THE UNIVERSITY OF CHICAGO

ROBUST ESTIMATION AND DISTRIBUTION-FREE INFERENCE FOR SUPERVISED LEARNING

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
YONGHOON LEE

CHICAGO, ILLINOIS
JUNE 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor Professor Rina Barber for the wonderful experience during my time as a PhD student. I have learned so much from her—not only did I learn how to do research, gain insights, and communicate both orally and in words, but she's also taught me how to be a caring and responsible person. I feel grateful for the opportunity to be advised by her and to take her courses, and I've enjoyed being part of her group. I also feel thankful for all of the advice and guidance that she has provided me, spanning from research and my future career to being a caring TA.

I'd also like to thank Professor Chao Gao and Professor Rebecca Willett for their helpful feedback and interesting questions. I learned a lot from both of them. The two courses I took from Chao Gao were very intriguing, and collaborating with Rebecca Willet has been a great experience.

ABSTRACT

We discuss problems where we have limited access to the information of underlying distribution of training data, which can be caused by imperfect data or insufficient prior knowledge.

We first look into the binary classification problem, in the setting where the label observations are corrupted by noise. We establish that corruption acts as a form of regularization, and we compute precise upper bounds on estimation error in the presence of corruption. Our results suggest that the presence of corrupted data points is beneficial up to a small fraction of the total sample, scaling with the square root of the sample size.

Next, we study the regression problem in the distribution-free setting. We show that there are three regimes in terms of the possibility of meaningful inference, which are characterized by the 'effective support size' of the feature distribution. Our result implies that there exists a counterintuitive in-between regime where we can still expect to obtain meaningful inference for a future input even when it is unlikely to have a value we have observed before.

We also develop distribution-free methods for predictive inference with hierarchically structured datasets. For the special case where we have i.i.d. repeated measurements, we propose to bound the expected squared conditional miscoverage rate in order to have a better control of the conditional coverage, and extend existing methods to construct distribution-free prediction sets that achieve the bound.

CHAPTER 1

INTRODUCTION

We consider problems where we only have limited information about the distribution of a dataset, which appear generally in many real-world problems due to the complexity of physical/societal phenomena, missing or noisy observations in the data collection process, insufficient prior knowledge, etc. In such settings, it is often desired to develop methodologies that are robust to the uncertainties we have, or are valid generally under weak distributional assumptions. In this work, we look into the following problems.

Supervised learning with imperfect data

Consider problems in supervised learning where we have a training dataset $(X_1, Y_1), \cdots, (X_n, Y_n)$ and our goal is to fit a regression function, to find the optimal classifier, or to construct a procedure for the prediction of the label Y_{n+1} for a future input X_{n+1}, etc. The basic idea in any method for data analysis, from linear regression to deep neural networks, is that the training data provide the information of the true distribution of (X, Y). However, in many real world problems it is often not easy to obtain perfect data where each observation is actually from the true underlying distribution due to many possible reasons, such as a noisy measuring procedure or technological limitations. Therefore, it is important to understand the effect of such imperfection and develop methods that account for it.

In Chapter 2, we look into the binary classification problem, where the goal is to make use of the training data $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n) \in \mathbb{R}^d \times \{-1, 1\}$ to fit a classifier $\hat{f} : \mathbb{R}^d \to \{-1, 1\}$, where $\hat{f}(X)$ accurately estimates the label Y for a new input X.[1] In particular, we discuss the case where we have corrupted training data $(X_1, \tilde{Y}_1), (X_2, \tilde{Y}_2), \cdots, (X_n, \tilde{Y}_n)$ due to the noise in the label observations, which prevents us from having direct access to

1. The paper corresponding to the work discussed in this chapter was published in Electronic Journal of Statistics 16(1): 1367-1392

the true distribution of (X, Y). It has been reported in many works that such noise can severly lower the quality of the resulting estimator, and a number of methods have been proposed to provide a noise-robust classifier or to 'correct' the classifier. Such methods can be useful in cases where we have some additional information about the distribution of the data so that we can choose/tune the adjustment procedure, but it is also possible that the modification or correction is not very effective or even makes the quality of the classifier worse or too conservative. In this work, we study the behavior of the classifier from the corrupted dataset, instead of making any modification or adjustment to the procedure. Our main result states that the noise in the label can work as regularization, in the sense that the classifier from the noisy data behaves like a regularized classifier. This implies that the corruption can be beneficial in some cases, and that it might be better not to apply any correction in such cases. We support this observation by Theorem 1 and relevant simulations.

Distribution-free inference

Now consider problems where our goal is beyond providing estimates—we might want to construct a classifier with some inferential guarantees, or we can aim to construct a prediction interval for Y_{n+1} given X_{n+1}. For such uncertainty quantification problems, we need stronger information on the distribution of the data, but there are fundamental limits on the amount of information that the n observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ can provide. Many statistical methods resolve this issue by making assumptions—e.g., linear regression assumes linearity of the mean function as well as the normality of noise, while nonparametric methods have weaker assumptions such as smoothness of the mean function. They are useful when there are sufficient reasons to believe that the assumptions hold, but it also implies that we cannot rely on such methods when the assumptions cannot be verified. Distribution-free inference aims to provide methodologies that are valid without distributional assumptions on the data. Several methods have been developed for predictive inference—e.g., conformal prediction

([Vovk et al., 2005]) and jackknife+ ([Barber et al., 2021b]), but many questions on the possible targets and the usefulness of distribution-free methods still remain unanswered.

In Chapter 3, we study regression problem, where we have training data $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ and the goal is to construct a distribution-free confidence interval for the conditional mean $\mathbb{E}[Y \mid X]$ at a new input X_{n+1}.[2] In the setting where the features X are continuously distributed, recent work has established that any confidence interval for $\mathbb{E}[Y \mid X]$ must have non-vanishing width, even as sample size tends to infinity. At the other extreme, if X takes only a small number of possible values, then inference on $\mathbb{E}[Y \mid X]$ is trivial to achieve. We study the problem in between these two extremes. We find that there are several distinct regimes in between the finite setting and the continuous setting, where vanishing-width confidence intervals are achievable if and only if the effective support size of the distribution of X is smaller than the square of the sample size.

In Chapter 4, we explore the setting where our dataset has a hierarchical structure, meaning that we have an exchangeable set of groups of exchangeable observations. We derive extensions of distribution-free methods that work under this hierarchical exchangeability. In a special case where we have independent repeated measurements, we show that we can aim beyond the marginal coverage guarantee, and propose a stronger guarantee which provides better control of conditional miscoverage rates. We construct distribution-free prediction sets that meet this guarantee and illustrate their performance with simulations.

2. The paper corresponding to the work discussed in this chapter was published in Advances in Neural Information Processing Systems 34 (NeurIPS 2021)

CHAPTER 2

BINARY CLASSIFICATION WITH CORRUPTED LABELS

2.1 Introduction

Consider a classification problem, where our goal is to predict a binary label $Y \in \{\pm 1\}$ using information captured by a feature vector $X \in \mathbb{R}^d$. Based on n training data points $(X_1, Y_1), \ldots (X_n, Y_n)$, the objective is to fit a classifier $\widehat{f} : \mathbb{R}^d \to \{\pm 1\}$ to this data, mapping a new test feature vector X to a predicted label $+1$ or -1.

In many settings, inherent noise in the measurement process can introduce corruption into the observed labels Y_i. For example, consider a medical application where features X_i for patient i determine their likelihood of having a particular disease, and $Y_i \in \{\pm 1\}$ indicates presence or absence of the disease. Imperfect diagnostic tests might mean that the observed label may differ from the true label Y_i. Writing $\widetilde{Y}_i \in \{\pm 1\}$ to denote the observed label, we might have $\mathbb{P}\left\{\widetilde{Y}_i = -1 \;\middle|\; Y_i = +1\right\} > 0$ (if the diagnostic test has a nonzero rate of false negatives) and similarly $\mathbb{P}\left\{\widetilde{Y}_i = +1 \;\middle|\; Y_i = -1\right\} > 0$ (indicating false positives).

2.1.1 Setting and notation

We begin by introducing some basic notation and definitions that we will use throughout. Consider the following model for the triples $(X, Y, \widetilde{Y})$, where as before, $X \in \mathbb{R}^d$ denotes the feature vector, $Y \in \{\pm 1\}$ is the true label (which we do not observe), and $\widetilde{Y} \in \{\pm 1\}$ is the observed label (which may be corrupted, i.e., may differ from the true label):

$$X \sim P_X \quad \text{(a distribution on } \mathbb{R}^d\text{)},$$
$$Y|X = \begin{cases} +1, & \text{with prob. } \eta(X), \\ -1, & \text{with prob. } 1 - \eta(X), \end{cases}$$

$$\widetilde{Y}|X,Y = \begin{cases} -Y, & \text{with prob. } \rho, \\ Y, & \text{with prob. } 1-\rho. \end{cases}$$

Here $\eta(x)$ denotes the probability of a positive (true) label,

$$\eta(x) = \mathbb{P}\left\{Y = +1 \mid X = x\right\},$$

while ρ denotes the probability that the observed label is corrupted, assumed to be identical across all data points (the "homogeneous noise" setting).

In the classification problem, our goal is to define a classification rule that, given a feature vector $x \in \mathbb{R}^d$, outputs a predicted label $+1$ or -1. The misclassification rate is minimized by predicting $+1$ or -1 depending on whether $\eta(x)$ is above or below 0.5, respectively. In a real data setting where $\eta(x)$ is unknown, the classification problem is typically addressed by fitting some function $f(x) \in \mathbb{R}$ and then predicting the label $\text{sign}(f(x))$. We can interpret $f(x)$ as containing information about both our prediction for the label (via the sign) and our confidence in this prediction (via the magnitude—values $f(x) \approx 0$ indicate uncertainty).

Given a possible choice of the function f, the misclassification rate on the training data set $\{(X_i, Y_i) : i = 1, \ldots, n\}$ is therefore given by the empirical 0-1 loss,

$$\widehat{\mathcal{L}}_n^{0/1}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{f(X_i) \cdot Y_i \leq 0\right\},$$

while

$$\widetilde{\mathcal{L}}_n^{0/1}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{f(X_i) \cdot \widetilde{Y}_i \leq 0\right\}$$

measures misclassification on the *corrupted* training data set $\{(X_i, \widetilde{Y}_i) : i = 1, \ldots, n\}$. Our goal is to ensure a low "true" misclassification rate, i.e., for predicting the label Y for a new

point with features X, that is,

$$\mathcal{L}^{0/1}(f) = \mathbb{P}\left\{f(X) \cdot Y \leq 0\right\},$$

where (X, Y) is a new data point drawn from the same distribution as the original training data—that is, $X \sim P_X$, and $Y|X$ is a label in $\{\pm 1\}$ with probabilities determined by $\eta(X)$.

Since the zero/one loss is challenging to optimize, it is standard to use a *surrogate loss function* $\ell : \mathbb{R} \to \mathbb{R}_+$, typically chosen to be continuous, convex, and monotone nonincreasing. For example, a logistic surrogate loss is given by

$$\ell(t) = \log(1 + e^{-t}),$$

while the hinge loss is given by

$$\ell(t) = \max\{0, 1 - t\}.$$

Given a sample of n data points, $(X_1, Y_1), \ldots, (X_n, Y_n)$, we then define the *empirical risk*

$$\widehat{\mathcal{L}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i) \cdot Y_i),$$

which is the average surrogate loss on the data set $\{(X_i, Y_i) : i = 1, \ldots, n\}$, and the *corrupted empirical risk*

$$\widetilde{\mathcal{L}}_n^{\rho}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i) \cdot \widetilde{Y}_i),$$

which is the average surrogate loss on the *corrupted* data set $\{(X_i, \widetilde{Y}_i) : i = 1, \ldots, n\}$. We will also write

$$\mathcal{L}(f) = \mathbb{E}\left[\ell(f(X) \cdot Y)\right],$$

the "true" risk of a function f, with expectation taken over a data point (X, Y) drawn from

the same distribution as before, i.e., $X \sim P_X$, and label $Y|X$ drawn with probabilities determined by $\eta(X)$.

2.1.2 Summary of questions and results

The key question of this work is to compare the performance of the empirical risk minimizer,

$$\widehat{f} = \text{argmin}_{f \in \mathcal{F}} \widehat{\mathcal{L}}_n(f),$$

and its corrupted counterpart,

$$\widetilde{f} = \text{argmin}_{f \in \mathcal{F}} \widetilde{\mathcal{L}}_n^{\rho}(f),$$

where the minimization is taken over some predefined class of functions $\mathcal{F}$ (for example, linear functions of x). That is, how does the presence of corrupted labels affect the performance of the empirical risk minimizer? In particular, we emphasize that the surrogate loss function is unchanged—we do not adjust ℓ or attempt to "correct" for the presence of corruption (this is in contrast to much of the existing literature, which we review below).

Our findings can be summarized as follows. First, we find that corruption mimics regularization—in particular, for a fixed function $f \in \mathcal{F}$, the corrupted empirical risk $\widetilde{\mathcal{L}}_n^{\rho}(f)$ is a *biased* estimate of the true risk $\mathcal{L}(f)$, but acts as an *unbiased* estimate of a penalized version of this risk,

$$\mathcal{L}(f) + \lambda \mathsf{R}(f)$$

where $\lambda > 0$ is a penalty parameter depending on the corruption level ρ, while the regularization function is given by

$$\mathsf{R}(f) = \mathbb{E}\left[\frac{\ell(f(X)) + \ell(-f(X))}{2}\right],$$

the expected loss of the function f under a completely random label.

While adding a penalty introduces bias into our estimator, it also serves to reduce variance, and for limited sample size n, this reduction in variance may outweigh the bias. Our second finding is therefore that, in some settings, corruption may lead to reduced risk for finite sample size, since it is effectively acting as a regularizer and can substantially reduce variance.

2.1.3 Prior work

The problem of learning a classifier in the presence of corrupted labels has been studied in many works in the recent literature. Here we give a very brief overview of the settings and types of results considered. Consider the more general model

$$\begin{aligned} X &\sim P_X \quad \text{(a distribution on } \mathbb{R}^d\text{)}, \\ Y|X &= \begin{cases} +1, & \text{with prob. } \eta(X), \\ -1, & \text{with prob. } 1-\eta(X), \end{cases} \\ \widetilde{Y}|X,Y &= \begin{cases} -Y, & \text{with prob. } \rho(X,Y), \\ Y, & \text{with prob. } 1-\rho(X,Y). \end{cases} \end{aligned}$$

Here $\eta(x)$ denotes the probability of a positive (true) label as before, while $\rho(x,y)$ denotes the probability that the observed label is corrupted,

$$\rho(x,y) = \mathbb{P}\left\{\widetilde{Y} \neq Y \;\middle|\; X = x, Y = y\right\},$$

which now may depend on x and/or y.

[Frénay et al., 2014] and [Frenay and Verleysen, 2014] provide overviews of recent works on this problem. They categorize the existing methods to three types: label noise-robust

models, data cleaning methods, and label noise-tolerant learning algorithms.

The *homogeneous noise* setting assumes that $\rho(x, y) \equiv \rho$ for all x, y—that is, there is a constant probability for each label to be corrupted. This is the setting we study in the present work. Under this setting, [Long and Servedio, 2010] study boosting algorithms and discuss negative consequences of label noise. [Van Rooyen et al., 2015] consider ERM method and propose a label noise-robust loss function. [Manwani and Sastry, 2013] discuss the noise-tolerance property of risk minimization. [Blanco et al., 2020] propose robust algorithms that apply relabeling and clustering to SVM.

The *class-dependent noise* setting assumes that $\rho(x, y) = \rho_y$ for all x, y—that is, the probability of corrupting a positive label ($Y = +1$ but $\widetilde{Y} = -1$) is constant with respect to the feature vector x, and similarly for a negative label, but these two probabilities may differ. For example, in our earlier medical example, the diagnostic test might have different false positive and false negative rates, but these rates themselves are constant across patients (i.e., independent of features such as age that might be included in the X vector). [Liu and Tao, 2016], [Scott et al., 2013], and [Blanchard et al., 2016] study the consistency of the classifier under corruption, while [Reeve and Kaban, 2019a] focus on the minimax optimal learning rate of the corrupted estimator. Some recent works try correction of the loss function or the observed labels; see [Natarajan et al., 2018], [van Rooyen and Williamson, 2018], [Patrini et al., 2017], and [Lin and Bradic, 2021]. Other recent works focus on studying or developing label noise-robust methods; see [Natarajan et al., 2013], [Patrini et al., 2016], [Reeve and Kaban, 2019b], [Bootkrajang and Kabán, 2012], and [Bootkrajang and Kabán, 2014].

Finally, the *general* setting—where $\rho(x, y)$ might vary with x—is studied by [Cannings et al., 2019]. In particular, they examine a setting where the corrupted labels $\widetilde{Y}_i$ are more "clean" than the original labels Y_i, in the sense that the corruption mechanism defined by $\rho(x, y)$ acts to denoise labels near the decision boundary (i.e., $\eta(x) \approx 0.5$) Specifically, suppose that, for values x with $\eta(x)$ slightly higher than 0.5, we have $\rho(x, +1) < \rho(x, -1)$

(that is, a label $Y_i = -1$ that "should" instead be positive, has a greater chance of being flipped to $\widetilde{Y}_i = +1$), and similarly if $\eta(x)$ is slightly lower than 0.5 then $\rho(x, +1) > \rho(x, -1)$. In this case, the $\widetilde{Y}_i$'s carry strictly more information for estimating the decision boundary, as compared to the Y_i's; this setting is therefore fundamentally different from the one we consider here, where homogeneous noise creates strictly noisier labels. [Menon et al., 2016] consider a similar general setting where they show that any consistent algorithm for noise free setting is also consistent under noisy labels under appropriate assumptions. Recent discussions on the noise-tolerence and the robustness of the corrupted classification under this setting can be found in [Ghosh et al., 2015] and [Cheng et al., 2020].

2.2 Main results

2.2.1 Intuition: corruption acts as regularization

The key idea for studying the corrupted estimator through the framework of regularization, is to find a regularizer that matches the expected behavior of the corruption. In order to do this, we first find a different representation of the corruption variables: define

$$R_i \overset{\text{iid}}{\sim} \text{Bernoulli}(2\rho) \text{ and } Z_i \overset{\text{iid}}{\sim} \text{Uniform}\{\pm 1\},$$

drawn independently from each other and independently of the clean data. Then let

$$\widetilde{Y}_i = (1 - R_i) \cdot Y_i + R_i \cdot Z_i.$$

That is, R_i determines whether the label Y_i will be replaced by a random sign, and Z_i provides this random sign. Examining this construction we can see that this yields the same distribution of the corrupted labels as the original definition. We can then write the

corrupted loss as

$$\widetilde{\mathcal{L}}_n^{\rho}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i)\cdot \widetilde{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}(1-R_i)\cdot \ell(f(X_i)\cdot Y_i) + \sum_{i=1}^{n} R_i \cdot \ell(f(X_i)\cdot Z_i).$$

Next, we treat f as fixed, and then condition on the clean data and marginalize over the distribution of the R_i's and Z_i's:

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\mathcal{L}}_n^{\rho}(f) \;\middle|\; X_{1:n}, Y_{1:n}\right] \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[1-R_i\right]\cdot \ell(f(X_i)\cdot Y_i) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[R_i\right]\cdot \mathbb{E}\left[\ell(f(X_i)\cdot Z_i) \mid X_i\right] \\
&\qquad\qquad = (1-2\rho)\cdot \widehat{\mathcal{L}}_n(f) + \rho \cdot \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(f(X_i)) + \ell(-f(X_i))\Big).
\end{aligned}$$

Recall the definition of the regularizer,

$$\mathsf{R}(f) = \mathbb{E}\left[\frac{\ell(f(X)) + \ell(-f(X))}{2}\right],$$

the expected loss of f on purely random labels. We can also consider an empirical version,

$$\widehat{\mathsf{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{\ell(f(X_i)) + \ell(-f(X_i))}{2}.$$

We therefore see that

$$\mathbb{E}\left[\widetilde{\mathcal{L}}_n^{\rho}(f) \;\middle|\; (X_i, Y_i), i = 1, \ldots, n\right] = (1-2\rho)\cdot\Big(\widehat{\mathcal{L}}_n(f) + \lambda\widehat{\mathsf{R}}_n(f)\Big),$$

where $\lambda = \frac{2\rho}{1-2\rho}$. Finally, for any fixed function f, we have

$$\mathbb{E}\left[\widehat{\mathcal{L}}_n(f) + \lambda\widehat{\mathsf{R}}_n(f)\right] = \mathcal{L}(f) + \lambda\mathsf{R}(f),$$

by definition. Therefore, we can view the corrupted empirical risk minimizer $\widetilde{f}$ as a sample estimate of the minimizer of the penalized loss $\mathcal{L}(f) + \lambda \mathsf{R}(f)$.

To summarize our findings so far, we have seen that $\widetilde{f} = \operatorname{argmin}_{f \in \mathcal{F}} \widetilde{\mathcal{L}}_n^\rho(f)$ can be described in two ways:

- Fixing the training data $\{(X_i, Y_i) : i = 1, \ldots, n\}$ and taking an expectation over the corruption mechanism (the R_i's and Z_i's above), we see that $\widetilde{\mathcal{L}}_n^\rho(f)$ has (conditional) expected value $\widehat{\mathcal{L}}_n(f) + \lambda \widehat{\mathsf{R}}_n(f)$, a penalized empirical risk.
- Taking expectations over both the original data and the random corruption, $\widetilde{\mathcal{L}}_n^\rho(f)$ has expected value $\mathcal{L}(f) + \lambda \mathsf{R}(f)$, a penalized true risk.

2.2.2 Results for the linear setting

Next, we will examine the implications of this relationship between corruption and regularization, on the goals of minimizing risk. From this point on, we will restrict our discussion to the setting where $\mathcal{F}$ consists of *linear* functions,

$$\mathcal{F} = \{x \mapsto w^\top x : w \in \mathbb{R}^d\},$$

in order to be able to achieve precise results. Consequently we will shift our notation from functions f to vectors w. Specifically, for each $w \in \mathbb{R}^d$ we will define the population-level loss and regularized loss,

$$\mathcal{L}(w) = \mathbb{E}\left[\ell(X^\top w \cdot Y)\right] \quad \text{and} \quad \widetilde{\mathcal{L}}^\rho(w) = \mathbb{E}\left[\ell(X^\top w \cdot Y)\right] + \frac{2\rho}{1-2\rho} \cdot \mathsf{R}(w),$$

where

$$\mathsf{R}(w) = \mathbb{E}\left[\frac{\ell(X^\top w) + \ell(-X^\top w)}{2}\right] = \frac{\mathcal{L}(w) + \mathcal{L}(-w)}{2},$$

as well as the empirical loss and empirical corrupted loss,

$$\widehat{\mathcal{L}}_n(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top w \cdot Y_i) \quad \text{and} \quad \widetilde{\mathcal{L}}_n^\rho(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top w \cdot \widetilde{Y}_i).$$

We will also define population-level minimizers

$$w_* = \text{argmin}_{w\in\mathbb{R}^d} \mathcal{L}(w) \quad \text{and} \quad \widetilde{w}_*^\rho = \text{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w), \tag{2.1}$$

and empirical minimizers

$$\widehat{w}_n = \text{argmin}_{w\in\mathbb{R}^d} \widehat{\mathcal{L}}_n(w) \quad \text{and} \quad \widetilde{w}_n^\rho = \text{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w), \tag{2.2}$$

whenever these minimizers exist. (Note that, in some settings, the loss or its empirical or corrupted counterpart may have no minimizer—for example, logistic loss, where the positive and negative labels can be perfectly separated.) For each of the four minimization problems, if the minimizer exists but is not unique, our results will apply to any minimizer (e.g., $\widetilde{w}_*^\rho$ denotes any element of the set $\text{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$, etc).

