, \ldots, \theta_K \sim \pi(\cdot \mid \mathcal{D}).$$

However, it is impractical to compute scores $s(\cdot; \mathcal{D}_{n+1}^y)$ for all values of y (since we cannot draw samples from each possible posterior distribution). This motivates an 'add-one-in' (AOI) approximation to the PPD, using importance weighting (Fong and Holmes, 2021; Deliu and Liseo, 2025): after sampling $\theta_k \sim \pi(\cdot \mid \mathcal{D}_n)$, the appropriate importance weight for adding (X_{n+1}, y) into the training set is proportional to $f_{\theta_k}(y \mid X_{n+1})$, and we therefore have the

prediction set

$$\widehat{C}_{n,\alpha}^{\text{AOI-PPD}}(X_{n+1}) = \Bigg\{ y \in \mathcal{Y} : -\sum_{k=1}^{K} f_{\theta_k}(y \mid X_{n+1}) \cdot f_{\theta_k}(y \mid X_{n+1}) \leq Q_{(1-\alpha)(1+\frac{1}{n})} \left(\left\{ -\sum_{k=1}^{K} f_{\theta_k}(y \mid X_{n+1}) \cdot f_{\theta_k}(Y_i \mid X_i) \right\}_{i=1}^{n} \right) \Bigg\},$$

where $\theta_1, \ldots, \theta_K \sim \pi(\cdot \mid \mathcal{D}_n)$.

4.2.1 Unifying notation

To summarize these examples, we introduce a common notation that encompasses all of these various approximations to full conformal prediction. Write

$$s_{\text{approx}}(z; \mathcal{D} + \{z'\})$$

which computes a score for data point z, as compared to a dataset $\mathcal{D}$ and an additional data point z'. The idea is that this is an approximation to the score computed on a dataset that includes z',

$$s_{\text{approx}}(z; \mathcal{D} + \{z'\}) \approx s(z; \mathcal{D} \cup \{z'\}).$$

However, $s_{\text{approx}}(z; \mathcal{D} + \{z'\}$ is required to be symmetric in $\mathcal{D}$ but *not* in the combined dataset $\mathcal{D} \cup \{z'\}$. We then define a prediction set

$$\widehat{C}_{n,\alpha}^{\text{approx}}(X_{n+1}) = \Bigg\{ y \in \mathcal{Y} : s_{\text{approx}}(Z_{n+1}^y; \mathcal{D}_n + \{Z_{n+1}^y\}) \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n})} \left(\left\{ s_{\text{approx}}(Z_i; \mathcal{D}_n + \{Z_{n+1}^y\}) \right\}_{i=1}^{n} \right) \Bigg\}. \tag{4.4}$$

Each of the four examples described above can be expressed as an instance of this new notation, by choosing the score s_{approx} as the following:

- **Example 0: Deletion.** We define $s_{\text{approx}}(z; \mathcal{D} + \{z'\}) = s(z; \mathcal{D})$, i.e., the data point

z' is not included when computing the score.

- **Example 1: Rounding.** Let $z' = (x', y')$, and let $[y']$ denote the value of y' after rounding to the grid. Then $s_{\text{approx}}(z; \mathcal{D} + \{z'\}) = s(z; \mathcal{D} \cup \{(x', [y'])\})$, i.e., the response y' is rounded to $[y']$ when computing the score.

- **Example 2: One-step update.** Let $z = (x, y)$. Then, writing $\text{Update}(\cdot)$ to denote a one-step model update rule (e.g., gradient descent) as before, we have $s_{\text{approx}}(z; \mathcal{D} + \{z'\}) = \ell\Big(z, \text{Update}(\hat{f}_{\mathcal{D}}, z')\Big)$.

- **Example 3: Approximating the Bayesian PPD.**[1] Let $z = (x, y)$ and $z' = (x', y')$. We then define $s_{\text{approx}}(z; \mathcal{D} + \{z'\}) = -\sum_{k=1}^{K} f_{\theta_k}(y \mid x) \cdot f_{\theta_k}(y' \mid x')$, where $\theta_1, \ldots, \theta_K \sim \pi(\cdot \mid \mathcal{D})$.

4.2.2 Other related works

Beyond the four classes of methods considered here, prior work on approximating full conformal prediction has mainly taken two forms: exploiting model-specific structure to compute full conformal sets more efficiently, and proving validity of cheaper approximations under algorithmic stability. For example, Lei (2019), Ndiaye and Takeuchi (2019), Ndiaye and Takeuchi (2023), and Cherubin et al. (2021) develop efficient exact or approximate full conformal algorithms in structured settings such as penalized regression, nearest-neighbor, and kernel methods. Relatedly, cross-conformal prediction (Vovk, 2015) and the jackknife+/CV+ methods of Barber et al. (2021) are methods closely related to ours; as we show later, leave-one-out cross conformal is a special case of our framework. More recently, Ndiaye (2022) and Lee and Zhang (2025) study computationally efficient conformal procedures under stability

1. We remark that this example uses a *randomized* score, since the score function depends on random samples drawn from the posterior. Conformal prediction theory extends naturally to the setting of randomized scores (Vovk et al., 2005). The results of our work hold for this setting as well, as we will show in the Appendix.

assumptions. These works are complementary to ours: they recover validity through stability assumptions, whereas our tournament based corrected methods achieve finite-sample validity without requiring any assumptions, with stability used only later to sharpen the guarantee.

4.3 Tournament correction

In this section we formally introduce our method. We mainly discuss how a slight modification to the score function s_{approx} can be used to develop methods that come with $1-2\alpha$ coverage guarantees. At a high level, the idea is that we will now treat *two* data points approximately in the training process—the test point $n+1$ along with one training point i, rather than treating only the test point approximately as in the examples above. Extending our earlier notation, we consider an approximate score of the form

$$s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) \tag{4.5}$$

which computes the score at data point z when the dataset $\mathcal{D}$ and the data points $\{z', z''\}$ are used to train the model possibly in asymmetric ways. As before, we require $s_{\text{approx}}(z; \mathcal{D} + \{z', z''\})$ to be symmetric in $\mathcal{D}$, and now also require $s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = s_{\text{approx}}(z; \mathcal{D} + \{z'', z'\})$.

Using this score we can define a prediction set

$$\widehat{C}^{\text{tourn}}_{n,\alpha}(X_{n+1}) := \left\{ y : \sum_{i=1}^{n} \mathbf{1}\Big\{ s_{\text{approx}}(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) > s_{\text{approx}}(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) \Big\} < (1-\alpha)(n+1) \right\}, \tag{4.6}$$

where $\mathcal{D}_{n\setminus i} = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\}$ i.e., all the training data points with the i-th point removed. As compared to an approximation to full conformal (any approximation of the form given in (4.4)), this general framework provides a modification to this approximation,

which we refer to as the *tournament correction* (we will explain the idea behind this name, below).

4.3.1 Examples

Next, we describe how each of the approximate methods stated in Section 4.2 can be modified using this framework.

- **Example 0: Deletion.** $s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = s(z; \mathcal{D})$ i.e., we remove both the data points z' and z''. The key difference from the approximate deletion method is that in the approximate method, to find the score at the training and the test point only the test point was removed and this introduced asymmetry in the training and the test scores, whereas, now when we compare the scores at Z^y_{n+1} and Z_i we remove both the i-th point and the test point. This method is exactly equivalent to the leave-one-out cross-conformal method proposed by Vovk (2015); Vovk et al. (2018).

- **Example 1: Rounding.** If $z' = (x', y')$ and $z'' = (x'', y'')$, define $s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = s(z; \mathcal{D} \cup \{(x', [y'])\} \cup (x'', [y'']))$. Again in this case, the difference from the approximate method is that in the tournament-corrected method (4.6), both the training point Z_i and the hypothesized test point Z^y_{n+1} are rounded, when computing the scores.

- **Example 2: One-step update.** As before, writing Update$(\cdot)$ to denote a one-step model update rule (e.g., gradient descent), we have $s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = \ell\Big(z, \text{Update}(\hat{f}_{\mathcal{D}}, \{z', z''\})\Big)$. As compared to the approximate method, in the tournament-corrected method (4.6) both the training point Z_i and the hypothesized test point Z^y_{n+1} are added into the trained model via a one-step update.

- **Example 3: Bayesian posterior predictive density.** Let $z = (x, y)$, $z' = (x', y')$, and $z'' = (x'', y'')$. We then define $s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = -\sum_{k=1}^{K} f_{\theta_k}(y \mid x) \cdot f_{\theta_k}(y' \mid$

$x')f_{\theta_k}(y'' \mid x'')$, where $\theta_1, \ldots, \theta_K \sim \pi(\cdot \mid \mathcal{D})$. Thus, in the tournament-corrected method (4.6), both the training point Z_i and the hypothesized test point Z^y_{n+1} are added into the posterior with the add-one-in weight-based approximation.[2]

4.3.2 Theoretical guarantee

In practice, we expect that the tournament-corrected method (4.6) will result in approximately $1-\alpha$ coverage, since it provides an approximation to the full conformal prediction set. Theoretically, however, we would need assumptions to ensure that the approximation is accurate. Nonetheless, the tournament correction ensures that, in the worst case, the miscoverage can increase by at most a factor of 2.

Theorem 4.3.1 (Validity)**.** *Let $\mathcal{D}_{n+1} = \{Z_1, \ldots, Z_{n+1}\}$ be an exchangeable dataset where $Z_i = (X_i, Y_i)$. For any score s_{approx} of the form* (4.5)*, the conformal prediction set defined in* (4.6) *satisfies the marginal coverage guarantee*

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1 - 2\alpha.$$

For the special case of leave-one-out cross-conformal (which is equivalent to our tournament-corrected method, in the setting of Example 0, as described above), this coverage result is established by Barber et al. (2021).

The broader result of Theorem 4.3.1 is proved via a generalization of the tournament matrix style argument used in Barber et al. (2021) (the complete proof is provided in Appendix C.1.1). The tournament matrix arises in the following way: we can see that for each training point $i \in [n]$, the test point $n+1$ plays a 'game' against this training point, where both Z_i and Z^y_{n+1} are removed from the training set (i.e., we compute scores of the form $s_{\text{approx}}(\cdot\,; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\})$), and then the winner of the 'game' is determined by

2. See Appendix C.5 for a more formal view of the validity of this randomized score function.

which point, Z_i or Z^y_{n+1}, has the higher score. The prediction set is then constructed by examining how many 'games' the test point wins, within this tournament, which is why we refer to this methodology as a *tournament correction*.

A tradeoff: validity versus computation. The result of Theorem 4.3.1 shows that the tournament correction allows us to restore theoretical validity when computing an approximation to the full conformal prediction set. However, we remark that this comes at a computational cost, since the tournament correction requires training models with each data point $i \in [n]$ deleted in turn. In some cases, like the Bayesian example though this can be avoided by using rejection sampling as discussed in Appendix C.5. We can view this as a tradeoff between statistical validity and computational cost; see Section 4.6 for further discussion.

4.4 Experiments: coverage and length

In this section, we study the empirical behavior of the approximate conformal methods from Section 4.2 and their tournament-corrected counterparts from Section 4.3. Our goals are twofold:

- In practice, existing approximations often work well, providing (nearly) $1-\alpha$ coverage empirically. In such scenarios, we aim to show that the tournament correction leads to minimal changes in the existing methods: the resulting prediction sets are nearly the same.

- On the other hand, the tournament correction ensures that coverage can never decrease below $1-2\alpha$, even in a worst-case scenario; in contrast, informal approximations to full conformal prediction do not offer any comparable distribution-free guarantees. Therefore, we aim to show that, in certain challenging settings, existing approximate

methods can substantially lose coverage, while the corresponding tournament-corrected methods remain reliable.

Taken together, these experiments show that the proposed correction is broadly useful: it does not lead to overly conservative results in scenarios where existing approximations are already accurate, but in regimes where existing approximations break down, the corrected method continues to provide valid uncertainty quantification.

4.4.1 Simulations

We first compare each approximation of full conformal (Examples 0, 1, 2, and 3) to its tournament-corrected version in a simulated setting, where data is generated from a Gaussian linear model: the training and test data points are drawn as

$$X_i \overset{\text{i.i.d.}}{\sim} N(0, I_p), \qquad Y_i = X_i^\top \beta^\star + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1).$$

The coefficient vector $\beta^\star$ is drawn uniformly from the sphere satisfying $\|\beta^\star\|_2 = \sqrt{10}$. Details on the implementation of each of the methods is provided in Section C.2.

For each method, we report empirical coverage and average prediction-set length, over 100 independent trials.

Figure 4.1 show that for smaller values of p, both the approximate and the corrected methods have similar performance in terms of both length and coverage. As $p \to n$ and we move towards more unstable regimes, all the approximate methods perform very poorly with coverage falling down far below the target threshold. These show that as the underlying model becomes unstable, the uncorrected approximations are no longer reliable, whereas all the tournament-corrected methods maintain a level of coverage close to $1 - \alpha = 0.9$.

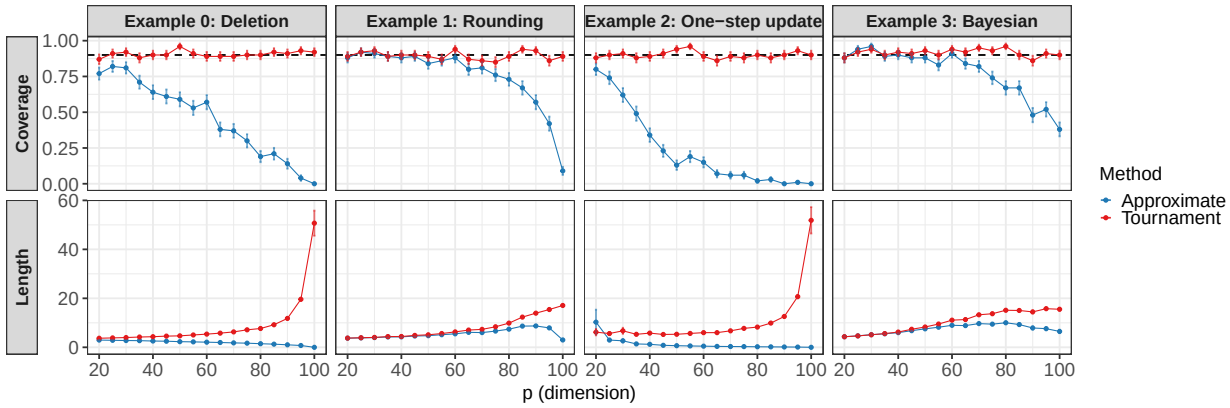

Figure 4.1: Empirical coverage and average prediction-set length for approximate and tournament-corrected conformal methods for $n = 100$, with dimension varying over $p = 20, 25, \ldots, 100$. A dashed line indicates the nominal coverage level $1 - \alpha = 0.9$. Results are averaged over 100 independent trials, with standard error bars shown.

4.4.2 Real-data examples

As we have seen in the simulations, in settings where the existing approximate methods exhibit the desired coverage level empirically, the tournament-corrected method behaves nearly identically: that is, the tournament-corrected method does not result in an overly conservative prediction set in scenarios where no correction is needed. In this next experiment, we study the same question in the setting of real data. Concretely, we replicate two experiments where approximate full conformal methods have been used in practice, and observe that the tournament-corrected method returns nearly identical results in both cases. Both experiments use the Diabetes dataset (Efron et al., 2004).

First, we consider the rounding approximation (Example 1). For implementing full conformal with rounding (i.e., the uncorrected approximation), we follow the implementation of Lee and Zhang (2025) (which studies full conformal as a baseline comparison for a different proposed method), using the residual score (4.3) for a model trained via ridge-regularized Huber regression. Second, we consider the Bayesian PPD add-one-in approximation (Example 3). We replicate an experiment conducted by Fong and Holmes (2021), using a Gaussian linear regression likelihood with a prior over the parameters.

For both real data examples, we implement the approximate conformal prediction method exactly as in the respective papers (using the same choice of coverage level $1-\alpha$ as used in those papers), and then implement the tournament-corrected version of the same method. Results are shown in Table 4.1. All implementation details are provided in Appendix C.3. We see that, in both cases, the uncorrected approximation shows approximately the desired level of coverage. Moreover, in both cases, the resulting coverage and prediction set length is nearly identical for the approximate method (4.4) and the tournament-corrected method (4.6), confirming that in practice, the tournament correction typically is not overly conservative when existing approximations are already providing the desired coverage level empirically.

Table 4.1: Comparison of the length and coverage of approximate and tournament-corrected versions on the `diabetes` data. Standard errors are shown in parentheses.

Method Based on paper	$1-\alpha$	Type	Coverage	Length
Rounding (Lee and Zhang, 2025)	0.9	Approximate	0.8879 (0.0037)	2.8244 (0.0045)
		Tournament-corrected	0.9008 (0.0036)	2.8862 (0.0046)
Bayesian PPD (Fong and Holmes, 2021)	0.8	Approximate	0.7986 (0.0054)	1.8552 (0.0085)
		Tournament-corrected	0.7986 (0.0055)	1.8552 (0.0084)

4.5 Coverage under stability

The guarantee of Theorem 4.3.1 ensures that coverage cannot be worse than $1-2\alpha$, but in practice we expect to see coverage $\approx 1-\alpha$, since the tournament-corrected method approximates the full conformal prediction set. In this section, we show that this can be proved with an additional assumption: if the score function resulting from fitted model is *stable* enough in the sense that it is not too sensitive to small changes in the way the data is used to train the model, then a 2ϵ-inflated prediction set which we define below satisfies a stronger coverage guarantee.

Assumption 4.5.1 (stability)**.** *There exist* $\epsilon \geq 0$ *and* $\nu \in [0,1]$ *such that for every* $i \in [n]$*,*

$$\mathbb{P}\left(\left|s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z_{n+1}\}\Big) - s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_n + \{Z_{n+1}\}\Big)\right| \leq \epsilon\right) \geq 1 - \nu.$$

The probability above is taken with respect to the distribution of $Z_1, \ldots, Z_{n+1}$. Essentially, the assumption requires that changing the role of a single training point Z_i (i.e., whether it is included in the training data $\mathcal{D}_n$, or is instead added to $\mathcal{D}_{n\setminus i}$ via an approximation) is not likely to lead to a large change in the score at Z_{n+1}. Assumptions of this flavor are common in learning theory and have also been used recently to justify computationally efficient approximations to full conformal prediction (Bousquet and Elisseeff, 2002; Hardt et al., 2016; Ndiaye, 2022; Lee and Zhang, 2025).

Next, define an inflated version of the tournament-corrected set:

$$\widehat{C}^{\text{tourn},2\epsilon}_{n,\alpha}(X_{n+1}) := \Bigg\{ y : \sum_{i=1}^{n} \mathbf{1}\Big\{ s_{\text{approx}}\Big(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z^y_{n+1}, Z_i\}\Big) > s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}\Big) + 2\epsilon \Big\} < (1-\alpha)(n+1) \Bigg\}. \quad (4.7)$$

Theorem 4.5.2 (Coverage under stability)**.** *If* D_{n+1} *is an exchangeable dataset, then under Assumption 4.5.1, the* 2ϵ*-inflated interval as defined in* (4.7) *satisfies the coverage guarantee*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}^{\text{tourn},2\epsilon}_{n,\alpha}(X_{n+1})\right) \geq 1 - \alpha - 3\sqrt{\nu}.$$

When ν is small, the bound $1 - \alpha - 3\sqrt{\nu}$ is therefore close to the nominal $1-\alpha$ level, providing a stronger coverage guarantee as compared to the $1-2\alpha$ guarantee of Theorem 4.3.1. The proof of this theorem is given in Appendix C.1.2.

4.5.1 Stability comparison

Theorem 4.5.2 shows that the tournament-corrected methods admit a stronger coverage guarantee under a single stability condition, namely Assumption 4.5.1. A natural question is whether the original approximate prediction sets in (4.4) can also enjoy a comparable near-$(1-\alpha)$ guarantee: in other words, if we are willing to assume stability, is it still necessary to apply the tournament correction?

In fact, we will now see that the tournament correction continues to be useful. This is because the type of stability assumption needed for the (uncorrected) approximation (4.4) to guarantee coverage, is much stronger than the type of stability required by Assumption 4.5.1. Specifically, in order for $\widehat{C}_{n,\alpha}^{\text{approx}}(X_{n+1})$ to lead to a similar coverage guarantee as Theorem 4.5.2, we would need to require that

$$\begin{aligned}
&\mathbb{P}\left(\left|s_{\text{approx}}(Z_{n+1}; \mathcal{D}_n + \{Z_{n+1}\}) - s(Z_{n+1}; \mathcal{D}_{n+1})\right| \leq \epsilon\right) \geq 1-\nu, \\
&\mathbb{P}\left(\left|s_{\text{approx}}(Z_i; \mathcal{D}_n + \{Z_{n+1}\}) - s(Z_i; \mathcal{D}_{n+1})\right| \leq \epsilon\right) \geq 1-\nu.
\end{aligned} \tag{4.8}$$

These conditions would lead to a coverage result for an inflated interval, analogous to the result of Theorem 4.5.2 (see Appendix C.4 for details).

However, the first condition in (4.8) is substantially more restrictive than Assumption 4.5.1: it requires that training only approximately on Z_{n+1}, will not lead to a large change in the score for Z_{n+1}, i.e., *for the same data point that was removed*. For example, if we use rounding (so Y_{n+1} is replaced by a value on the grid), an overparameterized method is likely to substantially change its prediction at X_{n+1}—and thus this type of stability condition would not hold. (See Barber et al. (2021) for further discussion of the difference in these stability conditions for the specific setting of Example 0, i.e., where the approximation is carried out via data point deletion).

To illustrate this difference empirically, for each of the four different methods, we estimate the stability properties of the score function. Specifically, for each method, we plot the

values (ν, ϵ) for which the conditions (4.8) hold, in order to assess whether the uncorrected approximation would likely have approximate coverage; we also plot the values of (ν, ϵ) for which Assumption 4.5.1 holds, in order to assess whether the tournament-corrected approximation would have $\approx 1 - \alpha$ coverage as in Theorem 4.5.2. The data is drawn from the same distribution as in Section 4.4.1 and the fitted model and all the implementational details are the same as Section 4.4.1. For this simulation, we consider $n = 100, p = 20$ which is a relatively stable regime and report the values over 200 iterations.

Even then across all four methods, the tournament-corrected methods exhibit stability curves that lie substantially below the curves corresponding to the uncorrected approximations. This indicates that Assumption 4.5.1 holds for small values of ϵ and ν, leading to favorable coverage properties (via Theorem 4.5.2) for the tournament-corrected method; in contrast, the first stability condition in (4.8) only holds at much larger values of ϵ and ν, meaning that the uncorrected approximations may not lead to coverage.

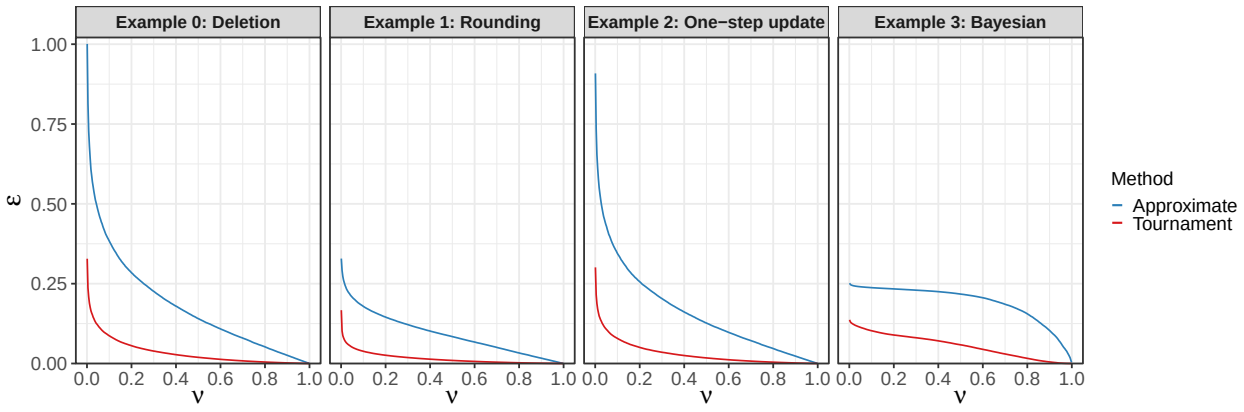

Figure 4.2: Comparison of the stability properties for the uncorrected approximation (see (4.8)) versus the tournament-corrected approximation (see Assumption 4.5.1).

4.6 Discussion

In this work, we develop a principled way of approximating full conformal prediction with provable distribution-free coverage guarantees. While there exist quite a few methods that

are used as an approximation of full conformal, such as using a grid of $\mathcal{Y}$ values or adding in the test point via a one-step model update instead of refitting the model for each value of y, in general such approximations do not offer assumption-free theoretical coverage guarantees, and can fail in unstable regimes. Our proposed method offers a *tournament correction* of these approximations, which ensures at least $1 - 2\alpha$ coverage with no additional assumptions.

Limitations and future work. It is important to note that implementing the tournament correction will typically increase computational cost: namely, n times the cost incurred by their approximate counterparts. In Appendix C.5, we discuss how we can get around this issue for the Bayesian case by using rejection sampling techniques, however, in the general case this may not always be possible. To address this, we may also define K-fold versions of the tournament correction (analogous to K-fold cross-conformal prediction, for the case of data deletion, i.e., Example 0). We leave the question of such extensions to future work.

CHAPTER 5

CONDITIONING ON POSTERIOR SAMPLES FOR FLEXIBLE FREQUENTIST GOODNESS-OF-FIT TESTING

This chapter is based on the paper Bhaduri et al. (2026) (joint work with Ritwik Bhaduri, Rina Foygel Barber and Lucas Janson). Tests of *goodness of fit* are used in nearly every domain where statistics is applied. One powerful and flexible approach is to sample artificial data sets that are exchangeable with the real data under the null hypothesis (but not under the alternative), as this allows the analyst to conduct a valid test using *any* test statistic they desire. Such sampling is typically done by conditioning on either an exact or approximate sufficient statistic, but existing methods for doing so have significant limitations, which either preclude their use or substantially reduce their power or computational tractability for many important models. In this work, we propose to condition on samples from a Bayesian posterior distribution, which constitute a very different type of approximate sufficient statistic than those considered in prior work. Our approach, *approximately co-sufficient sampling via Bayes*, considerably expands the scope of this flexible type of goodness-of-fit testing. We prove the approximate validity of the resulting test, and demonstrate its utility on three common null models where no existing methods apply, as well as its outperformance on models where existing methods do apply.

5.1 Introduction

Goodness-of-fit (GoF) testing refers to the problem of testing whether a particular family of distributions (the "model") is consistent with the observed data: for instance, does the data follow a Gaussian distribution? GoF testing is heavily studied in statistics and has applications across domains, including biology (Guo and Thompson, 1992), economics (Cowell et al., 2009), astronomy (Acharya and Kashyap, 2024), and finance (Frezza, 2014). In this

work we will consider *parametric* null hypotheses of the form

$$H_0 : X \sim f_\theta \quad \text{for some } \theta \in \Theta, \tag{5.1}$$

where $\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ represents a parametric family of densities. This is hypothesis testing with a *composite* null hypothesis space (i.e., the model whose goodness-of-fit is being tested). In GoF testing, it is common to leave the alternative hypothesis unspecified in general: we can interpret the test as asking whether the model appears to fit the data, or not. However, in specific settings, it may be the case that we design a test with a particular alternative in mind, as we will see in the examples developed later on. A key challenge of GoF testing problems is that often, any alternative hypothesis of interest would typically be very high-dimensional or even infinite-dimensional: for instance, if the data does not follow a Gaussian distribution (the null), then perhaps it instead follows some heavier-tailed distribution (a nonparametric alternative). In such settings, any powerful test statistic is often too complex to permit any theoretical calculation of its null distribution, even asymptotically.

If the null hypothesis were simple, i.e., if $\Theta = \{\theta_0\}$, then any function T of the data X could be used as a test statistic and converted to a valid p-value by fixing a positive integer M, generating i.i.d. samples $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ from f_{θ_0}, and computing

$$\text{pval}_T\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) = \frac{1 + \sum_{m=1}^{M} \mathbb{1}\{T(\widetilde{X}^{(m)}) \geq T(X)\}}{M+1}. \tag{5.2}$$

To see why the above p-value is valid, note that the numerator of (5.2) is the rank of $T(X)$ among $T(X), T(\widetilde{X}^{(1)}), \ldots, T(\widetilde{X}^{(M)})$, and under H_0, these $M+1$ random variables are exchangeable. Unlike standard parametric tests like the likelihood ratio, Wald, and score tests, which each prescribe a specific test statistic based on an alternative hypothesis space that must satisfy certain strong regularity conditions, the absolute flexibility in the choice of test statistic function T in (5.2) allows the user to leverage any domain knowledge, prior

information, and believed structure under the alternative (no matter what it is) to make the test as powerful as possible.

The appealing strategy of the previous paragraph works because H_0 is a simple null—it contains just one distribution, and so the analyst knows what distribution to sample from in order to get exchangeable copies of the data. The idea of *co-sufficient sampling* extends this strategy to *composite* null hypotheses by conditioning on a sufficient statistic for the unknown parameter under the null model, rendering the data distribution conditionally parameter-free under H_0 so the analyst knows once again what *conditional* distribution to sample from in order to get exchangeable copies of the data. Exact co-sufficient sampling (Bartlett, 1937; Engen and Lillegård, 1997; Agresti, 1992; Stephens, 2012) can only be fruitfully applied for a very narrow class of parametric null models, motivating approximate versions (Barber and Janson, 2022; Zhu and Barber, 2023; Xie and Huang, 2025; Awan and Cai, 2020) that apply to a much broader class of models. Yet these existing works' limitations in scope, power, and computational efficiency still prevent the idea of co-sufficient sampling from realizing its full methodological potential in terms of generality and performance. Section 5.2 will review in more detail the landscape of existing methods based on the idea of co-sufficient sampling, as well as their shortcomings and other related work.

Our contributions. This work presents a novel approach to approximate co-sufficient sampling that uses draws from a Bayesian posterior distribution as the approximate sufficient statistic that is conditioned on; we refer to our method as *approximate co-sufficient sampling via Bayes* (aCSS-B). We prove approximate exchangeability for aCSS-B's samples, and as a corollary, prove approximate type-I error control when the p-value (5.2) is constructed with its samples. Note that while our method uses Bayesian sampling as a tool, these results provide frequentist guarantees that do not rely on an assumed prior. We demonstrate its performance on a suite of examples. The main advantages of aCSS-B over prior work are that it applies considerably more broadly than previous methods and, even when previous methods

apply, aCSS-B is often more powerful, and may be more straightforward to implement and more computationally efficient. We demonstrate aCSS-B's improved generality for three canonical parametric null models in which no prior methods apply: a group-sparse linear model (5.4.5), a low-rank matrix model (5.4.4), and a linear spline model (5.4.6) (and to complement these findings, additional experiments show that aCSS-B performs well relative to existing methods on examples where prior methods do apply).

5.2 Background

5.2.1 Overview: inference via approximate exchangeability

Our objective is to construct a valid and powerful test of the composite parametric null hypothesis (5.1), and we reduce this problem to one of sampling (approximately) exchangeable copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ of the data X under H_0, as once this is accomplished (5.2) provides a valid p-value (for any test statistic function T). Note that for such a p-value to also provide *powerful* inference, we need sufficient "diversity" among the sampled copies; for instance, if we set $\widetilde{X}^{(m)} = X$ for all m, then trivially the copies would be exchangeable, but the p-value (5.2) would be deterministically equal to 1 for any choice of test statistic (i.e., valid but powerless). Next, we review some background on methods that use this type of strategy for inference.

5.2.2 Co-sufficient sampling

Co-sufficient sampling (CSS) (Bartlett, 1937; Engen and Lillegård, 1997; Agresti, 1992; Stephens, 2012) identifies a sufficient statistic $S(X)$ for the null model H_0 and then samples the copies $\widetilde{X}^{(m)}$ i.i.d. (and independent of X) from the conditional distribution of $X \mid S(X)$, denoted $f_\theta(\cdot \mid S(X))$, which by definition of sufficiency is a known conditional distribution (does not depend on θ) under H_0. It is then immediate that $X, \widetilde{X}^{(1)}, ..., \widetilde{X}^{(M)}$ are conditionally i.i.d., and hence exchangeable, under H_0.

However, for this method to be powerful, $S(X)$ should not contain too much information about X, otherwise conditioning on it will force all the sampled copies to be so similar to X that the p-value (5.2) has little or no power under the alternative (with the extreme worst case being $S(X) = X$). Unfortunately, in many cases there does not exist any such sufficient statistic; see Barber and Janson (2022) for examples where this issue arises, which include logistic regression, curved exponential families (even as simple as two independent Gaussians with equal means and unequal variances), heavy-tailed distributions, and latent variable models.

5.2.3 Approximate co-sufficient sampling

Approximate co-sufficient sampling (aCSS) was introduced in Barber and Janson (2022) to address the limitations of CSS testing. The key idea is to identify a statistic $S(X)$ that is *nearly* sufficient, in the sense that $f_\theta(\cdot \mid S(X))$ depends very weakly on θ, while still not revealing too much information about X. Then using copies $\widetilde{X}^{(m)}$ sampled from $f_{\hat{\theta}}(\cdot \mid S(X))$, for some estimator $\hat{\theta}$, results in approximate validity due to approximate sufficiency of $S(X)$ and high power due to the the randomness left in the distribution of X conditioned on $S(X)$. In particular, Barber and Janson (2022) propose using, for both $S(X)$ and $\hat{\theta}$, the maximizer of the null log likelihood plus a random linear perturbation. They prove approximate validity under conditions resembling those for asymptotic normality of the maximum likelihood estimator (MLE), and empirically demonstrate aCSS's high power on a range of examples.

Several extensions of the aCSS method have been proposed to handle more complex settings—in particular, settings where due to high dimensionality or non-regularity of the null model, the MLE is inconsistent. These extensions enable adding regularization, via either constraints or a penalty on θ, in addition to the random perturbation to the log likelihood in defining $S(X)$ and $\hat{\theta}$ in aCSS. Concretely, Zhu and Barber (2023) develop a version of aCSS that allows for linear constraints $a_i^\top \theta \leq b_i$ (or regularization via a corresponding penalty

function, $\max_i a_i^\top \theta$, which includes settings such as the Lasso). Xie and Huang (2025) extend further to allow for a group-wise penalty of the form $\sum_j \rho(\|\theta_{G_j}\|_2)$, where ρ is smooth (e.g., the group Lasso); their work also allows regularization via nonlinear constraints, $G_i(\theta) \leq b_i$, for functions G_i that are either smooth or are an ℓ_p norm.

Limitations of existing aCSS methods. The existing methods in the aCSS family—the original method proposed by Barber and Janson (2022), and the extensions to regularized versions of aCSS developed by Zhu and Barber (2023) and Xie and Huang (2025)—all suffer from certain limitations in scope. A key limitation is that these methods all require the null model to be open and convex (i.e., $\Theta \subseteq \mathbb{R}^d$ is open and convex), excluding models with any kind of structural constraint such as sparsity or low rank. While structures such as sparsity can sometimes be encoded via regularization, this is a special case and is not true for many other types of structure: for instance, none of the existing variants can encode a low-rank matrix constraint (we will consider such an example further, below). Finally, in practice even when one of these methods can be applied to a given problem, it can be computationally expensive and sensitive to tuning parameters, and can have low power.

5.2.4 Additional related work

There is an enormous literature on GoF testing dating back many decades, with classical examples including the χ^2, score, likelihood ratio, and Wald tests (see, e.g., GoF textbooks such as D'Agostino, 2017). GoF testing continues to be a subject of contemporary research, with recent papers considering challenging parametric or nonparametric null hypotheses and innovative tests (see, e.g., Candès et al., 2018; Berrett et al., 2020; Lundborg et al., 2022; Ramdas et al., 2022; Gangrade et al., 2023; Sen and Sen, 2014; Saha and Ramdas, 2024; Chwialkowski et al., 2016). What sets CSS and aCSS type methods, including the method proposed in this work, apart from the rest of the GoF testing literature is their flexibility in the choice of test statistic: CSS and aCSS methods are *wrapper* methods in the sense that

they can wrap around *any* test statistic, in principle enabling them to achieve high power for many different alternative distributions via unrestricted alternative-specific choices of test statistic.

A popular way to perform resampling based tests is a parametric bootstrap style procedure, in which resamples are generated from $f_{\widehat{\theta}}$, where $\widehat{\theta}$ is an estimator of θ_0. Although such a procedure generally does not enjoy finite-sample validity guarantees and is known to fail in some settings (see, e.g., Barber and Janson (2022)), it often performs well empirically and is widely used in practice in many fields such as in propensity-score analyses (Adusumilli, 2018) and population genetics (Emerson et al., 2001). One work closely related to the aCSS literature is Awan and Cai (2020), which proposes a sampling method that approximates co-sufficient sampling. However, they do not prove their sampling can be used for approximately valid testing, and it is unclear how to use their approximation error bounds to do so.

A final point is that the method we propose relies heavily on ideas from Bayesian sampling such as Markov chain Monte Carlo (MCMC) (Chib and Greenberg, 1995; Casella and George, 1992) and the Laplace approximation (Shun and McCullagh, 1995), though we emphasize that our problem statement, method, and guarantees remain purely frequentist in nature.

5.3 Main results

5.3.1 Method

At a high level, any p-value of the form (5.2) relies on constructing copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ that are, at least approximately, exchangeable with the data X under the null hypothesis of goodness-of-fit. Like the original aCSS method and many related approaches, our goal in aCSS-B is to provide a mechanism for sampling these copies.

To quantify the notion of approximate exchangeability we introduce the following definition.

Definition 5.3.1 (Distance to exchangeability)**.** *For any integer $k \geq 1$ and any set of random variables $A_1, \ldots, A_k$ with a joint distribution, define*

$$\mathrm{d}_{\text{exch}}\ (A_1, \ldots, A_k) = \inf \left\{ \mathrm{d}_{\mathrm{TV}} \left((A_1, \ldots, A_k), (B_1, \ldots, B_k) \right) : B_1, \ldots, B_k \text{ are exchangeable } \right\}.$$

Here d_{TV} denotes the total variation distance, and the infimum is taken over all sets of k random variables $B_1, \ldots, B_k$ with an exchangeable joint distribution.

This definition allows us to guarantee approximate validity of the p-value. Indeed, by construction, we have

$$\mathbb{P}\left(\mathrm{pval}_T \left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \right) \leq \alpha \right) \leq \alpha + \mathrm{d}_{\text{exch}}\ \left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \right), \tag{5.3}$$

for any $\alpha \in [0, 1]$. This is because, for any exchangeable random variables $(B_0, B_1, \ldots, B_M)$, it holds that $\mathbb{P}(\mathrm{pval}_T(B_0, B_1, \ldots, B_m) \leq \alpha) \leq \alpha$, and therefore,

$$\mathbb{P}\left(\mathrm{pval}_T \left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \right) \leq \alpha \right) \leq \alpha + \mathrm{d}_{\mathrm{TV}} \left(\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \right), (B_0, B_1, \ldots, B_M) \right).$$

To construct copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ that are approximately exchangeable with X, under the null hypothesis of goodness-of-fit, the aCSS-B procedure operates as follows: after sampling the data X (assumed to be drawn from the density f_{θ_0}, for some unknown $\theta_0 \in \Theta$), we first define a prior density π on Θ and generate B draws from the corresponding posterior, denoted by $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$. We then estimate the distribution of $X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$ (note that we cannot compute this conditional distribution exactly, since θ_0 is unknown), and sample the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ from this estimated distribution.

Now we turn to calculating the required components of the procedure. Formally, we will assume that each density f_θ in our parametric family is a density with respect to some common base measure $\nu_{\mathcal{X}}$ on $\mathcal{X}$, and will choose a prior with density π with respect to a

base measure ν_Θ on Θ. The posterior distribution of $\theta \mid X$ is then defined by the density

$$\pi(\theta \mid x) = \frac{\pi(\theta) f_\theta(x)}{\bar{f}_\pi(x)}, \tag{5.4}$$

which is again a density with respect to ν_Θ, where

$$\bar{f}_\pi(x) = \int_\Theta \pi(\theta) f_\theta(x) \, \mathsf{d}\nu_\Theta(\theta)$$

denotes the marginal density of X, under the prior $\theta \sim \pi$. (Formally, to ensure that future quantities will be well-defined, we will assume from this point on that the support of $f_\theta(x)$ is contained in the support of $\bar{f}_\pi(x)$, for every θ—that is, $\bar{f}_\pi(x)$ is positive for any value x we might observe. For instance, if f_θ has the same support for every θ, then this assumption is satisfied.)

Next, a straightforward calculation shows that the conditional distribution of $X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$ has density

$$g_\pi^{\theta_0}(x \mid \widehat{\theta}_{1:B}) \propto f_{\theta_0}(x) \cdot \prod_{b=1}^{B} \frac{f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)} \tag{5.5}$$

with respect to $\nu_\mathcal{X}$. Since θ_0 is unknown, however, we will replace $f_{\theta_0}(x)$ with $\bar{f}_\pi(x)$, so that the copies $\widetilde{X}^{(m)}$ are sampled according to density

$$g_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}}. \tag{5.6}$$

These steps describe the process of generating the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ for the aCSS-B procedure; see Algorithm 1 for a complete definition of the method.

Before presenting our finite-sample theoretical results, we first give some intuition for why the aCSS-B procedure can be expected to provide approximately exchangeable copies of the data X.

Intuition: the Bayesian setting. Suppose that we are in a true Bayesian setting, where

Algorithm 1: aCSS-B method

Given: prior density π on Θ, and test statistic $T : \mathcal{X} \to \mathbb{R}$.

Data: Observe data $X \sim f_{\theta_0}$.

1 Generate B posterior samples,

$$\widehat{\theta}_1, \ldots, \widehat{\theta}_B \mid X \overset{\text{i.i.d.}}{\sim} \pi(\cdot \mid X).$$

2 Generate M copies of the data,

$$\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)} \mid X, \widehat{\theta}_{1:B} \overset{\text{i.i.d.}}{\sim} g_\pi(\cdot \mid \widehat{\theta}_{1:B}),$$

where the density $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is defined as

$$g_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}}.$$

3 Compute the p-value

$$\text{pval}_T\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) = \frac{1 + \sum_{m=1}^M \mathbb{1}\{T(\widetilde{X}^{(m)}) \geq T(X)\}}{M+1}.$$

the unknown parameter θ_0 was in fact drawn from the prior density π. In that case, (5.5) gives the conditional density of $X \mid (\theta_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)$. But after marginalizing over θ_0, (5.6) is the conditional density of $X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$—this is the exact, rather than approximate, conditional distribution for X, and thus drawing the copies $\widetilde{X}^{(m)}$ from this density leads to exact exchangeability, and validity of the test.

Intuition: sufficiency of the posterior. We will now consider an alternative viewpoint on the intuition behind the method, without assuming a Bayesian setting—that is, we return to the setting of a fixed θ_0. After observing B draws from the posterior (for a large B), we have approximately observed the posterior density of $\theta \mid X$ with respect to the base measure ν_Θ, which is computed in (5.4). The following standard result, proved for completeness in Appendix D.2.1, tells us that $\pi(\cdot \mid X)$, which we view as statistic of the data (i.e., a map from the data X to this density function), is in fact a minimal sufficient statistic.

Proposition 5.3.2. *Let $X \sim f_\theta$ for the parametric family $\{f_\theta : \theta \in \Theta\}$. Fix any prior on Θ, with a positive density π (relative to some base measure ν_Θ). Then the posterior density $\pi(\cdot \mid X)$ is a minimal sufficient statistic for X.*

In other words, by conditioning on $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$, we are conditioning on an approximation to the posterior distribution, which can be approximated arbitrarily well by the empirical measure of its samples (Vapnik and Chervonenkis, 2013). Since the posterior distribution is a sufficient statistic, this means that the copies $\widetilde{X}^{(m)}$ will be approximately exchangeable with X (for large B), since we have nearly removed the effect of the unknown parameter θ_0.

Comparison to aCSS and its extensions. The core idea of our method is similar to the aCSS method of Barber and Janson (2022) (and the regularized extensions of aCSS proposed by Zhu and Barber (2023) and Xie and Huang (2025)), since our idea is to condition on an approximately sufficient statistic for X in order to generate the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$. However, our method makes a very different choice of information to condition on: while aCSS and its regularized variants each condition on a noisy version of the MLE for θ_0 given X (or, a noisy version of the regularized MLE), we instead condition on a large number B of draws from the posterior distribution of $\theta \mid X$. This different approximate sufficient statistic only requires us to sample from the posterior distribution of $\theta \mid X$, as opposed to requiring optimization of the likelihood, which can be computationally more expensive in many problems (Ma et al., 2019). We will also see in our empirical examples below that aCSS-B offers greater flexibility and is applicable to a wider range of problems.

5.3.2 Theoretical Guarantees

As we have seen in Proposition 5.3.2, the posterior density $\pi(\cdot \mid X)$ is a minimal sufficient statistic for the data X; this explains why, as $B \to \infty$, we can expect that the aCSS-B method should be valid. In practice, of course, we run aCSS-B with a finite set of posterior samples—leading to approximately exchangeable $(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$—thereby yielding an

approximately valid test.

In this section, we examine the behaviour of the method when B is finite, to understand the role of B in the validity of the aCSS-B approach, for which we will first need a few definitions. Below, let π_0 denote any prior density on Θ, with respect to the same base measure ν_Θ. For intuition, we should think of π_0 as being concentrated near the unknown true parameter value θ_0.

Definition 5.3.3 (Prior concentration)**.** *For any prior with density π_0, define*

$$\epsilon(\pi_0) = \mathrm{d_{TV}}\left(f_{\theta_0}, \bar{f}_{\pi_0}\right), \tag{5.7}$$

where as before, $\bar{f}_{\pi_0}$ denotes the density of the marginal likelihood of X when we draw $\theta \sim \pi_0$, i.e.,

$$\bar{f}_{\pi_0}(x) = \int_\Theta f_\theta(x)\pi_0(\theta)\, \mathsf{d}\nu_\Theta(\theta).$$

To give a concrete example, we can consider taking π_0 to be the restriction of π to some small neighborhood around θ_0—that is, π_0 is the distribution of $\theta \sim \pi$ conditional on the event $\|\theta - \theta_0\|_2 \leq r$. If r is sufficiently small, we would expect $\epsilon(\pi_0)$ to be small as well.

Definition 5.3.4 (Posterior sensitivity)**.** *Given observed data $X \in \mathcal{X}$, let $\pi(\cdot \mid X)$ (respectively, $\pi_0(\cdot \mid X)$) denote the posterior distribution of θ under the prior $\theta \sim \pi$ (respectively, $\theta \sim \pi_0$). Define*

$$\Delta(\pi_0) = \mathbb{E}_{\theta_0}\left[\mathrm{d}_{\chi^2}\left(\pi_0(\cdot \mid X) \,\|\, \pi(\cdot \mid X)\right)^{1/2}\right], \tag{5.8}$$

where d_{χ^2} denotes the χ^2 divergence between distributions, and where the expected value is taken with respect to $X \sim f_{\theta_0}$.

Note that, if the data X carries strong information for inferring the parameter θ, we might expect that the posterior is not affected too much by the choice of prior—and therefore

$\Delta(\pi_0)$ would not be too large, even for some π_0 that is strongly concentrated near θ_0 (so that $\epsilon(\pi_0) \approx 0$).

We emphasize that this prior π_0 will appear in the upper bound on the type-I error in Theorem 5.3.5, but does *not* need to be specified for running the algorithm.

With the above definitions, we state the main result, which bounds the distance to exchangeability—and therefore, the type-I error of the aCSS-B procedure.

Theorem 5.3.5. *After observing the data X, let $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ be sampled as in Algorithm 1, for some positive prior density π on Θ. Then, if $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$,*

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\},$$

where the infimum is taken over all densities π_0 on Θ with respect to base measure ν_Θ such that the support of $\bar{f}_{\pi_0}(x)$ contains the support of $f_\theta(x)$ for all θ.

The proof of this theorem is given in Appendix D.2.2. Theorem 5.3.5 states that the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ are approximately exchangeable with X. In particular, applying (5.3), this implies that for any predefined test statistic $T : \mathcal{X} \to \mathbb{R}$ and rejection threshold $\alpha \in [0, 1]$, the p-value satisfies

$$\mathbb{P}\left(\mathrm{pval}_T\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \alpha\right) \leq \alpha + \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\}.$$

5.3.3 Challenges: sampling from the posterior, and sampling the copies

As described in Algorithm 1, the application of aCSS-B to any hypothesis testing problem requires two sampling steps—first, drawing B independent samples $\widehat{\theta}_b$ from the posterior distribution, and second, sampling M independent copies $\widetilde{X}^{(m)}$ from $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. Both of these steps may be computationally challenging in practice, and in this section we briefly describe some solutions.

First we consider sampling the posterior draws $\widehat{\theta}_b$. In practice, it is often only possible to sample from the posterior with MCMC techniques, and sampling exactly i.i.d. draws is computationally infeasible; this is a standard challenge in Bayesian statistics. In our implementation, we employ MCMC techniques such as Gibbs sampling or Metropolis–Hastings (depending on the specific setting), along with adequate thinning to reduce autocorrelation between samples (Riabiz et al., 2022) and burn-in (Stewart and Johnson, 2009) so that the samples are more representative of the target distribution. These techniques, and the extent to which they approximate i.i.d. sampling from the posterior sampling, are well-studied in the Bayesian literature (Gelman et al., 1995; Gagniuc, 2017; Barber, 2012), so we do not explore their theoretical properties further here.

Our second challenge is the problem of sampling the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ i.i.d. from $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. This is again often infeasible to perform exactly—and unlike the problem of sampling from the posterior, this is no longer a standard challenge in Bayesian statistics, since it is not typical to condition on more than one draw from the posterior. Therefore, this requires careful treatment, to ensure that any approximations we take do not invalidate our finite-sample type-I error guarantees. In Appendix D.1, we provide a theoretical analysis of this challenge. Specifically, we treat two key issues that arise: first, that MCMC sampling techniques introduce dependence among the copies, and second, that the marginal density $\bar{f}_\pi(x) = \int_\Theta \pi(\theta) f(x;\theta)\, \mathrm{d}\theta$ (which appears in the denominator of $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$) may be hard to calculate exactly in some problem settings. Our results in Appendix D.1 establish type-I error control even when we use approximate strategies for sampling the copies, and bound the increase in type-I error when we use an approximation of $\bar{f}_\pi(x)$.

5.4 Experiments

5.4.1 Simulation Design

In this section, we examine the performance of aCSS-B in five simulated examples.[1] The first two examples are in settings where aCSS or its regularized extensions can be applied, and the remaining three examples represent setups that are outside the scope of these methods.

Across all five examples, we implement aCSS-B with $B = 25$ posterior draws, and with $M = 300$ copies $\widetilde{X}^{(m)}$ sampled for running the test. Informally, we find aCSS-B to be easy to tune—we (successfully) use the same default value of B for all five examples, and observe that standard, default priors work well out-of-the-box in each setting. For each example in the following subsections, details for the process of sampling the posterior draws $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$, and for sampling the copies $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$, can be found in the corresponding subsection of Appendix D.3.

For each experiment, after computing a p-value, we test the null hypothesis at the level $\alpha = 0.05$. We compute the type-I error rate for the setting where the null hypothesis is true (that is, $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$), and compute power for settings where the null is false, across a range of different signal strengths. In each example we compare this power against an oracle, which is given access to θ_0—that is, the oracle can calculate the null distribution of the test statistic $T(X)$ by simply sampling from f_{θ_0}. Results are reported after averaging over 500 independent trials, with standard error bars shown in the figures.

1. Code for reproducing all experiments is available at `https://github.com/Ritwik-Bhaduri/aCSS-B/`.

5.4.2 Logistic Regression

The model. We consider the logistic regression model as in Barber and Janson (2022). Here $X_i \in \{0,1\}$ is binary and the covariate $Z_i \in \mathbb{R}^d$ is treated as fixed; the likelihood function is

$$f(x;\theta) = \prod_{i=1}^{n} \left(\frac{e^{Z_i^\top \theta}}{1+e^{Z_i^\top \theta}}\right)^{x_i} \cdot \left(\frac{1}{1+e^{Z_i^\top \theta}}\right)^{1-x_i} \tag{5.9}$$

with parameter $\theta \in \Theta = \mathbb{R}^d$, with dimension $d = 5$ and number of observations $n = 100$. (We can interpret $f(x;\theta)$ as a density with respect to the base measure $\nu_{\mathcal{X}}$ on $\mathcal{X} = \{0,1\}^n$ that places mass 1 on each point $x \in \mathcal{X}$, i.e., the counting measure.) The true parameter vector is given by $\theta_0 = 0.2 \cdot \vec{1}_d$. We test a conditional independence hypothesis by considering a variable $Y_i \in \mathbb{R}$ whose conditional distribution given Z_i is independent of X_i under the null hypothesis, but is dependent under the alternative:

$$Y \mid (X,Z) \sim a\left(b(Z) + \beta_0^\top Z \cdot \mathbb{1}_{X=0} + \beta_1^\top Z \cdot \mathbb{1}_{X=1}\right),$$

where $a(t) = t + 0.5t^3$, $b(z) = 0.5\sum_{j=1}^{5}\left(z_j\right)_+$, $\beta_0 = c \cdot \vec{e}_1$, and $\beta_1 = c \cdot \vec{e}_5$, where $\vec{e}_j$ is the jth basis vector and where $c \in \{0, 0.1, 0.2, \ldots, 1\}$ indicates the signal strength (with $c = 0$ corresponding to the null hypothesis). As in Barber and Janson (2022, Example 1, Section 4.5.2), we use a test statistic based on sliced inverse regression; see that paper for more details.

Can we apply existing aCSS methods? Barber and Janson (2022) apply aCSS to this problem, and we will compare aCSS-B against the exact same implementation of aCSS as used for this problem in that paper.

Choice of prior for implementing aCSS-B. We choose the prior density π as

$$\pi(\theta) = \prod_{j=1}^{d} \phi(\theta_j; 0, 1),$$

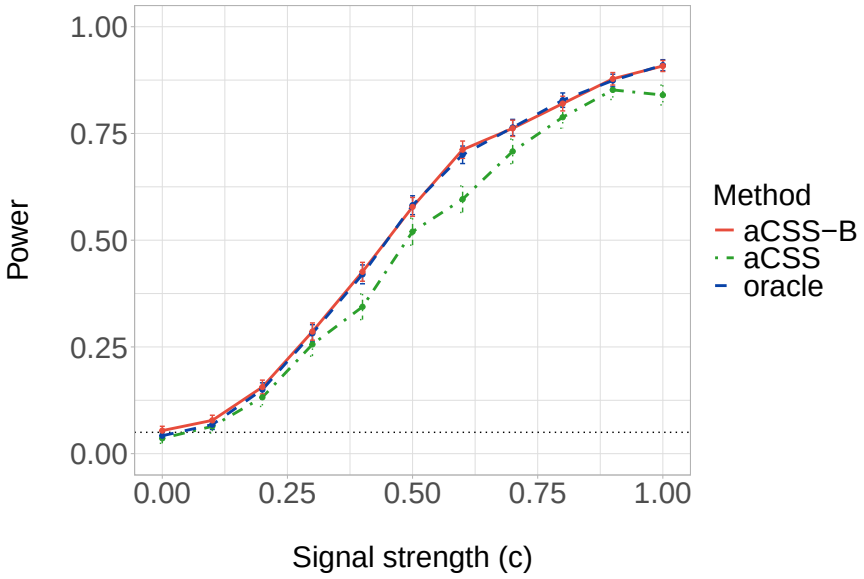

Figure 5.1: Power comparison between aCSS-B, aCSS, and an oracle for the logistic regression model of Section 5.4.2.

where $\phi\left(\cdot; \mu_j, \sigma^2\right)$ is the density of the normal distribution with mean μ and variance σ^2—that is, π is a standard Gaussian prior on the parameter vector $\theta \in \mathbb{R}^d$.

Results. Figure 5.1 compares the performance of aCSS-B to that of aCSS and the oracle. For this example, the oracle consists of sampling the copies $\widetilde{X}^{(m)}$ from the logistic model specified in (5.9), independently of Y. First, we see that all three methods result in a type-I error level of 5% under the null (i.e., signal strength $c = 0$). Under the alternative (signal strength $c > 0$), the methods show similar power, but we can see that aCSS has slightly lower power than the oracle, while aCSS-B appears to have power equal to that of the oracle.

5.4.3 Mixture of Gaussians

The model. This example is taken from Zhu and Barber (2023). The null data distribution is given by sampling n i.i.d. draws from a mixture of two Gaussians, so that the likelihood

function for the data $X = (X_1, \ldots, X_n)$ is given by

$$f(x;\theta) = \prod_{i=1}^{n} \left(w_1 \phi(x_i; \mu_1, \sigma_1^2) + (1 - w_1)\phi(x_i; \mu_2, \sigma_2^2) \right).$$

This family of distributions is therefore parametrized by $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \in \Theta$, where

$$\Theta = (0,1) \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+ \subseteq \mathbb{R}^5.$$

Therefore, the GoF test can be interpreted as testing the null hypothesis that the underlying model is a Gaussian mixture with 2 components. We will consider an alternative that the data is instead drawn from a Gaussian mixture with > 2 components. In our simulations, we take $n = 200$ and the data is generated from the following mixture

$$p\mathcal{N}(0, 0.01) + \frac{1-p}{2}\mathcal{N}(0.4, 0.01) + \frac{1-p}{2}\mathcal{N}(-0.4, 0.01) \tag{5.10}$$

where $p = 0$ corresponds to the null hypothesis being true, while $p > 0$ corresponds to the alternative. As in Zhu and Barber (2023, Section 6.1.1), we use a test statistic based on k-means clustering with $k = 2$ versus $k = 3$; see that paper for more details.

Can we apply existing aCSS methods? For this problem, aCSS cannot be applied because the likelihood maximization problem is degenerate (since σ_1^2 or σ_2^2 can be arbitrarily close to zero, leading to a likelihood that can approach infinity). Instead, Zhu and Barber (2023) apply a regularized extension of aCSS (which we will refer to as reg-aCSS) to this problem, placing constraints that bound σ_1^2, σ_2^2 away from zero, but find low power compared to the oracle. We will compare aCSS-B against the exact same implementation of aCSS-B as used for this problem in that paper.

Choice of prior for implementing aCSS-B. We assume the following prior distributions:

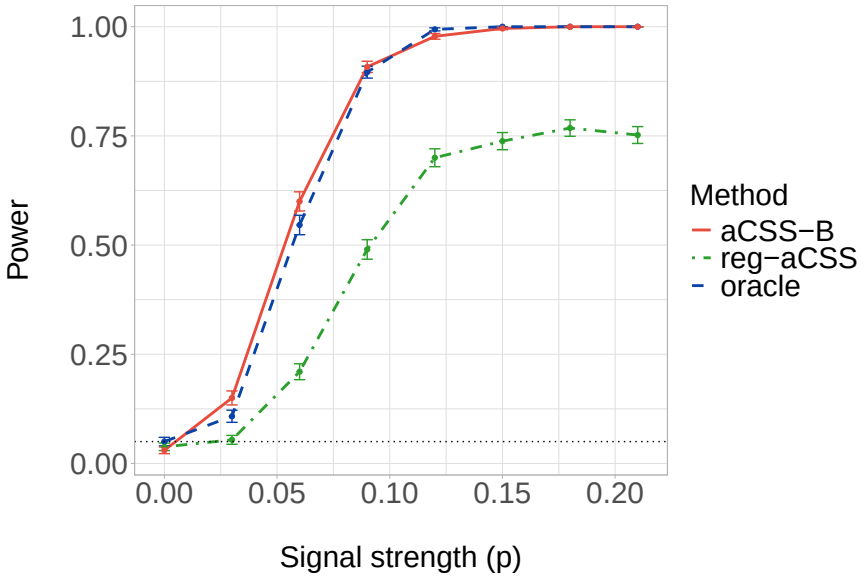

Figure 5.2: Power comparison between aCSS-B, reg-aCSS, and an oracle for the Gaussian mixture model of Section 5.4.3.

$w_1 \sim \text{Beta}(2, 2)$, and

$$\sigma_j^2 \sim \text{Inv-Gamma}(1, 0.5), \quad \mu_j \mid \sigma_j^2 \sim \mathcal{N}(0, \sigma_j^2)$$

independently for each $j = 1, 2$.

Results. Figure 5.2 compares the performance of aCSS-B to that of reg-aCSS and the oracle. For this example, the oracle consists of sampling the entries of $\widetilde{X}^{(m)}$ i.i.d. from the two-component mixture specified in (5.10) if we set $p = 0$. All three methods result in a type-I error level of 5% under the null (i.e., mixture weight $p = 0$). Under the alternative (mixture weight $p > 0$), aCSS-B enjoys similar power as the oracle across nearly all values of p, while reg-aCSS shows lower power.

5.4.4 Rank-1 matrix

Next we turn to examples that are beyond the scope of aCSS and its regularized extensions. Our first such example lies in the setting of low-rank matrix data.

The model. We assume that the data $X \in \mathbb{R}^{n \times n}$ (with $n = 10$) is generated as

$$X = A + W,$$

where A is a fixed matrix representing the underlying signal, while W is noise, with $W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.25)$. Under the null, the signal A is a rank-1 matrix. Therefore the GoF test can be represented with the family

$$f_\theta(x) = \prod_{i=1}^{n} \prod_{j=1}^{n} \phi(X_{ij} - A_{ij}; 0, 0.25),$$

parametrized by $\theta = A \in \Theta \subseteq \mathbb{R}^{n \times n}$, where Θ is the space of rank-1 matrices. Under the alternative, the rank of the underlying signal is > 1.

In order to generate the data, we define $A_0 = U_1 V_1^\top + c U_2 V_2^\top$ where $U = [U_1, U_2], V = [V_1, V_2] \in \mathbb{R}^{n \times 2}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. We vary c, with $c = 0$ corresponding to the null—note that $\text{rank}(A) = 1$ under the null and $\text{rank}(A) = 2$ under the alternative. The test statistic $T(X)$ is defined as the second largest eigenvalue of $X^\top X$.

Can we apply existing aCSS methods? In this example, the null parameter space is $\Theta = \{A \in \mathbb{R}^{n \times n} : \text{rank}(A) = 1\}$. The challenging nature of the rank-1 constraint means that none of the existing aCSS methods can be applied—specifically, Barber and Janson (2022) would require a convex and open null model space, while the regularized forms of aCSS (Zhu and Barber, 2023; Xie and Huang, 2025) allow constraints on the parameter θ but only limited types of constraints are allowed, which again cannot encompass a rank-1 restriction. (We could instead consider reparametrizing as $A = uv^\top$ and taking $\theta = (u, v)$, but the assumptions of aCSS are again violated—in this case, because the existing aCSS theory requires strong convexity of the log-likelihood at the MLE, which cannot hold under this reparametrization due to nonidentifiability.) However, aCSS-B can be applied here with a suitable choice of prior.

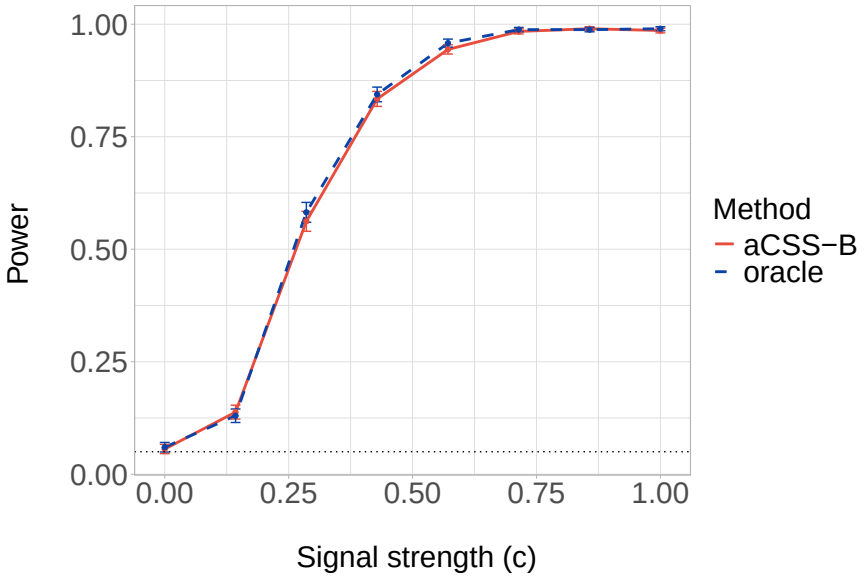

Figure 5.3: Power comparison between aCSS-B and an oracle for the rank-1 matrix model of Section 5.4.4.

Choice of prior for implementing aCSS-B. To specify the prior distribution for the null, we introduce two vectors $U, V \in \mathbb{R}^n$ such that $A = UV^\top$ and assume independent multivariate standard Gaussian priors for U and V, that is, $U, V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\vec{0}_n, \mathbb{I}_n)$.

Results. Figure 5.3 compares the performance of aCSS-B to that of the oracle. For this example, the oracle consists of sampling $\widetilde{X}^{(m)} = A_0 + W$ where $A_0 = U_1 V_1^\top$ and W has i.i.d standard normal entries (recall that the data is generated with mean $A = U_1 V_1^\top + cU_2 V_2^\top$, where $0 \leq c \leq 1$, so this choice of A_0 represents a rank-1 approximation to the true distribution). We can see that the aCSS-B method achieves type-I error control at level 5% under the null (i.e., when A has rank 1), as desired, and has nearly the same power as the oracle under the alternative.

5.4.5 Group-sparse regression

Our next example studies a group-sparse linear regression model.

The model. We consider the following model for data $X \in \mathbb{R}^n$, given covariates $Z \in \mathbb{R}^{n \times d}$

(which we treat as fixed):

$$X = Z\beta + \epsilon \text{ where } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{5.11}$$

where we chose $n = 100$ and $d = 50$. Given the partition $\{1, \ldots, d\} = I_1 \cup \cdots \cup I_{10}$ into $G = 10$ groups of equal size, $|I_g| = 5$, our null hypothesis is that only one group is "active" in the regression: that is, the active set $A = \{g : \beta_{I_g} \neq (0, \ldots, 0)\}$ has size $|A| = 1$ while under the alternative $|A| > 1$.

To generate the data, we draw the entries of Z independently from $\mathcal{N}(0,1)$ and choose two group indices $g_1 \neq g_2$ uniformly at random from the 10 groups. The coefficients are generated as follows:

$$\beta_{I_{g_1}} = \vec{1}_{|I_{g_1}|},\ \beta_{I_{g_2}} = c\,\vec{1}_{|I_{g_2}|},\quad \beta_j = 0\ \forall\ j \in [d] \backslash (I_{g_1} \cup I_{g_2}). \tag{5.12}$$

For the null, we consider $c = 0$ so that there is only one active group (namely, g_1), and for the alternative we take $c \in (0\ ,\ 0.3]$ so that there are two active groups (g_1 and g_2). We refer to c as the signal strength.

To define the test statistic, we first compute an estimate $\widehat{\beta}$ of the regression coefficients by fitting 5-fold cross-validated group LASSO from the R package `gglasso` (Wu and Lange, 2020) on the data. The test statistic is then defined as

$$T(X) := \frac{\sum_{g \neq \widehat{g}} ||\widehat{\beta}_{I_g}||_\infty}{||\widehat{\beta}_{I_{\widehat{g}}}||_\infty} \text{ where } \widehat{g} = \underset{g \in [G]}{\arg\max}\, ||\widehat{\beta}_{I_g}||_\infty,$$

which measures the sum of the largest estimated coefficients *outside* of the top selected group $\widehat{g}$ relative to the largest one within $\widehat{g}$.

Can we apply existing aCSS methods? In this example, the null parameter space is

$$\Theta = \{(\beta_1, \ldots, \beta_d) \in \mathbb{R}^d : \beta_{I_g} = \vec{0}_{d_g} \text{ for all } g \neq g^\star,\ \text{for some } g^\star \in \{1, \ldots, G\}\}.$$

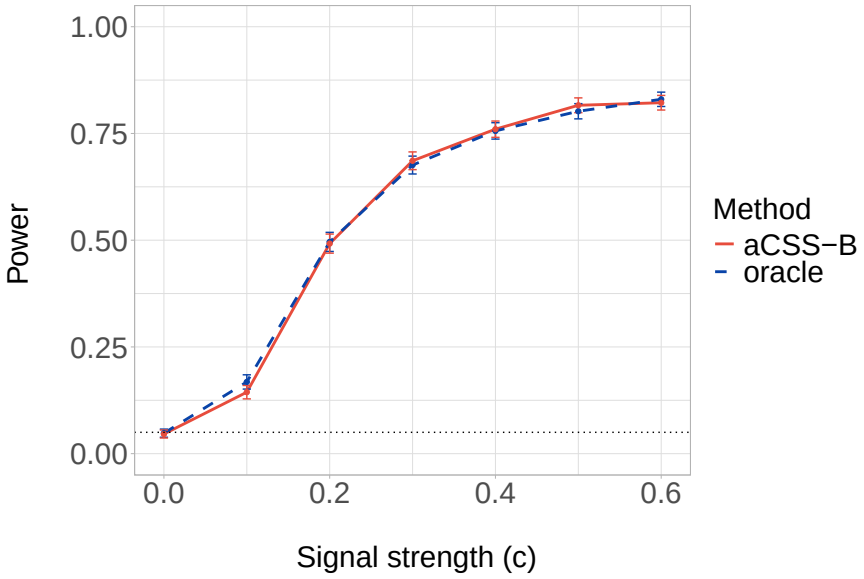

Figure 5.4: Power comparison between aCSS-B and an oracle for the group sparse model of Section 5.4.5.