It is well-known that regularization may help reduce risk, even at the cost of increasing bias due to the influence of the regularization function. As discussed earlier, since corruption mimics regularization, in many settings we empirically observe that corruption reduces the risk—that is, $\mathcal{L}(\widetilde{w}_n^\rho) < \mathcal{L}(\widehat{w}_n)$, even though the corruption introduces bias. We will next study why this phenomenon occurs, by establishing bounds on the loss $\mathcal{L}(\widetilde{w}_n^\rho)$ of the corrupted estimator.

Theoretical results

We begin by defining our assumptions. First, we require some conditions on the loss function ℓ:

Assumption 1. *The loss function ℓ is nonnegative, nonincreasing, convex, and L-Lipschitz. Furthermore, ℓ is strictly decreasing on negative values, with*

$$\ell(t) \geq \ell(0) + \gamma|t| \text{ for all } t \leq 0$$

for some $\gamma > 0$, and has a subexponential decay for positive values,

$$\ell(t) \leq c_1 e^{-c_2 t} \text{ for all } t \geq 0,$$

for some $c_1, c_2 > 0$.

The last two conditions ensure that the loss function enacts a strong penalty if $X^\top w$ predicts the sign of Y incorrectly (i.e., $\ell(t)$ is large for $t < 0$), but decays quickly if $X^\top w$ predicts the sign of Y correctly (i.e., $\ell(t)$ is small for $t > 0$). These conditions are satisfied by many well-known examples, for instance:

- The logistic loss $\ell_t = \log(1 + e^{-t})$ satisfies Assumption 1 with $\gamma = \frac{1}{2}$ and $L = c_1 = c_2 = 1$.
- The hinge loss $\ell_t = (1 - t)_+$ satisfies Assumption 1 with $L = \gamma = c_1 = c_2 = 1$.

We will also need some weak assumptions on the distribution of the feature vector X:

Assumption 2. *For some $a_0, a_1, a_2 > 0$, it holds that*

$$\mathbb{E}\left[e^{a_0|X^\top u|^2}\right] \leq a_1$$

and

$$\mathbb{E}\left[e^{-t|X^\top u|}\right] \leq \frac{a_2}{t} \text{ for all } t > 0.$$

for all unit vectors $u \in \mathbb{S}^{d-1}$.

For example, this assumption is satisfied by any multivariate Gaussian distribution with mean μ and covariance Σ, with the parameters a_0, a_1, a_2 depending on $\|\mu\|$ and on the largest and smallest eigenvalues of Σ, but not on the dimension d.

Under these assumptions, our main result establishes a bound on the loss of the corrupted estimator $\widetilde{w}_n^\rho$.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Let $n \geq 2$ and fix any $\alpha > 0$. Suppose $\rho \in (0, \frac{1}{2})$ satisfies*

$$\rho \geq C \cdot \frac{d \log n}{n}.$$

Then with probability at least $1 - n^{-\alpha}$, the set $\operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ is nonempty, and for all $\widetilde{w}_n^\rho \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ it holds that

$$\mathcal{L}(\widetilde{w}_n^\rho) \leq \inf_{w \in \mathbb{R}^d} \mathcal{L}(w) + C' \left[\rho^{1/2} + \rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}} \right].$$

Here C, C' depend only on α and on the constants in Assumptions 1 and 2, but not on n, d, or ρ.

We can see an immediate tradeoff in the upper bound in Theorem 1. The $\rho^{1/2}$ term acts as an "approximation error", where a large corruption proportion ρ leads to a potentially large gap between the loss of the regularized estimator, $\mathcal{L}(\widetilde{w}_*^\rho)$, and the minimum possible loss without regularization, $\inf_{w \in \mathbb{R}^d} \mathcal{L}(w)$. On the other hand, the $\rho^{-1/2} \cdot \sqrt{\frac{d \log n}{n}}$ term is the "estimation error", which is large when the corruption proportion ρ is small (i.e., insufficient regularization). The resulting upper bound on risk is minimized when the corruption level scales as $\rho \asymp \left(\frac{d \log n}{n}\right)^{1/2}$, leading to an upper bound on excess risk scaling as $\asymp \left(\frac{d \log n}{n}\right)^{1/4}$. This suggests that even a very small fraction of corrupted entries can lead to a reduced risk. In contrast, the uncorrupted minimization problem may not behave well under these weak assumptions—for instance, if the labels are perfectly linearly separable (as might be the case if, e.g., $Y|X$ follows a logistic regression with very high signal strength), then a minimizer

does not even exist (i.e., $\operatorname{argmin}_{w\in\mathbb{R}^d} \widehat{\mathcal{L}}_n(w)$ is empty).

Of course, the result of Theorem 1 is an upper bound on the loss, and may be loose for certain examples; the value of ρ that minimizes the upper bound (i.e., $\rho \asymp \left(\frac{d\log n}{n}\right)^{1/2}$) might not be the same as the value of ρ that minimizes the loss itself. In particular, the result can be viewed as a "worst case" bound that holds even when the unregularized loss has no minimizer (such as logistic regression with perfectly separable labels, as mentioned above); in problems where this is not the case, regularization is not as critical, and a smaller value of ρ (or even $\rho = 0$) may perform better.

Proof of Theorem 1

Our first step is to examine some properties of the regularized population minimizer $\widetilde{w}_*^\rho$ and its empirical counterpart, the corrupted estimator $\widetilde{w}_n^\rho$.

Lemma 1. *Suppose Assumptions 1 and 2 hold. Fix any $\rho \in (0, \frac{1}{2})$. Then $\operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ is nonempty, and any $\widetilde{w}_*^\rho \in \operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ must satisfy $\|\widetilde{w}_*^\rho\| \le C_0\rho^{-1/2}$ and*

$$\mathcal{L}(\widetilde{w}_*^\rho) \le \inf_{w\in\mathbb{R}^d} \mathcal{L}(w) + C_1\rho^{1/2}.$$

Moreover, for any $\alpha > 0$, if $n \ge 2$ and $\rho \ge C \cdot \frac{d\log n}{n}$ then with probability at least $1 - n^{-\alpha}$ it holds that $\operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ is nonempty, that any $\widetilde{w}_n^\rho \in \operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ must satisfy $\|\widetilde{w}_n^\rho\| \le C_0\rho^{-1/2}$, and that

$$\sup_{\|w\|\le C_0\rho^{-1/2}} \left|\widetilde{\mathcal{L}}_n^\rho(w) - \widetilde{\mathcal{L}}^\rho(w)\right| \le C_2\rho^{-1/2}\sqrt{\frac{d\log n}{n}}.$$

Here C, C_0, C_1, C_2 depend on α and on the constants in Assumptions 1 and 2, but not on n, d, or ρ.

Now we prove the theorem. By Lemma 1, with probability at least $1-n^{-\alpha}$, for any $\widetilde{w}_*^\rho \in \operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ and all $\widetilde{w}_n^\rho \in \operatorname{argmin}_{w\in\mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ it holds that $\mathcal{L}(\widetilde{w}_*^\rho) \le \inf_{w\in\mathbb{R}^d} \mathcal{L}(w) +$

$C_1\rho^{1/2}$ and that

$$\max\left\{\left|\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho) - \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho)\right|,\ \left|\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho)\right|\right\} \le C_2\rho^{-1/2}\sqrt{\frac{d\log n}{n}}.$$

From now on, we assume that these events all hold. Then we have

$$\begin{aligned}
\widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) &= \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) + \left(\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho) - \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho)\right) + \left(\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho)\right) + \left(\widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho)\right) \\
&\le \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) + \left(\widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_n^\rho) - \widetilde{\mathcal{L}}_n^\rho(\widetilde{w}_*^\rho)\right) + 2C_2\rho^{-1/2}\cdot\sqrt{\frac{d\log n}{n}} \\
&\le \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) + 2C_2\rho^{-1/2}\cdot\sqrt{\frac{d\log n}{n}} \quad \text{by optimality of } \widetilde{w}_n^\rho \\
&\le \inf_{w\in\mathbb{R}^d}\mathcal{L}(w) + C_1\rho^{1/2} + 2C_2\rho^{-1/2}\cdot\sqrt{\frac{d\log n}{n}} \\
&\le \inf_{w\in\mathbb{R}^d}\mathcal{L}(w) + \frac{C'}{2}\left[\rho^{1/2} + \rho^{-1/2}\cdot\sqrt{\frac{d\log n}{n}}\right],
\end{aligned}$$

where we set $C' = \max\{2C_1, 4C_2\}$. Next, by definition of $\widetilde{\mathcal{L}}^\rho$, we have

$$\begin{aligned}
\widetilde{\mathcal{L}}^\rho(\widetilde{w}_n^\rho) - \inf_{w\in\mathbb{R}^d}\mathcal{L}(w) &= (1-\rho)\cdot\big[\mathcal{L}(\widetilde{w}_n^\rho) - \inf_{w\in\mathbb{R}^d}\mathcal{L}(w)\big] + \rho\cdot\big[\mathcal{L}(-\widetilde{w}_n^\rho) - \inf_{w\in\mathbb{R}^d}\mathcal{L}(w)\big] \\
&\ge \frac{1}{2}\big[\mathcal{L}(\widetilde{w}_n^\rho) - \inf_{w\in\mathbb{R}^d}\mathcal{L}(w)\big]
\end{aligned}$$

where the last step holds since $\rho \le \frac{1}{2}$. Therefore,

$$\mathcal{L}(\widetilde{w}_n^\rho) \le \inf_{w\in\mathbb{R}^d}\mathcal{L}(w) + C'\left[\rho^{1/2} + \rho^{-1/2}\cdot\sqrt{\frac{d\log n}{n}}\right],$$

which completes the proof of the theorem.

Another perspective on the regularizer

The results above suggest that the main source of possible improvements by corruption is the shrinkage induced by the corruption (or, at the population level, by the regularizer $\mathsf{R}(w)$). In particular, the results of Lemma 1 show that, in the linear setting, the corruption (or the regularizer) lead to an upper bound on $\|w\|$. We will now examine this connection more closely.

The following lemma verifies that, up to constants, $\mathsf{R}(w)$ is equivalent to $\|w\|$. In a sense, then, we can view regularization with $\mathsf{R}(w)$ as effectively placing a penalty on $\|w\|$.

Lemma 2. *Suppose Assumptions 1 and 2 hold. Then it holds that*

$$\max\{c_L \cdot \|w\|, \ell(0)\} \leq \mathsf{R}(w) \leq c_U \cdot \|w\| + \ell(0) \text{ for all } w \in \mathbb{R}^d,$$

where c_L, c_U depend only on the constants in Assumptions 1 and 2.

Proof. In the calculations (2.3) and (2.4) appearing in the proof of Lemma 1, we will see that Assumption 2 implies that

$$\frac{\log 2}{2a_2} \leq \mathbb{E}\left[|X^\top u|\right] \leq \sqrt{\frac{a_1}{a_0}}$$

for all unit vectors $u \in \mathbb{R}^d$. For any $w \in \mathbb{R}^d$, for the lower bound, we have

$$\begin{aligned}\mathsf{R}(w) = \mathbb{E}\left[\frac{\ell(|X^\top w|) + \ell(-|X^\top w|)}{2}\right] \geq \mathbb{E}\left[\frac{\ell(-|X^\top w|)}{2}\right] \geq \mathbb{E}\left[\frac{\ell(-|X^\top w|) - \ell(0)}{2}\right] \\ \geq \frac{\gamma}{2} \cdot \mathbb{E}\left[|X^\top w|\right] \geq \frac{\gamma \log 2}{4a_2} \cdot \|w\|,\end{aligned}$$

and furthermore

$$\mathsf{R}(w) = \mathbb{E}\left[\frac{\ell(|X^\top w|) + \ell(-|X^\top w|)}{2}\right] \geq \ell(0)$$

by convexity of ℓ. For the upper bound, we have

$$\begin{aligned}\mathsf{R}(w) &= \mathbb{E}\left[\frac{\ell(|X^\top w|) + \ell(-|X^\top w|)}{2}\right] \\ &= \ell(0) + \mathbb{E}\left[\frac{\ell(-|X^\top w|) - \ell(0)}{2}\right] + \mathbb{E}\left[\frac{\ell(|X^\top w|) - \ell(0)}{2}\right] \\ &\leq \ell(0) + \mathbb{E}\left[\frac{\ell(-|X^\top w|) - \ell(0)}{2}\right] \leq \ell(0) + \frac{L}{2} \cdot \mathbb{E}\left[|X^\top w|\right] \leq \ell(0) + \frac{L}{2}\sqrt{\frac{a_1}{a_0}} \cdot \|w\|.\end{aligned}$$

□

2.3 Simulations

Now we empirically investigate the effect of corruption through a simulation. We generate the data $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ in the following way: choosing dimension $d = 50$, we draw

$$\begin{aligned} X_i &\sim \mathcal{N}(0, \mathbf{I}_d) \\ Y_i \mid X_i &= \begin{cases} +1, & \text{with probability } \frac{\exp\{3X_{i1}+0.5(X_{i2})^3\}}{1+\exp\{3X_{i1}+0.5(X_{i2})^3\}}, \\ -1, & \text{with probability } \frac{1}{1+\exp\{3X_{i1}+0.5(X_{i2})^3\}}, \end{cases}\end{aligned}$$

independently for each $i = 1, \ldots, n$. The corrupted labels $\{\widetilde{Y}_i\}_{1 \leq i \leq n}$ are generated as

$$\widetilde{Y}_i \mid X_i, Y_i = \begin{cases} -Y_i, & \text{with prob. } \rho, \\ Y_i, & \text{with prob. } 1-\rho, \end{cases}$$

independently for each $i = 1, \ldots, n$. We run the experiment at a small and large sample size, $n = 400$ and $n = 2000$, and at a range of values of the corruption probability, $\rho \in \{0, 0.01, 0.02, \ldots, 0.2\}$. For each sample size n and corruption level ρ, we run 100 independent trials of the experiment, we choose the logistic loss function $\ell(t) = \log(1+e^{-t})$, and compute

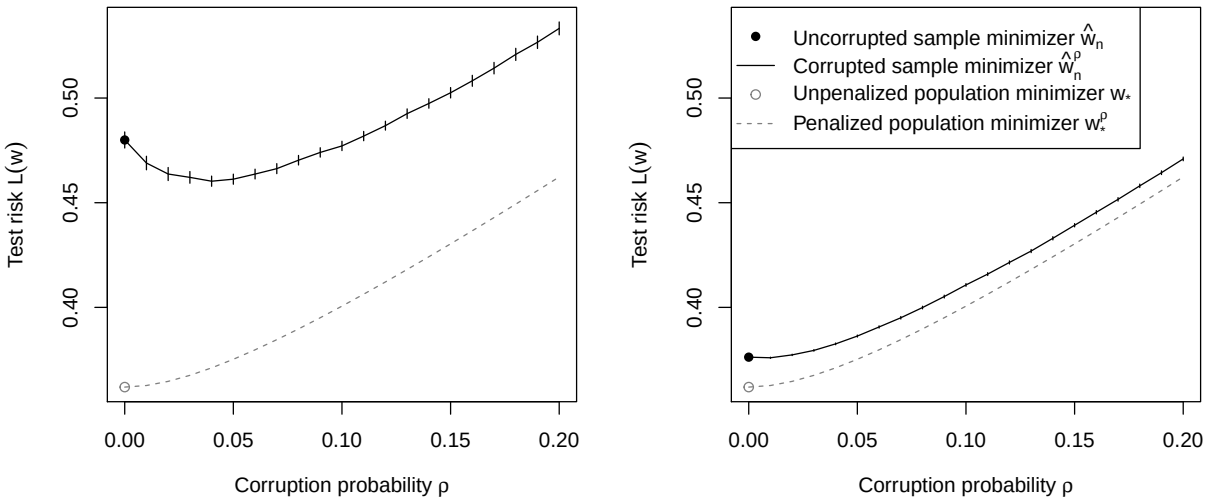

Figure 2.1: Risks of the original classifier $\widehat{w}_n$, the corrupted classifier $\widetilde{w}_n^\rho$, the optimal classifier w_*, and the population-level corrupted classifier $\widetilde{w}_*^\rho$ on the test set, with sample size $n = 400$ (left) and $n = 2000$ (right). For the sample estimators $\widehat{w}_n$ and $\widetilde{w}_n^\rho$, the figure displays the mean over 100 independent trials, with standard error bars. See Section 2.3 for further details.

the corrupted empirical minimizer $\widetilde{w}_n^\rho$ defined in (2.2) and the penalized population-level minimizer $\widetilde{w}_*^\rho$ as in (2.1) (which reduces to the uncorrupted empirical minimizer $\widehat{w}_n$ and the unpenalized population-level minimizer w_*, respectively, in the case $\rho = 0$). Note that the data generating distribution does not follow the logistic regression model (due to the cubic term), and so the logistic loss simply acts as a surrogate for the 0-1 loss (i.e., it does not correspond to a likelihood for some well-specified model).

Figure 2.1 shows the performance of the corrupted estimator $\widetilde{w}_n^\rho$ and its population-level version $\widetilde{w}_*^\rho$, across the range of corruption values $\rho \in \{0, 0.01, 0.02, \ldots, 0.2\}$, at each sample size $n \in \{400, 2000\}$; the result at $\rho = 0$ is highlighted in each case, as it corresponds to the uncorrupted estimator $\widehat{w}_n$ and to the corresponding population-level minimizer w_*. Overall, the plots illustrate how corruption acts as regularization—for the smaller sample size $n = 400$, we see that a small amount of corruption substantially reduces the test risk of the empirical

minimizer $\widetilde{w}_n^\rho$, while for the larger sample size $n = 2000$ the uncorrupted estimator $\widehat{w}_n$ achieves good performance and we no longer see any noticeable improvement from corruption. For the population-level minimizers, on the other hand, increasing regularization always leads to an increase in risk, as expected.

2.4 Discussion

In this chapter, we have shown that the corruption of labels has a regularization-type effect on binary classification problems, leading to a possibility of an improvement of the fitted classifier in terms of test risk. Unlike many prior works that apply adjustment or correction to achieve consistency or robustness of the estimator, our result implies that corruption itself can be beneficial without any adjustment to the estimation process, and thus it could be better in some cases to simply fit the corrupted dataset without any modification on the methods—in particular, this means that we do not need to know or estimate the corruption mechanism, as would be the case for a procedure that corrects for the corruption. For the fitting of linear classifiers using empirical risk minimization under homogeneous noise, Theorem 1 provides an explanation for the possibility of corruption being beneficial, illustrating the tradeoff between loss approximation and the estimation.

We can expect a similar tradeoff for more general settings where the noise is not homogeneous, or where different estimation methods are applied; in general, it is intuitive that a small amount of corruption can reduce the chance of overfitting, especially when the inherent noise level is low, and that this benefit may outweigh the low bias that is introduced. As an example of a broader setting where this type of phenomenon may be useful, we can consider a setting where some data points are known to be "clean" while others are potentially corrupted; while we might expect that performance could be improved by removing or down-weighting the latter data points in order to avoid or reduce the effect of corruption, our findings instead suggest that the presence of the non-"clean" data might even be beneficial.

The question of corrupted labels, with its possible risks and benefits, is studied only in a very specific setting in our work (i.e., linear prediction rules in low dimensions), and many open questions remain. First, noting that the corrupted loss can be thought as another surrogate of 0-1 loss, we may ask how corruption affects the prediction performance of the estimator in terms of misclassification rate, i.e., 0-1 risk. Second, do similar phenomena occur in the high-dimensional regime, $d \gg n$ or $d \propto n$? In particular, we have seen that homogeneous corruption mimics an ℓ_2 penalty in the low-dimensional setting; however, the same is not immediately true in high dimensions, since these results rely on concentration type arguments that would no longer hold (and, in particular, for $d \gg n$, in general both the uncorrupted data $\{(X_i, Y_i)\}_{1 \le i \le n}$ and the corrupted data $\{(X_i, \widetilde{Y}_i)\}_{1 \le i \le n}$ are perfectly linearly separable, so we cannot expect good performance without some additional constraints or regularization). Finally, since the key phenomenon underlying our results is the way that homogeneous corruption mimics ℓ_2 regularization (and therefore, corruption induces shrinkage in the resulting estimator), this does not explain any potential benefits from corruption if we instead use methods such as a k-nearest-neighbor estimator, or other methods where there is no notion of shrinkage; is corruption beneficial more broadly, by reducing the chance of overfitting in a more general sense? We leave these questions for future work.

2.5 Appendix

2.5.1 Proof of Lemma 1

We first verify that $\widetilde{\mathcal{L}}^\rho$ is β-Lipschitz, where $\beta = L\sqrt{\frac{a_1}{a_0}}$. For any $w \neq w' \in \mathbb{R}^d$ we have

$$\begin{aligned}
\left|\widetilde{\mathcal{L}}^\rho(w) - \widetilde{\mathcal{L}}^\rho(w')\right| &= \left|\mathbb{E}\left[\ell(X^\top w \cdot \widetilde{Y}) - \ell(X^\top w' \cdot \widetilde{Y})\right]\right| \\
&\le \mathbb{E}\left[\left|\ell(X^\top w \cdot \widetilde{Y}) - \ell(X^\top w' \cdot \widetilde{Y})\right|\right] \\
&\le \mathbb{E}\left[L \cdot \left|X^\top w \cdot \widetilde{Y} - X^\top w' \cdot \widetilde{Y}\right|\right] \quad \text{since } \ell \text{ is } L\text{-Lipschitz by Assumption 1}
\end{aligned}$$

$$
\begin{aligned}
&= L\mathbb{E}\left[\left|X^\top(w-w')\right|\right] \quad \text{since } \widetilde{Y} \in \{\pm 1\} \\
&= L\|w-w'\| \cdot \mathbb{E}\left[|X^\top u|\right] \quad \text{where } u = \frac{w-w'}{\|w-w'\|} \\
&\leq \beta \cdot \|w-w'\|,
\end{aligned}
$$

where the last inequality follows from Assumption 2 via the calculation

$$
a_1 \geq \mathbb{E}\left[e^{a_0|X^\top v|^2}\right] \geq a_0 \cdot \mathbb{E}\left[|X^\top v|^2\right] \geq a_0 \cdot \mathbb{E}\left[|X^\top v|\right]^2. \tag{2.3}
$$

We therefore have that $\widetilde{\mathcal{L}}^\rho$ is β-Lipschitz. Note that the above argument also holds for $\rho = 0$, implying that $\mathcal{L}$ is also β-Lipschitz.