As for the rank-1 constraint in the previous example, aCSS and its regularized extensions cannot be applied to this problem because of the challenging nature of the group-sparsity constraint.

Choice of prior for implementing aCSS-B. For the prior π, we consider a discrete uniform distribution for the active group and choose a standard Gaussian prior for the coefficients of the active group, while the rest of the coefficients are set to zero:

$$\begin{aligned} g^\star &\sim \text{Unif}(\{1,\ldots,\text{G}\}), \\ \beta_{I_{g^\star}} &\sim \mathcal{N}(5 \cdot \vec{1}_{|I_{g^\star}|}, \mathbb{I}_{|I_{g^\star}|}), \\ \beta_{I_g} &= \vec{0}_{|I_g|} \;\forall\; g \neq g^\star. \end{aligned} \tag{5.13}$$

Results. Figure 5.4 compares the performance of aCSS-B to that of the oracle. For this example, the oracle consists of sampling $\widetilde{X}^{(m)}$ from a linear model as in (5.11) where the coefficient vector β defined in (5.12) is redefined to have a single active group by setting $\beta_{I_{g_2}}$ to be zero (while keeping the original coefficients in the first active group, $\beta_{I_{g_1}}$). We can see

that the aCSS-B method achieves type-I error control at level 5% under the null (i.e., when the coefficient vector indeed has only one nonzero group), and the power is essentially the same as the oracle under the alternative.

5.4.6 Linear spline regression model

Our final example is in a nonlinear regression setting, where X follows a linear spline model given covariates Z.

The model. Consider a linear spline model with k knots $t_1 < \cdots < t_k \in \mathbb{R}$, with $n = 50$ observations,

$$X_i = \mu_i + \epsilon_i \text{ for } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.25), \tag{5.14}$$

where

$$\mu_i = \beta_{0j} + \beta_{1j} Z_i \text{ if } t_{j-1} \leq Z_i < t_j.$$

(Here for convenience we define $t_0 = -\infty$ and $t_{k+1} = \infty$; as in previous examples the covariates $Z_i \in \mathbb{R}$ are treated as fixed.) The parameters $\beta_{ij} \in \mathbb{R}$ are constrained so that the mean function is continuous in $\mathbb{R}$. The null hypothesis corresponds to the case where we have exactly $k = 1$ knot, while under the alternative we have $k = 2$.

To generate data from this distribution, we generate $Z_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(-5, 5)$ and choose two knots as $t_1 = -1.67$ and $t_2 = 1.67$. The coefficients are generated as

$$\beta_{01} = 1, \beta_{11} = -1, \beta_{21} = 1, \beta_{31} = 1 - c \tag{5.15}$$

and the other intercepts β_{02}, β_{03} are chosen so that the mean function is continuous. We consider a sequence of equally spaced values for c from 0 to 1.8—note that $c = 0$ corresponds to the null, since the slopes of the second and third segment are the same and thus effectively

we only have $k = 1$ knot. Under the alternative $c > 0$, there are $k = 2$ knots, with larger values of c denoting further deviation from the null.

The test statistic $T(X)$ is the residual sum of squares from a linear spline model with one knot, which we fit using the `segmented` (Muggeo, 2023) package in `R`.

Can we apply existing aCSS methods? In this example, the null parameter space is

$$\Theta = \Big\{ (\beta_{01}, \beta_{11}, \ldots, \beta_{0,k+1}, \beta_{1,k+1}, t_1, \ldots, t_k) \in \mathbb{R}^{3k+2} : \\ \beta_{0j} + \beta_{1j} t_j = \beta_{0(j+1)} + \beta_{1(j+1)} t_j \text{ for all } j = 1, \ldots, k,\ t_1 < \cdots < t_k \Big\}.$$

The above constraints, which stem from the continuity of the mean function at the knots in the linear spline model, cannot be accommodated by the existing aCSS methods. However, aCSS-B can still be applied, through a carefully constructed prior as we will see below.

Choice of prior for implementing aCSS-B. While we can choose priors which respect the constraints in this problem, sampling from the resulting posterior distribution would be complicated. However, we can avoid this problem by using the following reparametrization:

$$X = \gamma_0 + \gamma_1 Z + \sum_{j=1}^{k} \gamma_{1+j} b_j(Z) + \epsilon, \tag{5.16}$$

where $b_j(Z) = (Z - t_j)_+$ is applied element-wise to $Z = (Z_1, \ldots, Z_n)$, for all $j = 1, \ldots, k$. In this reparametrization, $\gamma = (\gamma_0, \ldots, \gamma_{k+1}) \in \mathbb{R}^{k+2}$ is unconstrained, and the knots $t_1, \ldots, t_k$ do not need to be ordered. Therefore, the parameter space becomes $\Theta = \mathbb{R}^{2k+2}$. (We note, however, that existing aCSS methods nonetheless cannot be applied, even with this reparametrization—this is because the log-likelihood is no longer differentiable with respect to the parameters t_j.)

On these unconstrained parameters, we choose the standard Gaussian priors:

$$\gamma_0, \gamma_1, \gamma_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),$$

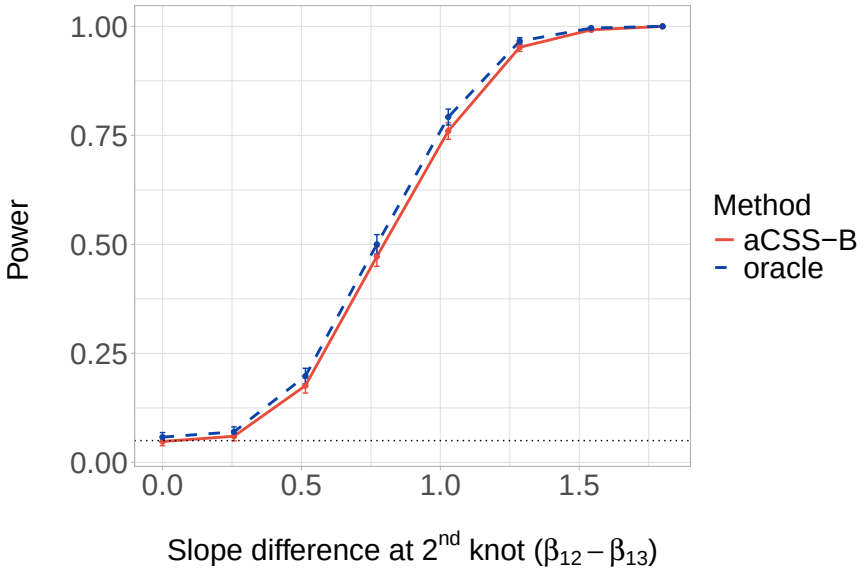

Figure 5.5: Power comparison between aCSS-B and an oracle for the linear spline model of Section 5.4.6.

$$t_1 \sim \mathcal{N}(0, 1),$$

for the null model with $k = 1$ knots.

Results. Figure 5.5 illustrates the performance aCSS-B as compared to the oracle. For this example, the oracle consists of sampling $\widetilde{X}^{(m)}$'s from the model given by (5.14) and (5.15) with $c = 0$. We see that the aCSS-B method achieves type-I error control at level 5% under the null, and the power is nearly the same as the oracle under the alternative.

5.5 Conclusion

This chapter introduces aCSS-B, a flexible resampling-based method for goodness-of-fit testing that uses finite number of posterior samples as the choice of an approximately sufficient statistic. This makes the approach applicable in settings where exact sufficient statistics are unavailable, or where existing aCSS methods are difficult to implement. The main practical advantage of aCSS-B is its flexibility. The experiments show that it can be applied not only

in classical settings, such as logistic regression and Gaussian mixtures, where previous aCSS methods are already applicable, but also in more structured models, such as rank-one matrix models, group-sparse regression, and spline regression, which are beyond the scope of existing aCSS-based methods.

There are several directions for future work. First, the choice of the number of posterior samples B affects both validity and power: larger values of B improve approximate exchangeability, but may reduce the diversity of the resampled copies and therefore decrease power. Although the default choice $B = 25$ works well across all of our examples, a more rigorous, theoretically motivated choice of B would be useful. Second, the current theory assumes i.i.d. sampling from the posterior, which may only be achieved approximately in practice when using MCMC methods. Extending the guarantees to account for such computational approximations is another important direction.

APPENDIX A

APPENDIX TO CHAPTER 2: GROUP-WEIGHTED CONFORMAL PREDICTION

A.1 Additional results and calculations

A.1.1 Additional proofs

Proof of Theorem 2.5.1. When we sample training data i.i.d. from the distribution P, as defined in (2.3), this is equivalent to first sampling group sizes $n_1, \ldots, n_K$ via the Multinomial$(p_1, \ldots, p_K)$ distribution, then drawing the appropriate number of samples from each Π_k to define our training set. In other words, conditional on $n_1, \ldots, n_K$, the distribution of the training and test data is identical to that assumed in Theorem 2.4.1. Therefore, applying Theorem 2.4.1, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; n_1, \ldots, n_K\right\} \geq 1 - \alpha - \max_{k:n_k>0}\{q_k/n_k\}.$$

Taking an expected value to marginalize over $n_1, \ldots, n_K$ completes the proof of the first part.

To complete the proof we need to verify that $\mathbb{E}\left[\max_{k:n_k>0}\{q_k/n_k\}\right] \leq 4n^{-1}\max_k\{q_k/p_k\}$, under the assumption $\min_k p_k \geq \frac{8\log n}{n}$. Let $\mathcal{E}$ be the event that $n_k \geq np_k(1-1/\sqrt{2})$ holds for all $k \in [K]$. On the event $\mathcal{E}$, we have

$$\max_{k:n_k>0}\{q_k/n_k\} \leq \max_k\left\{\frac{q_k}{np_k(1-1/\sqrt{2})}\right\} = (1-1/\sqrt{2})^{-1}n^{-1}\max_k\{q_k/p_k\},$$

while on $\mathcal{E}^c$ we have $\max_{k:n_k>0}\{q_k/n_k\} \leq \max_k q_k \leq 1 \leq \max_k\{q_k/p_k\}$. Therefore,

$$\mathbb{E}\left[\max_{k:n_k>0}\{q_k/n_k\}\right] \leq \max_k\{q_k/p_k\}\left[(1-1/\sqrt{2})^{-1}n^{-1} + \mathbb{P}\{\mathcal{E}^c\}\right].$$

Next, applying the multiplicative Chernoff bound to the random variable $n_k \sim \text{Binomial}(n, p_k)$, for each k we have

$$\mathbb{P}\left\{n_k \le np_k(1-1/\sqrt{2})\right\} \le e^{-(1-(1-1/\sqrt{2}))^2 \cdot np_k/2} = e^{-np_k/4} \le e^{-8\log n/4} = n^{-2},$$

since $p_k \ge \frac{8\log n}{n}$ by assumption. Thus by the union bound, $\mathbb{P}\{\mathcal{E}^c\} \le Kn^{-2} \le 1/(8n)$, where for the last step, we use the bound $K \le (\min_k p_k)^{-1} \le \frac{n}{8\log n} \le n/8$ (we can assume $n \ge 3$ since otherwise the last bound in the theorem holds trivially). Combining these calculations proves the bound. □

Proof of Corollary 2.5.2. First, condition on $n_1, \ldots, n_K$, and let Q' be the distribution placing uniform weights on all observed groups, i.e.,

$$\begin{cases} X^0 \sim \text{Uniform}(k : n_k \ge 1), \\ (X^1, Y) \mid (X^0 = k) \sim \Pi_k, \end{cases} \tag{A.1}$$

As in the proof of Theorem 2.5.1, applying Theorem 2.4.1 to a test point drawn from Q', we have

$$\mathbb{P}_{Q'}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; n_1, \ldots, n_K\right\} \ge 1 - \alpha - \max_{k:n_k>0}\{q_k/n_k\} = 1 - \alpha - \frac{1}{\widehat{K}\min_{k:n_k>0} n_k}.$$

Next, continuing to condition on $n_1, \ldots, n_K$, observe that Q is equal to a mixture placing weight $\frac{\widehat{K}}{K}$ on Q' and $1 - \frac{\widehat{K}}{K}$ on the remaining groups. Thus, for a test point (X_{n+1}, Y_{n+1}) drawn from Q, we now have

$$\begin{aligned} \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; n_1, \ldots, n_K\right\} &\ge \frac{\widehat{K}}{K} \cdot \mathbb{P}_{Q'}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; n_1, \ldots, n_K\right\} \\ &\ge \frac{\widehat{K}}{K}\left(1 - \alpha - \frac{1}{\widehat{K}\min_{k:n_k>0} n_k}\right) = (1-\alpha)\cdot\frac{\widehat{K}}{K} - \frac{1}{K\min_{k:n_k>0} n_k}. \end{aligned}$$

Taking an expected value to marginalize over $n_1, \ldots, n_K$, then,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq \mathbb{E}\left[(1-\alpha) \cdot \frac{\widehat{K}}{K} - \frac{1}{K \min_{k:n_k>0} n_k}\right]$$
$$= (1-\alpha) - \mathbb{E}\left[\frac{1}{K \min_{k:n_k>0} n_k} + (1-\alpha)\frac{\sum_k \mathbf{1}_{n_k=0}}{K}\right].$$

Next, we add the assumption that $\min_k p_k \geq \frac{8 \log n}{n}$. As in the proof of Theorem 2.5.1, let $\mathcal{E}$ be the event that $n_k \geq np_k(1-1/\sqrt{2})$ holds for all $k \in [K]$, with $\mathbb{P}\{\mathcal{E}^c\} \leq 1/(8n)$ as before. On the event $\mathcal{E}^c$ we have

$$\frac{1}{K \min_{k:n_k>0} n_k} + (1-\alpha)\frac{\sum_k \mathbf{1}_{n_k=0}}{K} \leq \frac{1}{K} + \frac{K-1}{K} = 1,$$

while on the event $\mathcal{E}$, we have $\sum_k \mathbf{1}_{n_k=0} = 0$, and $\frac{1}{K \min_{k:n_k>0} n_k} \leq \frac{1}{K \min_k np_k(1-1/\sqrt{2})} = \frac{1}{Kn \min_k p_k} \cdot \frac{1}{1-1/\sqrt{2}}$. Therefore,

$$\mathbb{E}\left[\frac{1}{K \min_{k:n_k>0} n_k} + (1-\alpha)\frac{\sum_k \mathbf{1}_{n_k=0}}{K}\right] \leq \frac{1}{Kn \min_k p_k} \cdot \frac{1}{1-1/\sqrt{2}} + \frac{1}{8n} \leq \frac{4}{Kn \min_k p_k},$$

where the last step holds since $K \min_k p_k \leq 1$. This completes the proof. □

Proof of Corollary 2.5.3. First we consider an equivalent representation of the problem. Suppose we have $m+1$ test points $(X_{n+1}, Y_{n+1}), (X_{n+2}, Y_{n+2}), \ldots, (X_{n+m+1}, Y_{n+m+1})$ sampled i.i.d. from Q; we will examine the coverage guarantee for the first test point (X_{n+1}, Y_{n+1}) as before, while taking the remaining m test points as our source of information for estimating the q_k's—that is, we will define

$$\tilde{q}_k = \frac{\sum_{i=1}^m \mathbf{1}\{X^0_{n+1+i} = k\}}{m}.$$

Now let $\tilde{Q} = \sum_{k=1}^{K} \tilde{q}_k \cdot \widehat{P}_{\text{score}}^{(k)}$, and let $\tilde{Q}^* = \sum_{k=1}^{K} \tilde{q}_k^* \cdot \widehat{P}_{\text{score}}^{(k)}$ where

$$\tilde{q}_k^* = \frac{\sum_{i=1}^{m+1} \mathbf{1}\{X_{n+i}^0 = k\}}{m+1}$$

is the empirical frequency of group k in the entire test set (i.e., the test point of interest, (X_{n+1}, Y_{n+1}), along with the m test points used to construct the $\tilde{q}_k$'s). Note that by definition we have

$$\widehat{q} = \text{Quantile}_{1-\alpha}(\tilde{Q}).$$

Furthermore,

$$\tilde{q}_k^* = \frac{m}{m+1}\tilde{q}_k + \frac{\mathbf{1}\{X_{n+1}^0 = k\}}{m+1},$$

and in particular, this means that we can calculate

$$\mathrm{d_{TV}}\left(\tilde{Q}, \tilde{Q}^*\right) = \frac{1}{2}\sum_{k=1}^{K} |\tilde{q}_k - \tilde{q}_k^*| = \frac{1}{2}\sum_{k=1}^{K} \frac{1}{m+1}\left|\tilde{q}_k - \mathbf{1}\{X_{n+1}^0 = k\}\right| \leq \frac{1}{m+1}.$$

Thus, $\widehat{q} = \text{Quantile}_{1-\alpha}(\tilde{Q}) \geq \text{Quantile}_{1-\alpha'}(\tilde{Q}^*)$ where $\alpha' = \alpha + \frac{1}{m+1}$, which further implies that

$$\widehat{C}_n^*(x) := \left\{y \in \mathcal{Y} : s(x,y) \leq \text{Quantile}_{1-\alpha'}(\tilde{Q}^*)\right\} \subseteq \widehat{C}_n(x)$$

. Therefore,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n^*(X_{n+1})\right\}.$$

But by symmetry, since the $\tilde{q}_k$'s are constructed as a symmetric function of the $m+1$ test points, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n^*(X_{n+1})\right\} = \mathbb{P}\left\{Y_{n+i} \in \widehat{C}_n^*(X_{n+i})\right\}$$

for all $i = 1, \ldots, m+1$. Taking the average, then,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n^*(X_{n+1})\right\} = \frac{1}{m+1}\sum_{i=1}^{m+1} \mathbb{P}\left\{Y_{n+i} \in \widehat{C}_n^*(X_{n+i})\right\}$$
$$= \mathbb{E}\left[\frac{1}{m+1}\sum_{i=1}^{m+1} \mathbf{1}\left\{Y_{n+i} \in \widehat{C}_n^*(X_{n+i})\right\}\right].$$

By the tower law we can write this as

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n^*(X_{n+1})\right\} = \mathbb{E}\left[\mathbb{E}\left[\frac{1}{m+1}\sum_{i=1}^{m+1} \mathbf{1}\left\{Y_{n+i} \in \widehat{C}_n^*(X_{n+i})\right\} \,\middle|\, X_{n+1}^0, \ldots, X_{n+m+1}^0\right]\right].$$

Equivalently, sampling an index I uniformly from $\{n+1, \ldots, n+m+1\}$,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n^*(X_{n+1})\right\} = \mathbb{E}\left[\mathbb{P}\left\{Y_{n+I} \in \widehat{C}_n^*(X_{n+I}) \,\middle|\, X_{n+1}^0, \ldots, X_{n+m+1}^0\right\}\right].$$

But conditional on the group assignments $X_{n+1}^0, \ldots, X_{n+m+1}^0$, we can observe that (X_{n+I}, Y_{n+I}) is a draw from the distribution $\tilde{Q}^*$. Applying Theorem 2.5.1 (with $\tilde{Q}^*$ in place of Q), then,

$$\mathbb{P}\left\{Y_{n+I} \in \widehat{C}_n^*(X_{n+I}) \,\middle|\, X_{n+1}^0, \ldots, X_{n+m+1}^0\right\} \geq 1 - \alpha' - \max_{k:n_k>0}\{\tilde{q}_k^*/n_k\}$$

and so after marginalizing,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n^*(X_{n+1})\right\} \geq 1 - \alpha' - \mathbb{E}\left[\max_{k:n_k>0}\{\tilde{q}_k^*/n_k\}\right].$$

Finally, we have

$$\mathbb{E}\left[(\tilde{q}_1, \ldots, \tilde{q}_K) \mid (\tilde{q}_1^*, \ldots, \tilde{q}_K^*), (n_1, \ldots, n_K)\right] = (\tilde{q}_1^*, \ldots, \tilde{q}_K^*),$$

so by Jensen's inequality,

$$\mathbb{E}\left[\max_{k:n_k>0}\{\tilde{q}_k/n_k\} \,\middle|\, (\tilde{q}_1^*, \ldots, \tilde{q}_K^*), (n_1, \ldots, n_K)\right] \geq \max_{k:n_k>0}\{\tilde{q}_k^*/n_k\}.$$

This completes the proof. □

A.1.2 Is the coverage guarantee tight?

In this section, we examine whether the coverage guarantees obtained in this work are tight, and whether less conservative versions of the method could be defined.

We will first examine the question of unobserved groups. Note that the GWCP method defined in (2.5) will compute a quantile $\widehat{q} = +\infty$, and consequently an infinite prediction interval $\widehat{C}(X_{n+1}) = \mathcal{Y}$, if the proportion of unobserved groups is too large, i.e., if $\sum_{k:n_k=0} q_k > \alpha$. The following proposition shows that this is unavoidable.

Proposition A.1.1. *Let $\mathcal{Y} = \mathbb{R}$. Let $\widehat{C}_n$ satisfy the distribution-free coverage guarantee*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\} \geq 1 - \alpha$$

with respect to data $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ drawn as in (2.8) *and test point (X_{n+1}, Y_{n+1}) drawn as in* (2.4)*, for* any *distributions $\Pi_1, \ldots, \Pi_K$. If $\sum_{k:n_k=0} q_k > \alpha$ then*

$$\mathbb{E}[\mathrm{Leb}(\widehat{C}_n(X_{n+1}))] = \infty.$$

Proof of Proposition A.1.1. Fix any $y \in \mathbb{R}$. We begin by defining an alternative distribution of the data: for each group k let

$$\tilde{\Pi}_k = \begin{cases} \Pi_k, & n_k > 0, \\ (\Pi_k)_{X^1} \times \delta_y, & n_k = 0, \end{cases}$$

where $(\Pi_k)_{X^1}$ denotes the marginal of X^1 for $(X^1, Y) \sim \Pi_k$. That is, if $n_k = 0$, then the value X^1 is drawn exactly as before, while Y is set to be equal to the constant value y.

Now let $(\tilde{X}_i, \tilde{Y}_i)$ denote data drawn with these new distributions $\tilde{\Pi}_k$ in place of the original

Π_k's, and let $\tilde{C}_n$ denote the prediction interval when constructed on data drawn from this new distribution. Note that the training data and test feature, $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n), \tilde{X}_{n+1}$, has an identical joint distribution as the original data, $(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}$, by construction, and therefore,

$$\tilde{C}_n(\tilde{X}_{n+1}) \stackrel{\mathrm{d}}{=} \widehat{C}_n(X_{n+1}),$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. Therefore,

$$\mathbb{P}\{y \in \widehat{C}_n(X_{n+1})\} = \mathbb{P}\{y \in \tilde{C}_n(\tilde{X}_{n+1})\}.$$

Moreover, if $\tilde{X}^0_{n+1} = k$ for some k with $n_k = 0$, then $\tilde{Y}_{n+1} = y$ almost surely, by definition of $\tilde{\Pi}_k$. Therefore,

$$\begin{aligned}
\mathbb{P}\{y \in \tilde{C}_n(\tilde{X}_{n+1})\} &\geq \mathbb{P}\{\tilde{Y}_{n+1} \in \tilde{C}_n(\tilde{X}_{n+1}), \tilde{Y}_{n+1} = y\} \\
&\geq \mathbb{P}\{\tilde{Y}_{n+1} \in \tilde{C}_n(\tilde{X}_{n+1}), \tilde{X}^0_{n+1} = k \text{ for some } k \text{ with } n_k = 0\} \\
&\geq \mathbb{P}\{\tilde{X}^0_{n+1} = k \text{ for some } k \text{ with } n_k = 0\} - \mathbb{P}\{\tilde{Y}_{n+1} \notin \tilde{C}_n(\tilde{X}_{n+1})\}.
\end{aligned}$$

And, $\mathbb{P}\{\tilde{Y}_{n+1} \notin \tilde{C}_n(\tilde{X}_{n+1})\} \leq \alpha$ (since the prediction interval satisfies the distribution-free coverage guarantee—and in particular, must have coverage at least $1-\alpha$ with respect to this newly defined distribution), while $\mathbb{P}\{\tilde{X}^0_{n+1} = k \text{ for some } k \text{ with } n_k = 0\} = \sum_{k:n_k=0} q_k$. So, we have

$$\mathbb{P}\{y \in \tilde{C}_n(\tilde{X}_{n+1})\} \geq \sum_{k:n_k=0} q_k - \alpha > 0.$$

Finally, the claim that $\mathbb{E}[\mathrm{Leb}(\widehat{C}_n(X_{n+1}))] = \infty$ follows directly from integration. For any k with $n_k = 0$,

$$\mathbb{E}[\mathrm{Leb}(\widehat{C}_n(X_{n+1}))] \geq q_k \cdot \mathbb{E}[\mathrm{Leb}(\widehat{C}_n(X_{n+1})) \mid X^0_{n+1} = k]$$

$$= q_k \cdot \mathbb{E}\left[\mathbb{E}[\mathrm{Leb}(\widehat{C}_n(X_{n+1})) \mid X^0_{n+1}] \mid X^0_{n+1} = k\right],$$

and so it suffices to show that $\mathbb{E}[\mathrm{Leb}(\widehat{C}_n(x))] = \infty$ for any x with $x^0 = k$. We have

$$\mathbb{E}[\mathrm{Leb}(\widehat{C}_n(x))] = \mathbb{E}\left[\int_{y\in\mathbb{R}} \mathbf{1}_{y\in\widehat{C}_n(x)}\, \mathsf{d}y\right] = \int_{y\in\mathbb{R}} \mathbb{E}\left[\mathbf{1}_{y\in\widehat{C}_n(x)}\right]\, \mathsf{d}y$$
$$= \int_{y\in\mathbb{R}} \mathbb{P}\{y \in \widehat{C}_n(x)\}\, \mathsf{d}y \geq \int_{y\in\mathbb{R}} (\sum_{k:n_k=0} q_k - \alpha)\, \mathsf{d}y = \infty,$$

where the second step holds by the Fubini–Tonelli theorem, and the last step holds since $\sum_{k:n_k=0} q_k - \alpha > 0$ by assumption. □

Next we consider the term $\max_k\{q_k/n_k\}$ in our coverage guarantee, Theorem 2.4.1, for the case of fixed group sizes n_k. Can this error term be improved? The following counterexample establishes that this term is actually essentially optimal.

Proposition A.1.2. *Let $\mathcal{Y} = \mathbb{R}$. Let $(1-\alpha)K$ be an integer, let $q_k \equiv 1/K$, and fix any $n_1, \ldots, n_K \geq 1$. Without loss of generality assume $n_1 = \min_k n_k$. Define a score function $s(x, y) = y$, and let Π_k have marginal distribution $(\Pi_k)_Y$ on the response Y given by*

$$Y \sim \begin{cases} \mathrm{Unif}[0,1], & k = 1, \\ \mathrm{Unif}[-1,0], & k = 2, \ldots, (1-\alpha)K, \\ \mathrm{Unif}[1,2], & k = (1-\alpha)K+1, \ldots, K. \end{cases}$$

Let $q_k \equiv 1/K$. Then the GWCP prediction interval (2.5) *has coverage*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(X_{n+1})\} = 1 - \alpha - \frac{1}{K(\min_k n_k + 1)}.$$

Since $q_k \equiv 1/K$ in this example, the excess loss of coverage is therefore equal to $\frac{1}{K(\min_k n_k+1)} = \max_k\{q_k/(n_k+1)\}$, which (up to at most a factor of 2) matches the

error term $\max_{k:n_k>0}\{q_k/n_k\}$, appearing in Theorem 2.4.1. Therefore, this establishes that the guarantee of this theorem is fairly tight.

Proof of Proposition A.1.2. Since $q_k \equiv 1/K$, the weighted empirical distribution $\widehat{P}_{\text{score}}$ is simply the average of the $\widehat{P}^{(k)}$'s. Since we have scores lying in $[-1, 0]$ for $(1-\alpha)K - 1$ many groups, in $[0, 1]$ for one group (namely, group $k = 1$), and in $[1, 2]$ for αK many groups, by definition of the quantile we can see that $\widehat{q}$ is equal to the maximum score from group $k = 1$—that is, $\widehat{q}$ is the maximum of n_1 many draws from the Unif$[0, 1]$ distribution. In particular, this implies that $\mathbb{E}[\widehat{q}] = \frac{n_1}{n_1+1}$ by properties of the uniform distribution.

Next, by construction of the score function, we have $\widehat{C}(X_{n+1}) = (-\infty, \widehat{q}]$, and so

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(X_{n+1})\} = \mathbb{P}\{Y_{n+1} \le \widehat{q}\} = \sum_{k=1}^{K} q_k \mathbb{P}\{Y_{n+1} \le \widehat{q} \mid X^0_{n+1} = k\}.$$

Recall that $\widehat{q}$ must lie in $[0, 1]$ by construction. Therefore, if the test group is $X^0_{n+1} = k$ for some $k = 2, \ldots, (1-\alpha)K$, we have $Y_{n+1} \le 0 \le \widehat{q}$ almost surely, and if the test group is $X^0_{n+1} = k$ for some $k > (1-\alpha)K$, we have $Y_{n+1} > 1 \ge \widehat{q}$ almost surely. On the other hand if $X^0_{n+1} = 1$ then Y_{n+1} is drawn from the Unif$[0, 1]$ distribution, and so

$$\mathbb{P}\{Y_{n+1} \le \widehat{q} \mid X^0_{n+1} = 1\} = \mathbb{E}\left[\mathbb{P}\{Y_{n+1} \le \widehat{q} \mid \widehat{q}; X^0_{n+1} = 1\} \mid X^0_{n+1} = 1\right] = \mathbb{E}[\widehat{q}] = \frac{n_1}{n_1+1}.$$

We thus have

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(X_{n+1})\} = \frac{(1-\alpha)K - 1}{K} + \frac{1}{K} \cdot \frac{n_1}{n_1+1} = 1 - \alpha - \frac{1}{K(n_1+1)}.$$

□

A.1.3 Which data should be used to estimate the weights?

Recall that in Section 2.5.4, we showed an empirical comparison between our new bounds and the coverage guarantees established by Lei and Candès (2021)'s existing result (2.2). In addition to the more rapid convergence of the lower bound in our new result, there is another interesting difference as compared to existing work. Specifically, existing theory (including (Lei and Candès, 2021)'s result, but also related works such as Candès et al. (2023); Gui et al. (2022); Yin et al. (2022)) applies to an estimated weight function $\widehat{w}(x)$ that was defined *independently* of the calibration set. For example, if a pretraining data set was used for defining the score function (e.g., fitting a model $\widehat{f} : \mathcal{X} \to \mathbb{R}$, so that the score function is given by the residual score, $s(x, y) = |y - \widehat{f}(x)|$), the weight function $\widehat{w}(x)$ might also be trained ahead of time by estimating the covariate distribution P_X using this pretraining data set. In contrast, the weights used in our proposed construction for the GWCP method specifically use the *calibration* set's proportions—recall from Section 2.3.1 that we can think of GWCP as essentially running WCP with weight function $\widehat{w}(X) = q_k/(n_k/n)$ for $X^0 = k$, where n_k is the number of *calibration* data points in group k.

For both types of results, the particular choice is essential to the theory: Lei and Candès (2021)'s result relies on $\widehat{w}$ being independent of the calibration set, while in contrast, our proofs rely on using the n_k's observed in the calibration set. Which approach is the "right" one for achieving the best performance—that is, for the least loss of coverage?

To study this, we perform a simulation that compares the different options. We consider prediction intervals of the form

$$\widehat{q} = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n} \frac{\widehat{w}(X_i)}{\sum_{i'=1}^{n} \widehat{w}(X_{i'})} \cdot \delta_{s_i}\right) \text{ where } \widehat{w}(X_i) = \widehat{w}_k \text{ for } X_i^0 = k, \tag{A.2}$$

i.e., the weight $\widehat{w}(X_i)$ placed on data point i depends only on the group k to which this data point belongs. (As discussed above, in Section 2.3.1, we can view this as a slightly less

conservative version of the WCP method.) We will test three different options for defining the weights:

- Estimate weights using the pretraining data set, $\{(X_i^*, Y_i^*)\}_{i\in[n_*]}$:[1]

$$\widehat{w}_k = q_k/\widehat{p}_k \text{ where } \widehat{p}_k = \frac{\sum_{i=1}^{n_*} \mathbf{1}_{X_i^{*0}=k}}{n_*}.$$

- Estimate weights using the calibration data set, $\{(X_i, Y_i)\}_{i\in[n]}$:

$$\widehat{w}_k = q_k/\widehat{p}_k \text{ where } \widehat{p}_k = \frac{\sum_{i=1}^{n} \mathbf{1}_{X_i^{0}=k}}{n}.$$

- Use the oracle weights,

$$\widehat{w}_k = q_k/p_k,$$

  where p_k is the true probability of group k under the training distribution P.

The distribution of the data is defined as follows: we have $K = 5$ groups, with $q_k \equiv 0.2$. We set $n = 100$ and $n_* = 100$. The data distribution is given by $(X, Y) \in \{1, 2, 3, 4, 5\} \times \mathbb{R}$, where $Y \mid X = k \sim \mathcal{N}(\theta_k, 1)$. The score function is fixed as $s(x, y) = |y|$.

We consider three settings for the group probabilities p_k in the training distribution P, and the means θ_k:

- Setting 1: the groups are equally represented in the training data, with $(p_1, p_2, p_3, p_4, p_5) = (0.2, 0.2, 0.2, 0.2, 0.2)$, and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (20, 15, 10, 5, 0)$.

- Setting 2: the group distributions are nonuniform, with $(p_1, p_2, p_3, p_4, p_5) = (0.4, 0.25, 0.2, 0.1, 0.05)$, and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (20, 15, 10, 5, 0)$. Note that in this setting, the more common groups tend to have higher values of the score.

1. Since the estimated weight function $\widehat{w}$ is not well-defined on the rare event that any group has zero observations in the pretraining set, we discard any trials for which this occurs.

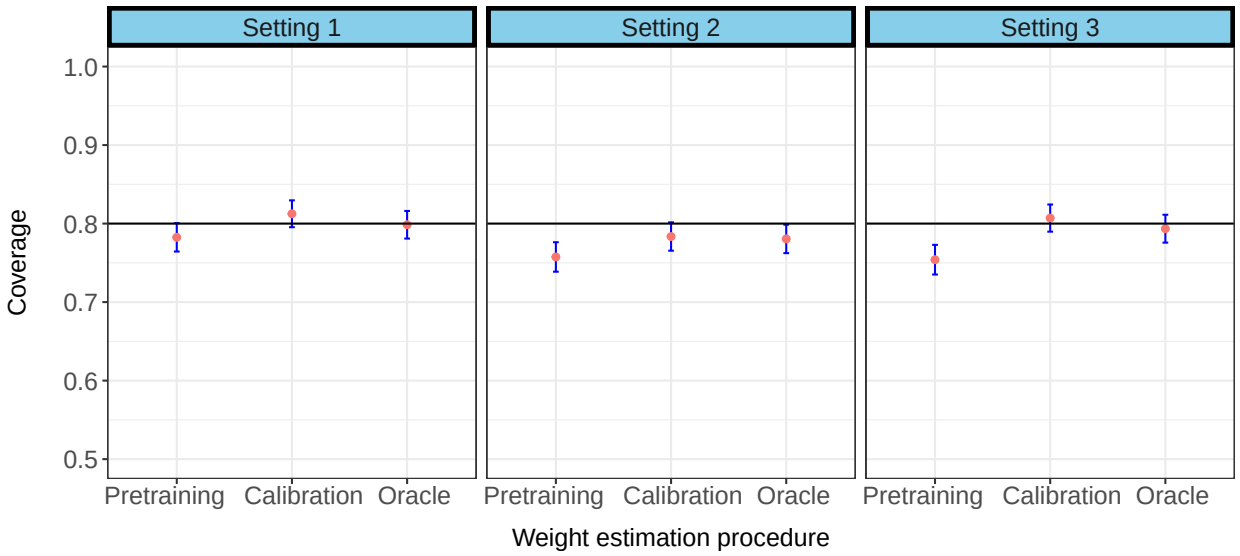

Figure A.1: Marginal coverage of the weighted conformal prediction interval (A.2), implemented with weight function $\widehat{w}$ estimated using the pretraining data or using the calibration data, or, using the the "oracle" weights that rely on knowledge of the true distribution. The horizontal line marks the target coverage level $1-\alpha = 80\%$. Results are averaged over 2000 independent trials, and the figure displays the mean coverage rate along with a 95% confidence interval. See Appendix A.1.3 for details.

- Setting 3: the group distributions are nonuniform, with $(p_1, p_2, p_3, p_4, p_5) = (0.4, 0.25, 0.2, 0.1, 0.05)$, $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (0, 5, 10, 15, 20)$. Note that in this setting, the more common groups tend to have lower values of the score.

The results are displayed in Figure A.1. We can see that in each setting, using the pretraining data for estimating the weights tends to result in undercoverage, while using the calibration data does not. (The slight undercoverage of the oracle method, in Setting 3, is due to the fact that we are not adding the correction term to our weighted conformal interval—that is, the difference between (2.6) and (2.7).)

Why does using the calibration data for estimating weights, lead to better coverage? Since we have chosen $n = n_* = 100$ (i.e., $\widehat{w}$ is estimated using 100 data points, for both options), this cannot be explained by a difference in the accuracy of estimating the true proportions p_k. Instead, we can understand this difference in performance as follows. When $\widehat{w}$ is estimated using the calibration data, the randomness in this estimate is *self-correcting*: for instance, if

for a particular group k we have a slightly larger n_k by random chance (i.e., $n_k > np_k$), then data points from this group are overrepresented when calculating the quantile $\widehat{q}$. But, the larger value of n_k means that the weight $\widehat{w}_k$ is slightly smaller, and so these data points are also downweighted. In other words, there is a sort of cancellation effect, where the errors in the $\widehat{w}_k$'s are actually helpful for achieving the desired coverage level. In contrast, if we use the pretraining set to estimate the weights, the noise in the n_k's is independent from the noise in the $\widehat{w}_k$'s, and so there is no such cancellation effect.

To summarize, in this work we have used a weight function $\widehat{w}$ trained on the calibration set rather than the pretraining set, which is unusual relative to much of the literature—but this simulation demonstrates that this choice has better performance, as is supported by our stronger theoretical guarantees.

A.1.4 A corrected GWCP for exact coverage

In this section, we provide a corrected version of the GWCP prediction interval that is guaranteed to provide coverage at level $1-\alpha$. The goal is to construct a minimally more conservative version of GWCP to remove the error term in the coverage guarantee. The corrected interval is given by

$$\widehat{C}_n(x) = \{y \in \mathcal{Y} : s(x,y) \leq \widehat{q}_k\} \text{ for any } x \in \mathcal{X} \text{ with } x^0 = k, \tag{A.3}$$

where $\widehat{q}_k = \text{Quantile}_{1-\alpha_k}\left(\sum_{k=1}^K q_k \widehat{P}^{(k)}_{\text{score}}\right)$ for

$$\alpha_k = \begin{cases} \alpha - \frac{q_k}{n_k}, & n_k > 0, \\ \alpha, & n_k = 0. \end{cases}$$

In other words, at a test point X_{n+1} in group k, if $n_k > 0$ then the threshold for the score $s(X_{n+1}, y)$ is adjusted to be a bit more conservative; the adjustment is large only for groups

with small observed sample size n_k. (In the degenerate case that $\alpha_k < 0$ for some k, we should interpret this by returning $\widehat{q}_k = +\infty$.)

Theorem A.1.3. *Suppose the training data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ are sampled from the fixed-group-size model, as in (2.8), while the test point (X_{n+1}, Y_{n+1}) is drawn independently from the distribution Q as defined in (2.4). Then, for any fixed (i.e., pretrained) score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the corrected GWCP prediction interval $\widehat{C}_n$ defined in (A.3) satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha.$$

Proof of Theorem A.1.3. Following the same notation and terminology as the proof of Theorem 2.4.1, we rewrite the coverage probability as

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} = \sum_{k=1}^{K} q_k \mathbb{P}\left\{s(X_{k,0}, Y_{k,0}) \leq \text{Quantile}_{1-\alpha_k}(\widehat{P}^{k,0}_{\text{score}})\right\},$$

i.e., this is the same calculation as in (2.10) except with α_k in place of α for each group k. The next step is again the same: by symmetry, we can rewrite this as

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{q_k}{n_k} \mathbb{P}\left\{s(X_{k,i}, Y_{k,i}) \leq \text{Quantile}_{1-\alpha_k}(\widehat{P}^{k,i}_{\text{score}})\right\}.$$

Now consider any group k. If $n_k = 0$ then $\widehat{P}^{k,i}_{\text{score}} = \widehat{P}_{\text{score}}$ and $\alpha_k = \alpha$, and thus, $\text{Quantile}_{1-\alpha_k}(\widehat{P}^{k,i}_{\text{score}}) = \text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}})$. If instead $n_k > 0$, then by construction, for any i we have $\mathrm{d_{TV}}(\widehat{P}_{\text{score}}, \widehat{P}^{k,i}_{\text{score}}) \leq q_k/n_k = \alpha - \alpha_k$, and therefore, $\text{Quantile}_{1-\alpha_k}(\widehat{P}^{k,i}_{\text{score}}) \geq \text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}})$. (This holds also for the case $\alpha_k < 0$ since then the quantile on the left-hand side is defined as $+\infty$.) Combining both cases, then,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{q_k}{n_k} \mathbb{P}\left\{s(X_{k,i}, Y_{k,i}) \leq \text{Quantile}_{1-\alpha}(\widehat{P}_{\text{score}})\right\} \geq 1 - \alpha,$$

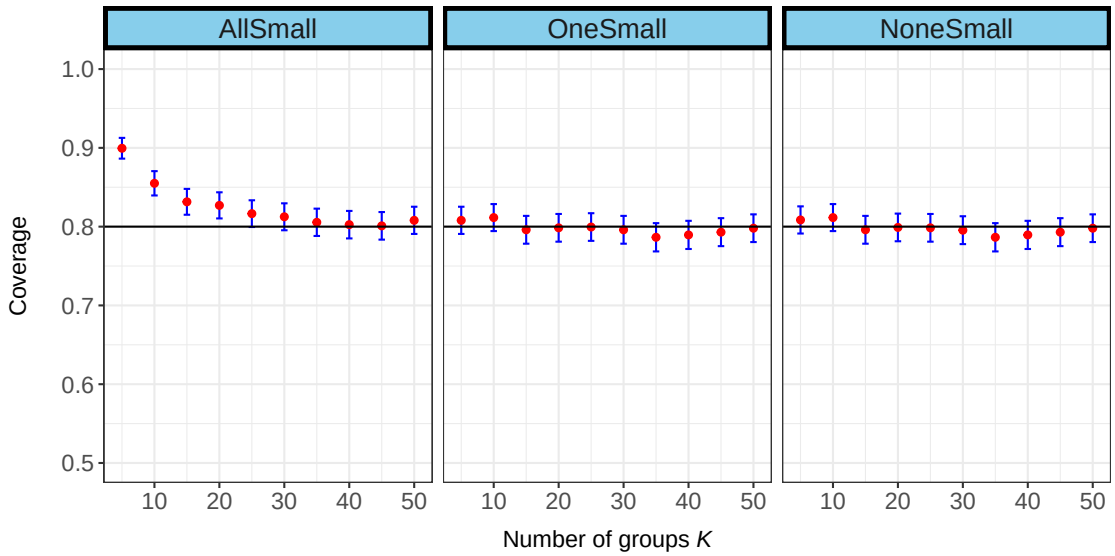

Figure A.2: Simulation results for the setting of fixed group sizes, in three regimes for the group sizes $n_1, \ldots, n_K$, with the corrected GWCP prediction interval (A.3) in place of the original version (2.5). The horizontal line marks the target coverage level $1 - \alpha = 80\%$. Results are averaged over 2000 independent trials, and the figure displays the mean coverage rate along with a 95% confidence interval. See Appendix A.1.4 for details.

where (as in the proof of Theorem 2.4.1) the last step holds by definition of the weighted empirical distribution $\widehat{P}_{\text{score}}$ (see, e.g., Harrison (2012, Lemma A.1)). □

To complete this section, we repeat the simulation in Section 2.4.2, with the corrected GWCP (A.3) in place of the original version (2.5). The simulation settings are exactly the same as in Section 2.4.2. The results are displayed in Figure A.2. In this figure, coverage is always at least $1-\alpha$, as guaranteed by Theorem A.1.3. In particular, comparing to Figure 2.2, we can see that the loss of coverage incurred when $\min_k n_k = 1$ is no longer present—but instead, the method is somewhat too conservative for the `AllSmall` setting, for smaller values of K.

APPENDIX B

APPENDIX TO CHAPTER 3: CONFORMAL PREDICTION WITH MACRO-COVERAGE GUARANTEES

B.1 Proof of Theorem 3.2.1

Proof. Conditional on $\mathcal{H}$, the group memberships $g(Y_1), \ldots, g(Y_n)$, the counts N_k, the set $\mathcal{K}_+$, and the weights $w(k)$ are fixed.

Let

$$\Delta := \max_{k \in \mathcal{K}_+} \frac{w(k)}{N_k}.$$

If $1 - \alpha - \Delta \leq 0$, then the result is immediate because coverage is nonnegative. Thus, assume $1 - \alpha - \Delta > 0$.

We first prove the result in the case where all groups are observed, so that $\mathcal{K}_+ = \mathcal{K}$.

For each $k \in \mathcal{K}$, let

$$(X_0^{(k)}, Y_0^{(k)})$$

be an independent "ghost" sample from the distribution of $(X, Y) \mid g(Y) = k$, independent of the calibration data, and define

$$S_0^{(k)} := s(X_0^{(k)}, Y_0^{(k)}).$$

Since $\mathcal{K}_+ = \mathcal{K}$, the weighted empirical score distribution is

$$\widehat{P}_{g,w} = \sum_{k \in \mathcal{K}} w(k) \frac{1}{N_k} \sum_{j=1}^{N_k} \delta_{S_j^{(k)}}.$$

By the definition of (g, w) macro-coverage,

$$\mathrm{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) = \sum_{k \in \mathcal{K}} w(k) \mathbb{P}\left(S_0^{(k)} \leq \mathrm{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) \mid \mathcal{H} \right).$$

Fix $k \in \mathcal{K}$ and $j \in \{1, \ldots, N_k\}$. Define the ghost-replaced weighted empirical distribution

$$\widehat{P}_{g,w}^{k,j} := \widehat{P}_{g,w} + \frac{w(k)}{N_k} \left(\delta_{S_0^{(k)}} - \delta_{S_j^{(k)}} \right).$$

That is, $\widehat{P}_{g,w}^{k,j}$ is obtained from $\widehat{P}_{g,w}$ by replacing the jth calibration score in group k by the independent ghost score from the same group.

Conditional on $\mathcal{H}$, the scores

$$S_0^{(k)}, S_1^{(k)}, \ldots, S_{N_k}^{(k)}$$

are exchangeable. Therefore,

$$\mathbb{P}\left(S_0^{(k)} \leq \mathrm{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) \mid \mathcal{H} \right) = \mathbb{P}\left(S_j^{(k)} \leq \mathrm{Quantile}_{1-\alpha}(\widehat{P}_{g,w}^{k,j}) \mid \mathcal{H} \right).$$

Averaging over $j = 1, \ldots, N_k$ and summing over $k \in \mathcal{K}$ gives

$$\mathrm{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) = \sum_{k \in \mathcal{K}} w(k) \frac{1}{N_k} \sum_{j=1}^{N_k} \mathbb{P}\left(S_j^{(k)} \leq \mathrm{Quantile}_{1-\alpha}(\widehat{P}_{g,w}^{k,j}) \mid \mathcal{H} \right). \tag{B.1}$$

For every $k \in \mathcal{K}$ and $j \in \{1, \ldots, N_k\}$, the distributions $\widehat{P}_{g,w}$ and $\widehat{P}_{g,w}^{k,j}$ differ only by moving mass $w(k)/N_k$ from $S_j^{(k)}$ to $S_0^{(k)}$. Hence

$$d_{\mathrm{TV}}\left(\widehat{P}_{g,w}, \widehat{P}_{g,w}^{k,j} \right) \leq \frac{w(k)}{N_k} \leq \Delta.$$

Therefore,

$$\text{Quantile}_{1-\alpha}(\widehat{P}^{k,j}_{g,w}) \geq \text{Quantile}_{1-\alpha-\Delta}(\widehat{P}_{g,w}).$$

Using this in (B.1),

$$\begin{aligned}\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) &\geq \sum_{k\in\mathcal{K}} w(k)\frac{1}{N_k}\sum_{j=1}^{N_k} \mathbb{P}\left(S_j^{(k)} \leq \text{Quantile}_{1-\alpha-\Delta}(\widehat{P}_{g,w}) \mid \mathcal{H}\right) \\ &= \mathbb{E}\left[\sum_{k\in\mathcal{K}} w(k)\frac{1}{N_k}\sum_{j=1}^{N_k} \mathbf{1}\left\{S_j^{(k)} \leq \text{Quantile}_{1-\alpha-\Delta}(\widehat{P}_{g,w})\right\} \middle| \mathcal{H}\right].\end{aligned}$$

By the definition of the weighted empirical distribution and of the quantile,

$$\sum_{k\in\mathcal{K}} w(k)\frac{1}{N_k}\sum_{j=1}^{N_k} \mathbf{1}\left\{S_j^{(k)} \leq \text{Quantile}_{1-\alpha-\Delta}(\widehat{P}_{g,w})\right\} \geq 1-\alpha-\Delta$$

almost surely. Therefore, in the all-observed case,

$$\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) \geq 1-\alpha-\Delta.$$

We now handle the general case, where some groups may have $N_k = 0$. Let

$$W_+ := \sum_{k\in\mathcal{K}_+} w(k)$$

be the total weight assigned to observed groups.

If $\text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) = \infty$, then $\mathcal{C}(x) = \mathcal{Y}$ for all $x \in \mathcal{X}$ and the macro-coverage guarantee is satisfied trivially.

If $\text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) < \infty$, then we must have $W_+ > 0$. Define normalized weights on the observed groups by

$$\widetilde{w}(k) := \frac{w(k)}{W_+} \qquad \text{for } k \in \mathcal{K}_+,$$

and define the observed-groups weighted empirical score distribution

$$\widehat{P}_+ := \sum_{k\in\mathcal{K}_+} \widetilde{w}(k)\frac{1}{N_k}\sum_{j=1}^{N_k}\delta_{S_j^{(k)}}.$$

By construction,

$$\widehat{P}_{g,w} = W_+\widehat{P}_+ + (1-W_+)\delta_\infty \quad \text{and} \quad \text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) = \text{Quantile}_{\frac{1-\alpha}{W_+}}(\widehat{P}_+).$$

Let

$$1-\widetilde{\alpha} := \frac{1-\alpha}{W_+}, \qquad \text{equivalently} \qquad \widetilde{\alpha} = 1-\frac{1-\alpha}{W_+}.$$

Applying the all-observed result above to the group set $\mathcal{K}_+$, the weights $\widetilde{w}(k)$, and the miscoverage level $\widetilde{\alpha}$, we obtain

$$\sum_{k\in\mathcal{K}_+}\widetilde{w}(k)\mathbb{P}\left(S_0^{(k)} \le \text{Quantile}_{\frac{1-\alpha}{W_+}}(\widehat{P}_+) \mid \mathcal{H}\right) \ge \frac{1-\alpha}{W_+} - \max_{k\in\mathcal{K}_+}\frac{\widetilde{w}(k)}{N_k}.$$

Using

$$\text{Quantile}_{\frac{1-\alpha}{W_+}}(\widehat{P}_+) = \text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}),$$

and multiplying both sides by W_+ gives

$$\sum_{k\in\mathcal{K}_+} w(k)\mathbb{P}\left(S_0^{(k)} \le \text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) \mid \mathcal{H}\right) \ge 1-\alpha - W_+\cdot\max_{k\in\mathcal{K}_+}\frac{\widetilde{w}(k)}{N_k}.$$

Since

$$W_+\cdot\max_{k\in\mathcal{K}_+}\frac{\widetilde{w}(k)}{N_k} = \max_{k\in\mathcal{K}_+}\frac{w(k)}{N_k} = \Delta,$$

we have

$$\sum_{k\in\mathcal{K}_+} w(k)\mathbb{P}\left(S_0^{(k)} \le \text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) \mid \mathcal{H}\right) \ge 1-\alpha-\Delta.$$

Finally,

$$\begin{aligned}\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) &= \sum_{k \in \mathcal{K}} w(k) \mathbb{P}\left(S_0^{(k)} \leq \text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) \mid \mathcal{H}\right) \\ &\geq \sum_{k \in \mathcal{K}_+} w(k) \mathbb{P}\left(S_0^{(k)} \leq \text{Quantile}_{1-\alpha}(\widehat{P}_{g,w}) \mid \mathcal{H}\right) \\ &\geq 1 - \alpha - \Delta.\end{aligned}$$

This proves the theorem. □

B.2 Proof of Proposition 3.2.2

Proposition B.2.1 (Formal version of Proposition 3.2.2)**.** *For $t \in \mathbb{R}$, define*

$$\tilde{\mathcal{C}}_t(x) = \{y \in \mathcal{Y} : \tilde{s}(x, y) \geq t\} \quad \textit{where} \quad \tilde{s}(x, y) = \frac{w(g(y))}{\rho(g(y))} \cdot p(y|x). \tag{B.2}$$

If there exists t_α such that $\text{MacroCov}_{g,w}(\tilde{\mathcal{C}}_{t_\alpha}) = 1 - \alpha$, then $\tilde{\mathcal{C}}_{t_\alpha}$ is the solution to

$$\min_{\mathcal{C}: \mathcal{X} \mapsto 2^{\mathcal{Y}}} \mathbb{E}[|\mathcal{C}(X)|] \quad \text{s.t.} \quad \text{MacroCov}_{g,w}(\mathcal{C}) \geq 1 - \alpha. \tag{B.3}$$

Remark B.2.2. If there does not exist t_α satisfying the equality exactly, the optimal set is still of the thresholded form above but must be combined with randomization to achieve the optimal solution with exact (g, w) macro-coverage of $1 - \alpha$. Shao (2008, Theorem 6.1) can be used to characterize such a randomized solution.

Proof. First, observe that for any grouping function g and weight function w, we can rewrite $\text{MacroCov}_{g,w}$ in terms of the identity grouping function $g_I(y) = y$ and some weighting function $\tilde{w} : \mathcal{Y} \to \mathbb{R}_{\geq 0}$:

$$\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) = \sum_{k \in \mathcal{K}} w(k)\, \mathbb{P}\Big(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid g(Y_{n+1}) = k, \mathcal{H}\Big)$$

$$= \sum_{k \in \mathcal{K}} w(k) \sum_{y: g(y)=k} \frac{p(y)}{\rho(k)} \mathbb{P}\Big(Y \in \mathcal{C}(X) \mid Y = y, \mathcal{H}\Big)$$

$$= \sum_{y \in \mathcal{Y}} \tilde{w}(y) \, \mathbb{P}\Big(Y \in \mathcal{C}(X) \mid Y = y, \mathcal{H}\Big) \tag{B.4}$$

$$\text{for} \quad \tilde{w}(y) := \frac{w(g(y)) \cdot p(y)}{\rho(g(y))}.$$

In order to find the smallest prediction set $\mathcal{C}$ satisfying $\text{MacroCov}_{g,w}(\mathcal{C} \mid \mathcal{H}) \geq 1 - \alpha$, we use a result from Ding et al. (2026). We restate the relevant part here for completeness.

Proposition 6 of Ding et al. (2026). Let $\omega : \mathcal{Y} \to [0, 1]$ be a non-negative weighting function summing to one. For $t \in \mathbb{R}$, define

$$\tilde{\mathcal{C}}_t(x) = \{y \in \mathcal{Y} : \tilde{s}(x, y) \geq t\} \quad \text{where} \quad \tilde{s}(x, y) = \frac{\omega(y)}{p(y)} \cdot p(y|x) \tag{B.5}$$

and $p(y|x)$ denotes the conditional probability of Y given $X = x$ and $p(y)$ is the marginal probability of Y. Let $\alpha \in [0, 1]$. If there exists t_α such that $\text{MacroCov}_\omega(\tilde{\mathcal{C}}_{t_\alpha}) = 1 - \alpha$, then $\tilde{\mathcal{C}}_{t_\alpha}$ is the optimal solution to

$$\min_{\mathcal{C}: \mathcal{X} \mapsto 2^{\mathcal{Y}}} \mathbb{E}[|\mathcal{C}(X)|] \quad \text{subject to} \sum_{y \in \mathcal{Y}} \omega(y) \mathbb{P}(y \in \mathcal{C}(X) \mid Y = y) \geq 1 - \alpha. \tag{B.6}$$

The representation of (g, w) macro-coverage given in (B.4) allows us to apply this proposition, which directly yields the desired result. □

B.3 Proofs of Propositions 3.2.3 and 3.2.4

Proof of Proposition 3.2.3. Let $\mathcal{C}_j$ refer to the prediction set produced by Algorithm 1 for the j-th coverage constraint. By Theorem 3.2.1, each $\mathcal{C}_j$ satisfies $\text{MacroCov}_{g_j, w_j} \geq 1 - \alpha_j$. When the true label is in any of these sets, it is also in the union of such sets, so the set union simultaneously satisfies all of the coverage conditions. Because all $\mathcal{C}_j$ are constructed

by applying the same fixed score function s and comparing against a threshold, the set union can be obtained by taking the max of the score thresholds implied by each constraint. $\square$

Next, we first state the formal version of Proposition 3.2.4 and then prove it.

Proposition B.3.1 (Formal version of Proposition 3.2.4)**.** *Suppose there exist* $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m_{\geq 0}$*, not identically zero, and* $t > 0$ *such that the deterministic prediction set*

$$\mathcal{C}_{\lambda,t}(x) = \left\{ y \in \mathcal{Y} : \frac{p(y \mid x)}{p(y)} \sum_{j=1}^{m} \lambda_j \tilde{w}_j(y) \geq t \right\}$$

is feasible, meaning that

$$\mathrm{MacroCov}_{g_j,w_j}(\mathcal{C}_{\lambda,t}) \geq 1 - \alpha_j, \qquad j = 1, \ldots, m,$$

and satisfies,

$$\lambda_j \left[\mathrm{MacroCov}_{g_j,w_j}(\mathcal{C}_{\lambda,t}) - (1 - \alpha_j) \right] = 0, \qquad j = 1, \ldots, m. \tag{B.7}$$

Then $\mathcal{C}_{\lambda,t}$ *is an optimal solution to*

$$\min_{\mathcal{C}:\mathcal{X}\to 2^{\mathcal{Y}}} \mathbb{E}|\mathcal{C}(X)| \quad \text{s.t.} \quad \mathrm{MacroCov}_{g_j,w_j}(\mathcal{C}) \geq 1 - \alpha_j, \qquad j = 1, \ldots, m.$$

Remark B.3.2. The proposition is stated for deterministic prediction sets. If the condition above is not satisfied by any $\mathcal{C}_{\lambda,t}$, then an optimal relaxed solution may require randomization on the boundary in order to satisfy the active constraints exactly.

Proof. Let $c(x, y) = \mathbf{1}\{y \in \mathcal{C}(x)\}$. We prove the result by considering the more general relaxed problem in which

$$c : \mathcal{X} \times \mathcal{Y} \to [0, 1]$$

is allowed to be a randomized set-inclusion rule. Deterministic prediction sets are recovered when c takes values in $\{0, 1\}$. Fix $j \in [m]$. For each group $k \in \mathcal{K}_j$, let $\rho_j(k) = \mathbb{P}(g_j(Y) = k)$.

For a relaxed rule c, the jth generalized macro-coverage functional is

$$\text{MacroCov}_{g_j,w_j}(\mathcal{C}) = \sum_{k \in \mathcal{K}_j} w_j(k)\mathbb{E}[c(X, Y) \mid g_j(Y) = k].$$

Expanding over labels inside each group gives

$$\begin{aligned}
\text{MacroCov}_{g_j,w_j}(c) &= \sum_{k \in \mathcal{K}_j} w_j(k) \sum_{y:g_j(y)=k} \mathbb{P}(Y = y \mid g_j(Y) = k)\mathbb{E}[c(X, y) \mid Y = y] \\
&= \sum_{k \in \mathcal{K}_j} w_j(k) \sum_{y:g_j(y)=k} \frac{p(y)}{\rho_j(k)}\mathbb{E}[c(X, y) \mid Y = y] \\
&= \sum_{y \in \mathcal{Y}} \frac{w_j(g_j(y))p(y)}{\rho_j(g_j(y))}\mathbb{E}[c(X, y) \mid Y = y] \\
&= \sum_{y \in \mathcal{Y}} \tilde{w}_j(y)\mathbb{E}[c(X, y) \mid Y = y].
\end{aligned}$$

Using Bayes' rule,

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)},$$

we can rewrite this as

$$\begin{aligned}
\text{MacroCov}_{g_j,w_j}(\mathcal{C}) &= \sum_{y \in \mathcal{Y}} \tilde{w}_j(y) \int c(x, y)p(x \mid y)\, dx \\
&= \int p(x) \sum_{y \in \mathcal{Y}} c(x, y)\tilde{w}_j(y)\frac{p(y \mid x)}{p(y)}\, dx. \qquad \text{(B.8)}
\end{aligned}$$

Similarly, the expected size is

$$\mathbb{E}|\mathcal{C}(X)| = \mathbb{E}\left[\sum_{y \in \mathcal{Y}} c(X, y)\right] = \int p(x) \sum_{y \in \mathcal{Y}} c(x, y)\, dx. \qquad \text{(B.9)}$$

Thus the relaxed oracle problem is the linear program

$$\min_{0\le c\le 1} \int p(x) \sum_{y\in\mathcal{Y}} c(x,y)\,dx \quad \text{s.t.} \int p(x) \sum_{y\in\mathcal{Y}} c(x,y)\tilde{w}_j(y)\frac{p(y \mid x)}{p(y)}\,dx \ge 1-\alpha_j, \quad j=1,\dots,m.$$

For multipliers $\mu_1,\dots,\mu_m \ge 0$, the Lagrangian is

$$\begin{aligned}\mathcal{L}(c,\mu) &= \mathbb{E}|\mathcal{C}(X)| + \sum_{j=1}^{m} \mu_j \left[(1-\alpha_j) - \mathrm{MacroCov}_{g_j,w_j}(c)\right] \\ &= \sum_{j=1}^{m} \mu_j(1-\alpha_j) + \int p(x) \sum_{y\in\mathcal{Y}} c(x,y) \left[1 - \frac{p(y \mid x)}{p(y)} \sum_{j=1}^{m} \mu_j \tilde{w}_j(y)\right] dx.\end{aligned}$$

For fixed μ, the Lagrangian is separable in (x,y). Therefore, minimizing it over $c(x,y) \in [0,1]$ can be done pointwise. The coefficient of $c(x,y)$ is

$$1 - \frac{p(y \mid x)}{p(y)} \sum_{j=1}^{m} \mu_j \tilde{w}_j(y).$$

Hence any minimizer of the Lagrangian satisfies

$$c_\mu(x,y) = \begin{cases} 1, & \dfrac{p(y \mid x)}{p(y)} \displaystyle\sum_{j=1}^{m} \mu_j \tilde{w}_j(y) > 1, \\ 0, & \dfrac{p(y \mid x)}{p(y)} \displaystyle\sum_{j=1}^{m} \mu_j \tilde{w}_j(y) < 1, \\ \text{any value in } [0,1], & \dfrac{p(y \mid x)}{p(y)} \displaystyle\sum_{j=1}^{m} \mu_j \tilde{w}_j(y) = 1. \end{cases} \tag{B.10}$$

Now take the deterministic set $\mathcal{C}_{\lambda,t}$ from the proposition statement and let

$$c_{\lambda,t}(x,y) = \mathbf{1}\{y \in \mathcal{C}_{\lambda,t}(x)\}.$$

Then $c_{\lambda,t}$ is a pointwise minimizer of the Lagrangian as given by equation (B.10) with $\mu_j = \frac{\lambda_j}{t}$.

In order to verify optimality, let c be any feasible relaxed rule. Since $c_{\lambda,t}$ minimizes $\mathcal{L}(\cdot, \mu)$,

$$\mathcal{L}(c_{\lambda,t}, \mu) \leq \mathcal{L}(c, \mu).$$

Since c is feasible,

$$(1 - \alpha_j) - \text{MacroCov}_{g_j, w_j}(c) \leq 0 \ \forall \ j,$$

and

$$\mathcal{L}(c, \mu) \leq \mathbb{E}\left[\sum_{y \in \mathcal{Y}} c(X, y)\right].$$

On the other hand, since $\mathcal{C}_{\lambda,t}$ satisfies condition (B.7),

$$\mathcal{L}(c_{\lambda,t}, \mu) = \mathbb{E}|\mathcal{C}_{\lambda,t}(X)|.$$

Combining these inequalities gives

$$\mathbb{E}|\mathcal{C}_{\lambda,t}(X)| \leq \mathbb{E}\left[\sum_{y \in \mathcal{Y}} c(X, y)\right]$$

for every feasible relaxed rule c. Thus $\mathcal{C}_{\lambda,t}$ is an optimal deterministic solution to the simultaneous coverage-constrained problem. □

B.4 Additional Experimental Results

Table B.1 is analogous to Table 3.3 in the main text, but for $\alpha = 0.05$.

Table B.1: Targeting generalized macro-coverage on Pl@ntNet at $\alpha = 0.05$. The left panel targets tail-focused macro-coverage; the right panel targets genus-level macro-coverage.

		Goal : $\text{MacroCov}_{\text{tail}} \geq 0.95$			Goal : $\text{GenusMacroCov} \geq 0.95$		
Method	Score	MarginalCov	$\text{MacroCov}_{\text{tail}}$	AvgSize	MarginalCov	GenusMacroCov	AvgSize
STANDARD	s_{softmax}	$0.951_{\pm 0.001}$	$0.785_{\pm 0.002}$	$5.9_{\pm 0.1}$	$0.951_{\pm 0.000}$	$0.939_{\pm 0.001}$	$5.9_{\pm 0.1}$
	$\hat{s}_{g,w}$	$0.949_{\pm 0.001}$	$0.899_{\pm 0.001}$	$3.8_{\pm 0.1}$	$0.950_{\pm 0.000}$	$\mathbf{0.963}_{\pm 0.000}$	$4.5_{\pm 0.0}$
CLASSWISE	s_{softmax}	$0.987_{\pm 0.000}$	$\mathbf{0.996}_{\pm 0.000}$	$263.4_{\pm 0.9}$	$0.970_{\pm 0.000}$	$\mathbf{0.970}_{\pm 0.001}$	$60.2_{\pm 1.1}$
LABEL-WEIGHTED	s_{softmax}	$0.993_{\pm 0.000}$	$\mathbf{0.950}_{\pm 0.001}$	$71.5_{\pm 1.9}$	$0.962_{\pm 0.001}$	$\mathbf{0.952}_{\pm 0.001}$	$8.1_{\pm 0.2}$
	$\hat{s}_{g,w}$	$0.984_{\pm 0.000}$	$\mathbf{0.951}_{\pm 0.001}$	$\mathbf{10.7}_{\pm 0.3}$	$0.930_{\pm 0.001}$	$\mathbf{0.952}_{\pm 0.001}$	$\mathbf{3.4}_{\pm 0.1}$

APPENDIX C

APPENDIX TO CHAPTER 4: APPROXIMATING FULL CONFORMAL PREDICTION: DISTRIBUTION FREE GUARANTEES VIA THE TOURNAMENT CORRECTION

C.1 Proofs of theorems

C.1.1 Proof of Theorem 4.3.1

The proof follows the core idea of proof of Barber et al. (2021, Theorem 1). We will construct a $(n+1) \times (n+1)$ matrix A, as a function of the data, such that A is a *tournament matrix*, meaning that $A_{ij} \in \{0, 1\}$, and $A_{ij} + A_{ji} \leq 1$, for all i, j (that is, A_{ij} can be viewed as an indicator of whether team i wins its game against team j—and thus we cannot have both $A_{ij} = 1$ and $A_{ji} = 1$). We will establish that the coverage event can be determined by the $(n+1)$st row of A, with

$$Y_{n+1} \notin \widehat{C}^{\text{tourn}}_{n,\alpha}(X_{n+1}) \iff \sum_{i=1}^{n+1} A_{n+1,i} \geq (1-\alpha)(n+1), \tag{C.1}$$

and that A satisfies a row/column-wise exchangeability property,

$$A \overset{\text{d}}{=} \Pi A \Pi^\top \text{ for any permutation matrix } \Pi \in \{0,1\}^{(n+1)\times(n+1)}. \tag{C.2}$$

As in the proof of Barber et al. (2021, Theorem 1), these properties are sufficient to prove that

$$\mathbb{P}(Y_{n+1} \in \widehat{C}^{\text{tourn}}_{n,\alpha}(X_{n+1})) \geq 1 - 2\alpha,$$

as desired: to summarize the proof idea, this is because

$$\sum_{j=1}^{n+1} \mathbf{1}\left\{\sum_{i=1}^{n+1} A_{j,i} \geq (1-\alpha)(n+1)\right\} \leq 2\alpha(n+1)$$

must hold *deterministically* for any tournament matrix (i.e., it is impossible for more than $2\alpha(n+1)$ of the teams to win $\geq (1-\alpha)(n+1)$ of their games in the tournament), while the exchangeability property (C.2) ensures that

$$\mathbb{P}\left(\sum_{i=1}^{n+1} A_{n+1,i} \geq (1-\alpha)(n+1)\right) = \mathbb{P}\left(\sum_{i=1}^{n+1} A_{j,i} \geq (1-\alpha)(n+1)\right) \text{ for all } j.$$

Now we construct the tournament matrix A, and verify that (C.1) and (C.2) both hold. First we define a matrix of scores, $S \in \mathbb{R}^{(n+1)\times(n+1)}$, with

$$\begin{cases} S_{ij} := s_{\text{approx}}(Z_i; \mathcal{D}_{[n+1]\setminus\{i,j\}} + \{Z_i, Z_j\}), & i \neq j, \\ S_{ii} := +\infty, & i = j. \end{cases}$$

By definition of the tournament-corrected prediction set (4.6),

$$Y_{n+1} \notin \widehat{C}^{\text{tourn}}_{n,\alpha}(X_{n+1}) \iff \sum_{i=1}^{n} \mathbf{1}\{S_{n+1,i} > S_{i,n+1}\} \geq (1-\alpha)(n+1).$$

Next define the matrix $A = A(S)$ with entries

$$A_{ij} = \mathbf{1}\{S_{ij} \geq S_{ji}\}.$$

By construction, A is a tournament matrix, and satisfies (C.1). Moreover, by exchangeability of the data together with the fact that $s(z; \mathcal{D} + \{z', z''\})$ is assumed to be symmetric in $\mathcal{D}$, we have

$$S \overset{\text{d}}{=} \Pi S \Pi^\top$$

for any permutation matrix Π. This row/column-wise exchangeability property is then inherited by $A = A(S)$, since

$$A(S) \stackrel{\mathrm{d}}{=} A(\Pi S \Pi^\top) = \Pi A(S) \Pi^\top.$$

This verifies (C.2), and thus completes the proof.