Now fix $t = C_0\rho^{-1/2}$ for any $C_0 > \sqrt{\frac{8c_1a_2^2}{c_2\gamma\log 2}}$. We will show that, for any $u \in \mathbb{S}^{d-1}$,

$$
\widetilde{\mathcal{L}}^\rho(t \cdot u) > \widetilde{\mathcal{L}}^\rho(0.5t \cdot u).
$$

First we calculate

$$
\mathbb{E}\left[|X^\top u| \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \geq \rho \cdot \mathbb{E}\left[|X^\top u|\right] \geq \rho \cdot \frac{\log 2}{2a_2}
$$

where the first inequality holds by definition of the distribution of the corrupted label $\widetilde{Y}$ (since $\mathbb{P}\left\{\widetilde{Y} = +1 \;\middle|\; X\right\} \in [\rho, 1-\rho]$ holds almost surely), while for the second inequality, by Jensen's inequality together with Assumption 2,

$$
e^{-2a_2\mathbb{E}\left[|X^\top u|\right]} \leq \mathbb{E}\left[e^{-2a_2|X^\top u|}\right] \leq \frac{a_2}{2a_2} = \frac{1}{2},
$$

so

$$
\mathbb{E}\left[|X^\top u|\right] \geq \frac{\log 2}{2a_2}. \tag{2.4}
$$

We also know that

$$\ell(-t \cdot |X^\top u|) - \ell(-0.5t \cdot |X^\top u|) \geq \gamma \cdot 0.5t \cdot |X^\top u|,$$

by Assumption 1, and so

$$\begin{aligned}
&\mathbb{E}\left[\left(\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \\
&\qquad\qquad \geq \mathbb{E}\left[\gamma \cdot 0.5t \cdot |X^\top u| \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \geq \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2}.
\end{aligned}$$

We therefore have

$$\begin{aligned}
&\widetilde{\mathcal{L}}^\rho(t \cdot u) - \widetilde{\mathcal{L}}^\rho(0.5t \cdot u) \\
&= \mathbb{E}\left[\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right] \\
&= \mathbb{E}\left[\left(\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} < 0\right\}\right] \\
&\qquad\qquad + \mathbb{E}\left[\left(\ell(t \cdot X^\top u \cdot \widetilde{Y}) - \ell(0.5t \cdot X^\top u \cdot \widetilde{Y})\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} \geq 0\right\}\right] \\
&\geq \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} + \mathbb{E}\left[\left(\ell(t \cdot |X^\top u|) - \ell(0.5t \cdot |X^\top u|)\right) \cdot \mathbb{1}\left\{X^\top u \cdot \widetilde{Y} \geq 0\right\}\right] \\
&\geq \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} - \mathbb{E}\left[\ell(0.5t \cdot |X^\top u|)\right] \\
&\geq \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} - c_1 \mathbb{E}\left[e^{-c_2 \cdot 0.5t \cdot |X^\top u|}\right] \text{ by Assumption 1} \\
&\geq \gamma \cdot 0.5t \cdot \rho \cdot \frac{\log 2}{2a_2} - \frac{c_1 a_2}{c_2 \cdot 0.5t} \text{ by Assumption 2} \\
&> 0 \text{ by definition of } t.
\end{aligned}$$

In particular, this implies that $\widetilde{\mathcal{L}}^\rho(tu) > \inf_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ for all $u \in \mathbb{S}^{d-1}$. Since $w \mapsto \widetilde{\mathcal{L}}^\rho(w)$ is continuous as shown above, this implies that $\widetilde{\mathcal{L}}^\rho(w)$ attains its infimum, and any $\widetilde{w}_*^\rho \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$ must satisfy $\|\widetilde{w}_*^\rho\| \leq t$.

Next we bound $\mathcal{L}(\widetilde{w}_*^\rho)$ for any $\widetilde{w}_*^\rho \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}^\rho(w)$. First note that the corrupted

risk can be written as

$$\widetilde{\mathcal{L}}^{\rho}(w) = (1-2\rho)\cdot\mathcal{L}(w) + 2\rho\cdot\mathsf{R}(w) = (1-\rho)\mathcal{L}(w) + \rho\mathcal{L}(-w). \tag{2.5}$$

Applying (2.5) with $w = \widetilde{w}_*^{\rho}$ we obtain

$$\widetilde{\mathcal{L}}^{\rho}(\widetilde{w}_*^{\rho}) = (1-\rho)\mathcal{L}(\widetilde{w}_*^{\rho}) + \rho\mathcal{L}(-\widetilde{w}_*^{\rho}),$$

and similarly applying (2.5) with $w = -\widetilde{w}_*^{\rho}$ we obtain

$$\widetilde{\mathcal{L}}^{\rho}(-\widetilde{w}_*^{\rho}) = (1-\rho)\mathcal{L}(-\widetilde{w}_*^{\rho}) + \rho\mathcal{L}(\widetilde{w}_*^{\rho}).$$

Since $\widetilde{\mathcal{L}}^{\rho}(\widetilde{w}_*^{\rho}) \le \widetilde{\mathcal{L}}^{\rho}(-\widetilde{w}_*^{\rho})$ by optimality of $\widetilde{w}_*^{\rho}$, and $\rho < \frac{1}{2}$ by assumption, this proves that $\mathcal{L}(\widetilde{w}_*^{\rho}) \le \mathcal{L}(-\widetilde{w}_*^{\rho})$ and therefore,

$$\mathcal{L}(\widetilde{w}_*^{\rho}) \le \widetilde{\mathcal{L}}^{\rho}(\widetilde{w}_*^{\rho}).$$

Next, fix any $w \in \mathbb{R}^d$. First consider the case that $\|w\| \le c\rho^{-1/2}$, where $c = \sqrt{\frac{c_1 a_2}{2\beta c_2}}$. Then

$$\begin{aligned}
\widetilde{\mathcal{L}}^{\rho}(\widetilde{w}_*^{\rho}) - \mathcal{L}(w) &\le \widetilde{\mathcal{L}}^{\rho}(w) - \mathcal{L}(w) \quad \text{by optimality of } \widetilde{w}_*^{\rho} \\
&= \rho\left(\mathcal{L}(-w) - \mathcal{L}(w)\right) \quad \text{by (2.5)} \\
&\le 2\rho\beta\cdot c\rho^{-1/2} \\
&= 2\beta c\rho^{1/2},
\end{aligned}$$

where the last inequality holds since $\mathcal{L}$ is β-Lipschitz.

Next consider the case that $\|w\| > c\rho^{-1/2}$. Let $u = w/\|w\|$ and $t = c\rho^{-1/2}$. Then by the reasoning above, we have

$$\widetilde{\mathcal{L}}^{\rho}(\widetilde{w}_*^{\rho}) - \mathcal{L}(tu) \le 2\beta c\rho^{1/2}.$$

Next, let $Z_u = X^\top u \cdot Y$, then we have

$$\begin{aligned}
&\mathcal{L}(tu) - \mathcal{L}(w) = \mathbb{E}\left[\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)\right] \\
&= \mathbb{E}\left[(\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)) \cdot \mathbb{1}\left\{Z_u > 0\right\}\right] + \mathbb{E}\left[(\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)) \cdot \mathbb{1}\left\{Z_u < 0\right\}\right] \\
&\leq \mathbb{E}\left[(\ell(t \cdot Z_u) - \ell(\|w\| \cdot Z_u)) \cdot \mathbb{1}\left\{Z_u > 0\right\}\right] \quad \text{since } \|w\| > t \text{ and } \ell \text{ is nonincreasing} \\
&\leq \mathbb{E}\left[\ell(t \cdot Z_u) \cdot \mathbb{1}\left\{Z_u > 0\right\}\right] \quad \text{since } \ell \text{ is nonnegative} \\
&\leq c_1 \mathbb{E}\left[e^{-c_2 t |X^\top u|}\right] \quad \text{by Assumption 1} \\
&\leq c_1 \cdot \frac{a_2}{c_2 t} \quad \text{by Assumption 2} \\
&= \frac{c_1 a_2}{c_2 c} \cdot \rho^{1/2}.
\end{aligned}$$

Therefore, for this second case, we have shown that

$$\widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) - \mathcal{L}(w) \leq \left(2\beta c + \frac{c_1 a_2}{c_2 c}\right) \cdot \rho^{1/2} = \sqrt{\frac{8\beta c_1 a_2}{c_2}} \cdot \rho^{1/2}.$$

Combining the two cases, we have shown that

$$\mathcal{L}(\widetilde{w}_*^\rho) \leq \widetilde{\mathcal{L}}^\rho(\widetilde{w}_*^\rho) \leq \mathcal{L}(w) + \sqrt{\frac{8\beta c_1 a_2}{c_2}} \cdot \rho^{1/2}$$

for all $w \in \mathbb{R}^d$, which proves the desired inequality with

$$C_1 = \sqrt{\frac{8\beta c_1 a_2}{c_2}}.$$

Now we turn to the corrupted estimator $\widetilde{w}_n^\rho$. First we will need a lemma to establish some concentration results.

Lemma 3. *Suppose Assumptions 1 and 2 hold. Fix any $\alpha > 0$, $\rho \in (0, \frac{1}{2})$, $t > 0$, and $r > 0$.*

Then with probability at least $1 - n^{-\alpha}$*, it holds that*

$$\inf_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\} \right\} \geq r_1 \rho - r_2 \cdot \frac{d \log n}{n} \tag{2.6}$$

and

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} e^{-t|X_i^\top u|} \right\} \leq \frac{r_3}{t} + r_4 \sqrt{\frac{d \log n}{n}} \tag{2.7}$$

and

$$\sup_{\|w\| \leq r} \left| \widetilde{\mathcal{L}}_n^\rho(w) - \widetilde{\mathcal{L}}^\rho(w) \right| \leq r_5 \cdot r \cdot \sqrt{\frac{d \log n}{n}}, \tag{2.8}$$

where $r_1, r_2, r_3, r_4, r_5 > 0$ *depend only on* α *and on the constants in Assumptions 1 and 2, and not on* n*,* d*,* r*, or* t*.*

We are now ready to prove the remainder of Lemma 1. First we bound $\|\widetilde{w}_n^\rho\|$. Define $C = \frac{2r_2}{r_1}$ and fix $t = C_0 \rho^{-1/2}$ for any $C_0 > \max\left\{ 2\sqrt{\frac{4c_1\left(2c_2^{-1} r_3\right)}{\gamma r_1}}, \frac{8c_1\left(C^{-1/2} r_4\right)}{\gamma r_1} \right\}$, which therefore satisfies

$$C_0 > \sqrt{\frac{4c_1\left(2c_2^{-1} r_3 + C_0 C^{-1/2} r_4\right)}{\gamma r_1}}.$$

We will show that, for any $u \in \mathbb{S}^{d-1}$,

$$\widetilde{\mathcal{L}}_n^\rho(t \cdot u) > \widetilde{\mathcal{L}}_n^\rho(0.5t \cdot u).$$

Then assuming $\rho \geq C \cdot \frac{d \log n}{n}$, the bound (2.6) in Lemma 3 implies that

$$\frac{1}{n} \sum_{i=1}^{n} |X_i^\top u| \cdot \mathbb{1}\left\{ X_i^\top u \cdot \widetilde{Y}_i < 0 \right\} = \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\} \geq \frac{r_1}{2} \cdot \rho,$$

for all $u \in \mathbb{S}^{d-1}$. Furthermore, since $t = C_0 \rho^{-1/2}$, the bound (2.7) in Lemma 3 (applied

with $0.5c_2 t$ in place of t) together with our assumption $\rho \geq C \cdot \frac{d \log n}{n}$ implies that

$$\frac{1}{n} \sum_{i=1}^{n} e^{-c_2 \cdot 0.5t |X_i^\top u|} \leq \frac{2c_2^{-1} r_3 + C_0 C^{-1/2} r_4}{t}$$

for all $u \in \mathbb{S}^{d-1}$. Following identical arguments as in the population case, we have

$$\widetilde{\mathcal{L}}_n^\rho(t \cdot u) - \widetilde{\mathcal{L}}_n^\rho(0.5t \cdot u) \geq \gamma \cdot 0.5t \cdot \rho \cdot r_1/2 - c_1 \cdot \frac{2c_2^{-1} r_3 + C_0 C^{-1/2} r_4}{t} > 0$$

for all $u \in \mathbb{S}^{d-1}$, where the last step holds by definition of t and of C_0. Since $\widetilde{\mathcal{L}}_n^\rho$ is continuous (because we have assumed the loss ℓ is continuous), as for the population case this again proves that $\widetilde{\mathcal{L}}_n^\rho(w)$ must attain its infimum, and that any $w \in \operatorname{argmin}_{w \in \mathbb{R}^d} \widetilde{\mathcal{L}}_n^\rho(w)$ must satisfy $\|w\| \leq t$.

Finally, the bound $\sup_{\|w\| \leq C_0 \rho^{-1/2}} \left| \widetilde{\mathcal{L}}_n^\rho(w) - \widetilde{\mathcal{L}}^\rho(w) \right| \leq C_2 \rho^{-1/2} \sqrt{\frac{d \log n}{n}}$ follows immediately from the bound (2.8) in Lemma 3, by setting $C_2 = C_0 r_5$.

2.5.2 Proof of Lemma 3

First, we prove (2.6). The distribution of $(X, \widetilde{Y})$ can equivalently be represented as

$$(X, \widetilde{Y}) = \big(X, (1 - R) \cdot Y + R \cdot Z\big),$$

where $R \sim \text{Bernoulli}(2\rho)$ is generated independently from (X, Y), and $Z \sim \text{Unif}\{\pm 1\}$ is generated independently from (X, Y, R). Let (X_i, Y_i, R_i, Z_i) generate the n i.i.d. data points. Furthermore, define

$$\bar{X} = X \cdot \min\left\{1, \frac{4\mathbb{E}\left[\|X\|\right]}{\|X\|}\right\}.$$

and

$$\bar{X}_i = X_i \cdot \min\left\{1, \frac{4\mathbb{E}\left[\|X\|\right]}{\|X_i\|}\right\}.$$

Then we can check that, for all $u \in \mathbb{S}^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\} \geq \frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -\bar{X}_i^\top u \cdot \tilde{Y}_i\right\} \geq \frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -\bar{X}_i^\top u \cdot R_i \cdot Z_i\right\}.$$

Define

$$\Delta = \sup_{u\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -\bar{X}_i^\top u \cdot R_i \cdot Z_i\right\} - \mathbb{E}\left[\max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\}\right]\right|.$$

We can verify that, since $\bar{X}, R, Z$ are independent, by definition of their distributions we have

$$\mathbb{E}\left[\max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\}\right] \geq \rho \cdot \mathbb{E}\left[|\bar{X}^\top u|\right].$$

Furthermore, by Jensen's inequality,

$$\begin{aligned}\exp\left\{-4a_2\mathbb{E}\left[|\bar{X}^\top u|\right]\right\} \leq \mathbb{E}\left[e^{-4a_2|\bar{X}^\top u|}\right] &\leq \mathbb{E}\left[e^{-4a_2|X^\top u|}\right] + \mathbb{P}\left\{\|X\| > 4\mathbb{E}\left[\|X\|\right]\right\} \\ &\leq \frac{a_2}{4a_2} + \frac{\mathbb{E}\left[\|X\|\right]}{4\mathbb{E}\left[\|X\|\right]} = \frac{1}{2},\end{aligned}$$

where the last inequality applies Assumption 2 together with Markov's inequality. Rearranging terms, then,

$$\mathbb{E}\left[|\bar{X}^\top u|\right] \geq \frac{\log 2}{4a_2}.$$

Therefore, combining everything we have shown so far, it holds deterministically that

$$\inf_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -X_i^\top u \cdot \tilde{Y}_i\right\}\right\} \geq \rho \cdot \frac{\log 2}{4a_2} - \Delta.$$

Now we need to bound Δ with high probability.

By the symmetrization inequality [Koltchinskii, 2011, Theorem 2.1] we have

$$\mathbb{E}\left[\Delta\right] \leq 2\mathbb{E}\left[\sup_{u\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot\max\left\{0,-\bar{X}_i^\top u\cdot R_i\cdot Z_i\right\}\right|\right],$$

where the last expectation is taken with respect to the i.i.d. data $(\bar{X}_i, \widetilde{Y}_i)$ as well as i.i.d. Rademacher random variables $\xi_i \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$. Since $t \mapsto \max\{0, -t\}$ is 1-Lipschitz, the contraction inequality [Koltchinskii, 2011, Theorem 2.2] verifies that

$$\mathbb{E}\left[\Delta\right] \leq 4\mathbb{E}\left[\sup_{u\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot\bar{X}_i^\top u\cdot R_i\cdot Z_i\right|\right].$$

Furthermore, deterministically we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot\bar{X}_i^\top u\cdot R_i\cdot Z_i\right| = \left|u^\top\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot R_i\cdot Z_i\cdot\bar{X}_i\right)\right| \leq \left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot R_i\cdot Z_i\cdot\bar{X}_i\right\|,$$

and so combining everything so far, we have shown that

$$\mathbb{E}\left[\Delta\right] \leq 4\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot R_i\cdot Z_i\cdot\bar{X}_i\right\|\right].$$

Moreover, we can see that $(\bar{X}_i, \xi_i\cdot Z_i)$ is equal in distribution to $(\bar{X}_i, \xi_i)$ (since $Z_i \in \{\pm 1\}$ while $\xi_i \sim \text{Unif}\{\pm 1\}$ is drawn independently from the data), and so

$$\mathbb{E}\left[\Delta\right] \leq 4\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot\bar{X}_i\cdot R_i\right\|\right].$$

Finally,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot\bar{X}_i\cdot R_i\right\|\right]^2 \leq \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot\bar{X}_i\cdot R_i\right\|^2\right] = \frac{1}{n^2}\sum_{j=1}^{d}\mathbb{E}\left[\left(\sum_{i=1}^{n}\bar{X}_{ij}R_i\xi_i\right)^2\right]$$

$$= \frac{1}{n^2} \sum_{j=1}^{d} \sum_{i=1}^{n} \mathbb{E}\left[\bar{X}_{ij}^2 R_i^2\right] = \frac{1}{n^2} \sum_{i=1}^{n} 2\rho \mathbb{E}\left[\|\bar{X}_i\|^2\right] \leq \frac{1}{n} \cdot 16\mathbb{E}\left[\|X\|\right]^2 \cdot 2\rho,$$

since by definition, it holds deterministically that $\|\bar{X}_i\| \leq 4\mathbb{E}\left[\|X\|\right]$, while $R_i \sim \text{Bernoulli}(2\rho)$ is independent from X_i. Combining everything so far,

$$\mathbb{E}\left[\Delta\right] \leq 4\sqrt{\frac{1}{n} \cdot 16\mathbb{E}\left[\|X\|\right]^2 \cdot 2\rho}.$$

Next, since for all $u \in \mathbb{S}^{d-1}$ we have

$$\mathbb{E}\left[\max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\}^2\right] \leq 2\rho \cdot \left(4\mathbb{E}\left[\|X\|\right]\right)^2$$

and

$$0 \leq \max\left\{0, -\bar{X}^\top u \cdot R \cdot Z\right\} \leq 4\mathbb{E}\left[\|X\|\right] \text{ almost surely,}$$

applying [Koltchinskii, 2011, Bousquet bound, Section 2.3] yields the concentration result

$$\mathbb{P}\left\{\Delta \leq \mathbb{E}\left[\Delta\right] + \sqrt{\frac{2\log(3n^\alpha) \cdot \left(2\rho \cdot 16\mathbb{E}[\|X\|]^2 + 4\mathbb{E}[\|X\|] \cdot 2\mathbb{E}[\Delta]\right)}{n}} + 4\mathbb{E}\left[\|X\|\right] \cdot \frac{\log(3n^\alpha)}{3n}\right\} \geq 1 - \frac{1}{3n^\alpha}.$$

Furthermore, Assumption 2 together with Jensen's inequality implies

$$e^{a_0 \mathbb{E}[\|X\|^2]/d} \leq e^{a_0 \max_{1\leq j\leq d} \mathbb{E}[|X_j|^2]} \leq \max_{1\leq j\leq d} \mathbb{E}\left[e^{a_0 |X_j|^2}\right] \leq a_1$$

and so $\mathbb{E}\left[\|X\|\right] \leq \mathbb{E}\left[\|X\|^2\right]^{1/2} \leq \sqrt{\frac{d \log a_1}{a_0}}$. Combined with our bound on $\mathbb{E}\left[\Delta\right]$, we can

verify that this bound can be relaxed to

$$\mathbb{P}\left\{\Delta \leq r'\left(\sqrt{\rho \cdot \frac{d\log n}{n}} + \frac{d\log n}{n}\right)\right\} \geq 1 - \frac{1}{3n^{\alpha}}$$

where r' is chosen appropriately as a function of α, a_0, and a_1. Therefore, we have shown that with probability at least $1 - \frac{1}{3n^{\alpha}}$,

$$\inf_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -X_i^{\top}u\cdot\tilde{Y}_i\right\}\right\} \geq \rho\cdot\frac{\log 2}{4a_2} - r'\left(\sqrt{\rho\cdot\frac{d\log n}{n}} + \frac{d\log n}{n}\right),$$

which is sufficient to verify (2.6) with r_1, r_2 chosen appropriately, since it holds that $\sqrt{\rho\cdot\frac{d\log n}{n}} \leq \frac{r''\rho}{2} + \frac{d\log n}{2r''n}$ for all $r'' > 0$.

Next we prove (2.7). Note that, comparing the two terms in the desired upper bound and noting that $1/t$ is only dominant if $t \leq \sqrt{\frac{n}{d\log n}}$, we can see that it suffices to prove the result for $t \leq \sqrt{\frac{n}{d\log n}}$, since $t \mapsto \sup_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-t|X_i^{\top}u|}\right\}$ is monotone nonincreasing in t.