Extension to randomized scores

We next extend the result of Theorem 4.3.1 to the setting of a randomized score function, to accommodate examples such as Bayesian PPD (Example 3).

Let $\xi_{n+1,i} \in [0,1]$ and $\xi_{i,n+1} \in [0,1]$ denote random variables (i.e., seeds used for randomization), drawn independently of the data. Specifically, we now define the tournament-corrected prediction set as

$$\begin{aligned}\widehat{C}^{\mathrm{tourn}}_{n,\alpha}(X_{n+1}) := \Bigg\{ y : \sum_{i=1}^{n} \mathbf{1}\Big\{ & s_{\mathrm{approx}}(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}; \xi_{n+1,i}) \\ & > s_{\mathrm{approx}}(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}; \xi_{i,n+1}) \Big\} < (1-\alpha)(n+1) \Bigg\},\end{aligned} \tag{C.3}$$

where now both the score $s_{\mathrm{approx}}(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}; \xi_{n+1,i})$ for the test point, and the score $s_{\mathrm{approx}}(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}; \xi_{i,n+1})$ for the ith training point, use the additional argument of a random seed (which permits randomization, e.g., a score constructed via random samples drawn from a posterior, as for Bayesian PPD).

We will need to assume that the random seeds $\xi_{n+1,i}$ and $\xi_{i,n+1}$ can be taken to be entries of a matrix $\xi \in [0,1]^{(n+1)\times(n+1)}$, satisfying row/column-wise exchangeability:

$$\xi \stackrel{\mathrm{d}}{=} \Pi \xi \Pi^\top \text{ for any permutation matrix } \Pi \in \{0,1\}^{(n+1)\times(n+1)}. \tag{C.4}$$

For instance, this holds if the $2n$ same random seeds $\xi_{n+1,1}, \ldots, \xi_{n+1,n}, \xi_{1,n+1}, \ldots, \xi_{n,n+1}$ are all equal (i.e., all equal to the same random variable $\xi \sim \text{Unif}[0,1]$), or alteratively, are i.i.d. draws from $\text{Unif}[0,1]$.

Theorem C.1.1. *In the setting of Theorem 4.3.1, now consider the randomized construction as in* (C.3)*. Assume also that the random seeds satisfy* (C.4) *and are independent of the data. Then*

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1 - 2\alpha.$$

Proof. The proof is essentially identical to that of Theorem 4.3.1. We define a score matrix S with entries

$$\begin{cases} S_{ij} := s_{\text{approx}}(Z_i; \mathcal{D}_{[n+1]\setminus\{i,j\}} + \{Z_i, Z_j\}; \xi_{ij}), & i \neq j, \\ S_{ii} := +\infty, & i = j, \end{cases}$$

and observe that S satisfies row/column-wise exchangeability by assumption on the data and the random seeds. The rest of the proof proceeds exactly as before. □

C.1.2 Proof of Theorem 4.5.2

This proof follows a similar argument as Barber et al. (2021, Theorem 5), which proves a related result for the setting of leave-one-out cross-conformal or cross-validation methods.

For $i = 1, \ldots, n$, define

$$R_i^y := s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n+1}^y \setminus \{Z_i\} + \{Z_i\}\Big),$$

and define

$$R_{n+1}^y := s_{\text{approx}}\Big(Z_{n+1}^y; \mathcal{D}_n + \{Z_{n+1}^y\}\Big).$$

Thus, when $y = Y_{n+1}$, writing $R_i := R_i^{Y_{n+1}}$, we have

$$R_i = s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n+1} \setminus \{Z_i\} + \{Z_i\}\Big), \qquad i = 1, \ldots, n+1.$$

Since $Z_1, \ldots, Z_{n+1}$ are exchangeable and $s_{\text{approx}}(z; \mathcal{D} + \{z\})$ is symmetric in $\mathcal{D}$, the scores $R_1, \ldots, R_{n+1}$ are exchangeable. Hence the oracle conformal set

$$\widetilde{C}_{\alpha_0}(X_{n+1}) := \left\{ y : \sum_{i=1}^{n} \mathbf{1}\{R_{n+1}^y > R_i^y\} < (1-\alpha_0)(n+1) \right\}$$

has coverage at least $1 - \alpha_0$, i.e.

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{R_{n+1} > R_i\} \geq (1-\alpha_0)(n+1) \right) \leq \alpha_0. \tag{C.5}$$

Now fix $i \in [n]$, and define

$$\Delta_i := \left| s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_{n+1}, Z_i\}\Big) - R_{n+1} \right|.$$

By Assumption 4.5.1,

$$\mathbb{P}(\Delta_i \leq \epsilon) \geq 1 - \nu.$$

Next define

$$\Delta_i' := \left| s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z_{n+1}\}\Big) - R_i \right|.$$

Let π be the permutation that swaps i and $n+1$, and set $\widetilde{Z}_j := Z_{\pi(j)}$. Since $Z_1, \ldots, Z_{n+1}$ are exchangeable,

$$(\widetilde{Z}_1, \ldots, \widetilde{Z}_{n+1}) \stackrel{d}{=} (Z_1, \ldots, Z_{n+1}).$$

Moreover, $\widetilde{Z}_{n+1} = Z_i$, $\widetilde{Z}_i = Z_{n+1}$, $\widetilde{\mathcal{D}}_{n\setminus i} = \mathcal{D}_{n\setminus i}$, and $\widetilde{\mathcal{D}}_n = \mathcal{D}_{n+1} \setminus \{Z_i\}$. Therefore

$$\left|s_{\text{approx}}\Big(\widetilde{Z}_{n+1}; \widetilde{\mathcal{D}}_{n\setminus i} + \{\widetilde{Z}_{n+1}, \widetilde{Z}_i\}\Big) - s_{\text{approx}}\Big(\widetilde{Z}_{n+1}; \widetilde{\mathcal{D}}_n + \{\widetilde{Z}_{n+1}\}\Big)\right| = \Delta_i'.$$

Applying Assumption 4.5.1 to the permuted sample yields

$$\mathbb{P}(\Delta_i' \leq \epsilon) \geq 1 - \nu.$$

If we let

$$G_i := \{\Delta_i \leq \epsilon\} \cap \{\Delta_i' \leq \epsilon\},$$

then by the union bound,

$$\mathbb{P}(G_i^c) \leq 2\nu. \tag{C.6}$$

Now suppose that G_i holds and $R_{n+1} \leq R_i$. Then

$$\begin{aligned} s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_{n+1}, Z_i\}\Big) &\leq R_{n+1} + \epsilon \\ &\leq R_i + \epsilon \\ &\leq s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z_{n+1}\}\Big) + 2\epsilon. \end{aligned}$$

Hence

$$\begin{aligned} \mathbf{1}\left\{s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_{n+1}, Z_i\}\Big) > s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z_{n+1}\}\Big) + 2\epsilon\right\} \\ \leq \mathbf{1}\{R_{n+1} > R_i\} + \mathbf{1}\{G_i^c\}. \end{aligned}$$

Summing over $i = 1, \ldots, n$,

$$\sum_{i=1}^{n} \mathbf{1}\left\{s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_{n+1}, Z_i\}\Big) > s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z_{n+1}\}\Big) + 2\epsilon\right\}$$

$$\leq \sum_{i=1}^{n} \mathbf{1}\{R_{n+1} > R_i\} + \sum_{i=1}^{n} \mathbf{1}\{G_i^c\}. \tag{C.7}$$

Set $\alpha_0 := \alpha + 2\sqrt{\nu}$. Using (C.7),

$$\begin{aligned}
&\mathbb{P}\Bigg(\sum_{i=1}^{n} \mathbf{1}\Big\{ s_{\text{approx}}\Big(Z_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_{n+1}, Z_i\}\Big) \\
&\qquad\qquad > s_{\text{approx}}\Big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z_{n+1}\}\Big) + 2\epsilon \Big\} \geq (1-\alpha)(n+1) \Bigg) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{R_{n+1} > R_i\} \geq (1-\alpha_0)(n+1) \right) + \mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{G_i^c\} \geq 2\sqrt{\nu}(n+1) \right).
\end{aligned}$$

The first term is at most α_0 by (C.5). For the second term, by (C.6) and Markov's inequality,

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{G_i^c\} \geq 2\sqrt{\nu}(n+1) \right) \leq \frac{\sum_{i=1}^{n} \mathbb{P}(G_i^c)}{2\sqrt{\nu}(n+1)} \leq \frac{2n\nu}{2\sqrt{\nu}(n+1)} \leq \sqrt{\nu}.$$

Therefore,

$$\mathbb{P}\left(Y_{n+1} \notin \widehat{C}_{n,\alpha}^{\text{infl},2\epsilon}(X_{n+1})\right) \leq \alpha + 2\sqrt{\nu} + \sqrt{\nu} = \alpha + 3\sqrt{\nu},$$

which completes the proof.

C.2 Simulation details of Section 4.4.1

We compare all the four approximation schemes for full conformal prediction: deletion, rounding, one-step update, and Bayesian add-one-in posterior predictive approximation. For each scheme, we evaluate both the original approximate method as described in Section 4.2 and the corresponding tournament-corrected method proposed here in Section 4.3. For the non-Bayesian methods, the fitted model is ordinary least squares without an intercept. When the design matrix is rank-deficient or underdetermined, we use the Moore–Penrose inverse and the score function is the absolute residual.

In all simulations, we use $n = 100$, $\alpha = 0.10$, and vary the dimension over $p \in$

$\{20, 25, 30, \ldots, 100\}$. For each value of p, we run 100 independent trials. so that $\|\beta^\star\|_2 = \sqrt{10}$.

For each method, empirical coverage is computed as

$$\widehat{\mathrm{Cov}} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{Y_{n+1}^{(b)} \in \widehat{C}^{(b)}(X_{n+1}^{(b)})\right\}, \qquad B = 100.$$

The average length is the average Lebesgue measure of the returned prediction set.

C.2.1 Example 0: Deletion

Approximate. For the deletion approximation, we use

$$s_{\text{approx}}^{\text{delete}}(z; \mathcal{D} + \{z'\}) = s(z; \mathcal{D}),$$

where, for $z = (x, y)$,

$$s(z; \mathcal{D}) = |y - x^\top \widehat{\beta}(\mathcal{D})|.$$

Here $\widehat{\beta}(\mathcal{D})$ is the ordinary least-squares fit on $\mathcal{D}$. Thus the uncorrected deletion method fits OLS once on $\mathcal{D}_n$, computes

$$R_i = s_{\text{approx}}^{\text{delete}}(Z_i; \mathcal{D}_n + \{Z_{n+1}^y\}) = |Y_i - X_i^\top \widehat{\beta}(\mathcal{D}_n)|, \qquad i \in [n],$$

and

$$s_{\text{approx}}^{\text{delete}}(Z_{n+1}^y; \mathcal{D}_n + \{Z_{n+1}^y\}) = |y - X_{n+1}^\top \widehat{\beta}(\mathcal{D}_n)|.$$

Tournament-corrected. For the tournament-corrected deletion method, we use

$$s_{\text{approx}}^{\text{delete}}(z; \mathcal{D} + \{z', z''\}) = s(z; \mathcal{D}) = |y - x^\top \widehat{\beta}(\mathcal{D})|.$$

Thus, for each $i \in [n]$, the pairwise comparison between Z_i and Z_{n+1}^y is computed using the leave-one-out fit $\widehat{\beta}(\mathcal{D}_{n\setminus i})$.

C.2.2 Example 1: Rounding

Constructing the grid. For the existing theory to work, the grid must be fixed before hand without looking at the data. This is so that the rounding operation treats the training and the test data symmetrically in the sense that all of them are rounded with respect to the same grid which is chosen independently of the training data. However, without looking at the Y values in our data, it is difficult to even decide what range of values to consider for our grid. Chen et al. (2018) show that if the grid is chosen just by looking at the range of the training Y values i.e. $\min_{i=1,\ldots,n} Y_i$ and $\max_{i=1,\ldots,n} Y_i$, and an equally spaced grid is chosen, then we would lose a coverage of $\frac{2}{n+1}$. Hence, for the rounding approximation, we use a grid of size $M = 10$. In each trial, the grid is constructed from the observed training responses. Let

$$Y_{\min} = \min_{1\le i\le n} Y_i, \qquad Y_{\max} = \max_{1\le i\le n} Y_i, \qquad R_Y = Y_{\max} - Y_{\min}.$$

The grid consists of 10 equally spaced points between

$$Y_{\min} - 0.02R_Y \qquad \text{and} \qquad Y_{\max} + 0.02R_Y.$$

We write $[y]$ for the nearest grid point to y. The real line is partitioned into the corresponding Voronoi cells, with cell boundaries given by the midpoints between adjacent grid values.

Approximate. For the uncorrected rounding method, the approximate score is

$$s^{\text{round}}_{\text{approx}}(z; \mathcal{D} + \{z'\}) = s\Big(z; \mathcal{D} \cup \{(x', [y'])\}\Big), \qquad z' = (x', y').$$

Using the absolute residual score, this is

$$s^{\text{round}}_{\text{approx}}(z; \mathcal{D} + \{z'\}) = |y - x^\top \widehat{\beta}(\mathcal{D} \cup \{(x', [y'])\})|.$$

Thus, for each grid point g_m, we fit OLS on

$$\mathcal{D}_n \cup \{(X_{n+1}, g_m)\}.$$

For candidate values y whose rounded value is $[y] = y^{(m)}$, the score at the candidate point is

$$s^{\text{round}}_{\text{approx}}(Z^y_{n+1}; \mathcal{D}_n + \{Z^y_{n+1}\}) = |y - X^\top_{n+1}\widehat{\beta}(\mathcal{D}_n \cup \{(X_{n+1}, y^{(m)})\})|,$$

and the training scores are

$$s^{\text{round}}_{\text{approx}}(Z_i; \mathcal{D}_n + \{Z^y_{n+1}\}) = |Y_i - X_i^\top \widehat{\beta}(\mathcal{D}_n \cup \{(X_{n+1}, y^{(m)})\})|.$$

Tournament-corrected. For the tournament-corrected rounding method, we use the corrected score

$$s^{\text{round}}_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = s\Big(z; \mathcal{D} \cup \{(x', [y'])\} \cup \{(x'', [y''])\}\Big),$$

where $z = (x, y)$ and $z' = (x', y')$ and $z'' = (x'', y'')$. With the absolute residual score,

$$s^{\text{round}}_{\text{approx}}(z; \mathcal{D} + \{z', z''\}) = |y - x^\top \widehat{\beta}(\mathcal{D} \cup \{(x', [y'])\} \cup \{(x'', [y''])\})|.$$

Therefore, for the pairwise comparison between Z_i and Z^y_{n+1}, the model is fit on

$$\mathcal{D}_{n\setminus i} \cup \{(X_i, [Y_i])\} \cup \{(X_{n+1}, [y])\}.$$

The score at Z_i is computed using the original response Y_i, but the model was trained after replacing Y_i by its rounded value $[Y_i]$.

C.2.3 Example 2: One-step update implementation

Approximate. In the simulation, the fitted model is linear, $f_\theta(x) = x^\top\theta$, and

$$\widehat{\theta}(\mathcal{D}) = \widehat{\beta}(\mathcal{D}) = \arg\min_{\theta\in\mathbb{R}^p} \sum_{(X_j,Y_j)\in\mathcal{D}} (Y_j - X_j^\top\theta)^2.$$

The step size is $\eta = 10$. For a single added point $z' = (x', y')$, the one-step update used in the simulation is

$$\mathrm{Update}_\eta(\widehat{f}_\mathcal{D}, z') = f_{\widetilde{\theta}},$$

where

$$\widetilde{\theta} = \widehat{\theta}(\mathcal{D}) + \frac{\eta}{|\mathcal{D}|+1} x' \left(y' - x'^\top \widehat{\theta}(\mathcal{D})\right).$$

This is the gradient-descent update for the squared-error loss, evaluated at the OLS solution on $\mathcal{D}$.

The uncorrected one-step score is

$$s^{\mathrm{one-step}}_{\mathrm{approx}}(z; \mathcal{D} + \{z'\}) = \left| y - \left[\mathrm{Update}_\eta(\widehat{f}_\mathcal{D}, z')\right](x) \right|, \qquad z = (x, y).$$

Thus, in the uncorrected method,

$$s^{\mathrm{one-step}}_{\mathrm{approx}}(Z^y_{n+1}; \mathcal{D}_n + \{Z^y_{n+1}\}) = \left| y - \left[\mathrm{Update}_\eta(\widehat{f}_{\mathcal{D}_n}, Z^y_{n+1})\right](X_{n+1}) \right|,$$

and

$$s^{\mathrm{one-step}}_{\mathrm{approx}}(Z_i; \mathcal{D}_n + \{Z^y_{n+1}\}) = \left| Y_i - \left[\mathrm{Update}_\eta(\widehat{f}_{\mathcal{D}_n}, Z^y_{n+1})\right](X_i) \right|.$$

The approximate one-step prediction set is computed exactly as a union of intervals. To do

this, let

$$\widehat{\beta} = \widehat{\beta}(\mathcal{D}_n), \qquad d = X_{n+1}^\top \widehat{\beta}, \qquad e_i = Y_i - X_i^\top \widehat{\beta}.$$

For $t = y - d$, define

$$\kappa_i = \frac{\eta}{n+1} X_i^\top X_{n+1}, \qquad \kappa_{n+1} = \frac{\eta}{n+1} \|X_{n+1}\|_2^2.$$

Then

$$s_{\text{approx}}^{\text{one-step}}(Z_{n+1}^y; \mathcal{D}_n + \{Z_{n+1}^y\}) = |1 - \kappa_{n+1}|\,|t|,$$

and

$$s_{\text{approx}}^{\text{one-step}}(Z_i; \mathcal{D}_n + \{Z_{n+1}^y\}) = |e_i - \kappa_i t|.$$

The only points where the rank of the test score among the training scores can change are the solutions of

$$|1 - \kappa_{n+1}|\,|t| = |e_i - \kappa_i t|, \qquad i \in [n],$$

together with $t = 0$. Equivalently, the breakpoints are

$$y = d, \qquad y = d + \frac{e_i}{\kappa_i + |1 - \kappa_{n+1}|}, \qquad y = d + \frac{e_i}{\kappa_i - |1 - \kappa_{n+1}|},$$

whenever the denominators are nonzero. The simulation sorts these breakpoints, evaluates the approximate conformal acceptance condition on each induced interval, and returns the union of accepted intervals. Coverage is evaluated by checking whether Y_{n+1} lies in this union, and length is the total length of the union.

Tournament-corrected. For the tournament-corrected one-step method, the update uses both points in the pairwise comparison. Namely, for $z = (x, y)$ and $z' = (x', y')$, we use

$$s_{\text{approx}}^{\text{one-step}}(z; \mathcal{D} + \{z', z''\}) = \left| y - \left[\text{Update}_\eta(\widehat{f}_{\mathcal{D}}, \{z', z''\}) \right](x) \right|,$$

where

$$\mathrm{Update}_\eta(\widehat{f}_{\mathcal{D}}, \{z', z''\}) = f_{\widetilde{\theta}},$$

and

$$\widetilde{\theta} = \widehat{\theta}(\mathcal{D}) + \frac{\eta}{|\mathcal{D}|+2}\left[x'\left(y' - x'^{\top}\widehat{\theta}(\mathcal{D})\right) + x''\left(y'' - x''^{\top}\widehat{\theta}(\mathcal{D})\right)\right].$$

For the comparison between Z_i and Z^y_{n+1}, we take $\mathcal{D} = \mathcal{D}_{n\setminus i}$. Let

$$\widehat{\beta}_{-i} = \widehat{\beta}(\mathcal{D}_{n\setminus i}), \qquad r_i = Y_i - X_i^{\top}\widehat{\beta}_{-i}, \qquad d_i = X_{n+1}^{\top}\widehat{\beta}_{-i}.$$

Also define

$$a_i = \frac{\eta}{n+1}X_i^{\top}X_{n+1}, \qquad g_i = \frac{\eta}{n+1}\|X_i\|_2^2, \qquad b = \frac{\eta}{n+1}\|X_{n+1}\|_2^2.$$

Writing $t_i = y - d_i$, the two scores in the pairwise comparison are

$$s^{\mathrm{one-step}}_{\mathrm{approx}}(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) = \left|(1-b)t_i - a_i r_i\right|,$$

and

$$s^{\mathrm{one-step}}_{\mathrm{approx}}(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) = \left|(1-g_i)r_i - a_i t_i\right|.$$

The corrected one-step candidate y is accepted if

$$\sum_{i=1}^{n} \mathbf{1}\left\{s^{\mathrm{one-step}}_{\mathrm{approx}}(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) > s^{\mathrm{one-step}}_{\mathrm{approx}}(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\})\right\} < (1-\alpha)(n+1).$$

The corrected one-step set is also computed exactly as a union of intervals. For each i, the breakpoints solve

$$\left|(1-b)(y-d_i) - a_i r_i\right| = \left|(1-g_i)r_i - a_i(y-d_i)\right|.$$

Equivalently, the two candidate breakpoints are

$$y = d_i + \frac{(1 - g_i + a_i)r_i}{1 - b + a_i}, \qquad y = d_i + \frac{(a_i + g_i - 1)r_i}{1 - b - a_i},$$

whenever the denominators are nonzero. The simulation sorts all such breakpoints, evaluates the tournament acceptance condition on each induced interval, and reports the total length of the accepted union. Coverage is evaluated by checking the same tournament condition at $y = Y_{n+1}$.

C.2.4 Example 3: Bayesian posterior predictive density

For the Bayesian experiment, we use the Gaussian linear-model likelihood

$$Y \mid X, \theta \sim N(X^\top \theta, \sigma_{\text{lik}}^2), \qquad \sigma_{\text{lik}} = 1.$$

The prior is

$$\theta \sim N(\mu_0, \Sigma_0), \qquad \mu_0 = 10 \cdot \mathbf{1}_p, \qquad \Sigma_0 = I_p.$$

Since the assumed likelihood and the prior are gaussian, conjugacy holds and we can draw the posterior samples from the closed-form gaussian posterior. We use $K = 100$ Monte Carlo samples.

Approximate. For the uncorrected add-one-in posterior predictive approximation, draw

$$\theta_1, \ldots, \theta_K \sim \pi(\cdot \mid \mathcal{D}_n).$$

Equivalently, the nonconformity score used in the comparisons is

$$s_{\text{approx}}^{\text{Bayes}}(z; \mathcal{D}_n + \{z'\}) = -\sum_{k=1}^{K} f_{\theta_k}(y \mid x) f_{\theta_k}(y' \mid x').$$

Therefore, for candidate y,

$$s^{\text{Bayes}}_{\text{approx}}(Z^y_{n+1}; \mathcal{D}_n + \{Z^y_{n+1}\}) = \sum_{k=1}^{K} f_{\theta_k}(y \mid X_{n+1})^2,$$

and

$$s^{\text{Bayes}}_{\text{approx}}(Z_i; \mathcal{D}_n + \{Z^y_{n+1}\}) = \sum_{k=1}^{K} f_{\theta_k}(Y_i \mid X_i) f_{\theta_k}(y \mid X_{n+1}).$$

Tournament-corrected. For the tournament-corrected Bayesian method, for each $i \in [n]$, we draw

$$\theta_{i,1}, \ldots, \theta_{i,K} \sim \pi(\cdot \mid \mathcal{D}_{n\setminus i}).$$

For the pairwise comparison between Z_i and Z^y_{n+1}, using samples from $\pi(\cdot \mid \mathcal{D}_{n\setminus i})$,

$$s^{\text{Bayes}}_{\text{approx}}(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) = \sum_{k=1}^{K} f_{\theta_{i,k}}(y \mid X_{n+1})^2 f_{\theta_{i,k}}(Y_i \mid X_i),$$

and

$$s^{\text{Bayes}}_{\text{approx}}(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}) = \sum_{k=1}^{K} f_{\theta_{i,k}}(Y_i \mid X_i)^2 f_{\theta_{i,k}}(y \mid X_{n+1}).$$

All Monte Carlo sums for the Bayesian methods are evaluated on the log scale using the log-sum-exp trick. Coverage is evaluated directly at the true response Y_{n+1}. The reported lengths are computed by an adaptive grid approximation. For both the methods the search range is taken as $[Y_{\min} - 0.02\,(Y_{\max} - Y_{\min})\,, Y_{\max} + 0.02\,(Y_{\max} - Y_{\min})]$. The adaptive search starts with a coarse grid and repeatedly refines the left and right endpoints by a shrink factor of 10, until the final grid resolution is 0.1. If no accepted point is found on the search range, the length is recorded as zero. Otherwise, the length is recorded as the difference between the estimated right and left endpoints of the accepted interval.

C.3 Real data implementation details

C.3.1 Rounding: diabetes data

We follow the diabetes-data implementation of Lee and Zhang (2025) as closely as possible. The dataset has $n = 442$ observations and $p = 10$ predictors. To mirror their preprocessing, we normalize the full dataset by

$$X \leftarrow \frac{X - \bar{X}}{s_X \sqrt{p}}, \qquad Y \leftarrow \frac{Y - \bar{Y}}{s_Y},$$

where $\bar{X}$ and s_X denote the columnwise empirical mean and population standard deviation of the full design matrix, and $\bar{Y}$ and s_Y are the empirical mean and population standard deviation of the full response vector. In Lee and Zhang (2025), the real-data experiments are repeated 100 times with test-set size= 100, miscoverage level $\alpha = 0.1$, and random seed 42. In our current experiment, we consider the $m = 100$ setting and average results over 100 random train/test splits.

The underlying prediction model is a no-intercept linear predictor

$$f_{\widehat{\beta}}(x) = x^\top \widehat{\beta},$$

where $\widehat{\beta}$ solves

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \phi_\epsilon\big(Y_i - X_i^\top \beta\big) + \frac{\lambda}{2} \|\beta\|_2^2 \right\}.$$

Here ϕ_ϵ is the Huber loss,

$$\phi_\epsilon(u) = \begin{cases} u^2, & |u| \le \epsilon, \\ 2\epsilon|u| - \epsilon^2, & |u| > \epsilon. \end{cases}$$

We use the same tuning parameters as Lee and Zhang (2025),

$$\epsilon = 1, \qquad \lambda = 2.$$

Their public code solves this optimization problem in `cvxpy`. In our R implementation, we solve the same convex objective using `optim` with the BFGS algorithm. The nonconformity score is the absolute residual

$$s(z; \mathcal{D}) = |y - f_{\widehat{\beta}}(x)|.$$

For the approximate rounding method, we use the same grid-search FullCP construction as in the public code of Lee and Zhang (2025). Given a training set and a test covariate X_{n+1}, define the candidate-response grid

$$\mathcal{Y}_{\text{grid}} = \left\{ y^{(1)}, \ldots, y^{(100)} \right\},$$

where the 100 grid points are equally spaced between

$$\min(Y_{\text{train}}) - \text{sd}(Y_{\text{train}}) \quad \text{and} \quad \max(Y_{\text{train}}) + \text{sd}(Y_{\text{train}}).$$

For each candidate $y \in \mathcal{Y}_{\text{grid}}$, we augment the training data with $Z_{n+1}^{y} = (X_{n+1}, y)$, refit the robust linear model on the augmented sample, and compute the $n+1$ absolute residual scores. We then include y whenever the test-point score is no larger than the empirical $(1-\alpha)$-quantile of these $n+1$ scores.

For the tournament-corrected rounding method, we use the same candidate grid $\mathcal{Y}_{\text{grid}}$, the same robust linear model, and the same absolute-residual score, but replace the approximate full conformal check by the corrected pairwise comparison from Section 4.3. Concretely, for

each candidate $y \in \mathcal{Y}_{\text{grid}}$ and each $i \in [n]$, we compute

$$s_{\text{approx}}\big(Z^y_{n+1}; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}\big) \qquad \text{and} \qquad s_{\text{approx}}\big(Z_i; \mathcal{D}_{n\setminus i} + \{Z_i, Z^y_{n+1}\}\big),$$

where the approximation is of rounding type, with Y_i rounded to its nearest grid value and y restricted to the same candidate grid.

Finally, for both methods we report empirical coverage and average prediction-set length over 100 random train/test splits.

C.3.2 Bayes: diabetes data

We use the diabetes regression data from the `lars` package, which contains $n = 442$ observations and $p = 10$ covariates. For each repetition, we randomly split the data into a 70% training set and a 30% test set. The covariates and response are standardized within each split using the training-set mean and standard deviation, and the same covariate standardization is then applied to the test set. All reported prediction-set lengths for this experiment are on the standardized response scale.

Following the sparse-regression experiment of Fong and Holmes (2021), we use a Bayesian lasso model. Conditional on the covariates, the likelihood is

$$Y_i \mid X_i, \beta_0, \beta, \sigma \sim N(\beta_0 + X_i^\top \beta, \sigma^2).$$

The regression coefficients are given independent Laplace priors,

$$\beta_j \mid b \sim \text{Laplace}(0, b), \qquad j = 1, \ldots, p,$$

with hyperprior

$$b \sim \text{Gamma}(1, 1).$$

For the noise scale, we use

$$\sigma \sim \text{HalfNormal}(1).$$

The intercept is not penalized. In the Stan implementation, we use a weak Gaussian prior

$$\beta_0 \sim N(0, 10^2),$$

which is effectively diffuse on the standardized response scale and is used for numerical stability. This is the same Bayesian lasso specification used for the conformal Bayes comparison, up to this weak regularization of the intercept.

Posterior sampling is performed using `rstan`. For each posterior fit, we run 4 chains with 3000 iterations per chain, of which 1000 are warmup iterations. Thus each fit produces 2000 post-warmup samples per chain. We use the same settings for both the full-data posterior used by the approximate Bayesian method and the leave-one-out posteriors used by the corrected method. The initialization is chosen deterministically: the coefficients are initialized at zero, the intercept at the empirical mean of the standardized training responses.

Prediction-set length is computed using the same grid-based approximation as in the conformal Bayes implementation. For each test point, we evaluate membership on a one-dimensional grid of 100 candidate response values,

$$\mathcal{G} = \{\min(Y_{\text{train}}) - 2, \ldots, \max(Y_{\text{train}}) + 2\},$$

where the response is on the standardized scale. If Δ denotes the grid spacing and $A \subseteq \mathcal{G}$ is the set of accepted grid points, then the reported length is $|A|\Delta$.

Coverage is not evaluated by rounding the true test response to the grid. Instead, for each test point we evaluate the conformal rank condition directly at the observed standardized value Y_{n+1}. The final result is over 50 independent train/test splits where for each repetition, we compute the average coverage and average length over the test points.

C.4 Coverage guarantee of approximate methods under stability

In this section we show that the uncorrected approximate prediction set $\widehat{C}_{n,\alpha}^{\text{approx}}$ can also satisfy a near-$(1-\alpha)$ coverage guarantee, but under a stronger stability condition. This result is analogous to the stability result for the naive method in Barber et al. (2021); in the deletion case, $s_{\text{approx}}(z; \mathcal{D} + \{z'\}) = s(z; \mathcal{D})$, the approximate method is exactly the naive method.

We define the 2ϵ-inflated version of the $\widehat{C}_{n,\alpha}^{\text{approx}}$ as

$$\widehat{C}_{n,\alpha}^{\text{approx},2\epsilon}(X_{n+1}) = \left\{ y : \sum_{i=1}^{n} \mathbf{1}\left\{ s_{\text{approx}}\Big(Z_{n+1}^{y}; \mathcal{D}_n + \{Z_{n+1}^{y}\}\Big) > s_{\text{approx}}\Big(Z_i; \mathcal{D}_n + \{Z_{n+1}^{y}\}\Big) + 2\epsilon \right\} < (1-\alpha)(n+1) \right\}.$$

Theorem C.4.1 (Coverage of the approximate method under stronger stability)**.** *Suppose that $\mathcal{D}_{n+1} = \{Z_1, \ldots, Z_{n+1}\}$ is exchangeable. Suppose also that the score s is symmetric in its dataset argument, and that the stability conditions in* (4.8) *hold for some $\epsilon \geq 0$ and $\nu \in [0,1]$. Then*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{approx},2\epsilon}(X_{n+1})\right) \geq 1 - \alpha - 3\sqrt{\nu}.$$

Proof. Write

$$A_{n+1} := s_{\text{approx}}(Z_{n+1}; \mathcal{D}_n + \{Z_{n+1}\}), \qquad A_i := s_{\text{approx}}(Z_i; \mathcal{D}_n + \{Z_{n+1}\}), \quad i \in [n],$$

and define the corresponding full-conformal scores

$$S_j := s(Z_j; \mathcal{D}_{n+1}), \qquad j = 1, \ldots, n+1.$$

By exchangeability of $\mathcal{D}_{n+1}$ and symmetry of s, the vector $(S_1, \ldots, S_{n+1})$ is exchangeable.

By definition of the 2ϵ-inflated approximate set,

$$Y_{n+1} \notin \widehat{C}_{n,\alpha}^{\text{approx},2\epsilon}(X_{n+1}) \quad \Longleftrightarrow \quad \sum_{i=1}^{n} \mathbf{1}\left\{A_{n+1} > A_i + 2\epsilon\right\} \geq (1-\alpha)(n+1).$$

Define the stability events

$$E_{n+1} := \left\{|A_{n+1} - S_{n+1}| \leq \epsilon\right\}, \qquad E_i := \left\{|A_i - S_i| \leq \epsilon\right\}, \quad i \in [n].$$

On the event $E_{n+1} \cap E_i$, if

$$A_{n+1} > A_i + 2\epsilon,$$

then

$$S_{n+1} \geq A_{n+1} - \epsilon > A_i + \epsilon \geq S_i.$$

Therefore, on the event E_{n+1},

$$\mathbf{1}\left\{A_{n+1} > A_i + 2\epsilon\right\} \leq \mathbf{1}\left\{S_{n+1} > S_i\right\} + \mathbf{1}\{E_i^c\}.$$

Summing over $i \in [n]$, we get, on E_{n+1},

$$\sum_{i=1}^{n} \mathbf{1}\left\{A_{n+1} > A_i + 2\epsilon\right\} \leq \sum_{i=1}^{n} \mathbf{1}\left\{S_{n+1} > S_i\right\} + \sum_{i=1}^{n} \mathbf{1}\{E_i^c\}.$$

$$\mathbb{P}\left(Y_{n+1} \notin \widehat{C}_{n,\alpha}^{\text{approx},2\epsilon}(X_{n+1})\right) \leq \mathbb{P}(E_{n+1}^c) + \mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{E_i^c\} > \sqrt{\nu}(n+1)\right) + \mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\left\{S_{n+1} > S_i\right\} \geq (1-\alpha-\sqrt{\nu})(n+1)\right).$$

By our stability assumptions, we have $\mathbb{P}(E_{n+1}^c) \leq \nu$ and therefore

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{E_i^c\} > \sqrt{\nu}(n+1)\right) \leq \frac{\sum_{i=1}^{n} \mathbb{P}(E_i^c)}{\sqrt{\nu}(n+1)} \leq \frac{n\nu}{\sqrt{\nu}(n+1)} \leq \frac{n}{n+1}\sqrt{\nu},$$

where the first step follows by Markov's inequality. Since $(S_1, \ldots, S_{n+1})$ is exchangeable, for any $t \in \mathbb{R}$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbf{1}\{S_{n+1} > S_i\} \geq (1-\alpha-\sqrt{\nu})(n+1)\right) \leq \alpha + \sqrt{\nu}.$$

Combining these, we obtain

$$\mathbb{P}\left(Y_{n+1} \notin \widehat{C}_{n,\alpha}^{\text{approx},2\epsilon}(X_{n+1})\right) \leq \nu + \sqrt{\nu} + \alpha + \sqrt{\nu} \leq \alpha + 3\sqrt{\nu},$$

and the result follows. □

C.5 Computational speedup for tournament-corrected Bayesian approximate PPD scores

In this section, we consider the question of computational cost for the Bayesian PPD method (Example 3). In particular, it is often costly to draw samples from posterior distributions; here we examine how this cost plays a role in the tournament-corrected method.

To implement the (uncorrected) approximation of Fong and Holmes (2021) (see Example 3 in Section 4.2), we need to sample K times from the posterior of θ given the dataset $\mathcal{D}_n$:

$$\theta_1, \ldots, \theta_K \sim \pi(\cdot \mid \mathcal{D}_n).$$

On the other hand, a naive implementation of the tournament-corrected method for this

example (see Example 3 in Section 4.3.1) requires sampling

$$\theta_{1,i}, \ldots, \theta_{K,i} \sim \pi(\cdot \mid \mathcal{D}_{n\setminus i}).$$

In this construction, K posterior samples are drawn for *each* leave-one-out dataset $\mathcal{D}_{n\setminus i}$. While the randomized coverage guarantee of Theorem C.1.1 ensures that this construction leads to $\geq 1-2\alpha$ coverage, this comes at a computational cost: an n-fold increase in sampling cost relative to the AOI-PPD method, i.e., the uncorrected approximation.

We now develop a different strategy that allows us to retain the coverage guarantee offered by the corrected method, while avoiding the n-fold increase in cost, incurred by drawing independent samples from each posterior $\pi(\cdot \mid \mathcal{D}_{n\setminus i})$. We will now see how we can use rejection sampling to avoid this increase in cost. We will need to assume that the likelihood is lower-bounded, $\inf_\theta f_\theta(y \mid x) \geq c(z)$ for some $c(z) > 0$, for each data point $z = (x, y)$. We will draw the posterior samples as follows:

Bayesian PPD with shared sampling:

- Define $k_i = 0$ for all $i \in [n]$.
- Repeat until $k_i = K$ for all $i \in [n]$:
  - Sample $\theta \sim \pi(\cdot \mid \mathcal{D}_n)$
  - For each $i \in [n]$ with $k_i < K$:
    - Sample $V \sim \text{Unif}[0, 1]$.
    - If $V \leq \frac{c(Z_i)}{f_\theta(Y_i \mid X_i)}$, then set $k_i \leftarrow k_i + 1$ and define $\theta_{k_i,i} = \theta$.

In other words, we use the same stream of candidate draws θ, and then perform rejection sampling for each $i \in [n]$ to define posterior samples $\theta_{k,i} \sim \pi(\cdot \mid \mathcal{D}_{n\setminus i})$. We now need to see how this construction can be used within the framework of our coverage results: it is not

immediately clear whether this sampling scheme might be treating the data points (training + test data) in a nonsymmetric way, violating exchangeability.

Our next observation is that we can view the scheme above in a different way. Imagine that we have access to the test point, so that we can use the entire dataset $\mathcal{D}_{n+1}$ for sampling. Then we proceed as follows:

Bayesian PPD with shared sampling—alternative representation:

- Define $k_i = 0$ for all $i \in [n]$.
- Repeat until $k_i = K$ for all $i \in [n]$:
  - Sample $\theta \sim \pi(\cdot \mid \mathcal{D}_{n+1})$ and $U, V_1, \ldots, V_n \overset{\text{d}}{\sim} \text{Unif}[0,1]$, independently.
  - For each $i \in [n]$ with $k_i < K$:
    - If $U \leq \frac{c(Z_{n+1})}{f_\theta(Y_{n+1}|X_{n+1})}$ and $V_i \leq \frac{c(Z_i)}{f_\theta(Y_i|X_i)}$, then set $k_i \leftarrow k_i + 1$ and define $\theta_{k_i,i} = \theta$.

But sampling $\theta \sim \pi(\cdot \mid \mathcal{D}_{n+1})$, and then proceeding only if the criterion $U \leq \frac{c(Z_{n+1})}{f_\theta(Y_{n+1}|X_{n+1})}$ is satisfied, is exactly equivalent to simply drawing $\theta \sim \pi(\cdot \mid \mathcal{D}_n)$ (i.e., this is rejection sampling, which removes the test data point Z_{n+1} from the posterior). In other words, this is exactly equivalent to the sampling scheme defined above.

Now we will see how to apply Theorem C.1.1. Let $\xi \sim \text{Unif}[0,1]$ be a random seed. Abusing notation, for any $\xi \in [0,1]$ we will write $\xi = (\xi', \xi'', \xi''')$ to denote 'splitting' the random seed into three i.i.d. Unif$[0,1]$ random seeds (e.g., by cycling through the random digits of ξ). Define a deterministic function $\texttt{Sample}_\pi(\mathcal{D}; \xi)$ such that

$$(\theta^{(1)}, \theta^{(2)}, \ldots) = \texttt{Sample}_\pi(\mathcal{D}; \xi)$$

returns an infinite stream of i.i.d. draws $\theta \sim \pi(\cdot \mid \mathcal{D})$ when the random seed is sampled as

$\xi \sim \text{Unif}[0,1]$. Similarly define a deterministic function $\texttt{Sample}_{\text{Unif}}(\xi)$ such that

$$(V^{(1)}, V^{(2)}, \dots) = \texttt{Sample}_{\text{Unif}}(\xi)$$

returns an infinite stream of i.i.d. draws $V \sim \text{Unif}[0,1]$ when the random seed is sampled as $\xi \sim \text{Unif}[0,1]$. Next define $\texttt{Sample}(\mathcal{D}, z, z'; \xi, K)$ as follows:

- Define $(\theta^{(1)}, \theta^{(2)}, \dots) = \texttt{Sample}_{\pi}(\mathcal{D} \cup \{z, z'\}; \xi')$;
- Define $(U^{(1)}, U^{(2)}, \dots) = \texttt{Sample}_{\text{Unif}}(\xi'')$ and $(V^{(1)}, V^{(2)}, \dots) = \texttt{Sample}_{\text{Unif}}(\xi''')$;
- Let $i_1 < \cdots < i_K$ be the first K indices i for which $U^{(i)} \le \frac{c(z)}{f_{\theta^{(i)}}(z)}$ and $V^{(i)} \le \frac{c(z')}{f_{\theta^{(i)}}(z')}$;
- Return $\theta^{(i_1)}, \dots, \theta^{(i_K)}$.

In other words, this returns a sample

$$(\theta_1, \dots, \theta_K) = \texttt{Sample}(\mathcal{D}, z, z'; \xi, K),$$

which is a deterministic function, but for a random seed $\xi \sim \text{Unif}[0,1]$ this is equivalent to a rejection sampling procedure (which draws from $\pi(\cdot \mid \mathcal{D} \cup \{z, z'\})$ and then uses rejection sampling to approximate removing data points z, z'). Note that, taking $\mathcal{D} = \mathcal{D}_{n\setminus i}$ and $z = Z_{n+1}, z' = Z_i$, this is exactly equivalent to sampling scheme (B) for each $i \in [n]$.

Finally, define the score

$$s_{\text{approx}}(z; \mathcal{D} + \{z', z''\}; \xi) = -\sum_{k=1}^{K} f_{\theta_k}(y \mid x) \cdot f_{\theta_k}(y' \mid x') \cdot f_{\theta_k}(y'' \mid x'') \tag{C.8}$$

where

$$(\theta_1, \dots, \theta_K) = \texttt{Sample}(\mathcal{D}, z', z''; \xi, K).$$

Now we apply this to the augmented dataset $\mathcal{D}_{n+1}$. Let $\xi', \xi''_1, \dots, \xi''_{n+1} \overset{\text{iid}}{\sim} \text{Unif}[0,1]$, and

define

$$\xi_{ij} = (\xi', \xi''_i, \xi''_j)$$

for $i, j \in [n+1]$. We can verify that the tournament-corrected method (with shared sampling, as defined above) is therefore equivalent to using the randomized score function (C.8) with this random seed construction. Since this randomized score function satisfies the conditions (C.4) required by Theorem C.1.1, this verifies theoretical validity.

APPENDIX D

APPENDIX TO CHAPTER 5: CONDITIONING ON POSTERIOR SAMPLES FOR FLEXIBLE FREQUENTIST GOODNESS-OF-FIT TESTING

D.1 Theoretical guarantees for sampling the copies

D.1.1 Dependence among the copies

In this section, we return to the question raised in Section 5.3.3: since sampling $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ i.i.d. from the distribution $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is often computationally infeasible, can we relax this sampling step without losing finite-sample type-I error control?

In Algorithm 1, we assume that, after observing the posterior draws $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$, the copies $\widetilde{X}^{(m)}$ are then sampled i.i.d. from the distribution $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. In practice, sampling from a complex distribution is often carried out via MCMC based strategies, and the resulting samples are only approximately i.i.d.—specifically, samples obtained by running a Markov chain will have dependence. While it is common in many sampling problems to assume mixing conditions for the Markov chain, in order to ensure that the resulting samples are approximately i.i.d., here we will use a different approach in order to ensure finite-sample validity.

Formally, define the joint sampling distribution of the copies, conditional on the data X and the posterior draws $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$, as

$$(\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid X, \widehat{\theta}_{1:B} \sim \widetilde{Q}_\pi^M(\cdot \mid X, \widehat{\theta}_{1:B}).$$

We will assume the following property of this joint distribution:

For any $\theta_1, \ldots, \theta_B \in \Theta$,

$$\text{if } X \sim g_\pi(\,\cdot \mid \theta_{1:B}) \text{ and } (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid X \;\sim\; \widetilde{Q}_\pi^M(\,\cdot \mid X, \theta_{1:B}),$$
$$\text{then the random vector } (X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \text{ is exchangeable.} \quad \text{(D.1)}$$

For example, the i.i.d. sampling strategy of Algorithm (1) satisfies this condition—we can define $\widetilde{Q}_\pi^M(\,\cdot \mid X, \theta_{1:B})$ as follows:

$$\widetilde{Q}_\pi^M(\,\cdot \mid X, \theta_{1:B}) \text{ is the distribution with joint density } g_\pi(x_1 \mid \theta_{1:B}) \cdot \ldots \cdot g_\pi(x_M \mid \theta_{1:B}). \quad \text{(D.2)}$$

However, even in settings where i.i.d. sampling is infeasible, we can nonetheless use MCMC strategies—e.g., the permuted serial sampler (Besag and Clifford, 1989)—to draw the copies from a joint distribution $\widetilde{Q}_\pi^M(\,\cdot \mid X, \theta_{1:B})$ that exactly satisfies this condition (see Barber and Janson (2022, Section 2.2.3) for more details on how to implement this in the setting of aCSS).

The aCSS-B method, with more general sampling strategy, is presented in Algorithm 2. Of course, the original version of the method, given in Algorithm 1, is simply a special case obtained by choosing the i.i.d. sampling strategy as in (D.2).

Our next result, proved in Appendix D.2.3, is a generalization of theorem 5.3.5, showing that as long as the copies are sampled from a distribution satisfying (D.1), the same bound on type-I error still holds.

Theorem D.1.1. *After observing the data X, let $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ be sampled as in Algorithm 2, for some positive prior density π on Θ. Then, if $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$,*

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\},$$

where the infimum is taken over all densities π_0 on Θ with respect to base measure ν_Θ such that the support of $\bar{f}_{\pi_0}(x)$ contains the support of $f_\theta(x)$ for all θ.

Algorithm 2: aCSS-B method (general case)

Given: Prior density π on Θ, and test statistic $T : \mathcal{X} \to \mathbb{R}$.
Data: $X \sim f_{\theta_0}$.

1 Generate B posterior samples,

$$\widehat{\theta}_1, \ldots, \widehat{\theta}_B \mid X \overset{\text{i.i.d.}}{\sim} \pi(\cdot \mid X).$$

2 Generate M copies of the data,

$$\left(\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \mid X, \widehat{\theta}_{1:B} \sim \widetilde{Q}_\pi^M\left(\cdot \mid X, \widehat{\theta}_{1:B}\right),$$

where the distribution $\widetilde{Q}_\pi^M(\cdot \mid X, \widehat{\theta}_{1:B})$ is chosen to satisfy (D.1).

3 Compute the p-value

$$\text{pval} = \frac{1 + \sum_{m=1}^M \mathbb{1}\{T(\widetilde{X}^{(m)}) \geq T(X)\}}{M+1}.$$

D.1.2 Estimating the distribution for the copies

Thus far, we have considered the setting where the density $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is computable exactly, even though drawing copies independently from this distribution may not be feasible. Next, we turn to the problem of computing this distribution itself. Recall that the density $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is defined as

$$\propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}},$$

where

$$\bar{f}_\pi(x) = \int_\Theta f_\theta(x)\pi(\theta)\, \mathsf{d}\nu_\Theta(\theta)$$

is the marginal density of X, integrated over the prior $\theta \sim \pi$. In many settings, evaluating the likelihood $f_\theta(x)$ at a single value θ is straightforward, but calculating the marginal likelihood can only be carried out approximately: the integral defining the marginal likelihood is rarely available in closed form except for simple conjugate models, and is well known to be intractable in practice in many settings (Chib, 1995; Gelman et al., 1995). A variety of numerical methods

have been proposed to estimate $\bar{f}_\pi(x)$, such as the Laplace approximation (Barber et al., 2016), importance sampling (Newton and Raftery, 1994), Chib's MCMC estimator (Chib, 1995; Chib and Jeliazkov, 2001), bridge sampling, and path sampling (Meng and Wong, 1996; Gelman and Meng, 1998). These techniques can be leveraged to provide an estimated marginal density, $\widehat{f}_\pi(x)$, that we can then use in place of $\bar{f}_\pi(x)$ in the aCSS-B method. (Note that $\widehat{f}_\pi(x)$ is treated as fixed—that is, we assume this estimate of the marginal distribution is computed independently of the data used in the aCSS-B procedure.)

Specifically, we can run Algorithm 1 with density $\widehat{g}_\pi(\cdot \mid \widehat{\theta}_{1:B})$, defined as

$$\widehat{g}_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\widehat{f}_\pi(x)^{B-1}},$$

in place of $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$. (As for $\bar{f}_\pi$ earlier, now we will need to assume positivity of $\widehat{f}_\pi(x)$ in order for this to be well-defined—that is, the support of $f_\theta(x)$ is contained in the support of $\widehat{f}_\pi(x)$, for every θ.) Or, more generally, if sampling i.i.d. copies is not feasible then we can run Algorithm 2 with a joint distribution $\widehat{Q}_\pi^M(\cdot \mid \widehat{\theta}_{1:B})$, satisfying

For any $\theta_1, \ldots, \theta_B \in \Theta$,

$$\text{if } X \sim \widehat{g}_\pi(\cdot \mid \theta_{1:B}) \text{ and } (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid X \sim \widehat{Q}_\pi^M(\cdot \mid X, \theta_{1:B}),$$
$$\text{then the random vector } (X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \text{ is exchangeable.} \quad \text{(D.3)}$$

Of course, this modification comes at the potential cost of additional type-I error inflation, since we are now sampling the copies from an approximation to the original distribution. The following theorem, proved in Appendix D.2.5, provides a bound on the type-I error inflation of aCSS-B, when we use an estimate $\widehat{f}_\pi$ in place of the exact marginal likelihood $\bar{f}_\pi$.

Theorem D.1.2. *After observing the data X, let $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}$ be sampled via Algorithm 2 except implemented with a joint distribution $\widehat{Q}_\pi^M(\cdot \mid X, \theta_{1:B})$ satisfying* (D.3)*, for some*

positive prior density π on Θ. Assume that

$$\mathbb{E}_{\theta_0}\left[\left|\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right|\right] \leq \delta_B.$$

Then, if $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$,

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \inf_{\pi_0}\left\{\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}\right\} + \delta_B,$$

where the infimum is taken over all densities π_0 on Θ with respect to base measure ν_Θ such that the support of $\bar{f}_{\pi_0}(x)$ contains the support of $f_\theta(x)$ for all θ.

In other words, the additional term in the bound, δ_B, is small whenever the estimated marginal density $\widehat{f}_\pi(x)$ is a sufficiently accurate approximation to $\bar{f}_\pi(x)$. (Note that the dependence on B implies that we require a more accurate approximation $\widehat{f}_\pi$ when B is large.)

D.2 Proofs

D.2.1 Proof of Theorem 5.3.2

Since π is assumed to be a positive density, note that $f_\theta(x) = \frac{\pi(\theta|x)}{\pi(\theta)} \cdot \bar{f}_\pi(x)$, where on the right-hand side, the first term depends on x only via the statistic $\pi(\,\cdot \mid x)$, and the second term depends on x only and not on θ. By the Fisher–Neyman factorization theorem (Casella and Berger, 2002, pg. 276), this implies that $\pi(\,\cdot \mid X)$ is a sufficient statistic of X. On the other hand, it is well known that the likelihood function $\theta \mapsto f_\theta(X)$ is a minimal sufficient statistic (see, e.g., Fraser (1963)). Since the posterior density function $\theta \mapsto \pi(\theta \mid X)$ depends on X only through the likelihood function $\theta \mapsto f_\theta(X)$, this means that the posterior is minimal sufficient.

D.2.2 Proof of Theorem 5.3.5

As explained in Section D.1.1, by defining $\widetilde{Q}_\pi^M(\cdot \mid X, \theta_{1:B})$ as in (D.2), the result of Theorem 5.3.5 is simply a special case of Theorem D.1.1—see Section D.2.3 for the proof of this more general theorem.

D.2.3 Proof of Theorem D.1.1

Let P_0 denote the joint distribution of $(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)$, which is given by

$$\begin{cases} X \sim f_{\theta_0}, \\ \{\widehat{\theta}_b\}_{b=1,\ldots,B} \mid X \overset{\text{iid}}{\sim} \pi(\cdot \mid X). \end{cases} \tag{D.4}$$

This distribution has the following joint density at $(x, \theta_1, \ldots, \theta_B)$:

$$f_{\theta_0}(x) \cdot \prod_{b=1}^{B} \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)}$$

(with respect to the base measure $\nu_{\mathcal{X}} \times \nu_\Theta \times \cdots \times \nu_\Theta$). The aCSS-B procedure can be equivalently characterized as

$$\begin{cases} (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim P_0, \\ (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widetilde{Q}_\pi^M(\cdot \mid X, \widehat{\theta}_{1:B}). \end{cases} \tag{D.5}$$

Our goal, then, is to bound the distance to exchangeability induced by this distribution.

Step 1: Defining P_1 to bound distance to exchangeability. We next define a joint distribution P_1, with joint density at $(x, \theta_1, \ldots, \theta_B)$ given by:

$$\frac{1}{B} \sum_{b=1}^{B} \frac{\pi_0(\theta_b)}{\pi(\theta_b)} \cdot \bar{f}_\pi(x) \cdot \prod_{b=1}^{B} \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)}.$$

(We can verify that this expression integrates to 1 by definition of $\bar{f}_\pi$, and therefore defines a valid joint density.) Note that, by construction, under the joint distribution P_1 we have

$$X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim g_\pi(\cdot \mid \widehat{\theta}_{1:B}),$$

where we recall that $g_\pi(\cdot \mid \widehat{\theta}_{1:B})$ is the distribution with density $\propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x)}{\bar{f}_\pi(x)^{B-1}}$. Therefore, if we consider the sampling distribution

$$\begin{cases} (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim P_1, \\ (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widetilde{Q}_\pi^M(\cdot \mid X, \widehat{\theta}_{1:B}), \end{cases} \tag{D.6}$$

then by (D.1) on $\widetilde{Q}_\pi^M(\cdot \mid X, \widehat{\theta}_{1:B})$, it holds that $(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ are exchangeable conditional on $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$, and consequently are also exchangeable after marginalizing over $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$.

Consequently, we can see that $\mathrm{d}_{\mathrm{exch}}(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ is bounded by the total variation distance between the sampling distributions given in (D.5) and in (D.6), which can be simplified to

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \mathrm{d}_{\mathrm{TV}}(P_0, P_1),$$

since the distributions (D.5) and (D.6) differ only in their first line. From this point on, then, we only need to bound this last total variation distance.

Step 2: Defining $P_{0.5}$. First, we define an intermediate distribution, $P_{0.5}$, with joint density at $(x, \theta_1, \ldots, \theta_B)$ given by

$$\frac{1}{B} \sum_{b=1}^B \frac{\pi_0(\theta_b)/\bar{f}_{\pi_0}(x)}{\pi(\theta_b)/\bar{f}_\pi(x)} \cdot f_{\theta_0}(x) \cdot \prod_{b=1}^B \frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)},$$

where

$$\bar{f}_{\pi_0}(x) = \int_\Theta f_\theta(x)\pi_0(\theta)\, \mathsf{d}\nu_\Theta(\theta)$$

is the marginal likelihood corresponding to the prior π_0. We can then write

$$\mathrm{d_{TV}}(P_0, P_1) \leq \mathrm{d_{TV}}(P_0, P_{0.5}) + \mathrm{d_{TV}}(P_{0.5}, P_1).$$

We now bound the remaining two terms separately.

Step 3: Bounding $\mathrm{d_{TV}}(P_{0.5}, P_1)$. Note that we can express the distribution $P_{0.5}$ as a mixture, $P_{0.5} = \frac{1}{B}\sum_{b=1}^{B} P_{0.5,b}$, where $P_{0.5,b}$ has joint density

$$f_{\theta_0}(x) \cdot \frac{f_{\theta_b}(x)\pi_0(\theta_b)}{\bar{f}_{\pi_0}(x)} \cdot \prod_{b' \neq b} \frac{f_{\theta_{b'}}(x)\pi(\theta_{b'})}{\bar{f}_{\pi}(x)},$$

which can be interpreted as follows:

$$\begin{cases} X \sim f_{\theta_0}, \\ \widehat{\theta}_b \mid X \ \sim \ \pi_0(\cdot \mid X), \\ \{\widehat{\theta}_{b'}\}_{b' \neq b} \mid X, \widehat{\theta}_b \ \overset{\text{iid}}{\sim} \ \pi(\cdot \mid X). \end{cases} \tag{D.7}$$

Similarly, we can express the distribution P_1 as a mixture, $P_1 = \frac{1}{B}\sum_{b=1}^{B} P_{1,b}$, where $P_{1,b}$ has joint density

$$\pi_0(\theta_b) \cdot f_{\theta_b}(x) \cdot \prod_{b' \neq b} \frac{f_{\theta_{b'}}(x)\pi(\theta_{b'})}{\bar{f}_{\pi}(x)},$$

which can be interpreted as follows:

$$\begin{cases} \widehat{\theta}_b \sim \pi_0, \\ X \mid \widehat{\theta}_b \ \sim \ f_{\widehat{\theta}_b}, \\ \{\widehat{\theta}_{b'}\}_{b' \neq b} \mid X, \widehat{\theta}_b \ \overset{\text{iid}}{\sim} \ \pi(\cdot \mid X). \end{cases}$$

Equivalently, by swapping the order in which we draw $\widehat{\theta}_b$ and X, this is the same as

$$\begin{cases} X \sim \bar{f}_{\pi_0}, \\ \widehat{\theta}_b \mid X \sim \pi_0(\cdot \mid X), \\ \{\widehat{\theta}_{b'}\}_{b' \neq b} \mid X, \widehat{\theta}_b \overset{\text{iid}}{\sim} \pi(\cdot \mid X). \end{cases} \tag{D.8}$$

By comparing (D.7) and (D.8), we can see that

$$\mathrm{d_{TV}}(P_{0.5,b}, P_{1,b}) = \mathrm{d_{TV}}(f_{\theta_0}, \bar{f}_{\pi_0}),$$

and moreover, $\mathrm{d_{TV}}(f_{\theta_0}, \bar{f}_{\pi_0}) \leq \epsilon(\pi_0)$ by our definition of $\epsilon(\pi_0)$ (see Theorem 5.3.3). Then

$$\mathrm{d_{TV}}(P_{0.5}, P_1) = \mathrm{d_{TV}}\left(\frac{1}{B}\sum_{b=1}^{B} P_{0.5,b}, \frac{1}{B}\sum_{b=1}^{B} P_{1,b}\right) \leq \frac{1}{B}\sum_{b=1}^{B} \mathrm{d_{TV}}(P_{0.5,b}, P_{1,b}) \leq \epsilon(\pi_0).$$

Step 4: Bounding $\mathrm{d_{TV}}(P_0, P_{0.5})$. By comparing (D.4) with (D.7), we can see that the marginal distribution of X is the same for both (i.e., $X \sim f_{\theta_0}$), while the conditional distribution of $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B) \mid X$ is different: under P_0, these B values are drawn i.i.d. from the posterior $\pi(\cdot \mid X)$, while under $P_{0.5}$, we instead draw one $\widehat{\theta}_b$ (with b selected uniformly at random) from $\pi_0(\cdot \mid X)$, and the remaining values $\{\widehat{\theta}_{b'}\}_{b' \neq b}$ are drawn i.i.d. from $\pi(\cdot \mid X)$. We will now need the following lemma:

Lemma D.2.1. *Let P and Q be probability measures on a measurable space $(\Omega, \mathcal{F})$ with $Q \ll P$ and let $B > 1$ be an integer. Let $Z = (Z_1, \ldots, Z_B)$, where $Z_1, \ldots, Z_B \overset{iid}{\sim} P$. Define a corrupted vector $Z' = \left(Z'_1, \ldots, Z'_B\right)$, where we draw a random $K \sim$ Uniform $\{1, \ldots, B\}$, and sample $Z'_K \sim Q$, and set $Z'_k = Z_k$ for all $k \neq K$. Then*

$$\mathrm{d_{TV}}\left(Z, Z'\right) \leq \frac{1}{2}\sqrt{\frac{\mathrm{d}_{\chi^2}(Q\|P)}{B}}.$$

Applying this Lemma, we then have

$$\mathrm{d_{TV}}\left((P_0)_{\widehat{\theta}_{1:B}|X}, (P_{0.5})_{\widehat{\theta}_{1:B}|X}\right) \leq \frac{1}{2}\sqrt{\frac{\mathrm{d}_{\chi^2}\big(\pi_0(\cdot \mid X) \,\|\, \pi(\cdot \mid X)\big)}{B}},$$

where $(P_0)_{\widehat{\theta}_{1:B}|X}$ denotes the conditional distribution of $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B) \mid X$ under the joint distribution P_0, and same for $P_{0.5}$. Therefore,

$$\mathrm{d_{TV}}(P_0, P_{0.5}) = \mathbb{E}\left[\mathrm{d_{TV}}\left((P_0)_{\widehat{\theta}_{1:B}|X}, (P_{0.5})_{\widehat{\theta}_{1:B}|X}\right)\right] \leq \frac{\Delta(\pi_0)}{2\sqrt{B}},$$

by definition of $\Delta(\pi_0)$ (see Theorem 5.3.4).

D.2.4 Proof of Theorem D.2.1

By construction, the distribution of Z' can be written as

$$\widetilde{Q} \;=\; \frac{1}{B}\sum_{b=1}^{B}\left(P^{b-1} \otimes Q \otimes P^{B-b}\right) \;\ll\; P^B.$$

If we denote $f = \frac{\mathsf{d}Q}{\mathsf{d}P}$, then for each b,

$$\frac{\mathsf{d}\left(P^{b-1} \times Q \times P^{B-b}\right)}{\mathsf{d}P^B}(x_1, \ldots, x_B) = f(x_b),$$

for P^B-almost-every $(x_1, \ldots, x_B)$. Hence we can calculate the Radon–Nikodym derivative

$$\frac{\mathsf{d}\widetilde{Q}}{\mathsf{d}P^B}(x_1, \ldots, x_B) = \frac{1}{B}\sum_{b=1}^{B} f(x_b),$$

for P^B-almost-every $(x_1, \ldots, x_B)$. By definition of total variation,

$$\mathrm{d_{TV}}(P^B, \widetilde{Q}) = \frac{1}{2}\int_{\Omega^B}\left|1 - \frac{1}{B}\sum_{b=1}^{B} f(x_b)\right| \mathsf{d}P^B(x).$$

Now define $\Delta(x) = 1 - f(x)$. Note that

$$\int_\Omega \Delta(x)\, \mathsf{d}P(x) = \int_\Omega (1 - f(x))\, \mathsf{d}P(x) = 0$$

and

$$\int_\Omega \Delta(x)^2\, \mathsf{d}P(x) = \int_\Omega (1 - f(x))^2\, \mathsf{d}P(x) = \mathrm{d}_{\chi^2}(Q\|P),$$

by definition of f. Then

$$\begin{aligned}\mathrm{d}_{\mathrm{TV}}(P^B, \widetilde{Q}) = \frac{1}{2B}\int_{\Omega^B}\left|\sum_{b=1}^B \Delta(x_b)\right| \mathsf{d}P^B(x) &\leq \frac{1}{2B}\left[\int_{\Omega^B}\left(\sum_{b=1}^B \Delta(x_b)\right)^2 \mathsf{d}P^B(x)\right]^{1/2} \\ &= \frac{1}{2B}\left[\sum_{b=1}^B\sum_{b'=1}^B \int_{\Omega^B} \Delta(x_b)\Delta(x_{b'})\, \mathsf{d}P^B(x)\right]^{1/2},\end{aligned}$$

by Cauchy–Schwarz. But for $b \neq b'$, we have

$$\int_{\Omega^B} \Delta(x_b)\Delta(x_{b'})\, \mathsf{d}P^B(x) = \left(\int_\Omega \Delta(x_b)\, \mathsf{d}P(x_b)\right)\cdot\left(\int_\Omega \Delta(x_{b'})\, \mathsf{d}P(x_{b'})\right) = 0,$$

while for the case $b = b'$ we have

$$\int_{\Omega^B} \Delta(x_b)^2\, \mathsf{d}P^B(x) = \mathrm{d}_{\chi^2}(Q\|P).$$

Therefore,

$$\mathrm{d}_{\mathrm{TV}}(P^B, \widetilde{Q}) \leq \frac{1}{2B}\left[\sum_{b=1}^B\sum_{b'=1}^B \mathrm{d}_{\chi^2}(Q\|P)\cdot \mathbb{1}_{b=b'}\right]^{1/2} = \frac{\sqrt{B\mathrm{d}_{\chi^2}(Q\|P)}}{2B},$$

which completes the proof.