We have

$$\sup_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-t|X_i^{\top}u|}\right\} \leq \sup_{u\in\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-t|\bar{X}_i^{\top}u|}\right\},$$

where, changing the definition of $\bar{X}$ and $\bar{X}_i$, we let

$$\bar{X} = X\cdot\min\left\{1, \frac{t\mathbb{E}\left[\|X\|\right]}{\|X\|}\right\}.$$

and analogously

$$\bar{X}_i = X_i\cdot\min\left\{1, \frac{t\mathbb{E}\left[\|X\|\right]}{\|X_i\|}\right\}.$$

Next fix $\epsilon > 0$, and take a covering $u_1, \ldots, u_M$ of $\mathbb{S}^{d-1}$ such that

$$\sup_{u\in\mathbb{S}^{d-1}}\left\{\min_{m=1,\ldots,M}\|u - u_m\|\right\} \leq \epsilon.$$

By [Lorentz et al., 1996, Chapter 15], for any $\epsilon > 0$ we can construct a set with this property of size $M \leq (3/\epsilon)^d$. Then for any $u \in \mathbb{S}^{d-1}$, if we find m such that $\|u - u_m\| \leq \epsilon$, we have

$$e^{-t|\bar{X}_i^\top u|} \leq e^{-t|\bar{X}_i^\top u_m|} + t\|\bar{X}_i\| \cdot \epsilon \leq e^{-t|\bar{X}_i^\top u_m|} + t^2 \mathbb{E}\left[\|X\|\right] \cdot \epsilon,$$

since $e^{-t|x|}$ is t-Lipschitz over $x \in \mathbb{R}$. Therefore,

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leq t^2 \mathbb{E}\left[\|X\|\right] \cdot \epsilon + \max_{m=1,\ldots,M} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|\bar{X}_i^\top u_m|} \right\}.$$

Next, for each m, by Hoeffding's inequality,

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|\bar{X}_i^\top u_m|} - \mathbb{E}\left[e^{-t|\bar{X}^\top u_m|}\right] > \sqrt{\frac{\log(3Mn^\alpha)}{2n}} \right\} \leq \frac{1}{3Mn^\alpha}.$$

Furthermore,

$$\mathbb{E}\left[e^{-t|\bar{X}^\top u_m|}\right] \leq \mathbb{E}\left[e^{-t|X^\top u_m|}\right] + \mathbb{P}\left\{\|X\| > t\mathbb{E}\left[\|X\|\right]\right\} \leq \frac{a_2 + 1}{t},$$

by applying Assumption 2 together with Markov's inequality. Therefore, combining everything, with probability at least $1 - \frac{1}{3n^\alpha}$,

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leq t^2 \mathbb{E}\left[\|X\|\right] \cdot \epsilon + \sqrt{\frac{\log(3 \cdot (3/\epsilon)^d \cdot n^\alpha)}{2n}} + \frac{a_2 + 1}{t}.$$

Since we have assumed that $t \leq n$, taking $\epsilon = n^{-2.5}$ we obtain

$$\sup_{u \in \mathbb{S}^{d-1}} \left\{ \frac{1}{n} \sum_{i=1}^n e^{-t|X_i^\top u|} \right\} \leq \frac{\mathbb{E}\left[\|X\|\right]}{\sqrt{n}} + \sqrt{\frac{\log(3 \cdot (3n^{2.5})^d \cdot n^\alpha)}{2n}} + \frac{a_2 + 1}{t},$$

which clearly satisfies (2.7) with r_3, r_4 chosen appropriately, since as shown before, $\mathbb{E}\left[\|X\|\right] \leq \sqrt{\frac{d \log a_1}{a_0}}$.

Finally we prove (2.8). We first bound the quantity in the expected value. We have

$$\mathbb{E}\left[\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]=\mathbb{E}\left[\sup_{\|w\|\le r}\left|\frac{1}{n}\sum_{i=1}^n\left(\ell(X_i^\top w\cdot\widetilde{Y}_i)-\mathbb{E}\left[\ell(X_i^\top w\cdot\widetilde{Y}_i)\right]\right)\right|\right]$$
$$\le 2\mathbb{E}\left[\sup_{\|w\|\le r}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\ell(X_i^\top w\cdot\widetilde{Y}_i)\right|\right],$$

by the symmetrization inequality [Koltchinskii, 2011, Theorem 2.1], where the last expectation is taken with respect to the i.i.d. data $(\bar{X}_i,\widetilde{Y}_i)$ as well as i.i.d. Rademacher random variables $\xi_i\overset{\text{iid}}{\sim}\text{Unif}\{\pm1\}$. Next, the contraction inequality [Koltchinskii, 2011, Theorem 2.2] verifies that

$$\mathbb{E}\left[\sup_{\|w\|\le r}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\ell(X_i^\top w\cdot\widetilde{Y}_i)\right|\right]\le 2L\mathbb{E}\left[\sup_{\|w\|\le r}\left|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i^\top w\cdot\widetilde{Y}_i\right|\right],$$

since ℓ is L-Lipschitz by Assumption 1. Furthermore, deterministically we have

$$\left|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i^\top w\cdot\widetilde{Y}_i\right|=\left|w^\top\left(\frac{1}{n}\sum_{i=1}^n\xi_i\cdot\widetilde{Y}_i\cdot X_i\right)\right|\le\|w\|\cdot\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot\widetilde{Y}_i\cdot X_i\right\|,$$

and so combining everything so far, we have shown that

$$\mathbb{E}\left[\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]\le 4Lr\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot\widetilde{Y}_i\cdot X_i\right\|\right].$$

Moreover, we can see that $(X_i,\xi_i\cdot\widetilde{Y}_i)$ is equal in distribution to (X_i,ξ_i) (since $\widetilde{Y}_i\in\{\pm1\}$ while $\xi_i\sim\text{Unif}\{\pm1\}$ is drawn independently from $(X_i,\widetilde{Y}_i)$), and so

$$\mathbb{E}\left[\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^\rho(w)-\widetilde{\mathcal{L}}^\rho(w)\right|\right]\le 4Lr\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\cdot X_i\right\|\right].$$

Finally,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot X_i\right\|\right]^2 \le \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\cdot X_i\right\|^2\right] = \frac{1}{n^2}\sum_{j=1}^{d}\mathbb{E}\left[\left(\sum_{i=1}^{n}X_{ij}\xi_i\right)^2\right]$$
$$= \frac{1}{n^2}\sum_{j=1}^{d}\sum_{i=1}^{n}\mathbb{E}\left[X_{ij}^2\right] = \frac{1}{n}\mathbb{E}\left[\|X\|^2\right] \le \frac{d}{n}\cdot\frac{\log a_1}{a_0},$$

since $\mathbb{E}\left[\|X\|^2\right] \le \frac{d\log a_1}{a_0}$ as calculated above. Therefore,

$$\mathbb{E}\left[\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^{\rho}(w) - \widetilde{\mathcal{L}}^{\rho}(w)\right|\right] \le \frac{4Lr\sqrt{\log a_1}}{\sqrt{a_0}}\cdot\sqrt{\frac{d}{n}}.$$

Next we prove that the quantity $\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^{\rho}(w) - \widetilde{\mathcal{L}}^{\rho}(w)\right|$ concentrates around its expectation. First, let $(X', \widetilde{Y}')$ be an i.i.d. draw from the distribution of $(X, \widetilde{Y})$. For $\lambda \ge 0$, we calculate

$$\mathbb{E}\left[\frac{1}{2}e^{\lambda\|X\widetilde{Y}-X'\widetilde{Y}'\|} + \frac{1}{2}e^{-\lambda\|X\widetilde{Y}-X'\widetilde{Y}'\|}\right] \le \mathbb{E}\left[e^{\lambda^2\|X\widetilde{Y}-X'\widetilde{Y}'\|^2/2}\right]$$
$$\le \mathbb{E}\left[e^{\lambda^2\cdot(\|X\widetilde{Y}\|^2+\|X'\widetilde{Y}'\|^2)}\right] = \mathbb{E}\left[e^{\lambda^2\cdot\|X\widetilde{Y}\|^2}\right]^2 = \mathbb{E}\left[e^{\lambda^2\cdot\|X\|^2}\right]^2$$
$$= \mathbb{E}\left[e^{\lambda^2\cdot\sum_{j=1}^{d}|X_j|^2}\right]^2 \le \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d}e^{d\lambda^2\cdot|X_j|^2}\right]^2,$$

by the AM–GM inequality. Applying Assumption 2, we then obtain

$$\mathbb{E}\left[\frac{1}{2}e^{\lambda\|X\widetilde{Y}-X'\widetilde{Y}'\|} + \frac{1}{2}e^{-\lambda\|X\widetilde{Y}-X'\widetilde{Y}'\|}\right] \le a_1^{\frac{2\lambda^2 d}{a_0}}$$

as long as $\lambda^2 \le a_0/d$. Following the proof of [Kontorovich, 2014, Theorem 1], since $\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^{\rho}(w) - \widetilde{\mathcal{L}}^{\rho}(w)\right|$ is a $\frac{Lr}{n}$-Lipschitz function of each data point product $X_i\cdot\widetilde{Y}_i$,

$$\mathbb{P}\left\{\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^{\rho}(w) - \widetilde{\mathcal{L}}^{\rho}(w)\right| - \mathbb{E}\left[\sup_{\|w\|\le r}\left|\widetilde{\mathcal{L}}_n^{\rho}(w) - \widetilde{\mathcal{L}}^{\rho}(w)\right|\right] > \frac{Lr}{n}\cdot\sqrt{\frac{8nd\log a_1\cdot\log(3n^{\alpha})}{a_0}}\right\}$$

$$\leq \exp\left\{\frac{2n\lambda^2 d\log a_1}{a_0} - \lambda\cdot\sqrt{\frac{8nd\log a_1\cdot\log(3n^\alpha)}{a_0}}\right\}.$$

Taking

$$\lambda = \frac{a_0}{4nd\log a_1}\cdot\sqrt{\frac{8nd\log a_1\cdot\log(3n^\alpha)}{a_0}}$$

(which clearly satisfies $\lambda \leq \sqrt{\frac{a_0}{d}}$ for sufficiently large n), this probability is bounded by $\frac{1}{3n^\alpha}$. (If instead n is not sufficiently large (i.e., $\lambda > \sqrt{\frac{a_0}{d}}$), then the guarantee (2.8) holds trivially.) Combining everything, and choosing r_5 appropriately, we have established (2.8).

CHAPTER 3

DISTRIBUTION-FREE INFERENCE FOR REGRESSION: DISCRETE, CONTINUOUS, AND IN BETWEEN

3.1 Introduction

Consider a regression problem, where our aim is to model the distribution of a response variable $Y \in \mathbb{R}$ based on the information carried by features $X \in \mathcal{X}$. Given training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we aim to build a fitted model to estimate the conditional distribution of $Y \mid X$, or some summary of this distribution such as the conditional mean or conditional median. In this type of setting, our goals are to simultaneously perform two tasks, estimation and inference—that is, we want to accurately estimate the conditional distribution, and we also want a reliable way of quantifying our uncertainty about this estimate.

To make this concrete, suppose the training data $\{(X_i, Y_i)\}$ are drawn i.i.d. from some unknown distribution P on $\mathbb{R}^d \times \mathbb{R}$, and we want to estimate the true conditional mean, $\mu_P(x) := \mathbb{E}[Y|X = x]$, of this distribution. Given the training data, we construct a fitted regression function $\widehat{\mu} : \mathbb{R}^d \to \mathbb{R}$ using any algorithm, for instance, a parametric method such as least squares or a nonparametric procedure such as a Gaussian kernel method. For many regression algorithms, assuming certain conditions on the underlying distribution P will ensure an accurate estimate of μ_P; however, unless we are able to verify these assumptions, we cannot be confident that the corresponding error rates will indeed lead to a valid confidence interval for μ_P. The goal of *distribution-free inference* is to provide inference guarantees—in this case, confidence intervals for $\mu_P(X_{n+1})$ at a newly observed feature vector X_{n+1}—that are valid universally over any underlying distribution P.

3.1.1 Our contributions

In this work, we study the problem of constructing a confidence interval $\widehat{C}_n(x)$ for $\mu_P(x)$, that satisfies the following property:

Definition 1. *An algorithm $\widehat{C}_n$ provides a distribution-free $(1-\alpha)$-confidence interval for the conditional mean if it holds that*

$$\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \text{ for all distributions } P \text{ on } (X,Y) \in \mathbb{R}^d \times [0,1].$$

Here the probability is taken with respect to the distribution of both the training data $(X_1,Y_1),\ldots,(X_n,Y_n)$ and the test point (X_{n+1},Y_{n+1}), all drawn i.i.d. from an arbitrary P.[1]

Recent work by [Vovk et al., 2005, Barber, 2020, Gupta et al., 2020] (studying the conditional mean of a binary response Y) and by [Medarametla and Candès, 2021] (studying the conditional median of a real-valued Y) proves that distribution-free coverage properties similar to Definition 1 lead to fundamental limits on the accuracy of inference. Writing P_X to denote the marginal distribution of X under P, these results show that if P_X is *nonatomic* (meaning that there are no point masses, i.e., $\mathbb{P}_{P_X}\{X=x\}=0$ for all points $x \in \mathbb{R}^d$), then any distribution-free confidence interval $\widehat{C}_n$ cannot have vanishing length as sample size n tends to infinity, regardless of the smoothness of P, or any other "nice" properties of this distribution. Specifically, these works show that if P_X is nonatomic, then $\widehat{C}_n$ must also be a valid predictive interval, i.e., must contain Y_{n+1} itself with probability $\geq 1-\alpha$. This implies that the length of $\widehat{C}_n$ cannot be vanishing, since Y_{n+1} is inherently noisy. An explicit lower bound on the length is proved in [Barber, 2020].

Our new results examine the possibility of constructing confidence intervals $\widehat{C}_n$ that are

1. In this definition and throughout our work, $\widehat{C}_n$ can be either a deterministic or randomized function of the training data; if the construction is randomized then the definition above should be interpreted as computing probability with respect to the distribution of the data and the randomization of the construction.

both distribution-free (Definition 1) and have vanishing length, when P_X may be discrete, nonatomic, or a mixture of the two. We find that the hardness of this problem can be characterized by the *effective support size* of P_X—essentially, how many points $x \in \mathbb{R}^d$ are needed to capture most of the mass of P_X (for example, if P_X is uniform over M points, then its effective support size is $\leq M$).

Our main theoretical results show that there are two regimes. If the effective support size is $\gg n^2$, then P_X essentially behaves like a nonatomic distribution because in a sample of size n, with high probability all the X values are observed at most once; in this regime, we find that the average length of $\widehat{C}_n(X_{n+1})$ is bounded away from zero, i.e., no distribution-free confidence interval can have vanishing length. If instead the effective support size is $\ll n^2$, then it becomes possible for $\widehat{C}_n(X_{n+1})$ to have vanishing length, and in particular, the minimum possible length scales as $\frac{M^{1/4}}{n^{1/2}}$ for effective support size M. Interestingly, vanishing length is possible even when M is larger than n, meaning that distribution-free inference for $\mathbb{E}[Y|X]$ is possible even if most X values were never observed in the training set.

3.1.2 Additional related work

The problem of distribution-free inference has been studied extensively in the context of *predictive inference*, where the goal is to provide a confidence band for the response value Y_{n+1} given a new feature vector X_{n+1}. The prediction problem is fundamentally different from the goal of covering the conditional mean. In particular, by splitting the data and using a holdout set, we can always empirically validate the coverage level of any constructed predictive band. Methods such as conformal prediction (see, e.g., [Vovk et al., 2005, Papadopoulos et al., 2002, Lei et al., 2018, Vovk et al., 2018]) or jackknife+ ([Barber et al., 2021b, Kim et al., 2020]) can ensure valid distribution-free predictive inference without the need to split the data set (thus avoiding reducing the sample size).

As mentioned earlier, [Vovk et al., 2005, Barber, 2020, Gupta et al., 2020, Medarametla and Candès, 2021] also study the problem of confidence intervals for the conditional mean or median of $Y|X$, establishing impossibility results on the setting of a nonatomic P_X. These results are connected to earlier results on the impossibility of adaptation to smoothness, in the nonparametric inference literature—specifically, if μ_P is β-Hölder smooth, then it is possible to build a confidence interval of length $\mathcal{O}(n^{-\frac{\beta}{2\beta+d}})$ if β is known (e.g., using k-nearest-neighbors with an appropriately chosen k), but this cannot be achieved when β is unknown (see, e.g., [Giné and Nickl, 2016, Section 8.3] for an overview of results of this type).

While the results above establish the challenges for distribution-free inference when the features X are nonatomic, at the other extreme we can consider scenarios where X has a discrete distribution. In this setting, the problem of estimating μ_P is related to the *discrete distribution testing*, where the aim is to test properties of a discrete distribution—for instance, we might wish to test equality of two distributions where we draw samples from each [Chan et al., 2014, Acharya et al., 2014, Diakonikolas and Kane, 2016, Canonne et al., 2015]; to test whether a sample is drawn from a known distribution P or not [Diakonikolas and Kane, 2016, Acharya et al., 2015, Valiant and Valiant, 2017, Diakonikolas et al., 2018], or drawn from any distribution belonging to a class $\mathcal{P}$ or not [Acharya et al., 2015, Canonne et al., 2018]; or to estimate certain characteristics of a distribution such as its entropy or support size [Valiant and Valiant, 2011b,a, Acharya et al., 2014]. The distribution-free confidence intervals we will construct in Section 3.3 are closely related to methods developed in this literature.

3.2 Main results: lower bound

Before presenting our main result, we begin with several definitions. For any distribution P_X on $X \in \mathbb{R}^d$, we first define the *effective support size* of P_X at tolerance level $\gamma \in [0, 1)$:

$$M_\gamma(P_X) = \min\left\{|\mathcal{X}| \,:\, \mathcal{X} \subset \mathbb{R}^d \text{ and } \mathbb{P}_{P_X}\left\{X \in \mathcal{X}\right\} \geq 1 - \gamma\right\},$$

where $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. In particular, if P_X is a distribution supported on M points, then $M_\gamma(P_X) \leq M$ for any γ. If instead P_X is nonatomic, then $M_\gamma(P_X) = \infty$ for all $\gamma > 0$. We note that, in many practical settings, the effective support size $M_\gamma(P_X)$ may be substantially smaller than the overall support size. For example, if $X \in \mathbb{R}^d$ measures d categorical covariates with m_j possible values for the jth covariate, then the support of P_X is potentially as large as $\prod_{j=1}^d m_j$, which will grow extremely rapidly with the dimension d even if each m_j is small; in real data, however, it may be the case that most combinations of covariate values are extremely unlikely, and so the effective support size $M_\gamma(P)$ would be substantially smaller, and might grow more slowly with d.

Next, for any distribution P on $(X, Y) \in \mathbb{R}^d \times [0, 1]$, we define

$$\sigma^2_{P,\beta} = \text{ the } \beta\text{-quantile of } \mathrm{Var}_P\left(Y|X\right), \text{ under the distribution } X \sim P_X.$$

With these definitions in place, our first main result establishes a lower bound on the expected length of any distribution-free confidence interval $\widehat{C}_n$. Let Leb denote the Lebesgue measure on $\mathbb{R}$.

Theorem 2. *Fix any $\alpha > 0$, and let $\widehat{C}_n$ be a distribution-free $(1-\alpha)$-confidence interval (i.e., satisfying Definition 1). Then for any distribution P on $\mathbb{R}^d \times \mathbb{R}$, for any $\beta > 0$ and*

$\gamma > \alpha + \beta$,

$$\mathbb{E}\left[\text{Leb}\left(\widehat{C}_n(X_{n+1})\right)\right] \geq \tfrac{1}{3}\sigma^2_{P,\beta}(\gamma-\alpha-\beta)^{1.5} \cdot \min\left\{\frac{\left(M_\gamma(P_X)\right)^{1/4}}{n^{1/2}}, 1\right\},$$

where the expected value is taken over data points $(X_i, Y_i) \overset{\text{iid}}{\sim} P$*, for* $i = 1, \ldots, n+1$.

3.2.1 Special cases

To help interpret this result, we now examine its implications in several special cases.

Uniform discrete features If P_X is a uniform distribution over M points, then for any $\gamma > 0$ the effective support size is $M_\gamma(P_X) = \lceil (1-\gamma)M \rceil$. Therefore, Theorem 2 implies that for any P with nonatomic marginal P_X,

$$\mathbb{E}\left[\text{Leb}\left(\widehat{C}_n(X_{n+1})\right)\right] \geq \tfrac{1}{3}\sigma^2_{P,\beta}(\gamma-\alpha-\beta)^{1.5}(1-\gamma)^{0.25} \cdot \min\left\{\frac{M^{1/4}}{n^{1/2}}, 1\right\}$$

for any $\beta \in (0, \gamma - \alpha)$. In particular, we see that $M \gg n^2$ implies a *constant* lower bound on the width of any distribution-free confidence interval, while $M \ll n^2$ allows for the possibility of a *vanishing* width for a distribution-free confidence interval.

Binary response If the response Y is known to be binary (i.e., $Y \in \{0, 1\}$), we might relax the requirement of distribution-free coverage to only include distributions of this type, i.e., we require

$$\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \text{ for all distributions } P \text{ on } \mathbb{R}^d \times \{0,1\}. \quad (3.1)$$

This condition is strictly weaker than Definition 1, where the coverage property is required to hold for all distributions P on $\mathbb{R}^d \times [0, 1]$, i.e., for a broader class of distributions. However, it turns out that relaxing the requirement does not improve the lower bound. Specifically, if we

have an algorithm to construct a confidence interval $\widehat{C}_n$ satisfying (3.1), then we can easily convert $\widehat{C}_n$ into a method that does satisfy Definition 1. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, for each $i = 1, \ldots, n$ draw a binary response $\tilde{Y}_i \sim \text{Bernoulli}(Y_i)$. Then we clearly have n i.i.d. draws from a distribution on $(X, \tilde{Y}) \in \mathbb{R}^d \times \{0, 1\}$, where $\mathbb{E}[\tilde{Y} \mid X] = \mathbb{E}[Y \mid X] = \mu_P(X)$. After running our algorithm to construct $\widehat{C}_n$ on the new data $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$, the binary distribution-free coverage property (3.1) satisfied by $\widehat{C}_n$ ensures that this modified procedure satisfies Definition 1.

To summarize, then, we see that the problem of distribution-free coverage is equally hard for the binary response case ($Y \in \{0, 1\}$) as for the more general bounded response case ($Y \in [0, 1]$).

Nonatomic features We now consider the setting where the marginal distribution of X is nonatomic, i.e., $\mathbb{P}_{P_X}\{X = x\} = 0$ for all x. (In particular, this includes the continuous case, where X has a continuous distribution on $\mathbb{R}^d$.) In this case, for any $\gamma > 0$ the effective support size is $M_\gamma(P_X) = \infty$. Therefore, Theorem 2 implies that for any P with nonatomic marginal P_X, for any $\beta \in (0, 1 - \alpha)$,

$$\mathbb{E}\left[\text{Leb}\left(\widehat{C}_n(X_{n+1})\right)\right] \geq \tfrac{1}{3}\sigma^2_{P,\beta}(1 - \alpha - \beta)^{1.5}.$$

In particular, this lower bound does not depend on n, and so the width of any distribution-free confidence interval is non-vanishing even for arbitrarily large sample size n (as long as $\sigma^2_{P,\beta} > 0$).

In case of a binary response, where P is a distribution on $\mathbb{R}^d \times \{0, 1\}$ with nonatomic marginal distribution P_X, [Barber, 2020] establishes that any distribution-free confidence interval for μ must satisfy a lower bound that is a function only of P and does not depend on n (and, in particular, does not vanish as $n \to \infty$). In this sense, our new result can be viewed as a generalization of this work, since the nonvanishing minimum length for nonatomic

P_X is a consequence of our result.