D.2.5 Proof of Theorem D.1.2

Recall from the proof of Theorem D.1.1, the joint distribution P_1 on $(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)$, which we defined to have the joint density

$$\frac{1}{B}\sum_{b=1}^{B}\frac{\pi_0(\theta_b)}{\pi(\theta_b)}\cdot\bar{f}_\pi(x)\cdot\prod_{b=1}^{B}\frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)}.$$

Now define a distribution P_2, with joint density

$$\frac{\left(\frac{\bar{f}_\pi(x)}{\widehat{f}_\pi(x)}\right)^{B-1}}{\mathbb{E}_{\bar{f}_{\pi_0}}\left[\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1}\right]}\cdot\frac{1}{B}\sum_{b=1}^{B}\frac{\pi_0(\theta_b)}{\pi(\theta_b)}\cdot\bar{f}_\pi(x)\cdot\prod_{b=1}^{B}\frac{f_{\theta_b}(x)\pi(\theta_b)}{\bar{f}_\pi(x)},$$

which differs from P_1 only in the presence of the first term. We can observe that, under P_2, the conditional distribution of $X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$ is given by

$$X \mid (\widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widehat{g}_\pi(\,\cdot \mid \widehat{\theta}_{1:B}),$$

where we recall that this joint density is defined as $\widehat{g}_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\widehat{f}_\pi(x)^{B-1}}$. Therefore, if we consider the sampling distribution

$$\begin{cases}(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim P_2, \\ (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widehat{Q}_\pi^M(\,\cdot \mid X, \widehat{\theta}_{1:B}),\end{cases} \tag{D.9}$$

then by (D.3) on $\widehat{Q}_\pi^M(\,\cdot \mid X, \widehat{\theta}_{1:B})$, it holds that $(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)})$ are exchangeable conditional on $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$, and consequently are also exchangeable after marginalizing over $(\widehat{\theta}_1, \ldots, \widehat{\theta}_B)$.

Now compare this to running aCSS-B with the approximated marginal, which can be

characterized by the sampling distribution

$$\begin{cases} (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim P_0, \\ (\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}) \mid (X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B) \sim \widehat{Q}_\pi^M(\,\cdot \mid X, \widehat{\theta}_{1:B}). \end{cases}$$

Comparing this with (D.9), we can see that

$$\mathrm{d}_{\mathrm{exch}}\left(X, \widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(M)}\right) \leq \mathrm{d}_{\mathrm{TV}}(P_0, P_2),$$

similarly to the proof of Section D.2.3. Next, we have already shown in the proof of Section D.2.3 that $\mathrm{d}_{\mathrm{TV}}(P_0, P_1) \leq \epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}$, so from this point on we only need to bound $\mathrm{d}_{\mathrm{TV}}(P_1, P_2)$.

By definition of the joint density for P_2 as compared to P_1, we can compute the Radon–Nikodym derivative as

$$\frac{\mathsf{d}P_2(x, \theta_1, \ldots, \theta_B)}{\mathsf{d}P_1(x, \theta_1, \ldots, \theta_B)} = \frac{\left(\frac{\bar{f}_\pi(x)}{\widehat{f}_\pi(x)}\right)^{B-1}}{\mathbb{E}_{\bar{f}_{\pi_0}}\left[\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1}\right]}.$$

If the denominator is ≤ 1, then we have

$$\mathrm{d}_{\mathrm{TV}}(P_1, P_2) = \mathbb{E}_{P_1}\left[\left(1 - \frac{\mathsf{d}P_2(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)}{\mathsf{d}P_1(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)}\right)_+\right] \leq \mathbb{E}_{P_1}\left[\left(1 - \left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1}\right)_+\right],$$

while if instead the denominator is ≥ 1 then

$$\mathrm{d}_{\mathrm{TV}}(P_1, P_2) = \mathbb{E}_{P_1}\left[\left(\frac{\mathsf{d}P_2(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)}{\mathsf{d}P_1(X, \widehat{\theta}_1, \ldots, \widehat{\theta}_B)} - 1\right)_+\right] \leq \mathbb{E}_{P_1}\left[\left(\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right)_+\right].$$

In either case, then,

$$\mathrm{d_{TV}}(P_1, P_2) \leq \mathbb{E}_{P_1}\left[\left|\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right|\right] = \mathbb{E}_{\theta_0}\left[\left|\left(\frac{\bar{f}_\pi(X)}{\widehat{f}_\pi(X)}\right)^{B-1} - 1\right|\right],$$

where the last step holds since under P_1, the marginal distribution of X is given by f_{θ_0}. This verifies that $\mathrm{d_{TV}}(P_1, P_2) \leq \delta_B$, which completes the proof.

D.3 Sampling details

D.3.1 Logistic Regression

This section gives the implementation details for the logistic regression experiment presented in Section 5.4.2. (For the comparison to aCSS, the implementation of aCSS is exactly as in Barber and Janson (2022, Example 1, Section 4.5.2).)

Sampling from the posterior: The posterior distribution is challenging to sample from directly, so we instead sample the draws $\widehat{\theta}_b$ using Metropolis–Hastings sampling.

We now give details. The (unnormalized) log-density of the posterior distribution is given by

$$\Psi(\theta; x) = \log f_\theta(x) + \log \pi(\theta),$$

where

$$f_\theta(x) = \prod_{i=1}^{n} \left(\frac{e^{Z_i^\top \theta}}{1 + e^{Z_i^\top \theta}}\right)^{x_i} \cdot \left(\frac{1}{1 + e^{Z_i^\top \theta}}\right)^{1-x_i},$$
$$\pi(\theta) = \prod_{j=1}^{d} \phi\left(\theta_j; 0, 1\right).$$

We will use a Laplace approximation to the posterior distribution as the proposal distribution. Define

$$\widehat{\theta} = \arg\max_\theta \Psi(\theta).$$

Since this does not admit a closed-form solution, in practice we solve this optimization problem numerically via `optim` in `R`. By a second-order Taylor expansion to Ψ around $\widehat{\theta}$, we have

$$\Psi(\theta) \approx \Psi_{\text{Laplace}}(\theta) := \Psi(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^\top \mathbb{H}(\theta - \widehat{\theta}),$$

where $\mathbb{H}$ is the Hessian of the negative log posterior evaluated at $\widehat{\theta}$:

$$\mathbb{H} = -\nabla^2 \Psi(\theta)\big|_{\theta=\widehat{\theta}} = \sum_{i=1}^{n} \frac{e^{Z_i^\top \widehat{\theta}}}{\left(1 + e^{Z_i^\top \widehat{\theta}}\right)^2} Z_i \, Z_i^\top \; + \; \mathbb{I}_d.$$

Therefore, $\Psi(\theta)$ is approximately equal to $\Psi_{\text{Laplace}}(\theta)$, which is the (unnormalized) log-density of a multivariate Gaussian distribution, $\mathcal{N}(\widehat{\theta}, \mathbb{H}^{-1})$. To sample the posterior draws $\widehat{\theta}_b$, we therefore run the Metropolis–Hastings algorithm (Owen, 2013), with proposal distribution $\mathcal{N}(\widehat{\theta}, \mathbb{H}^{-1})$. We use a burn-in of 500 steps, and then extract samples $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$ at every tenth step.

Sampling the copies: Next we give details on sampling the copies $\widetilde{X}^{(m)}$. Recall that our goal is to sample the copies i.i.d. from the distribution with density defined as in (5.6). We will first compute an approximation to the marginal density (recall Section D.1.2). Using the same Laplace approximation as above, we replace $\bar{f}_\pi(x)$ with

$$\widehat{f}_\pi(x) \propto \int \exp(\Psi_{\text{Laplace}}(\theta)) \, \mathrm{d}\theta = \int \exp(\Psi(\widehat{\theta})) \exp\left(-\frac{1}{2}(\theta - \widehat{\theta})^\top \mathbb{H}(\theta - \widehat{\theta})\right) \mathrm{d}\theta$$
$$= \exp(\Psi(\widehat{\theta})) \sqrt{\frac{(2\pi)^d}{\det(\mathbb{H})}}.$$

(Note that this right-hand side is, implicitly, a function of x, since $\widehat{\theta}$ and $\mathbb{H}$ both depend on x.) This then leads to an approximate density, $\widehat{g}_\pi(x \mid \widehat{\theta}_1, \ldots, \widehat{\theta}_B)$ (as in (5.6)) from which to sample the copies. This density then serves as the target density in a Gibbs sampling algorithm (Owen, 2013). We use the permuted serial sampler (Besag and Clifford, 1989), as

follows:

1. *Initialization.* Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X,$$

2. *Iterations.* For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, n$,

$$\text{sample } \widetilde{X}_i^{(t)} \sim \widehat{g}_\pi(\cdot \mid x_{-i}, \widehat{\theta}_{1:B}) \text{ using } x_{-i} = (\widetilde{X}_{<i}^{(t)}, \widetilde{X}_{>i}^{(t-1)}),$$

where $\widehat{g}_\pi(\cdot \mid x_{-i}, \widehat{\theta}_{1:B})$ is the conditional density induced by $\widehat{g}_\pi(x \mid \widehat{\theta}_1, \ldots, \widehat{\theta}_B)$. Similarly, for $t = m_0 - 1, \ldots, 0$, for each $i = n, \ldots, 1$,

$$\text{sample } \widetilde{X}_i^{(t)} \sim \widehat{g}_\pi(\cdot \mid x_{-i}, \widehat{\theta}_{1:B}) \text{ using } x_{-i} = (\widetilde{X}_{<i}^{(t+1)}, \widetilde{X}_{>i}^{(t)}).$$

Since each $\widetilde{X}_i^{(t)}$ is Bernoulli, we can compute the probability explicitly as

$$\mathbb{P}(X_i^{(t)} = 0) = \frac{\widehat{g}_\pi(0 \mid x_{-i}, \widehat{\theta}_{1:B})}{\widehat{g}_\pi(0 \mid x_{-i}, \widehat{\theta}_{1:B}) + \widehat{g}_\pi(1 \mid x_{-i}, \widehat{\theta}_{1:B})}$$

which allows us to draw from the conditional distribution.

3. *Output.* Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.[1]

D.3.2 Mixture of Gaussians

This section gives the implementation details for the mixture of Gaussians experiment presented in Section 5.4.3. (For the comparison to reg-aCSS, the implementation of reg-aCSS is exactly as in Zhu and Barber (2023, Section 6.1).)

1. For the serial sampler to return exchangeable copies, as required in (D.1) or (D.3), formally we would also need to randomly permute the set of copies produced by this algorithm—but since the p-value is invariant to shuffling the copies, it is unnecessary for this last step.

Sampling from the posterior: Define $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. To sample from the posterior of θ we introduce the latent variable $Z \in \{1, 2\}^n$.

$$X_i \mid Z_i = j \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad \mathbb{P}(Z_i = j) = w_j,$$

for each $j = 1, 2$, where for convenience we write $w_2 = 1 - w_1$. This makes sampling from the posterior of θ straightforward. The full conditional distributions are given by the following:

$$\mathbb{P}(Z_i = j \mid X, w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \frac{w_j \, \phi(x_i; \mu_j, \sigma_j^2)}{\sum_{\ell=1}^2 w_\ell \, \phi(x_i; \mu_\ell, \sigma_\ell^2)}, \quad i = 1, \ldots, n, \tag{D.10}$$

$$w_1 \mid Z \;\sim\; \text{Beta}\left(2 + n_1, 2 + n_2\right), \quad n_j = \sum_{i=1}^n \mathbf{1}\{Z_i = j\}, \tag{D.11}$$

$$\mu_j \mid \sigma_j^2, Z, X \;\sim\; \mathcal{N}\left(\frac{\sum_{i:z_i=j} x_i}{1 + n_j}, \frac{\sigma_j^2}{1 + n_j}\right), \tag{D.12}$$

$$\sigma_j^2 \mid \mu_j, Z, X \;\sim\; \text{Inv-Gamma}\left(\frac{3}{2} + \frac{n_j}{2}, \; \frac{1}{2} + \frac{1}{2} \sum_{i:Z_i=j} (x_i - \mu_j)^2 + \frac{1}{2}\mu_j^2\right). \tag{D.13}$$

These allow for efficient sampling using the Gibbs sampler.

We begin by initializing the two–component mixture as follows:

$$w^{(0)} = \left(w_1^{(0)}, w_2^{(0)}\right) = \left(\tfrac{1}{2}, \; \tfrac{1}{2}\right).$$

The latent allocations are then initialized by splitting the data at its sample median:

$$Z_i^{(0)} = \begin{cases} 1, & X_i \le \text{median}(X), \\ 2, & X_i > \text{median}(X), \end{cases} \qquad i = 1, \ldots, n.$$

Conditional on these initial assignments, we set each component's parameters to the empirical

moments of its cluster:

$$\mu_j^{(0)} = \frac{1}{n_j^{(0)}} \sum_{i:Z_i^{(0)}=j} X_i, \qquad \sigma_j^{2\,(0)} = \frac{1}{n_j^{(0)}} \sum_{i:Z_i^{(0)}=j} \left(X_i - \mu_j^{(0)}\right)^2, \quad n_j^{(0)} = \sum_{i=1}^{n} \mathbf{1}\{Z_i^{(0)} = j\}.$$

Thereafter, for each iteration t, we perform the following Gibbs-sampling updates:

1. Independently for each i, sample each $Z_i^{(t)} \mid X, w^{(t-1)}, \mu^{(t-1)}, \sigma^{2,(t-1)}$ according to (D.10).
2. Sample $w_1^{(t)} \;\big|\; Z^{(t)}$ according to (D.11), and set $w_2^{(t)} = 1 - w_1^{(t)}$.
3. For each $j = 1, 2$, draw $\mu_j^{(t)} \mid \sigma_j^{2,(t-1)}, Z^{(t)}, X$ and $\sigma_j^{2,(t)} \mid \mu_j^{(t)}, Z^{(t)}, X$ according to (D.12) and (D.13), respectively.

We discard the first 500 draws and extract the posterior samples $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$ at every tenth step.

Sampling the copies: To sample the copies, we again use an approximation of the marginal $\bar{f}_\pi(x)$ and sample from

$$\widehat{g}_\pi(x \mid \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^{B} f_{\widehat{\theta}_b}(x)}{\widehat{f}_\pi(x)^{\,B-1}}.$$

In order to approximate the marginal, we note that

$$\bar{f}_\pi(x) = \int f_\theta(x)\, \pi(\theta)\, d\theta,$$

and define the log of the unnormalized posterior as

$$\Psi(\theta; x) = \log f_\theta(x) + \log \pi(\theta).$$

For our choice of priors it takes the following form:

$$\Psi(\theta; x) = \sum_{i=1}^{n} \log\left(w_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) + (1 - w_1) \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right)$$

$$+ \log w_1 + \log(1 - w_1) + \sum_{j=1}^{2} \left[\log(0.5) - 2\log(\sigma_j^2) - \frac{0.5}{\sigma_j^2} - \tfrac{1}{2}\log(2\pi\sigma_j^2) - \frac{\mu_j^2}{2\sigma_j^2} \right].$$

In our setting since we have two components in the null mixture distribution, the posterior is invariant to the swap of the component labels and has *two* modes. A single Laplace approximation is therefore inaccurate. We approximate $\bar{f}_\pi(x)$ by a mixture of two normal distributions where the means of the two components are identified as the two local maxima of $\Psi(\theta; x)$; say $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$. Following the same idea as in the usual construction of a Laplace approximation, the variance of these individual Gaussian components would be the inverse of the determinant of the Hessian of $\Psi(\theta; x)$ at $\theta = \widetilde{\theta}_1$ and $\widetilde{\theta}_2$, respectively, where the Hessians at these two points are defined as

$$\mathbb{H}_1 = -\nabla^2 \Psi(\theta)\big|_{\theta=\widetilde{\theta}_1} \qquad \text{and} \qquad \mathbb{H}_2 = -\nabla^2 \Psi(\theta)\big|_{\theta=\widetilde{\theta}_2},$$

both computable in closed form. Our two–mode Laplace approximation to the marginal is then

$$\widehat{f}_\pi(x) \;=\; e^{L_1} + e^{L_2},$$

where

$$L_s \;=\; \Psi(\widetilde{\theta}_s; x) \;-\; \tfrac{1}{2}\,\mathrm{logdet}\,\mathbb{H}_s \;+\; \tfrac{5}{2}\log(2\pi), \qquad s = 1, 2$$

is the contribution to the integral from the s^{th} component of the normal mixture approximation to the (unnormalized) posterior. In order to obtain $\widehat{\theta}_s$, we use `optim` in `R` as follows:

1. *Reparametrization.* To convert the constrained optimization into an unconstrained one,

define the following reparameterization for $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_1^2)$:

$$\vartheta(\theta) = (z_1, \mu_1, s_1, \mu_2, s_2) \text{ where } z_1 = \log\left(\frac{w_1}{1-w_1}\right),\ s_1 = \log(\sigma_1^2),\ \text{and}\ s_2 = \log(\sigma_2^2).$$

2. *Initialization.* Apply k-means with $k = 2$ on the observed data and obtain two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ such that $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, n\}$ with the cluster centers c_1 and c_2 respectively. Define $\tau_1^2 = \frac{1}{|\mathcal{C}_1|-1}\sum_{i\in\mathcal{C}_1}(x_i - c_1)^2$ and $\tau_2^2 = \frac{1}{|\mathcal{C}_2|-1}\sum_{i\in\mathcal{C}_2}(x_i - c_2)^2$. Define the following two initializations:

   - $\vartheta_1^{(0)} = \vartheta\left(\frac{|\mathcal{C}_1|}{n}, c_1, \tau_1^2, c_2, \tau_2^2\right)$,
   - $\vartheta_2^{(0)} = \vartheta\left(\frac{|\mathcal{C}_2|}{n}, c_2, \tau_2^2, c_1, \tau_1^2\right)$.

3. *Optimization.* Starting from the two initial values $\vartheta_1^{(0)}$ and $\vartheta_2^{(0)}$, maximize Ψ in the unconstrained parameter space $\mathbb{R}^5$ using `optim` with the `BFGS` algorithm to obtain $\widehat{\vartheta}_1$ and $\widehat{\vartheta}_2$, respectively. Map these back to the original parameters and get $\widetilde{\theta}_s = \vartheta^{-1}(\widehat{\vartheta}_s)$ for $s \in \{1, 2\}$.

Thus $\widehat{f}_\pi(x)$ gives us an approximation to the marginal. We would like to sample the X_i's from the $\widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_\theta(x_i)}{\widehat{f}_\pi(x)}$. To do that we will be approximating $\widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B})$ using a two component normal distribution which will be used as a proposal in a Metropolis–Hastings algorithm. The rest of this section describes how to obtain this mixture of normal proposal from $\widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B})$.

Define the negative log-density as $\zeta(x_i) := -\log \widehat{g}_\pi(x_i \mid x_{-i}, \widehat{\theta}_{1:B})$. We consider a grid $\{\xi_j\}_{j=1}^K$ with $K = 20$, spanning $[a, b]$ where $a = \min_i X_i, b = \max_i X_i$ and evaluate $\zeta(\xi_j)$ for all $j = 1, \ldots, 20$. We fit a continuous piecewise–quadratic surrogate of the form

$$Q(\xi; c, \beta) \ = \ \beta_0 + \mathbb{1}\{\xi \le c\}\Big(\beta_{1L}(\xi - c) + \beta_{2L}(\xi - c)^2\Big) + \mathbb{1}\{\xi > c\}\Big(\beta_{1R}(\xi - c) + \beta_{2R}(\xi - c)^2\Big),$$

where $\beta = (\beta_0, \beta_{1L}, \beta_{2L}, \beta_{1R}, \beta_{2R})$, to these ζ evaluations by minimizing the objective

function

$$\mathrm{SSE}(c) := \min_{\beta} \sum_{j=1}^{K} \{\zeta(\xi_j) - Q(\xi_j; c, \beta)\}^2.$$

For each candidate value of c, the optimum $\widehat{\beta}$ can be obtained by a least squares algorithm and the optimum value c (say $\widehat{c}$) is chosen via a grid search on 400 different values. This yields coefficients $(\widehat{\beta}_0, \widehat{\beta}_{1L}, \widehat{\beta}_{2L}, \widehat{\beta}_{1R}, \widehat{\beta}_{2R})$ at breakpoint $\widehat{c}$.

For each arm $s \in \{L, R\}$, we map the quadratic to a Gaussian as follows:

$$v_s = \max\left\{\epsilon, \frac{1}{2\widehat{\beta}_{2s}}\right\}, \qquad m_s = \widehat{c} - \widehat{\beta}_{1s} v_s$$

with $\epsilon = 10^{-8}$. To obtain the mixture weights, we match the magnitude of Q at the two means m_L and m_R with that of a two-component mixture of normal distribution with means m_L and m_R and variances v_L and v_R by solving the following equation for p and q:

$$\begin{bmatrix} \phi(m_L; m_L, v_L) & \phi(m_L; m_R, v_R) \\ \phi(m_R; m_L, v_L) & \phi(m_R; m_R, v_R) \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} \exp\left(-Q\left(m_L; \widehat{c}, \widehat{\beta}\right)\right) \\ \exp\left(-Q\left(m_R; \widehat{c}, \widehat{\beta}\right)\right) \end{bmatrix}.$$

where $\phi(\,\cdot\,; m, v)$ denotes the pdf of a Gaussian distribution with mean m and variance v. Finally, we set the weight (of the L component) as

$$w = \frac{p}{p+q}.$$

So, the proposal x_{prop} is drawn from the two–component Gaussian mixture density

$$w\,\phi(\cdot; \mu_L, v_L) + (1-w)\,\phi(\cdot; \mu_R, v_R).$$

This yields the following Metropolis–Hastings algorithm with the permuted serial sampler:

1. *Initialization.* Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X.$$

2. *Iterations.* For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, n$ draw x_{prop} from the density $w\,\phi(\cdot; \mu_L, v_L) + (1-w)\,\phi(\cdot; \mu_R, v_R)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{1,\ \frac{g_\pi\left(x_{\text{prop}} \mid x_{-i}, \widehat{\theta}_{1:B}\right)}{\widehat{g}_\pi\left(X_i^{(t-1)} \mid x_{-i}, \widehat{\theta}_{1:B}\right)} \cdot \frac{w\,\phi(X_i^{(t-1)}; \mu_L, v_L) + (1-w)\,\phi(X_i^{(t-1)}; \mu_R, v_R)}{w\,\phi(x_{\text{prop}}; \mu_L, v_L) + (1-w)\,\phi(x_{\text{prop}}; \mu_R, v_R)}\right\}$$

and $x_{-i} = \left(\widetilde{X}_{<i}^{(t)},\ \widetilde{X}_{>i}^{(t-1)}\right)$. Similarly, for $t = m_0 - 1, \ldots, 0$, for each $i = n, \ldots, 1$, we draw x_{prop} from the density $w\,\phi(\cdot; \mu_L, v_L) + (1-w)\,\phi(\cdot; \mu_R, v_R)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

as before with the only difference that $x_{-i} = \left(\widetilde{X}_{<i}^{(t+1)},\ \widetilde{X}_{>i}^{(t)}\right)$.

3. *Output.* Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.

D.3.3 Rank-1 matrix

This section gives the implementation details for the rank-1 matrix experiment presented in Section 5.4.4.

Sampling from the posterior: Recall that our prior is the multivariate standard normal distribution, $U, V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\vec{0}_n, \mathbb{I}_n)$. While the posterior distribution of (U, V)—that is, the distribution of $(U, V) \mid X$—is challenging to work with, it is straightforward to compute the conditionals of the posterior distribution: we have

$$U \mid X, V \sim \mathcal{N}\left(\frac{4XV}{4\|V\|_2^2 + 1},\ \frac{1}{4\|V\|_2^2 + 1}\,\mathbb{I}_n\right) \tag{D.14}$$

and

$$V \mid X, U \sim \mathcal{N}\left(\frac{4X^\top U}{4\|U\|_2^2 + 1}, \frac{1}{4\|U\|_2^2 + 1}\mathbb{I}_n\right). \tag{D.15}$$

Thus, we can easily sample from the posterior using Gibbs sampling. We begin by initializing U, V via a rank-1 approximation to X: writing $u, v \in \mathbb{R}^n$ as the leading left and right singular vectors of X, respectively, we initialize with

$$U = \sqrt{n} \cdot u, \quad V = \sqrt{n} \cdot v,$$

and iterate the Gibbs sampler,

$$\begin{cases} \text{Sample } U \mid V, X \text{ according to distribution (D.14)}, \\ \text{Sample } V \mid U, X \text{ according to distribution (D.15)}. \end{cases}$$

We use a burn-in of 500 steps, and then extract samples $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$ at every tenth step.

Sampling the copies: As in earlier examples, we will need to use a Laplace approximation for the marginal distribution of X. However, in this specific setting, we will need to proceed a bit differently—this is because the negative log likelihood of $(U, V) \mid X$ is a non-convex function, and indeed has issues of identifiability, as, e.g., the likelihood takes equal values at (u, v) and $(-u, -v)$. Instead, we will first marginalize over U exactly, and then use a Laplace approximation for marginalizing over V.

First, we calculate the distribution of $X \mid V$, i.e., we marginalize over U: using X_j to denote the jth row of X, we have

$$X_j \mid V \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, 0.25\,\mathbb{I}_n + VV^\top\right).$$

The distribution of $X \mid V = v$ therefore has conditional density

$$\begin{aligned}
f_v(x) &= \frac{1}{(2\pi)^{n^2/2}\,\det(0.25\,\mathbb{I}_n + vv^\top)^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^n x_j^\top (0.25\,\mathbb{I}_n + vv^\top)^{-1} x_j} \\
&= \frac{1}{(2\pi)^{n^2/2}\left[(0.25)^n(1+4\|v\|_2^2)\right]^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^n x_j^\top \left(4\mathbb{I}_n - \frac{16}{1+4\|v\|_2^2} vv^\top\right) x_j} \\
&= \frac{4^{n^2/2}}{(2\pi)^{n^2/2}\,(1+4\|v\|_2^2)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2 + \frac{8}{1+4\|v\|_2^2}\|xv\|_2^2}.
\end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, and x_j now denotes the j^{th} row of $x \in \mathbb{R}^{n\times n}$. Writing $X = A\,\mathrm{diag}\{d\}\,B^\top$ as the singular value decomposition of X where $A, B \in \mathbb{R}^{n\times n}$, we can rewrite the marginal as follows:

$$\begin{aligned}
\bar{f}_\pi(x) &= \mathbb{E}_{V\sim\mathcal{N}(0,\mathbb{I}_n)}\left[f_V(x)\right] \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathbb{I}_n)}\left[\frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4\|V\|_2^2\right)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2 + 8\frac{\left\|A\cdot\mathrm{diag}\{d\}\cdot B^\top V\right\|_2^2}{1+4\|V\|_2^2}}\right] \\
&= \mathbb{E}_{V\sim\mathcal{N}(0,\mathbb{I}_n)}\left[\frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4\|V\|_2^2\right)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2 + 8\frac{\|\mathrm{diag}\{d\}V\|_2^2}{1+4\|V\|_2^2}}\right] \\
&= \mathbb{E}_{W_1,\ldots,W_n\overset{\text{iid}}{\sim}\chi_1^2}\left[\frac{4^{n^2/2}}{(2\pi)^{n^2/2}\left(1+4\sum_{i=1}^n W_i\right)^{n/2}} e^{-2\|x\|_{\mathrm{F}}^2 + 8\frac{\sum_{i=1}^n W_i d_i^2}{1+4\sum_{i=1}^n W_i}}\right],
\end{aligned} \tag{D.16}$$

where the second-to-last step holds due to rotational invariance of the standard normal

distribution and the ℓ_2 norm. In the last step, we reparametrize the integrand in (D.16) in terms of W instead of V (i.e., $W_i \sim \chi^2_1$ replaces V_i^2). This reparametrization addresses the rotational and sign invariances of f in terms of V, thereby avoiding complications in the integration arising from multimodality. Before proceeding with the Laplace approximation, we again reparametrize $t_i = \log W_i$; otherwise the χ^2_1 prior for W_i (and hence the unnormalized posterior integrand) is unbounded at 0, violating the regularity conditions for a Laplace approximation. Since $W_i \sim \chi^2_1$, denoting by $\pi_{W_i}(w_i)$ the χ^2_1 density at w_i, the change of variables gives

$$\pi(t_i) = \pi_{W_i}(e^{t_i}) \left| \frac{d}{dt_i} e^{t_i} \right| = \frac{1}{\sqrt{2\pi}}\, e^{t_i/2}\, e^{-e^{t_i}/2}, \qquad t_i \in \mathbb{R},\ i \in \{1, \ldots, n\},$$

hence

$$\log \pi(t) = \sum_{i=1}^{n} \left(-\tfrac{1}{2} \log(2\pi) + \tfrac{1}{2} t_i - \tfrac{1}{2} e^{t_i} \right), \quad t = (t_1, \ldots, t_n)^\top.$$

If we denote

$$S_1(t) := \sum_{i=1}^{n} e^{t_i}, \qquad S_d(t) := \sum_{i=1}^{n} d_i^2 e^{t_i},$$

from (D.16), the conditional density $f_t(x)$ is

$$f_t(x) = \frac{4^{n^2/2}}{(2\pi)^{n^2/2} \left(1 + 4S_1(t)\right)^{n/2}} \exp\left(-2\|x\|_F^2 + \frac{8\, S_d(t)}{1 + 4S_1(t)} \right).$$

Define

$$\begin{aligned} \Psi(t; x) &:= \log f_t(x) + \log \pi(t) \\ &= -\frac{n}{2} \log\Big(1 + 4S_1(t)\Big) + \frac{8\, S_d(t)}{1 + 4S_1(t)} + \sum_{i=1}^{n} \Big(\tfrac{1}{2} t_i - \tfrac{1}{2} e^{t_i} \Big) + \text{const}(x) \end{aligned}$$

where $\text{const}(x) = \frac{n^2}{2} \log 4 - \frac{n^2}{2} \log(2\pi) - 2\|x\|_F^2$ does not depend on t. Introduce the following

notations:

$$c(t) = 1 + 4S_1(t), \qquad N_i(t) = c(t)\, d_i^2 - 4S_d(t).$$

The gradient is then given by

$$\nabla_t \Psi(t;x) = \left(\frac{\partial \Psi}{\partial t_1}, \ldots, \frac{\partial \Psi}{\partial t_n}\right)^\top, \quad \text{where} \quad \frac{\partial \Psi}{\partial t_j} = \frac{1}{2} - \frac{1}{2}e^{t_j} - \frac{2n}{c(t)}e^{t_j} + \frac{8e^{t_j}}{c(t)^2}\Big(c(t)\, d_j^2 - 4S_d(t)\Big).$$

The Hessian of the negative log–posterior is $H(t;x) := -\nabla_t^2 \Psi(t;x)$ with

$$\begin{aligned} H_{jk}(t;x) &= -\Big[\frac{8n}{c(t)^2}e^{t_j}e^{t_k} + \frac{32}{c(t)^2}e^{t_j}e^{t_k}\,(d_j^2 - d_k^2) - \frac{64}{c(t)^3}e^{t_j}e^{t_k}\,N_j(t)\Big], \qquad j \neq k, \\ H_{jj}(t;x) &= \frac{1}{2}e^{t_j} + 2n\left(\frac{e^{t_j}}{c(t)} - \frac{4e^{2t_j}}{c(t)^2}\right) - \; 8e^{t_j}\left(\frac{N_j(t)}{c(t)^2} - \frac{8e^{t_j}N_j(t)}{c(t)^3}\right), \qquad j = k. \end{aligned}$$

We find $\hat{t} = \arg\max_t \Psi(t;x)$ by using `optim` in `R` and denote $\mathbb{H} := H(\hat{t};x)$. The Laplace approximation to the marginal is

$$\widehat{f}_\pi(x) = (2\pi)^{n/2}\, \det(\mathbb{H})^{-1/2}\, \exp\Big\{\Psi(\hat{t};x)\Big\}.$$

Using $\widehat{f}_\pi(x)$ we define $\widehat{g}_\pi\left(\,\cdot \mid \widehat{\theta}_{1:B}\right)$ as in Section D.1.2 which serves as the target density. Consider the density induced by $\widehat{g}_\pi(\cdot \mid \widehat{\theta}_{1:B})$ on each X_{ij}, i.e.

$$\widehat{g}_\pi(x_{ij} \mid x_{-ij}, \widehat{\theta}_{1:B}) \propto \frac{\prod_{b=1}^B f_{\widehat{\theta}_b}(x_{ij})}{\widehat{f}_\pi(x)^{B-1}},$$

where x_{-ij} denotes all entries of x except the (i,j)-th position. In order to sample each X_{ij} from that density, we use the Metropolis–Hastings algorithm. For the proposal density, we perform a Laplace approximation on this density. Consider $\zeta(x_{ij}) = \log \widehat{g}_\pi\left(x_{ij} \mid x_{-ij}, \widehat{\theta}_{1:B}\right)$, let

$$x_{ij}^\star = \arg\max_{x_{ij}} \zeta(x_{ij}),$$

and denote the Hessian at $x_{ij}^\star$ by $\zeta''(x_{ij}^\star)$. (Note that $x_{ij}^\star$ and $\zeta''(x_{ij}^\star)$ are implicitly functions of x_{-ij}). The proposal distribution is given by $\mathcal{N}\left(x_{ij}^\star, -\frac{1}{\zeta''(x_{ij}^\star)}\right)$. This optimization is performed numerically via `optim` in `R` and the resulting Metropolis–Hastings algorithm has average acceptance probability very close to 1. Putting all of these steps together, we obtain the following permuted serial sampler to sample $\widetilde{X}^{(m)}$'s:

1. *Initialization.* Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X,$$

2. *Iterations.* For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, m$ and $j = 1, \ldots, n$,

   draw $x_{\text{prop}} \sim \mathcal{N}\left(x_{ij}^\star,\ -\frac{1}{\zeta''(x_{ij}^\star)}\right)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_{ij}^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_{ij}^{(t-1)}, & \text{otherwise,} \end{cases}$$

   where

$$\alpha = \min\left\{1,\ \frac{\widehat{g}_\pi\left(x_{\text{prop}} \mid x_{-ij}, \widehat{\theta}_{1:B}\right)}{\widehat{g}_\pi\left(X_{ij}^{(t-1)} \mid x_{-ij}, \widehat{\theta}_{1:B}\right)} \cdot \frac{\phi\left((X_{ij}^{(t-1)} - x_{ij}^\star)\sqrt{-\zeta''(x_{ij}^\star)}\right)}{\phi\left((x_{\text{prop}} - x_{ij}^\star)\sqrt{-\zeta''(x_{ij}^\star)}\right)}\right\},$$
$$x_{-ij} = \left(\widetilde{X}_{<i,1:n}^{(t)},\ \widetilde{X}_{i,<j}^{(t)},\ \widetilde{X}_{i,>j}^{(t-1)},\ \widetilde{X}_{>i,1:n}^{(t-1)}\right).$$

   Similarly, for $t = m_0 - 1, \ldots, 0$, for each $j = n, \ldots, 1$ and $i = m, \ldots, 1$, we draw $x_{\text{prop}} \sim \mathcal{N}\left(x_{ij}^\star,\ -\frac{1}{\zeta''(x_{ij}^\star)}\right)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_{ij}^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_{ij}^{(t-1)}, & \text{otherwise,} \end{cases}$$

as before with the only difference that $x_{-ij} = \left(\widetilde{X}^{(t)}_{<i,1:n},\ \widetilde{X}^{(t)}_{i,<j},\ \widetilde{X}^{(t+1)}_{i,>j},\ \widetilde{X}^{(t+1)}_{>i,1:n}\right)$.

3. *Output.* Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.

D.3.4 Group-sparse regression

This section gives the implementation details for the group-sparse regression experiment presented in Section 5.4.5.

Sampling from the posterior. We restate the priors defined in Section 5.4.5 here with the general parameters as follows:

$$\begin{aligned} g^\star &\sim \text{Unif}(\{1,\ldots,\text{G}\}), \\ \beta_{I_{g^\star}} &\sim \mathcal{N}(\vec{0}_{|I_{g^\star}|}, I_{|I_{g^\star}|}), \\ \beta_{I_g} &= \vec{0}_{|I_g|} \ \forall\ g \neq g^\star. \end{aligned}$$

Under this prior, the posterior distribution of β can be derived with the following hierarchical structure: defining

$$D_g(X) = |A_g|^{-\frac{1}{2}} \exp\left(\frac{1}{2} b_g^\top A_g^{-1} b_g - \frac{1}{2} X^\top X\right)$$

for each $g \in [G]$, where

$$A_g = Z_{I_g}^\top Z_{I_g} + I_{d_g}, \quad b_g = Z_{I_g}^\top X,$$

we first sample the active group $g^\star$ as

$$\mathbb{P}(g^\star = g \mid X) \propto D_g(X),$$

then sample $\beta \mid g^\star, X$ as

$$\beta_{I_{g^\star}} \mid g^\star, X \sim \mathcal{N}\left(A_{g^\star}^{-1} b_{g^\star}, A_{g^\star}^{-1}\right),$$

$$\beta_{I_g} \mid g^\star, X = 0 \;\forall\; g \neq g^\star.$$

Sampling the copies: In this case, we can evaluate the marginal of X exactly. Up to normalizing constants,

$$\bar{f}_\pi(X) = \frac{1}{G}\sum_{g=1}^{G} \int_{\mathbb{R}^{d_g}} \exp\left(-\frac{1}{2}||X - Z\beta||^2 - \frac{1}{2}||\beta_{I_g}||^2\right) \mathtt{d}\beta_{I_g} = \frac{1}{G}\sum_{g=1}^{G} D_g(X).$$

In order to sample from the final sampling density $g_\pi(\,\cdot \mid \widehat{\theta}_{1:B})$ as in (5.6), we use a Laplace approximation as the proposal in a Metropolis–Hastings sampler. Our goal is to approximate the conditional sampling density of X_i, and we will update the X_i's one at a time. Since this conditional density is proportional to $g_\pi(X \mid \widehat{\theta}_{1:B})$, we define

$$\zeta(x_i) = -\frac{1}{2}\sum_{b=1}^{B}(x_i - Z_i^\top \widehat{\beta}_b)^2 - (B-1)\log\left(\frac{1}{G}\sum_{g=1}^{G} D_g(x)\right).$$

Taking the derivatives, we can find the maximum and also the Hessian which can then be used to perform a Laplace approximation.

$$\begin{aligned}
\zeta'(x_i) &= -\sum_{b=1}^{B}(x_i - Z_i^\top \widehat{\beta}_b) - (B-1)\frac{\sum_{g=1}^{G}\frac{\partial}{\partial x_i}D_g(x)}{\sum_{g=1}^{G} D_g(x)}, \\
\zeta''(x_i) &= \frac{B-1}{\left(\sum_{g=1}^{G} D_g(x)\right)^2}\left[\left(\sum_{g=1}^{G} D_g(x)\right)\left(\sum_{g=1}^{G}\frac{\partial^2}{\partial x_i^2}D_g(x)\right) - \left(\sum_{g=1}^{G}\frac{\partial}{\partial x_i}D_g(x)\right)^2\right],
\end{aligned}$$

where Z_i is the i^{th} row of the covariate matrix Z and

$$\begin{aligned}
\frac{\partial}{\partial x_i}D_g(x) &= \left[D_g(x)\left(Z_{I_g}A_g^{-1}b_g - x\right)\right]_i, \\
\frac{\partial^2}{\partial x_i^2}D_g(x) &= \left[D_g(x)\left(\frac{1}{\sigma^2}Z_{I_g}A_g^{-1}Z_{I_g}^\top - \mathbb{I}_n\right) + \left(Z_{I_g}A_g^{-1}b_g - x\right)D_g'(x)^\top\right]_{ii}.
\end{aligned}$$

The optimization is carried out numerically by `optim` in `R` to find the maximum $x_i^\star$ and

then the approximated Gaussian density $\mathcal{N}\left(x_i^\star, -\frac{1}{\zeta''(x_i^\star)}\right)$ is used as the proposal for our Metropolis–Hastings sampling. The acceptance probability in the Metropolis–Hastings stays close to 1. We use the permuted serial sampler, as follows:

1. *Initialization.* Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X,$$

2. *Iterations.* For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, n$ draw $x_{\text{prop}} \sim \mathcal{N}\left(x_i^\star, -\frac{1}{\zeta''(x_i^\star)}\right)$, $u \sim \text{Unif}(0,1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{1,\ \frac{g_\pi\left(x_{\text{prop}} \mid x_{-i}, \widehat{\theta}_{1:B}\right)}{g_\pi\left(X_i^{(t-1)} \mid x_{-i}, \widehat{\theta}_{1:B}\right)} \cdot \frac{\phi\left((X_i^{(t-1)} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\right)}{\phi\left((x_{\text{prop}} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\right)}\right\}$$

and $x_{-i} = \left(\widetilde{X}_{<i}^{(t)},\ \widetilde{X}_{>i}^{(t-1)}\right)$. Similarly, for $t = m_0 - 1, \ldots, 0$, for each $i = n, \ldots, 1$, we draw $x_{\text{prop}} \sim \mathcal{N}\left(x_i^\star,\ -\frac{1}{\zeta''(x_i^\star)}\right)$, $u \sim \text{Unif}(0,1)$ and set $\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$ as before with the only difference that $x_{-i} = \left(\widetilde{X}_{<i}^{(t+1)},\ \widetilde{X}_{>i}^{(t)}\right)$.

3. *Output.* Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.

D.3.5 Linear spline regression

This section gives the implementation details for the linear spline regression experiment presented in Section 5.4.6.

Sampling from the posterior: Following (5.16), with $k = 1$ (i.e., one knot) the linear spline model can be represented as

$$X = \gamma_0 + \gamma_1 Z + \gamma_2 b_1(Z) + \epsilon$$

which can be further written as:

$$X = h_t(Z)\gamma + \epsilon,$$

where

$$h_t(Z) = \left(\vec{1}_n, Z, b_1(Z)\right) \in \mathbb{R}^{n\times 3} \quad \text{and} \quad \gamma = (\gamma_0, \gamma_1, \gamma_2)^\top.$$

Let $Z_{(i)}$ denote the i-th order statistic of $Z_1, \ldots, Z_n$ and for notational convenience we shall denote $Z_{(0)} = -\infty$ and $Z_{(n+1)} = \infty$. We shall use $X_{(i)}$ to denote the response corresponding to covariate $Z_{(i)}$. Recall that we are using the prior distribution

$$\begin{aligned} \gamma_0, \gamma_1, \gamma_2 &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ t_1 &\sim \mathcal{N}(0,1). \end{aligned}$$

To draw the posterior samples, we use a Gibbs sampling algorithm where the conditional distributions are as follows. First, the conditional distribution of γ is given by

$$(\gamma \mid t, Z, X) \sim \mathcal{N}\left(\mu_\gamma, V_\gamma\right), \tag{D.17}$$
$$\text{where } V_\gamma = \left(4h_t(Z)^\top h_t(Z) + I_3\right)^{-1} \quad \text{and} \quad \mu_\gamma = V_\gamma\left(4h_t(Z)^\top X\right).$$

Next, the conditional distribution of t_1 has density

$$P(t_1 \mid \gamma, X, Z) \propto \phi\left(\frac{\|X - h_t(Z)\gamma\|}{\sigma}\right)\phi\left(\frac{t_1}{\tau_2}\right)$$

$$\propto \exp\left(-2\sum_{i=1}^{n}\left(X_i-\gamma_0-\gamma_1 Z_i-\gamma_2 b_1(Z_i)\right)^2\right)\exp\left(-\frac{t_1^2}{2}\right).$$

Recalling that $b_1(z)=(z-t_1)_+$, we note that in this distribution, for any i, when $t_1 \in (Z_{(i)}, Z_{(i+1)})$, the density $P(t_1 \mid \gamma, X, Z)$ is

$$\propto \exp\left(-2\sum_{i'=i+1}^{n}\left(X_{(i')}-\gamma_0-\gamma_1 Z_{(i')}-\gamma_2(Z_{(i')}-t_1)\right)^2\right)\exp\left(-\frac{t_1^2}{2}\right),$$

and is therefore proportional to the following Gaussian distribution:

$$P(t_1 \mid t_1 \in (Z_{(i)}, Z_{(i+1)}), \gamma, X, Z) \propto \mathcal{N}(\mu_{t,i}, \sigma_{t,i}^2), \tag{D.18}$$

$$\text{where,} \qquad \mu_{t,i} = \left(4(n-i)\,\gamma_2^2+1\right)^{-1}\left(4\gamma_2\sum_{i'=i+1}^{n}\left(X_{(i')}-\gamma_0-\gamma_1 Z_{(i')}-\gamma_2\, Z_{(i')}\right)\right),$$
$$\sigma_{t,i}^2 = \left(4(n-i)\,\gamma_2^2+1\right)^{-1}.$$

The probability that $t_1 \in (Z_{(i)}, Z_{(i+1)})$ can be obtained by evaluating the following integral

$$\begin{aligned}
&\mathbb{P}(t_1 \in (Z_{(i)}, Z_{(i+1)}) \mid \gamma, X, Z) \\
&\quad = \int_{t_1=Z_{(i)}}^{Z_{(i+1)}} e^{-2\sum_{i'=1}^{i}\left(X_{(i')}-\gamma_0-\gamma_1 Z_{(i')}\right)^2 - 2\sum_{i'=i+1}^{n}\left(X_{(i')}-\gamma_0-\gamma_1 Z_{(i')}-\gamma_2(Z_{(i')}-t_1)\right)^2 - \frac{1}{2}t_1^2}\,\mathsf{d}t_1 \\
&\quad =: w_i.
\end{aligned} \tag{D.19}$$

This integration can be solved by expressing the exponent in the integrand as a quadratic in t_1 and then using the Gaussian CDF. This enables us to sample from the posterior of t_1 by first sampling the interval in which t_1 lies using the above probability weights and subsequently drawing t_1 from a truncated normal distribution. We will represent this sampling distribution

concisely as follows:

$$t_1 \mid Z, X, \gamma \sim \sum_{i=0}^{n} w_i \, \mathrm{TN}(\mu_{t,i}, \sigma^2_{t,i}, Z_{(i)}, Z_{(i+1)}), \tag{D.20}$$

where $\mathrm{TN}(\mu, \sigma^2, a, b)$ is the truncated normal distribution—that is, the distribution $\mathcal{N}(\mu, \sigma^2)$ truncated to the interval $[a, b]$. This results in the following Gibbs sampling algorithm to sample from the posterior distribution of the parameters. We use the `R` package `segmented` (Muggeo, 2023) to find the position of the knots and initialize $t^{(0)}$ at that point. We initialize $\gamma^{(0)}$ as the coefficient vector from the regression of X on $h_{t^{(0)}}$. For $b = 1, \ldots, B$,

1. Generate $\gamma^{(b)}$ from $\gamma \mid t_1^{(b-1)}, Z, X$ as in Equation (D.17).
2. Generate $t_1^{(b)}$ from $t_1 \mid Z, X, \gamma^{(b)}$ as in Equation (D.20).

After the burn-in of $B_0 = 500$, we extract $B = 25$ posterior samples $\{t^{(b)}, \gamma^{(b)}\}$ at every tenth step.

Sampling the copies: Our first step is to approximate the marginal $\bar{f}_\pi(x)$. In this case, we first marginalize out γ from the distribution of $X \mid t, \gamma$ as follows:

$$\begin{aligned} X \mid t_1, \gamma &\sim \mathcal{N}\left(h_t(Z)\gamma, 0.25\mathbb{I}_n\right) \\ \implies X \mid t_1 &\sim \mathcal{N}\left(\vec{0}_n, h_t(Z)h_t(Z)^\top + 0.25\mathbb{I}_n\right). \end{aligned}$$

Now consider the joint density of (X, t_1). Since our prior on t_1 is $\mathcal{N}(0, 1)$, the joint density is given by

$$\Psi(x, t_1) = \frac{1}{\sqrt{(2\pi)^n |h_t(Z)h_t(Z)^\top + 0.25\mathbb{I}_n|}} e^{-\frac{1}{2}x^\top (h_t(Z)h_t(Z)^\top + 0.25\mathbb{I}_n)^{-1}x} \cdot \frac{1}{\sqrt{2\pi}} e^{-t_1^2/2}$$

and the marginal is then given by $\bar{f}_\pi(x) = \int \Psi(x, t_1) \, \mathsf{d}t_1$. Note that evaluating $\Psi(x, t_1)$ at a single value t_1 is simple, but integrating this quantity is challenging. We will therefore

take an approximation: first re-define $Z_{(0)} = -C$ and $Z_{(n+1)} = C$ (for a large value C chosen so that the tail mass of Ψ outside $[-C, C]$ is negligible), and let $Z_{(1)}, \ldots, Z_{(n)}$ be as before. We then define a grid $z_0 \leq \cdots \leq z_{K(n+1)}$ where $z_{Ki} = Z_{(i)}$ for each i, and the points $z_{K(i-1)}, \ldots, z_{Ki}$ form an equally spaced grid from $Z_{(i-1)}$ to $Z_{(i)}$ for each i. After evaluating $\Psi(z_0), \ldots, \Psi(z_{K(n+1)})$, we approximate the integral with the trapezoid rule:

$$\widehat{f}_\pi(x) = \sum_{i=1}^{n+1} \sum_{k=1}^{K} \frac{\Psi(x, z_{K(i-1)+k-1}) + \Psi(x, z_{K(i-1)} + k)}{2} \cdot \frac{Z_{(i)} - Z_{(i-1)}}{K}.$$

For our implementation we choose $C = 10$ and $K = 20$.

Using this we can define the approximate target density $\widehat{g}_\pi\left(\cdot \mid \widehat{\theta}_{1:B} \right)$ as in Section D.1.2 which serves as the target density. In order to sample from that density, we use the Metropolis–Hastings algorithm. Consider $\zeta(x_i) = \log \widehat{g}_\pi \left(x_i \mid x_{-i}, \widehat{\theta}_{1:B} \right)$,

$$x_i^\star = \arg\max_{x_i} \zeta(x_i),$$

and denote the Hessian at $x_i^\star$ by $\zeta''(x_i^\star)$. The proposal distribution is given by $\mathcal{N}\left(x_i^\star, -\frac{1}{\zeta''(x_i^\star)} \right)$. This optimization is performed numerically via `optim` in `R`. We use the permuted serial sampler, as follows:

1. *Initialization.* Draw $m_0 \in \{0, \ldots, M\}$ uniformly at random, and set

$$\widetilde{X}^{(m_0)} \leftarrow X.$$

2. *Iterations.* For $t = m_0 + 1, \ldots, M$, for each $i = 1, \ldots, n$ draw $x_{\text{prop}} \sim \mathcal{N}\left(x_i^\star, -\frac{1}{\Psi''(x_i^\star)} \right)$, $u \sim \text{Unif}(0, 1)$ and set

$$\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \leq \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$$

where

$$\alpha = \min\left\{1,\ \frac{\widehat{g}_\pi\Big(x_{\text{prop}} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)}{\widehat{g}_\pi\Big(X_i^{(t-1)} \mid x_{-i}, \widehat{\theta}_{1:B}\Big)} \cdot \frac{\phi\Big((X_i^{(t-1)} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\Big)}{\phi\Big((x_{\text{prop}} - x_i^\star)\sqrt{-\zeta''(x_i^\star)}\Big)}\right\}$$

and $x_{-i} = \big(\widetilde{X}_{<i}^{(t)},\ \widetilde{X}_{>i}^{(t-1)}\big)$. Similarly, for $t = m_0 - 1, \ldots, 0$, for each $i = n, \ldots, 1$, we draw $x_{\text{prop}} \sim \mathcal{N}\Big(x_i^\star,\ -\frac{1}{\zeta''(x_i^\star)}\Big)$, $u \sim \text{Unif}(0,1)$ and set $\widetilde{X}_i^{(t)} = \begin{cases} x_{\text{prop}}, & \text{if } u \le \alpha, \\ \widetilde{X}_i^{(t-1)}, & \text{otherwise,} \end{cases}$ as before with the only difference that $x_{-i} = \big(\widetilde{X}_{<i}^{(t+1)},\ \widetilde{X}_{>i}^{(t)}\big)$.

3. *Output.* Discard $\widetilde{X}^{(m_0)}$ and return copies $\widetilde{X}^{(0)}, \ldots, \widetilde{X}^{(m_0-1)}, \widetilde{X}^{(m_0+1)}, \ldots, \widetilde{X}^{(M)}$.

D.4 Sensitivity analysis

D.4.1 Sensitivity with respect to concentration of prior

In this section, we investigate the impact of the concentration of the prior on the performance of aCSS-B. While aCSS-B maintains validity and matches the oracle's power across our simulations under standard specifications, it is instructive to examine the performance under a wide range of priors; two extreme cases being a highly concentrated prior centred far from the true parameter, and conversely, an extremely diffuse prior. Notably, highly concentrated priors are seldom used in practice without substantial prior information; furthermore, such priors are generally inadvisable for use with aCSS-B.

To get some intuition about the behaviour of our method, recall that the type-I error inflation bound in Theorem 5.3.5 is given by the term $\epsilon(\pi_0) + \frac{\Delta(\pi_0)}{2\sqrt{B}}$. While the first term $\epsilon(\pi_0)$ does not directly depend on the chosen prior π, $\Delta(\pi_0)$, which is the expected distance between the posteriors $\pi_0(\cdot \mid X)$ and $\pi(\cdot \mid X)$, is substantially impacted by this choice. Since π_0 is highly concentrated around θ_0, if π concentrates around some other θ_1 far away from

θ_0, the posterior distribution $\pi(\cdot \mid X)$ might be substantially different from $\pi_0(\cdot \mid X)$ i.e. $\mathrm{d}_{\chi^2}(\pi_0(\cdot \mid X), \pi(\cdot \mid X))$ will be large making $\Delta \gg 0$. This would require using a very large B to ensure that $\frac{\Delta(\pi_0)}{2\sqrt{B}} \approx 0$. However, if B is too large, this will end up conditioning on too much information in the distribution of $\widetilde{X}$ which would cause this distribution to be highly concentrated around X and result in low power. Conversely, when π is flat or weakly informative, with a sufficient sample size n both the posteriors $\pi_0(\cdot \mid X)$ and $\pi(\cdot \mid X)$ are concentrated around the data-generating parameter θ_0, making $\Delta(\pi_0)$ small. This allows a much smaller B which results in a more powerful test.

To empirically investigate how the choice of prior affects the power and validity of our method, we choose two settings from the ones presented—logistic regression (Section 5.4.2) and group-sparse regression (Section 5.4.5)—applying our method with five different prior-parameter choices that control the prior's flatness.

For logistic regression, we vary the prior scale by considering five values of τ, the standard deviation in $\theta_j \sim \mathcal{N}(0, \tau^2)$: $\tau \in \{0.01, 0.1, 1, 10, 100\}$, with data generated under $\theta_0 = 0.2\,\mathbb{1}_d$. In practice, highly concentrated priors such as $\tau = 0.1$ or 0.01 are unlikely to be chosen by practitioners; we include them primarily to assess the behaviour of our method under extreme prior specifications. Across this range—from very concentrated to highly diffuse priors—the resulting aCSS-B procedures remain valid and achieve essentially unchanged power, closely tracking the oracle.

For group-sparse regression, recall the prior

$$
\begin{aligned}
&g^\star \sim \mathrm{Unif}(\{1, \ldots, G\}),\\
&\beta_{I_{g^\star}} \sim \mathcal{N}(5 \cdot \vec{1}_{|I_{g^\star}|}, \tau^2 I_{|I_{g^\star}|}),\\
&\beta_{I_g} = \vec{0}_{|I_g|} \;\forall\; g \neq g^\star.
\end{aligned} \tag{D.21}
$$

We vary the prior standard deviation over $\tau \in \{0.01, 0.1, 1, 10, 100\}$ and use a fixed null

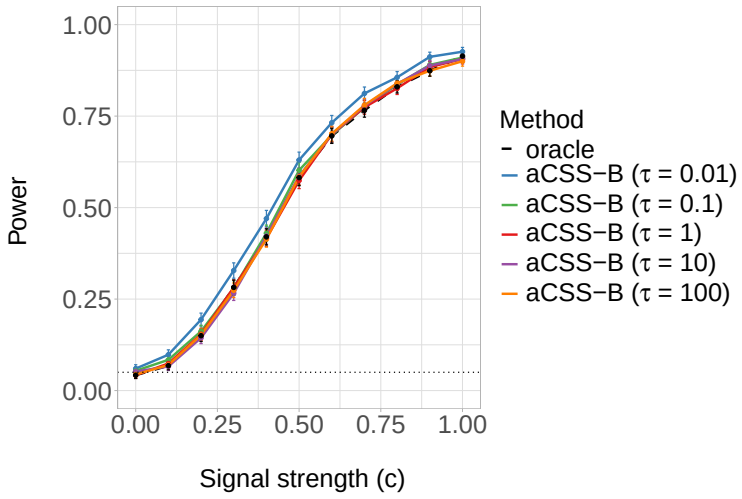

Figure D.1: Power of aCSS-B under different prior standard deviations in the logistic regression model of Section 4.1.

parameter β with

$$\beta_{I_{g_1}} = \vec{1}_{|I_{g_1}|},\ \beta_{I_{g_2}} = c\,\vec{1}_{|I_{g_2}|},\quad \beta_j = 0\ \forall\ j \in [d]\backslash(I_{g_1} \cup I_{g_2}). \tag{D.22}$$

which ensures that the true parameter is systematically separated from the prior mean $5 \cdot \vec{1}_{|I_{g_1}|}$—especially for concentrated priors. As in the logistic regression experiment, $\tau = 0.01$ and 0.1 correspond to highly concentrated and practically implausible prior choices.

Figure D.2 shows that aCSS-B remains valid for $\tau \in \{0.1, 1, 10, 100\}$ and achieves power comparable to the oracle, but becomes invalid under the extremely concentrated prior $\tau = 0.01$.

D.4.2 Sensitivity with respect to number of posterior samples

Section D.4.1 shows that in the group sparse regression example, when using $B = 25$ posterior samples, a highly concentrated (around the wrong parameter value) prior such as one with $\tau = 0.01$ results in substantial type-I error inflation by the aCSS-B method. Theorem 5.3.5 suggests that this should be mitigated when using a higher number of posterior samples,

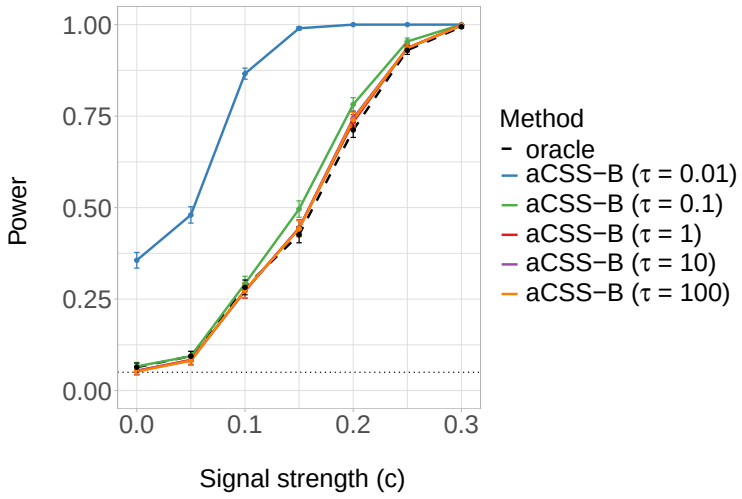

Figure D.2: Power of aCSS-B under different prior standard deviations for the active group in the group-sparse model of Section 4.4, with $B = 25$ posterior draws. Error bars show ± 1 standard error.

however, possibly at the cost of power. In this section, we empirically verify this phenomenon in more detail by considering two different values of the prior standard deviation $\tau = \{0.01, 1\}$ in the group sparse regression model. These two prior parameters respectively represent highly informative and reasonable choices of prior. The prior mean is 5 which is quite far away from the true data generating parameter 1. The data generating distribution and the prior distribution have been specified in detail in (D.21) and (D.22) respectively and we vary B in the range $\{25, 100, 250, 500, 1000\}$.

Figure D.3 illustrates that when the prior is overly concentrated far from the true data-generating parameter ($\tau = 0.01$), aCSS-B loses validity for most values of B. In this adversarial setting, B must be as large as 1000 to achieve validity comparable to the oracle test. Conversely, with a more reasonable prior ($\tau = 1$), $B = 25$ is sufficient to ensure validity. While increasing B beyond this point does not affect validity, as expected, an extremely large value of B does result in a reduction of statistical power.

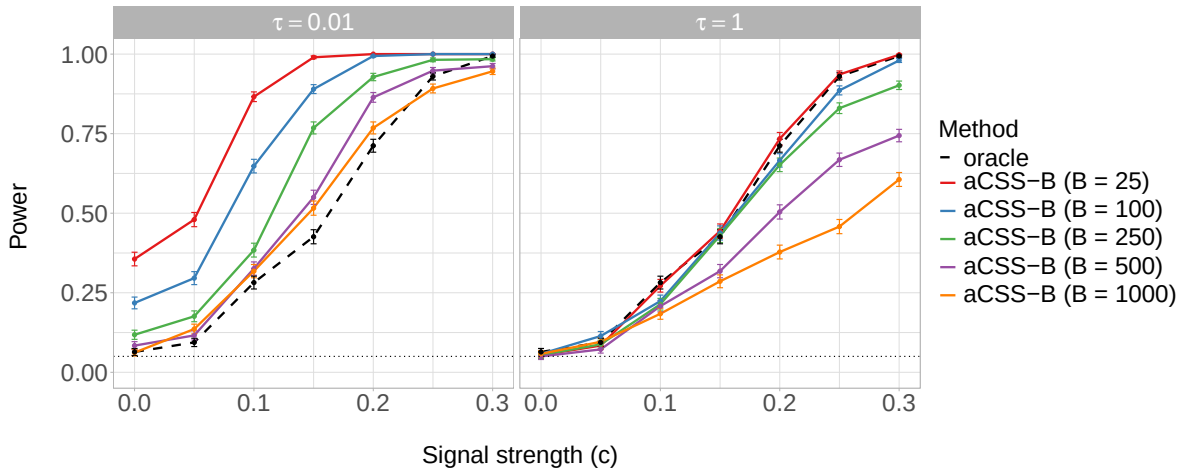

Figure D.3: Power of aCSS-B under different choices of B for two different prior standard deviations ($\tau = 0.01$ and 1) in the group-sparse model of Section 4.4. Error bars show ± 1 standard error.

REFERENCES

Anshuman Acharya and Vinay L Kashyap. Spectral fit residuals as an indicator to increase model complexity. *Research Notes of the AAS*, 8(1):1, 2024.

KARUN Adusumilli. Bootstrap inference for propensity score matching. Technical report, Working paper, 2018.

Alan Agresti. A survey of exact inference for contingency tables. *Statistical science*, 7(1): 131–153, 1992.

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

Jordan Awan and Zhanrui Cai. Approximate co-sufficient sampling for goodness-of-fit tests and synthetic data. *arXiv preprint arXiv:2006.02397*, 2020.

David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Rina Foygel Barber and Lucas Janson. Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *The Annals of Statistics*, 50(5):2514–2544, 2022.

Rina Foygel Barber, Mathias Drton, and Kean Ming Tan. Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 15–36. Springer, 2016.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Maurice Stevenson Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282, 1937.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):175–197, 2020.

Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76 (4):633–642, 1989.

Ritwik Bhaduri, Aabesh Bhattacharyya, Rina Foygel Barber, and Lucas Janson. Conditioning on posterior samples for flexible frequentist goodness-of-fit testing. *Biometrika*, page asag019, 2026.

Aabesh Bhattacharyya and Rina Foygel Barber. Group-weighted conformal prediction. *arXiv preprint arXiv:2401.17452*, 2024.

Aabesh Bhattacharyya, Tiffany Ding, and Rina Foygel Barber. Conformal prediction with macro-coverage guarantees. *arXiv preprint arXiv:2606.28598*, 2026a.

Aabesh Bhattacharyya, Boxuan Zhang, and Rina Foygel Barber. Approximating full conformal prediction: distribution free guarantees via the tournament correction. *arXiv preprint arXiv:2605.29200*, 2026b.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

George Casella and Roger L Berger. *Statistical Inference*. Thomson Learning, 2002.

George Casella and Edward I George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Wenyu Chen, Kelli-Jean Chun, and Rina Foygel Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1):e173, 2018.

Giovanni Cherubin, Konstantinos Chatzikokolakis, and Martin Jaggi. Exact optimization of conformal predictors via incremental and decremental learning. In *International Conference on Machine Learning*, pages 1836–1845. PMLR, 2021.

Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis–Hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American statistical association*, 96(453):270–281, 2001.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/chwialkowski16.html`.

Frank Cowell, Emmanuel Flachaire, and Sanghamitra Bandyopadhyay. Goodness-of-fit: An economic approach. *Economics Series Working Papers 444, University of Oxford, Department of Economics*, 2009.

Ralph B D'Agostino. *Goodness-of-fit-techniques*. Routledge, 2017.

Nina Deliu and Brunero Liseo. The interplay between bayesian inference and conformal prediction. *arXiv preprint arXiv:2510.26930*, 2025.

Tiffany Ding, Anastasios N Angelopoulos, Stephen Bates, Michael I Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *arXiv preprint arXiv:2306.09335*, 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. *International Conference on Learning Representations*, 2026.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, pages 1–12, 2022.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. 2004.

BC Emerson, KM Ibrahim, and GM Hewitt. Selection of evolutionary models for phylogenetic hypothesis testing using parametric methods. *Journal of Evolutionary Biology*, 14(4): 620–631, 2001.

Steinar Engen and Magnar Lillegård. Stochastic simulations conditioned on sufficient statistics. *Biometrika*, 84(1):235–240, 1997.

Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Edwin Fong and Chris C Holmes. Conformal bayesian computation. *Advances in Neural Information Processing Systems*, 34:18268–18279, 2021.

DAS Fraser. On sufficiency and the exponential family. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(1):115–123, 1963.

Massimiliano Frezza. Goodness of fit assessment for a fractal model of stock markets. *Chaos, Solitons & Fractals*, 66:41–50, 2014.

Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

Aditya Gangrade, Alessandro Rinaldo, and Aaditya Ramdas. A sequential test for log-concavity. *arXiv preprint arXiv:2301.03542*, 2023.

Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Pl@ntNet-300K: A plant image dataset with high label ambiguity and a long-tailed distribution. In *Advances in Neural Information Processing Systems*, 2021.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Yu Gui, Rohan Hore, Zhimei Ren, and Rina Foygel Barber. Conformalized survival analysis with adaptive cutoffs. *arXiv preprint arXiv:2211.01227*, 2022.

Sun Wei Guo and Elizabeth A Thompson. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, pages 361–372, 1992.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Matthew T Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: a robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Kiljae Lee and Yuan Zhang. Leave-one-out stable conformal prediction. *arXiv preprint arXiv:2504.12189*, 2025.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*, 2023.

Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4):749–764, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

David D Lewis. Evaluating text categorization i. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

Shuqi Liu, Jianguo Huang, and Luke Ong. Conformal prediction meets long-tail classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23828–23836, 2026.

Anton Rask Lundborg, Rajen D Shah, and Jonas Peters. Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850, 2022.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42): 20881–20885, 2019.

Javier Abad Martinez, Umang Bhatt, Adrian Weller, and Giovanni Cherubin. Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6631–6639, 2023.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Vito M. Muggeo. *segmented: Regression Models with Break-Points / Change-Points Estimation*, 2023. URL `https://CRAN.R-project.org/package=segmented`. R package version 1.6-4.

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

Eugene Ndiaye and Ichiro Takeuchi. Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32, 2019.

Eugene Ndiaye and Ichiro Takeuchi. Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1):151–176, 2023.

Michael A Newton and Adrian E Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1):3–26, 1994.

Art B. Owen. *Monte Carlo theory, methods and examples.* `https://artowen.su.domains/mc/`, 2013.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: European Conference on Machine Learning*, pages 345–356, 2002.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pages 27907–27947. PMLR, 2023.

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

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

Marina Riabiz, Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, and Chris J Oates. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1059–1081, 2022.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Aytijhya Saha and Aaditya Ramdas. Robust likelihood ratio tests for composite nulls and alternatives. *arXiv preprint arXiv:2408.14015*, 2024.

Arnab Sen and Bodhisattva Sen. Testing independence and goodness-of-fit in linear models. *Biometrika*, 101(4):927–942, 2014.

Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2008.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Janardhan R Doppa, and Yan Yan. Conformal prediction for class-wise coverage via augmented label rank calibration. *Advances in Neural Information Processing Systems*, 37:132133–132178, 2024.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):749–760, 1995.

Michael A Stephens. Goodness-of-fit and sufficiency: Exact and approximate tests. *Methodology and Computing in Applied Probability*, 14:785–791, 2012.

Leland T Stewart and Jerry D Johnson. Determining optimum burn-in and replacement times using Bayesian decision theory. *IEEE Transactions on Reliability*, 21(3):170–175, 2009.

Dharmesh Tailor, Alvaro HC Correia, Eric Nalisnick, and Christos Louizos. Approximating full conformal prediction for neural network regression with gauss-newton influence. *arXiv preprint arXiv:2507.20272*, 2025.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. *On the Uniform Convergence of the Frequencies of Occurrence of Events to Their Probabilities*, pages 7–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-41136-6. doi:10.1007/978-3-642-41136-6_2. URL `https://doi.org/10.1007/978-3-642-41136-6_2`.

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

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *conformal and probabilistic prediction and applications*, pages 37–51. PMLR, 2018.

Chunlei Wu and Kenneth Lange. *gglasso: Group Lasso Regularization*, 2020. URL `https://cran.r-project.org/package=gglasso`. R package version 1.5.

Jie Xie and Dongming Huang. A generalized framework for approximate co-sufficient sampling. *arXiv preprint arXiv:2506.12334*, 2025.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *arXiv preprint arXiv:2203.01761*, 2022.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2022.

Wanrong Zhu and Rina Foygel Barber. Approximate co-sufficient sampling with regularization. *arXiv preprint arXiv:2309.08063*, 2023.