3.2.2 Adding knowledge of P_X

One way we might try to weaken the notion of distribution-free coverage would be to allow assumptions about the marginal distribution P_X, while remaining assumption-free for the function μ_P determining the conditional mean. In other words, we might weaken Definition 1 to require coverage over all distributions P for which $P_X = P_X^*$, for a known P_X^* (or, all P for which P_X satisfies some assumed property). Interestingly, the lower bound in Theorem 2 remains the same even under this milder definition of validity—we will see in the proof that knowledge of P_X does not affect the lower bound, since the argument relies only on our uncertainty about the conditional distribution of $Y|X$.

3.2.3 Bounded or unbounded?

The lower bound established in Theorem 2 assumes distribution-free coverage for distributions with a bounded response Y—that is, Definition 1 requires coverage to hold for distributions where the response Y is supported on $[0, 1]$ (although no other assumptions are placed on P). Would it be possible for us to instead consider the general case, where P is an unknown distribution on $\mathbb{R}^d \times \mathbb{R}$? The following result shows that this more general question is not meaningful:

Proposition 1. *Suppose an algorithm $\widehat{C}_n$ satisfies*

$$\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \text{ for all distributions } P \text{ on } \mathbb{R}^d \times \mathbb{R}.$$

Then for all distributions P, for all $y \in \mathbb{R}$ it holds that

$$\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{y \in \widehat{C}_n(X_{n+1}))\right\} \geq 1-\alpha.$$

This means that if we require $\widehat{C}_n$ to have distribution-free coverage over distributions with *unbounded* response, then inevitably, *every point in the real line* is contained in the resulting confidence interval a substantial portion of the time. (In particular, $\widehat{C}_n(X_{n+1})$ will of course have infinite expected width.) Clearly an unbounded Y cannot result in any meaningful distribution-free inference, and for this reason we therefore restrict our attention to the setting where the response Y takes values in $[0, 1]$ (of course, these results can easily generalize to $Y \in [a, b]$ for any known $a < b$).

3.3 Main results: upper bound

We next construct an algorithm that, for certain "nice" distributions P, can achieve a confidence interval length that matches the rate of the lower bound. Our procedure requires two main ingredients as input:

1. A hypothesized ordered support set $\{x^{(1)}, x^{(2)}, \dots\} \subset \mathbb{R}^d$ for the marginal P_X, and
2. A hypothesized mean function $\mu : \mathbb{R}^d \to [0, 1]$.

One possible way of obtaining these inputs would be to use data splitting, where one portion of our data (combined with prior knowledge if available) is used to construct a hypothesized support set and mean function, and the second portion of the data is then used for constructing the confidence interval (note that the sample size n in our construction below refers to the size of this second part of the data, e.g., half of the total available sample size). Any algorithm can be applied for estimating μ, for example, logistic regression, nearest neighbors regression, or a neural network.

We emphasize that the coverage guarantee provided by our method does not rely in any way on the accuracy of these initial guesses—the constructed confidence interval will satisfy distribution-free validity (Definition 1) even if these initial parameters are chosen in a completely uninformed way. In particular, while the algorithm that fits μ might be able to

guarantee accuracy of μ under some assumptions placed on P, the validity of our inference procedure does not rely on these assumptions. However, the length of the resulting confidence interval will be affected, since high accuracy in these initial guesses can be expected to result in a shorter confidence interval. In particular, the hypothesized support set $\{x^{(1)}, x^{(2)}, \dots\}$ should aim to list the highest-probability values of X early in the list, while the hypothesized mean function μ should aim to be as close to the true conditional mean μ_P as possible. (Our theoretical results below will make these goals more precise.)

Given the hypothesized support and hypothesized mean function, to run our algorithm, we first choose parameters $\gamma, \delta > 0$ satisfying $\gamma + \delta < \alpha$, and then compute the following steps.

- **Step 1: estimate the effective support size.** First, we compute an upper bound on the support size needed to capture $1 - \gamma$ of the probability under P_X,

$$\widehat{M}_\gamma = \min\left\{ m : \sum_{i=1}^{n} \mathbb{1}\left\{X_i \in \{x^{(1)}, \dots, x^{(m)}\}\right\} \geq (1-\gamma)n + \sqrt{\frac{n\log(2/\delta)}{2}} \right\},$$

  or $\widehat{M}_\gamma = \infty$ if there is no m that satisfies the inequality. Applying the Hoeffding inequality to the Binom(n, γ) distribution, we see that $\mathbb{P}\left\{\widehat{M}_\gamma \geq M^*_\gamma(P_X)\right\} \geq 1 - \delta/2$, where

$$M^*_\gamma(P_X) = \min\left\{ m : \mathbb{P}_{P_X}\left\{X \in \{x^{(1)}, \dots, x^{(m)}\}\right\} \geq 1 - \gamma \right\}. \tag{3.2}$$

  (Note that $M^*_\gamma(P_X) \geq M_\gamma(P_X)$ by definition.)

- **Step 2: estimate error at each repeated X value.** Next, for each $m = 1, 2, \dots,$ let $n_m = \sum_{i=1}^{n} \mathbb{1}\left\{X_i = x^{(m)}\right\}$ denote the number of times $x^{(m)}$ was observed, and let

$$N_{\geq 2} = \sum_{m \geq 1} \mathbb{1}\left\{n_m \geq 2\right\} \tag{3.3}$$

  be the number of X values observed at least twice. For each m with $n_m \geq 2$, let

$\bar{y}_m = \frac{1}{n_m}\sum_{i=1}^n Y_i \cdot \mathbb{1}\left\{X_i = x^{(m)}\right\}$ and $s_m^2 = \frac{1}{n_m - 1}\sum_{i=1}^n (Y_i - \bar{y}_m)^2 \cdot \mathbb{1}\left\{X_i = x^{(m)}\right\}$ be the sample mean and sample variance of the corresponding Y values. Define

$$Z = \sum_{\substack{m=1,2,\ldots \\ \text{s.t. } n_m \geq 2}} (n_m - 1) \cdot \left((\bar{y}_m - \mu(x^{(m)}))^2 - n_m^{-1} s_m^2\right). \tag{3.4}$$

This construction is inspired by analogous statistics appearing in the literature for testing properties of discrete distributions—for instance, the work of [Chan et al., 2014]. To see the intuition behind this construction, we observe that $\{Y_i : X_i = x^{(m)}\}$ is a collection of i.i.d. observations with mean $\mu_P(x^{(m)})$. Therefore, conditional on n_m (with $n_m \geq 2$),

$$\mathbb{E}\left[\bar{y}_m\right] = \mu_P(x^{(m)}) \text{ and } \mathrm{Var}\left(\bar{y}_m\right) = n_m^{-1}\mathbb{E}\left[s_m^2\right],$$

and therefore $\mathbb{E}\left[(\bar{y}_m - \mu(x^{(m)}))^2 - n_m^{-1}s_m^2\right] = (\mu(x^{(m)}) - \mu_P(x^{(m)}))^2$ is an estimate of our error at this X value.

- **Step 3: construct the confidence interval.** Finally, we define our confidence interval. Let

$$\widehat{\Delta} = \sqrt{\frac{2\widehat{M}_\gamma + n}{n(n-1)}} \cdot \sqrt{4Z_+ + 8\sqrt{N_{\geq 2}/\delta} + 24/\delta},$$

where Z_+ denotes $\max\{Z, 0\}$. Then for each $x \in \mathbb{R}^d$, we define

$$\widehat{C}_n(x) = \left[\max\left\{0, \mu(x) - \frac{\widehat{\Delta}}{\alpha - \delta - \gamma}\right\}, \min\left\{1, \mu(x) + \frac{\widehat{\Delta}}{\alpha - \delta - \gamma}\right\}\right]. \tag{3.5}$$

We now verify that this construction yields a valid distribution-free confidence interval.

Theorem 3. *The confidence interval constructed in* (3.5) *is a distribution-free* $(1-\alpha)$*-confidence interval (i.e.,* $\widehat{C}_n$ *satisfies Definition 1).*

Next, we will see how this construction is able to match the rate of the lower bound established in Theorem 2—specifically, in a scenario where the hypothesized support set and mean function are "chosen well", i.e., are a good approximation to the true distribution P. For simplicity, we only consider the case where the marginal P_X is approximately uniform over some finite subset of the hypothesized support, and the hypothesized function μ has uniformly bounded error.

Theorem 4. *Suppose the distribution P on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ has marginal P_X that is supported on $\{x^{(1)}, \ldots, x^{(M)}\}$ and satisfies $\mathbb{P}_{P_X}\left\{X = x^{(m)}\right\} \leq \eta/M$ for all m, and suppose that P has conditional mean $\mu_P : \mathbb{R}^d \to \mathbb{R}$ that satisfies $\mathbb{E}_{P_X}\left[(\mu_P(X) - \mu(X))^2\right] \leq \mathrm{err}_\mu^2$. Then the confidence interval constructed in* (3.5) *satisfies*

$$\mathbb{E}\left[\mathrm{Leb}(\widehat{C}_n(X_{n+1}))\right] \leq c\left(\mathrm{err}_\mu + \frac{M^{1/4}}{n^{1/2}}\right),$$

where c depends only on the parameters α, δ, γ, η.

To see some concrete examples of where this upper bound might be small, suppose that μ is constructed via data splitting (i.e., our initial data set has sample size $2n$, and we use n data points to train μ and then the remaining n to construct the confidence interval). If μ is constructed via logistic regression, and the distribution P follows this model, then under standard conditions on P_X we would have $\mathrm{err}_\mu = \mathcal{O}(\sqrt{d/n})$; in a k-sparse regression setting where we use logistic lasso we might instead obtain $\mathrm{err}_\mu = \mathcal{O}(\sqrt{k\log(d)/n})$ [Negahban et al., 2012]. If instead μ is constructed via k-nearest neighbors, if $x \mapsto \mu_P(x)$ is β-Hölder smooth (and k is chosen appropriately), then as mentioned earlier we have $\mathrm{err}_\mu = \mathcal{O}(n^{-\beta/(\beta+d)})$ [Györfi et al., 2002, Giné and Nickl, 2016].

3.4 Discussion

Our main result of this chapter, Theorem 2, shows that the problem of constructing distribution-free confidence intervals for a conditional mean has hardness characterized by the effective support size $M_\gamma(P_X)$ of the feature distribution; distribution-free confidence intervals may have vanishing length if the sample size is at least as large as the square root of the effective support size, but must have length bounded away from zero if the sample size is smaller. The rate of the lower bound on length, scaling as $\min\{\frac{M_\gamma(P_X))^{1/4}}{n^{1/2}}, 1\}$, is achievable in certain settings—Theorems 3 and 4 establish that distribution-free confidence intervals may achieve this length if we have a good hypothesis μ for μ_P. Of course, the specific construction used for these matching bounds may not be optimal—both in terms of constant factors that may inflate its length, and in terms of the range of settings in which it is able (up to constants) to match the lower bound. Improving this construction to provide a practical and accurate algorithm is an important question for future work.

One counterintuitive implication of our result is that a meaningful distribution-free inference can be achieved even in the case $M_\gamma(P_X) \gg n$, where with high probability, the new observation X_{n+1} is a value that was never observed in the training set. The reason inference is possible in this regime is that the repeated X values in the training set provide some information we need to construct a meaningful confidence interval, and since the set of X values that are repeated is random, this leads to a coverage guarantee (recall that these repeated X values were central to the construction of our confidence interval in Section 3.3). An interesting possible application of this finding is for distribution-free calibration, where the aim is to cover within-bin averages of the form $\mu_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$ where $\mathbb{R}^d = \cup_{b=1,\ldots,B} \mathcal{X}_b$ is a partition into bins. [Gupta et al., 2020] study this problem in the distribution-free setting, and develop methods for guaranteeing coverage of *each* μ_b when the number of bins satisfies $B \ll n$; in contrast, the methods studied in our present work suggest that we may be able

to cover μ_b *on average* over all bins b in the regime $n \ll B \ll n^2$.

Generally, all inference methods must inherently involve a tradeoff between the strength of the guarantees, and the precision of the resulting answers. In this present work, we consider a universally strong guarantee (i.e., coverage of the conditional mean for *all* distributions P), which results in precise inference (i.e., vanishing-length confidence intervals) for only *some* distributions P, namely, those with effective support size $\ll n^2$. This tradeoff may not be desirable in practice, since in an applied setting we might instead prefer to relax the required coverage properties for more challenging distributions P in order to allow for more precise answers. In practice, we may be satisfied with a validity condition that yields weaker guarantees in a nonatomic setting, but still yields the stronger coverage guarantee in the achievable regime where $M_\gamma(P_X) \ll n^2$. In future work, we aim to study whether this more adaptive type of validity definition, which is weaker than distribution-free coverage, may enable us to build confidence intervals that have vanishing length even in the nonatomic setting.

3.5 Appendix

3.5.1 Proof of Proposition 1

To prove this proposition, we will consider replacing P with a distribution that places vanishing probability on some extremely large value.[2] Fix any distribution P, and any $y \in \mathbb{R}$. For any fixed $\epsilon > 0$, define a new distribution Q as follows:

$$\text{Draw } X \sim P_X, \text{ then draw } Y|X \sim (1-\epsilon)P_{Y|X} + \epsilon\delta_{\epsilon^{-1}y-(\epsilon^{-1}-1)\mu_P(X)},$$

2. Similar constructions are used in many related results in the literature—e.g., [Lei and Wasserman, 2014, Lemma 1] proves an analogous infinite-width result for the problem of prediction intervals required to be valid conditional on X_{n+1}, while here we are interested in confidence intervals but only require marginal validity.

where $P_{Y|X}$ is the conditional distribution of $Y|X$ under P, and δ_t denotes the point mass at t. Then we can trivially calculate that $\mathrm{d_{TV}}(P^n \times P_X, Q^n \times Q_X) \leq n\epsilon$. Therefore,

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

On the other hand, the distribution Q has conditional mean

$$\mu_Q(x) = (1-\epsilon)\mu_P(x) + \epsilon\left(\epsilon^{-1}y - (\epsilon^{-1}-1)\mu_P(x)\right) = y,$$

and so the conditional mean $\mu_Q(X_{n+1})$ is equal to y almost surely. Therefore,

$$\mathbb{P}_{Q^n \times Q_X}\left\{y \in \widehat{C}_n(X_{n+1})\right\} = \mathbb{P}_{Q^n \times Q_X}\left\{\mu_Q(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha,$$

where the last step holds since $\widehat{C}_n$ must satisfy distribution-free coverage and, therefore, must satisfy coverage with respect to Q. Since $\epsilon > 0$ is arbitrarily small, this completes the proof.

3.5.2 Proof of Theorem 2

To prove the theorem, we will need several supporting lemmas:

Lemma 4. *Let Q be any distribution on $[0,1]$ with variance σ^2. Then we can write Q as a mixture of two distributions Q_0, Q_1 on $[0,1]$ such that*

$$Q = 0.5Q_0 + 0.5Q_1 \text{ and } \mathbb{E}_{Q_1}[X] - \mathbb{E}_{Q_0}[X] \geq 2\sigma^2.$$

Lemma 5. *Let P_X be any distribution on $\mathbb{R}^d$, and let $\mathbb{R}^d = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots$ be a fixed partition. Define a distribution P_0 on $(X, Z) \in \mathbb{R}^d \times \{0,1\}$ as:*

Draw $X \sim P_X$, and draw $Z \sim$ Bernoulli(0.5), independently from X.

For any fixed sequence $a = (a_1, a_2, \dots)$ *of signs* $a_1, a_2, \dots \in \{\pm 1\}$*, and any fixed* $\epsilon_1, \epsilon_2, \dots \in [0, 0.5]$*, define a distribution* P_a *on on* $(X, Z) \in \mathbb{R}^d \times \{0, 1\}$ *as:*

$$\text{Draw } X \sim P_X, \text{ and conditional on } X, \text{ draw } Z|X \in \mathcal{X}_m \sim \text{Bernoulli}(0.5 + a_m \cdot \epsilon_m).$$

Finally define $\tilde{P}_0 = (P_0)^n$ *(i.e., n i.i.d. draws from* P_0*), and define* $\tilde{P}_1$ *as the mixture distribution:*

- *Draw* $A_1, A_2, \dots \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$.
- *Conditional on* $A_1, A_2, \dots$*, draw* $(X_1, Z_1), \dots, (X_n, Z_n) \overset{\text{iid}}{\sim} P_A$.

Then

$$\mathrm{d_{TV}}(\tilde{P}_0, \tilde{P}_1) \leq 2n \sqrt{\sum_{m=1}^{\infty} \epsilon_m^4 \cdot \mathbb{P}_{P_X}\{X \in \mathcal{X}_m\}^2}.$$

We are now ready to prove the theorem. Define $\mathcal{X}_1 = \{x \in \mathbb{R}^d : \mathbb{P}_{P_X}\{X = x\} > \frac{1}{M_\gamma(P_X)}\}$. We must have $|\mathcal{X}_1| < M_\gamma(P_X)$ since P_X is a probability measure, and therefore, by definition of the effective support size, we must have $\mathbb{P}_{P_X}\{X \in \mathcal{X}_1\} < 1 - \gamma$. On the set $\mathbb{R}^d \backslash \mathcal{X}_1$, any point masses of the distribution P_X must each have probability $\leq 1/M_\gamma(P_X)$, by definition of $\mathcal{X}_1$; P_X may also have a nonatomic component. Applying [Dudley et al., 2011, Proposition A.1], we can partition $\mathbb{R}^d \backslash \mathcal{X}_1$ into countably many sets, $\mathcal{X}_2 \cup \mathcal{X}_3 \cup \dots$, such that $\mathbb{P}_{P_X}\{X \in \mathcal{X}_m\} \leq 1/M_\gamma(P_X)$ for all $m \geq 2$. From this point on, write $p_m = \mathbb{P}_{P_X}\{X \in \mathcal{X}_m\}$.

For each x in the support of P_X, let $P_{Y|X=x}$ denote the conditional distribution of Y given $X = x$. By Lemma 4, we can construct distributions $P^1_{Y|X=x}$ and $P^0_{Y|X=x}$ such that

$$P_{Y|X=x} = 0.5P^1_{Y|X=x} + 0.5P^0_{Y|X=x} \text{ and } \mathbb{E}_{P^1_{Y|X=x}}[Y] - \mathbb{E}_{P^0_{Y|X=x}}[Y] \geq 2\sigma^2_P(x),$$

where $\sigma^2_P(x) = \mathrm{Var}(Y \mid X = x)$ is the variance of $P_{Y|X=x}$.

Next fix any $\epsilon \in (0, 0.5]$. For any vector $a = (a_1, a_2, \dots)$ of signs $a_1, a_2, \dots \in \{\pm 1\}$, define the distribution P_a over (X, Y) as follows:

- Draw $X \sim P_X$, i.e., the same as the marginal distribution of X under P.
- Conditional on X, draw Y as

$$Y \mid X = x \sim \begin{cases} P_{Y|X=x}, & \text{if } x \in \mathcal{X}_1, \\ (0.5 + a_m\epsilon) \cdot P^1_{Y|X=x} + (0.5 - a_m\epsilon) \cdot P^0_{Y|X=x}, & \text{if } x \in \mathcal{X}_m \text{ for } m \geq 2. \end{cases}$$

In other words, P_a differs from P in that, conditional on $X = x \in \mathcal{X}_m$ for any $m \geq 2$, the distribution of the variable Y is perturbed to be slightly more likely (if $a_m = +1$) or slightly less likely (if $a_m = -1$) to be drawn from $P^1_{Y|X=x}$ rather than from $P^0_{Y|X=x}$. Finally, we define a mixture distribution P_{mix} on $(X_1, Y_1), \ldots, (X_n, Y_n)$ as:

- Draw $A_1, A_2, \ldots \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$.
- Conditional on $A_1, A_2, \ldots$, draw $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P_A$.

Below, we will verify that we can apply Lemma 5 to obtain

$$\mathrm{d}_{\text{TV}}\left(P_{\text{mix}}, P^n\right) \leq 2n \sqrt{\sum_{m \geq 1} \epsilon_m^4 p_m^2} = 2n \sqrt{\sum_{m \geq 2} \epsilon^4 p_m^2} \leq \frac{2\epsilon^2 n}{\sqrt{M_\gamma(P_X)}}, \tag{3.6}$$

where the last step holds since $p_m \leq 1/M_\gamma(P_X)$ for all $m \geq 2$, by definition.

The remainder of the proof will center on the fact that, if ϵ is chosen to make the total variation distance between P^n and P_{mix} sufficiently small, then it is impossible to distinguish between data drawn from P^n or from P_A^n for a random A (i.e., from P_{mix}); since the conditional mean of $Y|X$ differs by $\mathcal{O}(\epsilon)$ between P and P_A, this means that our confidence interval for μ_P will need to have width at least $\mathcal{O}(\epsilon)$. For any P_a, since $\widehat{C}_n$ satisfies distribution-free coverage, we have

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

We can also calculate, for each $m \geq 2$ and each $x \in \mathcal{X}_m$ that lies in the support of P_X,

$$\begin{aligned}
\mu_{P_a}(x) &= (0.5 + a_m\epsilon)\mathbb{E}_{P^1_{Y|X=x}}[Y] + (0.5 - a_m\epsilon)\mathbb{E}_{P^0_{Y|X=x}}[Y] \\
&= 0.5\left(\mathbb{E}_{P^1_{Y|X=x}}[Y] + \mathbb{E}_{P^0_{Y|X=x}}[Y]\right) + a_m\epsilon\left(\mathbb{E}_{P^1_{Y|X=x}}[Y] - \mathbb{E}_{P^0_{Y|X=x}}[Y]\right) \\
&= \mu_P(x) + a_m\epsilon\Delta(x),
\end{aligned}$$

where we write $\Delta(x) = \left(\mathbb{E}_{P^1_{Y|X=x}}[Y] - \mathbb{E}_{P^0_{Y|X=x}}[Y]\right)$. In particular, if $X_{n+1} \notin \mathcal{X}_1$, then

$$\mu_{P_a}(X_{n+1}) \in \widehat{C}_n(X_{n+1}) \;\Rightarrow\; \{\mu_P(X_{n+1}) \pm \epsilon\Delta(X_{n+1})\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}_{(P_a)^n \times P_X}\left\{\{\mu_P(X_{n+1}) \pm \epsilon\Delta(X_{n+1})\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\} \\
&\geq \mathbb{P}_{(P_a)^n \times P_X}\left\{\mu_{P_a}(X_{n+1}) \in \widehat{C}_n(X_{n+1}) \text{ and } X_{n+1} \notin \mathcal{X}_1\right\} \\
&\geq \mathbb{P}_{(P_a)^n \times P_X}\left\{\mu_{P_a}(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} - \mathbb{P}_{(P_a)^n \times P_X}\{X_{n+1} \in \mathcal{X}_1\} \\
&\geq (1-\alpha) - (1-\gamma) = \gamma - \alpha.
\end{aligned}$$

Since this bound holds for all sign vectors $a = (a_1, a_2, \dots)$, and since P_{mix} is a mixture of distributions $(P_a)^n$, we therefore have

$$\mathbb{P}_{P_{\text{mix}} \times P_X}\left\{\{\mu_P(X_{n+1}) \pm \epsilon\Delta(X_{n+1})\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\} \geq \gamma - \alpha.$$

By our total variation bound above, therefore,

$$\mathbb{P}_{P^n \times P_X}\left\{\{\mu_P(X_{n+1}) \pm \epsilon\Delta(X_{n+1})\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\} \geq \gamma - \alpha - \frac{2\epsilon^2 n}{\sqrt{M_\gamma(P_X)}}. \tag{3.7}$$

Now fix some $\epsilon_0 \in [0, 0.5]$. We calculate

$$\begin{aligned}
&\mathrm{Leb}\left(\widehat{C}_n(X_{n+1})\right) = \int_{t\in\mathbb{R}} \mathbb{1}\left\{t \in \widehat{C}_n(X_{n+1})\right\}\, \mathsf{d}t \\
&\geq \int_{t\geq 0} \mathbb{1}\left\{\{\mu_P(X_{n+1}) \pm t\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\}\, \mathsf{d}t \\
&\geq \int_{t=0}^{\epsilon_0 \Delta(X_{n+1})} \mathbb{1}\left\{\{\mu_P(X_{n+1}) \pm t\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\}\, \mathsf{d}t \\
&= \int_{\epsilon=0}^{\epsilon_0} \mathbb{1}\left\{\{\mu_P(X_{n+1}) \pm \epsilon\Delta(X_{n+1})\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\} \cdot \Delta(X_{n+1})\, \mathsf{d}\epsilon \\
&\geq 2\sigma^2_{P,\beta} \int_{\epsilon=0}^{\epsilon_0} \mathbb{1}\left\{\sigma^2_P(X_{n+1}) \geq \sigma^2_{P,\beta} \text{ and } \{\mu_P(X_{n+1}) \pm \epsilon\Delta(X_{n+1})\} \cap \widehat{C}_n(X_{n+1}) \neq \emptyset\right\}\, \mathsf{d}\epsilon,
\end{aligned}$$

where the last step holds since $\Delta(X_{n+1}) \geq 2\sigma^2_P(X_{n+1})$ by Lemma 4. Applying (3.7), and since $\mathbb{P}\left\{\sigma^2_P(X_{n+1}) \geq \sigma^2_{P,\beta}\right\} \geq 1 - \beta$ by definition of $\sigma^2_{P,\beta}$, we have

$$\begin{aligned}
\mathbb{E}_{P^n \times P_X}\left[\mathrm{Leb}\left(\widehat{C}_n(X_{n+1})\right)\right] &\geq 2\sigma^2_{P,\beta} \int_{\epsilon=0}^{\epsilon_0} \left(\gamma - \alpha - \frac{2\epsilon^2 n}{\sqrt{M_\gamma(P_X)}}\right) - \beta\, \mathsf{d}\epsilon \\
&= 2\sigma^2_{P,\beta}\left[\epsilon_0(\gamma - \alpha - \beta) - \frac{2\epsilon_0^3 n}{3\sqrt{M_\gamma(P_X)}}\right].
\end{aligned}$$

Finally, choosing $\epsilon_0 = \min\left\{\left(\frac{(\gamma-\alpha-\beta)\sqrt{M_\gamma(P_X)}}{2n}\right)^{1/2}, 0.5\right\}$ yields the desired lower bound.

To complete the proof, we need to verify (3.6). To compare P and P_a, we can equivalently characterize these distributions as follows:

- Draw $X \sim P_X$.
- Conditional on X, draw $Z \mid X \in \mathcal{X}_m \sim$ Bernoulli(0.5) (for the distribution P, or for the distribution P_a if $m = 1$), or $Z \mid X \in \mathcal{X}_m \sim$ Bernoulli$(0.5 + a_m\epsilon)$ (for the distribution P_a if $m \geq 2$).

- Conditional on X, Z draw Y as

$$Y \mid X = x, Z = z \sim P^z_{Y|X=x}.$$

Define $\tilde{P}$ as the distribution over (X, Y, Z) induced by P, and $\tilde{P}_a$ as the distribution over (X, Y, Z) induced by P_a. Then the marginal distribution of (X, Y) under $\tilde{P}$ and under $\tilde{P}_a$ is given by P and by P_a, respectively.

Now consider comparing two distributions on triples $(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)$. We will compare $\tilde{P}^n$ versus the mixture distribution $\tilde{P}_{\text{mix}}$ defined as follows:

- Draw $A_1, A_2, \ldots \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$.
- Conditional on $A_1, A_2, \ldots$, draw $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \overset{\text{iid}}{\sim} \tilde{P}_A$.

Since in our characterization above, the distribution of $Y_1, \ldots, Y_n$ conditional on $X_1, \ldots, X_n$ and on $Z_1, \ldots, Z_n$ is the same for both, the only difference lies in the conditional distribution of $Z_1, \ldots, Z_n$ given $X_1, \ldots, X_n$. Therefore, we can apply Lemma 5 with $\epsilon_1 = 0$ and $\epsilon_2 = \epsilon_3 = \cdots = \epsilon$ to obtain

$$\mathrm{d_{TV}}\left(\tilde{P}_{\text{mix}}, \tilde{P}^n\right) \leq 2n \sqrt{\sum_{m \geq 2} \epsilon^4 p_m^2}.$$

Now let P_{mix} be the marginal distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ under $\tilde{P}_{\text{mix}}$. Noting that P^n is the marginal distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ under $\tilde{P}^n$, we therefore have

$$\mathrm{d_{TV}}\left(P_{\text{mix}}, P^n\right) \leq \mathrm{d_{TV}}\left(\tilde{P}_{\text{mix}}, \tilde{P}^n\right) \leq 2n \sqrt{\sum_{m \geq 2} \epsilon^4 p_m^2}.$$

3.5.3 Proof of Theorem 3

First, define $p_m = \mathbb{P}_{P_X}\left\{X = x^{(m)}\right\}$. The following lemma establishes some results on its support, expected value, and concentration properties of Z:

Lemma 6. *For Z and $N_{\geq 2}$ defined as in* (3.4) *and* (3.3)*, the following holds:*

$$\mathbb{E}[Z] = \sum_{m=1}^{\infty} (\mu(x^{(m)}) - \mu_P(x^{(m)}))^2 \cdot \big(np_m - 1 + (1-p_m)^n\big),$$

$$\mathbb{E}[Z \mid X_1, \ldots, X_n] = \sum_{m=1}^{\infty} (n_m - 1)_+ \cdot \big(\mu(x^{(m)}) - \mu_P(x^{(m)})\big)^2,$$

$$\mathrm{Var}\left(\mathbb{E}[Z \mid X_1, \ldots, X_n]\right) \leq 2\mathbb{E}[Z],$$

$$\mathrm{Var}\left(Z \mid X_1, \ldots, X_n\right) \leq N_{\geq 2} + 2\mathbb{E}[Z \mid X_1, \ldots, X_n].$$

In particular, the first part of the lemma will allow us to use $\mathbb{E}[Z]$ to bound the error in μ—here the calculations are similar to those in [Chan et al., 2014] for the setting of testing discrete distributions. Recalling the definition of $M^*_\gamma(P_X)$ given in (3.2), define

$$\Delta = \sqrt{\frac{2M^*_\gamma(P_X) + n}{n(n-1)}} \cdot \sqrt{\mathbb{E}[Z]}.$$

We have

$$\begin{aligned}
\sum_{m=1}^{M^*_\gamma(P_X)} p_m |\mu(x^{(m)}) - \mu_P(x^{(m)})| &= \sum_{m=1}^{M^*_\gamma(P_X)} \frac{p_m |\mu(x^{(m)}) - \mu_P(x^{(m)})|}{\sqrt{2 + np_m}} \cdot \sqrt{2 + np_m} \\
&\leq \sqrt{\sum_{m=1}^{M^*_\gamma(P_X)} \frac{p_m^2 (\mu(x^{(m)}) - \mu_P(x^{(m)}))^2}{2 + np_m}} \cdot \sqrt{\sum_{m=1}^{M^*_\gamma(P_X)} 2 + np_m} \\
&\leq \sqrt{\frac{\mathbb{E}[Z]}{n(n-1)}} \cdot \sqrt{2M^*_\gamma(P_X) + n} \\
&= \Delta,
\end{aligned}$$

where the next-to-last step holds by the following identity:

Lemma 7. *For all $n \geq 1$ and $p \in [0,1]$, $np - 1 + (1-p)^n \geq \frac{n(n-1)p^2}{2+np}$.*

Next, we will use Lemma 6 to relate Δ and $\widehat{\Delta}$. By Chebyshev's inequality, conditional

on $X_1, \ldots, X_n$, with probability at least $1 - \delta/4$ we have

$$Z \geq \mathbb{E}\left[Z \mid X_1, \ldots, X_n\right] - \sqrt{\frac{\mathrm{Var}\left(Z \mid X_1, \ldots, X_n\right)}{\delta/4}}$$
$$\geq \mathbb{E}\left[Z \mid X_1, \ldots, X_n\right] - \sqrt{\frac{N_{\geq 2} + 2\mathbb{E}\left[Z \mid X_1, \ldots, X_n\right]}{\delta/4}},$$

which can be relaxed to

$$\mathbb{E}\left[Z \mid X_1, \ldots, X_n\right] \leq 2Z + 4\sqrt{N_{\geq 2}/\delta} + 8/\delta.$$

Marginalizing over $X_1, \ldots, X_n$, this bound holds with probability at least $1 - \delta/4$. Moreover, again applying Chebyshev's inequality, with probability at least $1 - \delta/4$ we have

$$\mathbb{E}\left[Z \mid X_1, \ldots, X_n\right] \geq \mathbb{E}\left[Z\right] - \sqrt{\frac{\mathrm{Var}\left(\mathbb{E}\left[Z \mid X_1, \ldots, X_n\right]\right)}{\delta/4}} \geq \mathbb{E}\left[Z\right] - \sqrt{\frac{2\mathbb{E}\left[Z\right]}{\delta/4}},$$

which can be relaxed to

$$\mathbb{E}\left[Z\right] \leq 2\mathbb{E}\left[Z \mid X_1, \ldots, X_n\right] + 8/\delta.$$

Combining our bounds, then, we have $\mathbb{E}\left[Z\right] \leq 4Z + 8\sqrt{N_{\geq 2}/\delta} + 24/\delta$ with probability at least $1 - \delta/2$. Since $\mathbb{P}\left\{\widehat{M}_\gamma \geq M^*_\gamma(P_X)\right\} \geq 1 - \delta/2$ by Hoeffding's inequality, this implies that

$$\mathbb{P}\left\{\widehat{\Delta} \geq \Delta\right\} \geq 1 - \delta.$$

Now we verify the coverage properties of $\widehat{C}_n$. We have

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \notin \widehat{C}_n(X_{n+1})\right\} = \mathbb{P}\left\{|\mu_P(X_{n+1}) - \mu(X_{n+1})| > (\alpha - \delta - \gamma)^{-1}\widehat{\Delta}\right\}$$
$$\leq \mathbb{P}\left\{\widehat{\Delta} < \Delta\right\} + \mathbb{P}\left\{|\mu_P(X_{n+1}) - \mu(X_{n+1})| > (\alpha - \delta - \gamma)^{-1}\Delta\right\}$$
$$\leq \mathbb{P}\left\{\widehat{\Delta} < \Delta\right\} + \mathbb{P}\left\{X_{n+1} \notin \{x^{(1)}, \ldots, x^{(M^*_\gamma(P_X))}\right\}$$

$$
\begin{aligned}
&\quad + \sum_{m=1}^{M_\gamma^*(P_X)} \mathbb{P}\left\{X_{n+1} = x^{(m)},\ |\mu_P(X_{n+1}) - \mu(X_{n+1})| > (\alpha - \delta - \gamma)^{-1}\Delta\right\} \\
&\leq \delta + \gamma + \sum_{m=1}^{M_\gamma^*(P_X)} \mathbb{P}\left\{X_{n+1} = x^{(m)},\ |\mu_P(X_{n+1}) - \mu(X_{n+1})| > (\alpha - \delta - \gamma)^{-1}\Delta\right\} \\
&\leq \delta + \gamma + \sum_{m=1}^{M_\gamma^*(P_X)} p_m \mathbb{1}\left\{\left|\mu_P(x^{(m)}) - \mu(x^{(m)})\right| > (\alpha - \delta - \gamma)^{-1}\Delta\right\} \\
&\leq \delta + \gamma + \frac{\sum_{m=1}^{M_\gamma^*(P_X)} p_m \left|\mu_P(x^{(m)}) - \mu(x^{(m)})\right|}{(\alpha - \delta - \gamma)^{-1}\Delta} \\
&\leq \delta + \gamma + \frac{\Delta}{(\alpha - \delta - \gamma)^{-1}\Delta} = \alpha,
\end{aligned}
$$

which verifies the desired coverage guarantee.

3.5.4 Proof of Theorem 4

First, we have $\widehat{M}_\gamma \leq M$ almost surely by our assumption on P_X. Next we need to bound $\mathbb{E}[Z_+]$. We have

$$
\begin{aligned}
\mathbb{E}[Z_-] &\leq \mathbb{E}[(Z - \mathbb{E}[Z \mid X_1, \ldots, X_n])_-] \text{ since this conditional expectation is nonnegative} \\
&\leq \sqrt{\mathbb{E}\left[(Z - \mathbb{E}[Z \mid X_1, \ldots, X_n])^2\right]} \\
&= \sqrt{\mathbb{E}\left[\mathbb{E}\left[(Z - \mathbb{E}[Z \mid X_1, \ldots, X_n])^2 \mid X_1, \ldots, X_n\right]\right]} \\
&= \sqrt{\mathbb{E}\left[\mathrm{Var}\left(Z \mid X_1, \ldots, X_n\right)\right]} \\
&\leq \sqrt{\mathbb{E}\left[N_{\geq 2} + 2\mathbb{E}[Z \mid X_1, \ldots, X_n]\right]} \text{ by Lemma 6} \\
&= \sqrt{\mathbb{E}\left[N_{\geq 2}\right] + 2\mathbb{E}[Z]}.
\end{aligned}
$$

We then have

$$
\mathbb{E}[Z_+] = \mathbb{E}[Z] + \mathbb{E}[Z_-] \leq \mathbb{E}[Z] + \sqrt{2\mathbb{E}[Z] + \mathbb{E}\left[N_{\geq 2}\right]} \leq 1.5\mathbb{E}[Z] + 1 + \sqrt{\mathbb{E}\left[N_{\geq 2}\right]}.
$$

Next we need a lemma:

Lemma 8. *For all $n \geq 1$ and $p \in [0,1]$, $np - 1 + (1-p)^n \leq \frac{n^2p^2}{1+np}$.*

Combined with the calculation of $\mathbb{E}[Z]$ in Lemma 6, we have

$$\begin{aligned}
\mathbb{E}[Z] &\leq \sum_{m=1}^{M} (\mu(x^{(m)}) - \mu_P(x^{(m)}))^2 \cdot \frac{n^2 p_m^2}{1 + np_m} \\
&\leq \sum_{m=1}^{M} p_m \cdot (\mu(x^{(m)}) - \mu_P(x^{(m)}))^2 \cdot \frac{n^2 \cdot \eta/M}{1 + n \cdot \eta/M} \\
&= \frac{\eta n^2}{M + \eta n} \cdot \mathbb{E}_{P_X}\left[(\mu_P(X) - \mu(X))^2\right] \\
&\leq (\mathrm{err}_\mu)^2 \cdot \frac{\eta n^2}{M + \eta n},
\end{aligned}$$

since we have assumed that P_X is supported on $\{x^{(1)}, \ldots, x^{(M)}\}$ and that $\mathbb{P}_{P_X}\left\{X = x^{(m)}\right\} \leq \eta/M$ for all m, where we must have $\eta \geq 1$. Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[N_{\geq 2}\right] = \sum_{m=1}^{M} \mathbb{P}\{n_m \geq 2\} \leq \sum_{m=1}^{M} \mathbb{E}[(n_m - 1)_+] \\
&= \sum_{m=1}^{M} n \cdot \mathbb{P}_{P_X}\left\{X = x^{(m)}\right\} - 1 + \left(1 - \mathbb{P}_{P_X}\left\{X = x^{(m)}\right\}\right)^n \text{ by the proof of Lemma 6} \\
&\leq \sum_{m=1}^{M} n \cdot \eta/M - 1 + (1 - \eta/M)^n \\
&\leq \sum_{m=1}^{M} \frac{n^2(\eta/M)^2}{1 + n\eta/M} \text{ by Lemma 8} \\
&= \frac{\eta^2 n^2}{M + \eta n}.
\end{aligned}$$

We also have $N_{\geq 2} \leq M$ almost surely, and so combining these two bounds, $\mathbb{E}\left[N_{\geq 2}\right] \leq$

$\min\{\frac{\eta^2 n^2}{M}, M\}$. Combining everything, then,

$$\mathbb{E}\left[Z_+\right] \leq 1.5(\mathrm{err}_\mu)^2 \cdot \frac{\eta n^2}{M+\eta n} + 1 + \sqrt{\min\left\{\frac{\eta^2 n^2}{M}, M\right\}}.$$

Plugging these calculations into the definition of $\widehat{\Delta}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\widehat{\Delta}\right] &= \mathbb{E}\left[\sqrt{\frac{2\widehat{M}_\gamma + n}{n(n-1)}} \cdot \sqrt{4Z_+ + 8\sqrt{N_{\geq 2}/\delta} + 24/\delta}\right] \\
&\leq \mathbb{E}\left[\sqrt{\frac{2M+n}{n(n-1)}} \cdot \sqrt{4Z_+ + 8\sqrt{N_{\geq 2}/\delta} + 24/\delta}\right] \\
&\leq \sqrt{\frac{2M+n}{n(n-1)}} \cdot \sqrt{4\mathbb{E}\left[Z_+\right] + 8\sqrt{\mathbb{E}\left[N_{\geq 2}\right]/\delta} + 24/\delta} \\
&\leq \sqrt{\frac{2M+n}{n(n-1)}} \cdot \\
&\quad \sqrt{4\left(1.5(\mathrm{err}_\mu)^2 \cdot \frac{\eta n^2}{M+\eta n} + 1 + \sqrt{\min\left\{\frac{\eta^2 n^2}{M}, M\right\}}\right) + 8\sqrt{\min\left\{\frac{n^2}{M}, M\right\} \cdot 1/\delta} + 24/\delta} \\
&\leq \sqrt{\frac{2M+n}{n(n-1)}} \cdot \left[\sqrt{6(\mathrm{err}_\mu)^2 \cdot \frac{\eta n^2}{M+\eta n}} + \sqrt{4(1+2/\sqrt{\delta})\sqrt{\min\left\{\frac{\eta^2 n^2}{M}, M\right\}}} + \sqrt{4+24/\delta}\right].
\end{aligned}$$

We can assume that $M \leq n^2$ and $n \geq 2$ (as otherwise, the upper bound would be trivial, since we must have $\mathrm{Leb}(\widehat{C}_n(X_{n+1})) \leq 1$ by construction). If $M \geq n$, then $\frac{2M+n}{n(n-1)} \leq \frac{6M}{n^2}$ and the above simplifies to

$$\mathbb{E}\left[\widehat{\Delta}\right] \leq 6\sqrt{\eta} \cdot \mathrm{err}_\mu + \sqrt{\frac{6(4+24/\delta)M}{n^2}} + \sqrt{24\eta(1+2/\sqrt{\delta})}\sqrt[4]{\frac{M}{n^2}},$$

and since we assume $M \leq n^2$, we therefore have

$$\mathbb{E}\left[\widehat{\Delta}\right] \leq 6\sqrt{\eta} \cdot \mathrm{err}_\mu + \left(\sqrt{6(4+24/\delta)} + \sqrt{24\eta(1+2/\sqrt{\delta})}\right) \cdot \sqrt[4]{\frac{M}{n^2}}. \tag{3.8}$$

If instead $M < n$, then $\frac{2M+n}{n(n-1)} \leq \frac{6}{n}$ and the above bound on $\mathbb{E}\left[\widehat{\Delta}\right]$ simplifies to

$$\mathbb{E}\left[\widehat{\Delta}\right] \leq 6 \cdot \mathrm{err}_\mu + \sqrt{\frac{6}{n}} \cdot \left[\sqrt{4(1+2/\sqrt{\delta})\sqrt{M}} + \sqrt{4+24/\delta}\right],$$

which again yields the same bound (3.8) since $M \geq 1$ and $\eta \geq 1$. Finally, by definition of $\widehat{C}_n(X_{n+1})$, we have

$$\mathbb{E}\left[\mathrm{Leb}(\widehat{C}_n(X_{n+1}))\right] \leq \mathbb{E}\left[\widehat{\Delta}\right] \cdot \frac{2}{\alpha - \delta - \gamma},$$

which completes the proof for c chosen appropriately as a function of $\alpha, \delta, \gamma, \eta$.

3.5.5 Proofs of lemmas

Proof of Lemma 4

Let x_{med} be the median of Q. Define

$$q_< = \mathbb{P}_Q\{X < x_{\mathrm{med}}\}, \; q_> = \mathbb{P}_Q\{X > x_{\mathrm{med}}\},$$

and note that $q_<, q_> \in [0, 0.5]$. For $X \sim Q$, let $Q_<$ be the distribution of X conditional on $X < x_{\mathrm{med}}$ and let $Q_>$ be the distribution of X conditional on $X > x_{\mathrm{med}}$. Then we can write

$$Q = q_< \cdot Q_< + (1 - q_< - q_>) \cdot \delta_{x_{\mathrm{med}}} + q_> \cdot Q_>,$$

where δ_t denotes the point mass distribution at t. Now define

$$Q_0 = 2q_< \cdot Q_< + (1 - 2q_<) \cdot \delta_{x_{\mathrm{med}}}$$

and

$$Q_1 = 2q_> \cdot Q_> + (1 - 2q_>) \cdot \delta_{x_{\mathrm{med}}}.$$

Then clearly $Q = 0.5Q_0 + 0.5Q_1$. Next let μ_0, μ_1 be the means of these two distributions, satisfying $\frac{\mu_0+\mu_1}{2} = \mu$ where μ is the mean of Q, and let σ_0^2, σ_1^2 be the variances of these two distributions. By the law of total variance, we have

$$\begin{aligned}\sigma^2 &= \mathrm{Var}\left(0.5\delta_{\mu_0} + 0.5\delta_{\mu_1}\right) + \mathbb{E}\left[0.5\delta_{\sigma_0^2} + 0.5\delta_{\sigma_1^2}\right] \\ &= \frac{(\mu_1 - \mu_0)^2}{4} + 0.5\sigma_0^2 + 0.5\sigma_1^2.\end{aligned}$$

Next, Q_0 is a distribution supported on $[0, x_{\mathrm{med}}]$ with mean μ_0, so its variance is bounded as

$$\sigma_0^2 \le \mu_0(x_{\mathrm{med}} - \mu_0),$$

where the maximum is attained if all the mass is placed on the endpoints 0 or x_{med}. Similarly, Q_1 is a distribution supported on $[x_{\mathrm{med}}, 1]$ with mean μ_1, so its variance is bounded as

$$\sigma_1^2 \le (1 - \mu_1)(\mu_1 - x_{\mathrm{med}}).$$

Using the fact that $\frac{\mu_0+\mu_1}{2} = \mu$, we can simplify to

$$\begin{aligned}\sigma_0^2 + \sigma_1^2 &\le \mu_0(x_{\mathrm{med}} - \mu_0) + (1 - \mu_1)(\mu_1 - x_{\mathrm{med}}) \\ &= \mu(x_{\mathrm{med}} - \mu_0) + (1 - \mu)(\mu_1 - x_{\mathrm{med}}) - 0.5(\mu_1 - \mu_0)^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sigma^2 &= \frac{(\mu_1 - \mu_0)^2}{4} + 0.5\sigma_0^2 + 0.5\sigma_1^2 \le 0.5\mu(x_{\mathrm{med}} - \mu_0) + 0.5(1 - \mu)(\mu_1 - x_{\mathrm{med}}) \\ &= 0.5(2\mu - 1)x_{\mathrm{med}} - 0.5\mu\mu_0 + 0.5(1 - \mu)\mu_1 = 0.5(2\mu - 1)(x_{\mathrm{med}} - \mu) + 0.25(\mu_1 - \mu_0).\end{aligned}$$

Next, $|2\mu - 1| \le 1$ since $\mu \in [0,1]$, and $|x_{\text{med}} - \mu| \le 0.5|\mu_1 - \mu_0|$ since $\mu_0 \le x_{\text{med}} \le \mu_1$ and $\frac{\mu_0+\mu_1}{2} = \mu$. Therefore, $\sigma^2 \le 0.5(\mu_1 - \mu_0)$, proving the lemma.

Proof of Lemma 5

First we need a supporting lemma.

Lemma 9. *For any $N \ge 1$ and any $\epsilon \in [0, 0.5]$,*

$$\mathrm{d}_{\mathrm{KL}}\Big(0.5 \cdot \mathrm{Binom}(N, 0.5+\epsilon) + 0.5 \cdot \mathrm{Binom}(N, 0.5-\epsilon) \;\Big\|\; \mathrm{Binom}(N, 0.5)\Big) \le 8N(N-1)\epsilon^4.$$

Proof of Lemma 9. Let f_0 be the probability mass function of the $\mathrm{Binom}(N, 0.5)$ distribution, and let f_1 be the probability mass function of the mixture $0.5 \cdot \mathrm{Binom}(N, 0.5+\epsilon) + 0.5 \cdot \mathrm{Binom}(N, 0.5-\epsilon)$. Then we would like to bound $\mathrm{d}_{\mathrm{KL}}(f_1 \| f_0)$. We calculate the ratio

$$\begin{aligned}\frac{f_1(k)}{f_0(k)} &= \frac{0.5 \cdot \binom{N}{k}(0.5+\epsilon)^k(0.5-\epsilon)^{N-k} + 0.5 \cdot \binom{N}{k}(0.5-\epsilon)^k(0.5+\epsilon)^{N-k}}{\binom{N}{k}(0.5)^N} \\ &= \frac{(1+2\epsilon)^k(1-2\epsilon)^{N-k} + (1-2\epsilon)^k(1+2\epsilon)^{N-k}}{2}.\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}&\mathbb{E}_{\mathrm{B}(N,0.5)}\left[\left(\frac{f_1(X)}{f_0(X)}\right)^2\right] \\ &= \mathbb{E}_{\mathrm{B}(N,0.5)}\left[\left(\frac{(1+2\epsilon)^X(1-2\epsilon)^{N-X} + (1-2\epsilon)^X(1+2\epsilon)^{N-X}}{2}\right)^2\right] \\ &= \mathbb{E}_{\mathrm{B}(N,0.5)}\left[\frac{(1+2\epsilon)^{2X}(1-2\epsilon)^{2N-2X} + (1-2\epsilon)^{2X}(1+2\epsilon)^{2N-2X} + 2(1-4\epsilon^2)^N}{4}\right] \\ &= \frac{(1-2\epsilon)^{2N}\mathbb{E}_{\mathrm{B}(N,0.5)}\left[\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{2X}\right] + (1+2\epsilon)^{2N}\mathbb{E}_{\mathrm{B}(N,0.5)}\left[\left(\frac{1-2\epsilon}{1+2\epsilon}\right)^{2X}\right] + 2(1-4\epsilon^2)^N}{4}\end{aligned}$$

$$
\begin{aligned}
&= \frac{(1-2\epsilon)^{2N}\mathbb{E}_{\mathrm{Bern}(0.5)}\left[\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{2X}\right]^N + (1+2\epsilon)^{2N}\mathbb{E}_{\mathrm{Bern}(0.5)}\left[\left(\frac{1-2\epsilon}{1+2\epsilon}\right)^{2X}\right]^N + 2(1-4\epsilon^2)^N}{4} \\
&= \frac{(1-2\epsilon)^{2N}\left[0.5\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^2 + 0.5\right]^N + (1+2\epsilon)^{2N}\left[0.5\left(\frac{1-2\epsilon}{1+2\epsilon}\right)^2 + 0.5\right]^N + 2(1-4\epsilon^2)^N}{4} \\
&= \frac{\left[0.5(1+2\epsilon)^2 + 0.5(1-2\epsilon)^2\right]^N + \left[0.5(1-2\epsilon)^2 + 0.5(1+2\epsilon)^2\right]^N + 2(1-4\epsilon^2)^N}{4} \\
&= \frac{(1+4\epsilon^2)^N + (1-4\epsilon^2)^N}{2} \\
&= 1 + \sum_{k\geq 1}\binom{N}{2k}(4\epsilon^2)^{2k} \\
&= 1 + \sum_{k\geq 1}\frac{N(N-1)\ldots(N-2k+2)(N-2k+1)}{(2k)!}(4\epsilon^2)^{2k} \\
&\leq 1 + \sum_{k\geq 1}\frac{(N(N-1))^k}{2^k k!}(4\epsilon^2)^{2k} \\
&\leq e^{8\epsilon^4 N(N-1)},
\end{aligned}
$$

where $\mathrm{B}(N, 0.5)$ denotes Binomial$(N, 0.5)$. Applying Jensen's inequality, we then have

$$
\begin{aligned}
\mathrm{d}_{\mathrm{KL}}(f_1\|f_0) &= \sum_{k=0}^{n} f_1(k)\log\left(\frac{f_1(k)}{f_0(k)}\right) = \mathbb{E}_{f_1}\left[\log\left(\frac{f_1(X)}{f_0(X)}\right)\right] \leq \log\left(\mathbb{E}_{f_1}\left[\frac{f_1(X)}{f_0(X)}\right]\right) \\
&= \log\left(\mathbb{E}_{\mathrm{Binom}(N,0.5)}\left[\left(\frac{f_1(X)}{f_0(X)}\right)^2\right]\right) \leq \log\left(e^{8\epsilon^4 N(N-1)}\right) = 8\epsilon^4 N(N-1).
\end{aligned}
$$

□

Now we turn to the proof of Lemma 5. Let $p_m = \mathbb{P}\{X \in \mathcal{X}_m\}$ for each $m = 1, 2, \ldots$. Define a distribution P_0' on $(W, Z) \in \mathbb{N} \times \{0, 1\}$ as:

$$
\text{Draw } W \sim \sum_{m=1}^{\infty} p_m \delta_m, \text{ and draw } Z \sim \text{Bernoulli}(0.5), \text{ independently from } W.
$$

and for any signs $a_1, a_2, \cdots \in \{\pm 1\}$, define a distribution P'_a on $(W, Z) \in \mathbb{N} \times \{0, 1\}$ as:

$$\text{Draw } W \sim \sum_{m=1}^{\infty} p_m \delta_m, \text{ and conditional on } W, \text{ draw } Z|W = m \sim \text{Bernoulli}(0.5 + a_m \cdot \epsilon_m).$$

Then define $\tilde{P}'_0 = (P'_0)^n$ and define $\tilde{P}'_1$ as the following mixture distribution.

- Draw $A_1, A_2, \ldots \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}$.
- Conditional on $A_1, A_2, \ldots$, draw $(W_1, Z_1), \ldots, (W_n, Z_n) \overset{\text{iid}}{\sim} P'_A$.

Note that $(X_1, Z_1), \ldots, (X_n, Z_n) \sim \tilde{P}_0$ can be drawn by first drawing $(W_1, Z_1), \ldots, (W_n, Z_n) \sim \tilde{P}'_0$ and then drawing $X_i|W_i \sim P_{X|X \in \mathcal{X}_{W_i}}$ for each i. Similarly, $(X_1, Z_1), \ldots, (X_n, Z_n) \sim \tilde{P}_1$ is equivalent to first drawing $(W_1, Z_1), \ldots, (W_n, Z_n) \sim \tilde{P}'_1$ and then drawing $X_i|W_i \sim P_{X|X \in \mathcal{X}_{W_i}}$ for each i. This implies $\mathrm{d_{TV}}(\tilde{P}_1||\tilde{P}_0) \leq \mathrm{d_{TV}}(\tilde{P}'_1||\tilde{P}'_0)$.

Now we can calculate the probability mass function of $\tilde{P}'_0$ as

$$\tilde{P}'_0\big((w_1, z_1), \ldots, (w_n, z_n)\big) = \prod_{i=1}^{n} \left(p_{w_i} \cdot 0.5\right),$$

and for $\tilde{P}'_1$ as

$$\tilde{P}'_1\big((w_1, z_1), \ldots, (w_n, z_n)\big) = \mathbb{E}_{A_i \overset{\text{iid}}{\sim} \text{Unif}\{\pm 1\}} \left[\prod_{i=1}^{n} \left(p_{w_i} \cdot (0.5 + A_{w_i}\epsilon_m)^{z_i} \cdot (0.5 - A_{w_i}\epsilon_m)^{1-z_i}\right)\right].$$

Defining summary statistics

$$n_m = \sum_{i=1}^{n} \mathbb{1}\left\{w_i = m\right\} \text{ and } k_m = \sum_{i=1}^{n} \mathbb{1}\left\{w_i = m, z_i = 1\right\},$$

we can rewrite the above as

$$\tilde{P}'_0\big((w_1, z_1), \ldots, (w_n, z_n)\big) = \prod_{m=1}^{\infty} p_m^{n_m} \cdot 0.5^{n_m},$$

and

$$\begin{aligned}
&\tilde{P}_1'\big((w_1,z_1),\ldots,(w_n,z_n)\big)\\
&=\mathbb{E}_{A_i\overset{\text{iid}}{\sim}\text{Unif}\{\pm1\}}\left[\prod_{m=1}^{\infty} p_m^{n_m}\cdot(0.5+A_m\epsilon_m)^{k_m}\cdot(0.5-A_m\epsilon_m)^{n_m-k_m}\right]\\
&=\prod_{m=1}^{\infty} p_m^{n_m}\cdot\frac{1}{2}\sum_{a_m\in\{\pm1\}}(0.5+a_m\epsilon_m)^{k_m}\cdot(0.5-a_m\epsilon_m)^{n_m-k_m}
\end{aligned}$$

We then calculate

$$\begin{aligned}
&\mathrm{d}_{\mathrm{KL}}(\tilde{P}_1'||\tilde{P}_0')\\
&=\mathbb{E}_{\tilde{P}_1}\left[\log\left(\frac{\tilde{P}_1'\big((W_1,Z_1),\ldots,(W_n,Z_n)\big)}{\tilde{P}_0'\big((W_1,Z_1),\ldots,(W_n,Z_n)\big)}\right)\right]\\
&=\mathbb{E}_{\tilde{P}_1'}\left[\log\left(\frac{\prod_{m=1}^{\infty} p_m^{N_m}\cdot\frac{1}{2}\sum_{a_m\in\{\pm1\}}(0.5+a_m\epsilon_m)^{K_m}\cdot(0.5-a_m\epsilon_m)^{N_m-K_m}}{\prod_{m=1}^{\infty} p_m^{N_m}\cdot(0.5)^{N_m}}\right)\right]\\
&=\sum_{m=1}^{\infty}\mathbb{E}_{\tilde{P}_1'}\left[\log\left(\frac{\frac{1}{2}\sum_{a_m\in\{\pm1\}}(0.5+a_m\epsilon_m)^{K_m}\cdot(0.5-a_m\epsilon_m)^{N_m-K_m}}{(0.5)^{N_m}}\right)\right]\\
&=\sum_{m=1}^{\infty}\mathbb{E}_{\tilde{P}_1'}\left[\mathbb{E}_{\tilde{P}_1'}\left[\log\left(\frac{\frac{1}{2}\sum_{a_m\in\{\pm1\}}(0.5+a_m\epsilon_m)^{K_m}\cdot(0.5-a_m\epsilon_m)^{N_m-K_m}}{(0.5)^{N_m}}\right)\;\middle|\;N_m\right]\right],
\end{aligned}$$

where

$$N_m=\sum_{i=1}^{n}\mathbb{1}\left\{W_i=m\right\}\text{ and }K_m=\sum_{i=1}^{n}\mathbb{1}\left\{W_i=m,Z_i=1\right\},$$

Next, we calculate the conditional expectation in the last expression above. If $N_m=0$ then trivially it is equal to $\log(1)=0$. If $N_m\geq 1$, then under $\tilde{P}_1'$, we can see that

$$K_m\mid N_m\sim 0.5\cdot\text{Binom}(N_m,0.5+\epsilon_m)+0.5\cdot\text{Binom}(N_m,0.5-\epsilon_m),$$

and therefore,

$$\begin{aligned}
&\mathbb{E}_{\tilde{P}_1'}\left[\log\left(\frac{\frac{1}{2}\sum_{a_m\in\{\pm 1\}}(0.5+a_m\epsilon_m)^{K_m}\cdot(0.5-a_m\epsilon_m)^{N_m-K_m}}{(0.5)^{N_m}}\right)\,\middle|\,N_m\right] \\
&\quad = \mathrm{d_{KL}}\Big(0.5\cdot\mathrm{Binom}(N_m,0.5+\epsilon_m)+0.5\cdot\mathrm{Binom}(N_m,0.5-\epsilon_m)\;\Big\|\;\mathrm{Binom}(N_m,0.5)\Big) \\
&\quad \le 8N_m(N_m-1)\epsilon_m^4,
\end{aligned}$$

where the last step applies Lemma 9. Therefore,

$$\begin{aligned}
\mathrm{d_{KL}}(\tilde{P}_1'||\tilde{P}_0') &\le \sum_{m=1}^{\infty}\mathbb{E}_{\tilde{P}_1'}\left[8N_m(N_m-1)\epsilon_m^4\right] \\
&= 8\sum_{m=1}^{\infty}\epsilon_m^4\mathbb{E}_{\tilde{P}_1'}\left[N_m^2-N_m\right] \\
&= 8\sum_{m=1}^{\infty}\epsilon_m^4\left(\left(np_m(1-p_m)+n^2p_m^2\right)-np_m\right) \\
&= 8\cdot n(n-1)\sum_{m=1}^{\infty}\epsilon_m^4p_m^2,
\end{aligned}$$

since $N_m\sim\mathrm{Binom}(n,p_m)$ by definition. Applying Pinsker's inequality and $\mathrm{d_{TV}}(\tilde{P}_1||\tilde{P}_0)\le \mathrm{d_{TV}}(\tilde{P}_1'||\tilde{P}_0')$ completes the proof.

Proof of Lemma 6

Define

$$Z_m=\begin{cases}(n_m-1)\cdot\left((\bar{y}_m-\mu(x^{(m)}))^2-n_m^{-1}s_m^2\right), & n_m\ge 2,\\ 0, & n_m=0 \text{ or } 1.\end{cases}$$

Then $Z=\sum_{m=1}^{\infty}Z_m$. Now we calculate the conditional mean and variance. Conditional on $X_1,\ldots,X_n$, $\bar{y}_m$ and s_m^2 are the sample mean and sample variance of n_m i.i.d. draws from a distribution with mean $\mu_P(x^{(m)})$ and variance $\sigma_P^2(x^{(m)})$, supported on $[0,1]$, where we let $\sigma_P^2(x^{(m)})$ be the variance of the distribution of $Y|X=x^{(m)}$, under the joint distribution P.

For any m with $n_m \geq 2$, we therefore have

$$\begin{aligned}
&\mathbb{E}\left[\bar{y}_m \mid X_1, \ldots, X_n\right] = \mu_P(x^{(m)}), \\
&\operatorname{Var}\left(\bar{y}_m \mid X_1, \ldots, X_n\right) = n_m^{-1}\sigma_P^2(x^{(m)}) = \mathbb{E}\left[n_m^{-1}s_m^2 \;\middle|\; X_1, \ldots, X_n\right],
\end{aligned}$$

and so

$$\begin{aligned}
&\mathbb{E}\left[(\bar{y}_m - \mu(x^{(m)}))^2 - n_m^{-1}s_m^2 \;\middle|\; X_1, \ldots, X_n\right] \\
&\quad = n_m^{-1}\sigma_P^2(x^{(m)}) + (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 - n_m^{-1}\sigma_P^2(x^{(m)}) = (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2.
\end{aligned}$$

Next, we have $(n_1, \ldots, n_M) \sim \text{Multinom}(n, p)$, which implies that marginally $n_m \sim \text{Binom}(n, p_m)$ and so

$$\mathbb{E}\left[(n_m - 1)_+\right] = \mathbb{E}\left[n_m - 1 + \mathbb{1}\left\{n_m = 0\right\}\right] = np_m - 1 + (1 - p_m)^n.$$

Combining these calculations completes the proof for the expected value $\mathbb{E}[Z]$ and conditional expected value $\mathbb{E}[Z \mid X_1, \ldots, X_n]$.

Next, we calculate conditional and marginal variance. We have

$$\begin{aligned}
&\operatorname{Var}\left((\bar{y}_m - \mu(x^{(m)}))^2 - n_m^{-1}s_m^2 \;\middle|\; X_1, \ldots, X_n\right) \\
&= \operatorname{Var}\left((\bar{y}_m - \mu(x^{(m)}))^2 - n_m^{-1}s_m^2 - (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 \;\middle|\; X_1, \ldots, X_n\right) \\
&\leq \mathbb{E}\left[\left((\bar{y}_m - \mu(x^{(m)}))^2 - n_m^{-1}s_m^2 - (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2\right)^2 \;\middle|\; X_1, \ldots, X_n\right] \\
&= \mathbb{E}\left[\left((\bar{y}_m - \mu_P(x^{(m)}))^2 + 2(\bar{y}_m - \mu_P(x^{(m)}))(\mu_P(x^{(m)}) - \mu(x^{(m)})) - n_m^{-1}s_m^2\right)^2 \;\middle|\; X_1, \ldots, X_n\right] \\
&\leq 4\mathbb{E}\left[\left((\bar{y}_m - \mu_P(x^{(m)}))\right)^4 \;\middle|\; X_1, \ldots, X_n\right] \\
&\qquad + 2\mathbb{E}\left[\left(2(\bar{y}_m - \mu_P(x^{(m)}))(\mu_P(x^{(m)}) - \mu(x^{(m)}))\right)^2 \;\middle|\; X_1, \ldots, X_n\right]
\end{aligned}$$

$$+ 4\mathbb{E}\left[\left(n_m^{-1}s_m^2\right)^2 \;\middle|\; X_1,\ldots,X_n\right],$$

where the last step holds since $(a+b+c)^2 \le 4a^2+2b^2+4c^2$ for any a,b,c. Now we bound each term separately. First, we have

$$\begin{aligned}
&\mathbb{E}\left[\left((\bar{y}_m - \mu_P(x^{(m)}))\right)^4 \;\middle|\; X_1,\ldots,X_n\right] \\
&= \frac{1}{n_m^4} \sum_{\substack{i_1,i_2,i_3,i_4 \text{ s.t.} \\ X_{i_1}=X_{i_2}=X_{i_3}=X_{i_4}=x^{(m)}}} \mathbb{E}\left[\prod_{k=1}^{4}(Y_{i_k} - \mu_P(x^{(m)})) \;\middle|\; X_1,\ldots,X_n\right] \\
&= \frac{1}{n_m^4}\Bigg[n_m \cdot \mathbb{E}\left[(Y-\mu_P(x^{(m)}))^4 \;\middle|\; X = x^{(m)}\right] \\
&\qquad\qquad + 3n_m(n_m-1)\cdot \mathbb{E}\left[(Y-\mu_P(x^{(m)}))^2 \;\middle|\; X = x^{(m)}\right]^2\Bigg] \\
&\le \frac{1}{n_m^4}\left[n_m \cdot \sigma_P^2(x^{(m)}) + 3n_m(n_m-1)\cdot(\sigma_P^2(x^{(m)}))^2\right] \\
&\le \frac{1}{n_m^4}\left[n_m \cdot \tfrac{1}{4} + 3n_m(n_m-1)\cdot(\tfrac{1}{4})^2\right] = \frac{3n_m+1}{16n_m^3},
\end{aligned}$$

where the second step holds by counting tuples (i_1,i_2,i_3,i_4) where either all four indices are equal, or there are two pairs of equal indices (since otherwise, the expected value of the product is zero). Next,

$$\begin{aligned}
&\mathbb{E}\left[\left(2(\bar{y}_m - \mu_P(x^{(m)}))(\mu_P(x^{(m)}) - \mu(x^{(m)}))\right)^2 \;\middle|\; X_1,\ldots,X_n\right] \\
&= 4(\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 \mathbb{E}\left[(\bar{y}_m - \mu_P(x^{(m)}))^2 \;\middle|\; X_1,\ldots,X_n\right] \\
&= 4(\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 \cdot n_m^{-1}\sigma_P^2(x^{(m)}) \\
&\le n_m^{-1}(\mu_P(x^{(m)}) - \mu(x^{(m)}))^2.
\end{aligned}$$

Finally, since $s_m^2 \le \frac{n_m}{4(n_m-1)}$ holds deterministically,

$$\mathbb{E}\left[\left(n_m^{-1}s_m^2\right)^2 \;\middle|\; X_1,\ldots,X_n\right] \le n_m^{-2}\cdot\frac{n_m}{4(n_m-1)}\cdot\mathbb{E}\left[s_m^2 \;\middle|\; X_1,\ldots,X_n\right]$$
$$= n_m^{-2}\cdot\frac{n_m}{4(n_m-1)}\cdot\sigma_P^2(x^{(m)}) \le \frac{1}{16n_m(n_m-1)}.$$

Combining everything, then,

$$\mathrm{Var}\left((\bar{y}_m-\mu(x^{(m)}))^2 - n_m^{-1}s_m^2 \;\middle|\; X_1,\ldots,X_n\right)$$
$$\le 4\cdot\frac{3n_m+1}{16n_m^3} + 2\cdot n_m^{-1}(\mu_P(x^{(m)})-\mu(x^{(m)}))^2 + 4\cdot\frac{1}{16n_m(n_m-1)},$$

and so for $n_m \ge 2$,

$$\mathrm{Var}\left(Z_m \mid X_1,\ldots,X_n\right)$$
$$\le (n_m-1)^2\cdot\left[4\cdot\frac{3n_m+1}{16n_m^3} + 2\cdot n_m^{-1}(\mu_P(x^{(m)})-\mu(x^{(m)}))^2 + 4\cdot\frac{1}{16n_m(n_m-1)}\right]$$
$$\le 1 + 2(n_m-1)\cdot(\mu_P(x^{(m)})-\mu(x^{(m)}))^2 = 1 + 2\mathbb{E}\left[Z_m \mid X_1,\ldots,X_n\right].$$

If instead $n_m = 0$ or $n_m = 1$ then $Z_m = 0$ by definition, and so $\mathrm{Var}\left(Z_m \mid X_1,\ldots,X_n\right) = 0$. Therefore, in all cases, we have

$$\mathrm{Var}\left(Z_m \mid X_1,\ldots,X_n\right) \le \mathbb{1}\left\{n_m \ge 2\right\} + 2\mathbb{E}\left[Z_m \mid X_1,\ldots,X_n\right].$$

It is also clear that, conditional on $X_1,\ldots,X_n$, the Z_m's are independent, and so

$$\mathrm{Var}\left(Z \mid X_1,\ldots,X_n\right) = \sum_{m=1}^{\infty}\mathrm{Var}\left(Z_m \mid X_1,\ldots,X_n\right) \le N_{\ge 2} + 2\mathbb{E}\left[Z \mid X_1,\ldots,X_n\right].$$

Finally, we need to bound $\mathrm{Var}\left(\mathbb{E}\left[Z \mid X_1,\ldots,X_n\right]\right)$. First, we have

$$\mathrm{Var}\left(\mathbb{E}\left[Z_m \mid X_1,\ldots,X_n\right]\right) = \mathrm{Var}\left((n_m-1)_+\right)\cdot(\mu_P(x^{(m)})-\mu(x^{(m)}))^4$$

$$\leq \mathrm{Var}\left((n_m - 1)_+\right) \cdot (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2,$$

and we can calculate

$$\begin{aligned}
&\mathrm{Var}\left((n_m - 1)_+\right) \\
&= \mathrm{Var}\left(n_m + \mathbb{1}\left\{n_m = 0\right\}\right) \\
&= \mathrm{Var}\left(n_m\right) + \mathrm{Var}\left(\mathbb{1}\left\{n_m = 0\right\}\right) + 2\mathrm{Cov}\left(n_m, \mathbb{1}\left\{n_m = 0\right\}\right) \\
&= \mathrm{Var}\left(n_m\right) + \mathrm{Var}\left(\mathbb{1}\left\{n_m = 0\right\}\right) - 2\mathbb{E}\left[n_m\right]\mathbb{E}\left[\mathbb{1}\left\{n_m = 0\right\}\right] \text{ since } n_m \cdot \mathbb{1}\left\{n_m = 0\right\} = 0 \text{ a.s.} \\
&= np_m(1 - p_m) + (1 - p_m)^n\big(1 - (1 - p_m)^n\big) - 2np_m(1 - p_m)^n.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&2\mathbb{E}\left[(n_m - 1)_+\right] - \mathrm{Var}\left((n_m - 1)_+\right) \\
&= 2np_m - 2 + 2(1 - p_m)^n - np_m(1 - p_m) - (1 - p_m)^n\big(1 - (1 - p_m)^n\big) + 2np_m(1 - p_m)^n \\
&= np_m(1 + p_m) + (1 - p_m)^n\big(1 + 2np_m + (1 - p_m)^n\big) - 2 \\
&\geq 0,
\end{aligned}$$

where the last step holds since, defining $f(t) = nt(1 + t) + (1 - t)^n\big(1 + 2nt + (1 - t)^n\big)$, we can see that $f(0) = 2$ and $f'(t) \geq 0$ for all $t \in [0, 1]$. This verifies that

$$\begin{aligned}
\mathrm{Var}\left(\mathbb{E}\left[Z_m \mid X_1, \ldots, X_n\right]\right) &\leq \mathrm{Var}\left((n_m - 1)_+\right) \cdot (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 \\
&\leq 2\mathbb{E}\left[(n_m - 1)_+\right] \cdot (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 = 2\mathbb{E}\left[Z_m\right].
\end{aligned}$$

Next, for any $m \neq m'$,

$$\mathrm{Cov}\left(\mathbb{E}\left[Z_m \mid X_1, \ldots, X_n\right], \mathbb{E}\left[Z_{m'} \mid X_1, \ldots, X_n\right]\right)$$

$$= \mathrm{Cov}\left((n_m - 1)_+, (n_{m'} - 1)_+\right) \cdot (\mu_P(x^{(m)}) - \mu(x^{(m)}))^2 \cdot (\mu_P(x^{(m')}) - \mu(x^{(m')}))^2$$
$$\leq 0.$$

For the last step, we use the fact that $\mathrm{Cov}\left((n_m - 1)_+, (n_{m'} - 1)_+\right) \leq 0$, which holds since, conditional on n_m, we have $n_{m'} \sim \mathrm{Binom}\left(n - n_m, \frac{p_{m'}}{1-p_m}\right)$, and so the distribution of $n_{m'}$ is stochastically smaller whenever n_m is larger. Therefore,

$$\mathrm{Var}\left(\mathbb{E}\left[Z \mid X_1, \ldots, X_n\right]\right) \leq \sum_{m=1}^{\infty} \mathrm{Var}\left(\mathbb{E}\left[Z_m \mid X_1, \ldots, X_n\right]\right) \leq \sum_{m=1}^{\infty} 2\mathbb{E}\left[Z_m\right] = 2\mathbb{E}\left[Z\right].$$

Proofs of Lemma 7 and Lemma 8

Replacing p with $1 - s$, equivalently, we need to show that, for all $s \in [0, 1]$,

$$\frac{n(n-1)(1-s)^2}{2 + n(1-s)} \leq n(1-s) - 1 + s^n \leq \frac{n^2(1-s)^2}{1 + n(1-s)}.$$

After simplifying, this is equivalent to proving that

$$\frac{n(1-s)^2 + 2n(1-s)}{2 + n(1-s)} \geq 1 - s^n \geq \frac{n(1-s)}{1 + n(1-s)},$$

which we can further simplify to

$$\frac{n(1-s) + 2n}{2 + n(1-s)} \geq 1 + s + \cdots + s^{n-1} \geq \frac{n}{1 + n(1-s)} \tag{3.9}$$

by dividing by $1 - s$ (note that this division can be performed whenever $s < 1$, while if $s = 1$, then the desired inequalities hold trivially).

Now we address the two desired inequalities separately. For the left-hand inequality

in (3.9), define

$$h(s) = (2 + n(1-s)) \cdot (s + s^2 + \cdots + s^{n-1}) = ns + 2(s + s^2 + \cdots + s^{n-1}) - ns^n.$$

We calculate $h(1) = 2(n-1)$, and for any $s \in [0,1]$,

$$\begin{aligned} h'(s) = n + \sum_{i=1}^{n-1} 2is^{i-1} - n^2 s^{n-1} &\geq n + \sum_{i=1}^{n-1} 2is^{n-1} - n^2 s^{n-1} \\ &= n + s^{n-1}\left(\sum_{i=1}^{n-1} 2i - n^2\right) = n - ns^{n-1} \geq 0, \end{aligned}$$

where the first inequality holds since $s^{i-1} \geq s^{n-1}$ for all $i = 1, \ldots, n-1$, and the second inequality holds since $s^{n-1} \leq 1$. Therefore, $h(s) \leq h(1) = 2(n-1)$ for all $s \in [0,1]$, and so

$$\begin{aligned} 1 + s + \cdots + s^{n-1} &= \frac{(1 + s + \cdots + s^{n-1}) \cdot (2 + n(1-s))}{2 + n(1-s)} \\ &= \frac{2 + n(1-s) + h(s)}{2 + n(1-s)} \leq \frac{2 + n(1-s) + 2(n-1)}{2 + n(1-s)} = \frac{n(1-s) + 2n}{2 + n(1-s)}, \end{aligned}$$

as desired.

To verify the right-hand inequality in (3.9), we have

$$\begin{aligned} 1 + s + \cdots + s^{n-1} &= \frac{(1 + s + \cdots + s^{n-1}) \cdot (1 + n(1-s))}{1 + n(1-s)} \\ &= \frac{(n+1)(1 + s + \cdots + s^{n-1}) - n(s + s^2 + \cdots + s^n)}{1 + n(1-s)} \\ &= \frac{n + (1 + s + \cdots + s^{n-1}) - ns^n}{1 + n(1-s)} \\ &\geq \frac{n}{1 + n(1-s)}, \end{aligned}$$

where the last step holds since, for $s \in [0,1]$, we have $s^i \geq s^n$ for all $i = 0, 1, \ldots, n-1$.

CHAPTER 4

DISTRIBUTION-FREE INFERENCE WITH HIERARCHICAL DATA

4.1 Introduction

Consider a standard distribution-free prediction problem where we have training data $\{(X_i, Y_i)\}, i = 1, 2, \cdots, n$ and a new observation X_{n+1}, and the task is to construct $\widehat{C}_n$ such that

$$\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha \tag{4.1}$$

holds for any distribution P. Methods such as conformal prediction, invented by [Vovk et al., 2005], provide an answer to this problem with exchangeability as the only assumption.

Though the marginal coverage guarantee (4.1) guarantees an overall quality of the prediction set, it is often more desired to have a useful guarantee with conditional coverage

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1}\right\}, \text{ or } \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; (X_i, Y_i)_{1\leq i\leq n}, X_{n+1}\right\},$$

to ensure that the prediction is accurate conditional on the specific observations we have. However, achieving a useful distribution-free conditional coverage is a much more challenging problem. In case of nonatomic features, recent work by [Barber et al., 2021a], [Barber, 2020], [Medarametla and Candès, 2021] proves that there are limits on having a useful coverage for certain characteristics of the conditional distribution of $Y_{n+1}|X_{n+1}$. On the other hand, [Lee and Barber, 2021] illustrates the possibility of achieving a meaningful coverage for the conditional mean $\mathbb{E}\left[Y_{n+1} \mid X_{n+1}\right]$ in case of discrete/mixed distributions that allows repeats in the training data.

In this work, we look further into the setting where we have multiple observations for each individual. We begin by studying a more general setting where we have data with

hierarchical structure, which contains exchangeable groups of exchangeable measurements, and develop an extension of conformal prediction that works for such data structure, which we denote as hierarchical conformal prediction(HCP). In the special case where we have i.i.d repeated measurements, we show that we can have a better control of the conditional miscoverage rates by making use of the information the repeats provide.

4.1.1 Problem setting

Suppose we have hierarchical data $(\tilde{Z}_1, \tilde{Z}_2, \cdots, \tilde{Z}_n)$, where each group $\tilde{Z}_i = (Z_{i,1}, \cdots, Z_{i,K_i})$ consists of K_i measurements $Z_{i,k} \in \mathcal{Z}$. We assume that we have an exchangeable draw of distributions $\Pi_1, \Pi_2, \cdots, \Pi_n \overset{\text{exch}}{\sim} P_\Pi$, that the group sizes are drawn by $K_i \mid \Pi_i \sim P_{K|\Pi}$, and that $Z_{i,1} \cdots, Z_{i,K_i} \mid K_i, \Pi_i \overset{\text{exch}}{\sim} \Pi_i$.

In other words, we consider a hierarchical data with a hierarchical exchangeability structure, where the group size can be random. Given such training data, we aim to provide inference for a new data point Z_{n+1}, where we assume that Z_{n+1} is a draw from distribution Π_{n+1} which is exchangeable with $\Pi_1, \cdots, \Pi_n$.

A special case is the setting where we have data with repeated measurements, which we will discuss in Section 4.3. This is the setting where each $Z_{ij} = (X_i, Y_{ij})$. We formulate this special case as follows:

Suppose we have an i.i.d. training data $(X_i, K_i, \tilde{Y}_i), i = 1, 2, \cdots, n$ from an unknown distribution $P = P_X \times P_K \times P_{\tilde{Y}|K,X}$ on $\mathcal{X} \times \mathbb{N} \times \mathcal{Y}$, where $\tilde{Y}_i = (Y_{i,1}, \cdots, Y_{i,K_i})$ and $\mathcal{Y} = \mathbb{R} \cup \mathbb{R}^2 \cup \mathbb{R}^3 \cup \cdots$. $P_{\tilde{Y}|K,X}$ is given by

$$\tilde{Y} = (Y_1, \cdots, Y_K) \mid K, X \overset{\text{iid}}{\sim} P_{Y|X}.$$

In other words, we assume that we have multiple label observations for each individual i where the repeat number K_i can be random. We write the joint distribution of $(X_i, \tilde{Y}_i)$ as $P_{X,\tilde{Y}}$. Note that we have two types of exchangeability here: the exchangeability of

individuals $\{(X_i, K_i, \tilde{Y}_i) : i = 1, 2, \cdots, n\}$ and the exchangeability of $Y_{i,1}, \cdots, Y_{i,K_i}$ given X_i and K_i for each individual i. Given such training data $(X_1, \tilde{Y}_1), \cdots, (X_n, \tilde{Y}_n) \overset{\text{iid}}{\sim} P_{X,\tilde{Y}}$ with this hierarchical exchangeability and a new input X_{n+1} drawn from P_X independently from the training data, we aim to do inference on Y_{n+1}, a draw from $P_{Y|X=X_{n+1}}$, with a useful guarantee.

4.1.2 Related work

Distribution-free inference has received much attention recently, and a lot of efforts have been made to explore the usefulness as well as the limits of distribution-free methods, to put them into practical use, and to extend them to various applications.

Conformal prediction (see [Vovk et al., 2005, Papadopoulos et al., 2002, Lei et al., 2018] for example) provides a universal framework for distribution-free prediction. Split conformal prediction ([Vovk et al., 2005, Papadopoulos et al., 2002]) applies data splitting to reduce computational cost, where one uses one split of data to construct a nonconformity score function and then makes use of the exchangeability of the score values on the other split to construct the prediction set. These methods generally provide prediction sets with marginal coverage guarantee.

For the goal of having a useful bound for conditional coverage in the distribution-free setting, [Barber et al., 2021a] provides important impossibility results for the case the distribution of feature is nonatomic. Similarly, [Barber, 2020, Medarametla and Candès, 2021] discusses limits on having a useful confidence set for conditional mean or median, for the nonatomic feature. On the other hand, [Lee and Barber, 2021] proves that in case of discrete/mixed feature, we can make use of the repeated feature observations to attain a meaningful inference for the conditional mean.

The hierarchical structure setting was previously studied by [Dunn et al., 2022]. They introduce methods such as double conformal, pooling cdfs and subsampling, which we will

discuss further in the next section.

4.2 Hierarchical conformal prediction

We begin with introducing an extension of split conformal prediction for the hierarchical data. Let us write δ_t to denote the distribution with a unique point mass at t, and write $Q_\beta(\mathcal{D})$ to denote β-quantile of distribution $\mathcal{D}$, defined by

$$Q_\beta(\mathcal{D}) = \inf\left\{x : \mathbb{P}_{X\sim\mathcal{D}}\left\{X \le x\right\} \ge \beta\right\}.$$

Theorem 5 (Hierarchical conformal prediction)**.** *Let $s : \mathcal{Z} \to \mathbb{R}^+$ be any nonconformity score function, constructed independently of the data $\{\tilde{Z}_1, \cdots, \tilde{Z}_n\}$. Define*

$$\widehat{C}_n = \left\{ z \in \mathcal{Z} : s(z) \le Q_{1-\alpha}\left(\sum_{i=1}^{n}\sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{s(Z_{i,j})} + \frac{1}{n+1}\cdot\delta_\infty \right) \right\}.$$

Then

$$\mathbb{P}\left\{Z_{n+1} \in \widehat{C}_n\right\} \ge 1-\alpha$$

holds under any distribution P on $(K, \tilde{Z})$, where the probability is taken with respect to $\Pi_1, \cdots, \Pi_{n+1} \overset{\text{exch}}{\sim} P_\Pi$, $K_i \mid \Pi_i \sim P_{K|\Pi}$, $Z_{i,1}, \cdots, Z_{i,K_i} \mid \Pi_i, K_i \overset{\text{exch}}{\sim} \Pi_i$ and $Z_{n+1} \mid \Pi_{n+1}, \{(\tilde{Z}_i)\}_{1\le i\le n} \sim \Pi_{n+1}$. Moreover, if all the $s(Z_{i,j})$'s are distinct almost surely, it additionally holds that

$$\mathbb{P}\left\{Z_{n+1} \in \widehat{C}_n\right\} \le 1-\alpha+\frac{2}{n+1}.$$

Therefore, we have a weighted conformal type prediction set where the weights are determined by the group sizes.

Remark 1. *An analogous construction provides the extension of full conformal prediction (See Theorem 8 in the Appendix). However, the full conformal-based method is unlikely to be*

practical, as we need to repeat the construction of s for different candidates of $\tilde{Z}_{n+1}$ which can have an arbitrary size.

4.2.1 Comparison with existing methods

[Dunn et al., 2022] introduces multiple approaches for distribution-free inference with hierarhical data. Here we rewrite their methods in terms of a general score function s, and then compare with hierarchical conformal prediction. Throughout the section we write $S_{i,j}$ to denote $s(Z_{i,j})$. Note that in their setting the group sizes are nonrandom values, hence we write k_i to denote the size of group i in this section.

1. *Double conformal*

   This approach quantifies the within-group uncertainties for each group and then combine them to obtain the prediction set, in the setting where all the group sizes are equal. The prediction set is given by

   $$\widehat{C}_n = \left\{ z \in \mathcal{Z} : s(z) \in [\ell_{(\lfloor (n+1)(\alpha/4) \rfloor)}, u_{(\lfloor (n+1)(1-\alpha/4) \rfloor)}] \right\}, \tag{4.2}$$

   where for each $1 \leq i \leq n$, $\ell_i = S_{i,(\lfloor (k+1)(\alpha/2) \rfloor)}$ and $u_i = S_{i,(\lfloor (k+1)(1-\alpha/2) \rfloor)}$ and k is the shared group size. Here $S_{i,(j)}$ denotes the j-th order statistics of $S_i = (S_{i,1}, \cdots, S_{i,k})$.

   Double conformal provides a valid $(1-\alpha)$ coverage for $s(Z_{n+1})$, but can be overly conservative.

2. *Pooling CDFs*

   This method estimates the conditional cumulative distribution functions of each group, and then combine them to have an estimate of the marginal cdf. Write the empirical cdf for group i by

   $$\hat{F}_i(t) = \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbb{1} \left\{ S_{i,j} \leq t \right\},$$

then the prediction set is given by

$$\widehat{C}_n = \{z \in \mathcal{Z} : s(z) \in [\hat{q}(\alpha), \hat{q}(1-\alpha/2)]\}, \tag{4.3}$$

where

$$\hat{q}(\alpha) = \inf\left\{t \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \hat{F}_i(t) \geq \alpha\right\}.$$

The prediction set (4.3) provides an asymptotic $(1-\alpha)$ coverage for Z_{n+1} as the sample size n tends to infinity.

Note that this method and HCP provide similar prediction sets. To see this, observe that the pooled cdf $\frac{1}{n}\sum_{i=1}^{n} \hat{F}_i$ is the cdf of the distribution

$$\sum_{i=1}^{n}\sum_{j=1}^{k_i} \frac{1}{nk_i} \cdot \delta_{S_{i,j}},$$

while the distribution we have in HCP is

$$\sum_{i=1}^{n}\sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{S_{i,j}} + \frac{1}{n+1} \cdot \delta_{\infty}.$$

Therefore, in the case where the group sizes are fixed, HCP can be viewed as a 'correction' of the pooling cdfs method, where the additional point mass at $+\infty$ leads to a corrected pooled cdf that provides finite sample guarantee.

3. *Subsampling*

This is a simple approach where we first construct a non-hierarchical dataset with one measurement per group through subsampling, and then apply standard methods. The first step is to draw one sample Z_i from $\{Z_{i,1}, \cdots, Z_{i,k_i}\}$ at uniformly random to have

an exchangeable data set $Z_1, Z_2, \cdots, Z_n$, and then the prediction set is given by

$$\widehat{C}_n = \{z \in \mathcal{Z} : s(z) \in [S_{(\lfloor (n+1)(\alpha/2) \rfloor)}, S_{(\lfloor (n+1)(1-\alpha/2) \rfloor)}]\}, \tag{4.4}$$

where $S_i = s(Z_i)$ and $S_{(i)}$ denotes the i-th order statistic of $\{S_1, S_2, \cdots, S_n\}$.

The subsampling-based prediction set provides a valid $(1-\alpha)$ marginal coverage but has the issue of ignoring lots of observations. To relieve this issue, [Dunn et al., 2022] also provides an alternative method that makes use of multiple repeats of subsampling, but this method loses guarantee and instead provides $1-2\alpha$ coverage rate.

Hierarchical conformal prediction is different from these methods in the sense that it assumes the groups sizes to be exchangeable random variables, and that it provides a valid finite sample guarantee without issues such as conservativeness or loss of information. We next illustrate these comparisons further by experiments.

4.2.2 Simulations

Here we show some simulation results for the comparison of the performance of hierarchical conformal prediction (HCP) and the three methods proposed by [Dunn et al., 2022] that we reviewed in the previous section. For multiple combinations of (n, k), we repeated generating an i.i.d data of size n by

$$\begin{aligned}
&X_i \sim \text{Unif}([0, 5]), \\
&K_i \equiv k, \tilde{Y}_i = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,k}) | X_i = x \overset{\text{iid}}{\sim} \mathcal{N}(\mu(x), \sigma(x)), \\
&\mu(x) = 1 + x + 0.1 \cdot x^2, \\
&\sigma(x) = 1 + 0.5 \cdot x,
\end{aligned}$$

and applying the four methods with score function $s((x,y)) = |(y - \hat{\mu(x)})/\hat{\sigma}(x)|$, where $\hat{\mu}$ and $\hat{\sigma}$ were obtained by running linear regressions on a separate training data. We checked the conditional coverage rates and widths of the prediction sets(Figure 4.1), and also their marginal coverage rates (Table 4.1).

	$n = 50, k = 50$	$n = 500, k = 5$	$n = 500, k = 50$
HCP	0.8163 (0.0010)	0.8018 (0.0010)	0.8012 (0.0009)
Pooling CDFs	0.7891 (0.0015)	0.7874 (0.0015)	0.7718 (0.0012)
Double conformal	0.9214 (0.0006)	0.9691 (0.0003)	0.9154 (0.0005)
Subsampling	0.8057 (0.0027)	0.7993 (0.0013)	0.8004 (0.0012)

Table 4.1: Marginal coverage rates of hierarchical conformal prediction(HCP), double conformal, pooling CDFs, and subsampling, with standard errors.

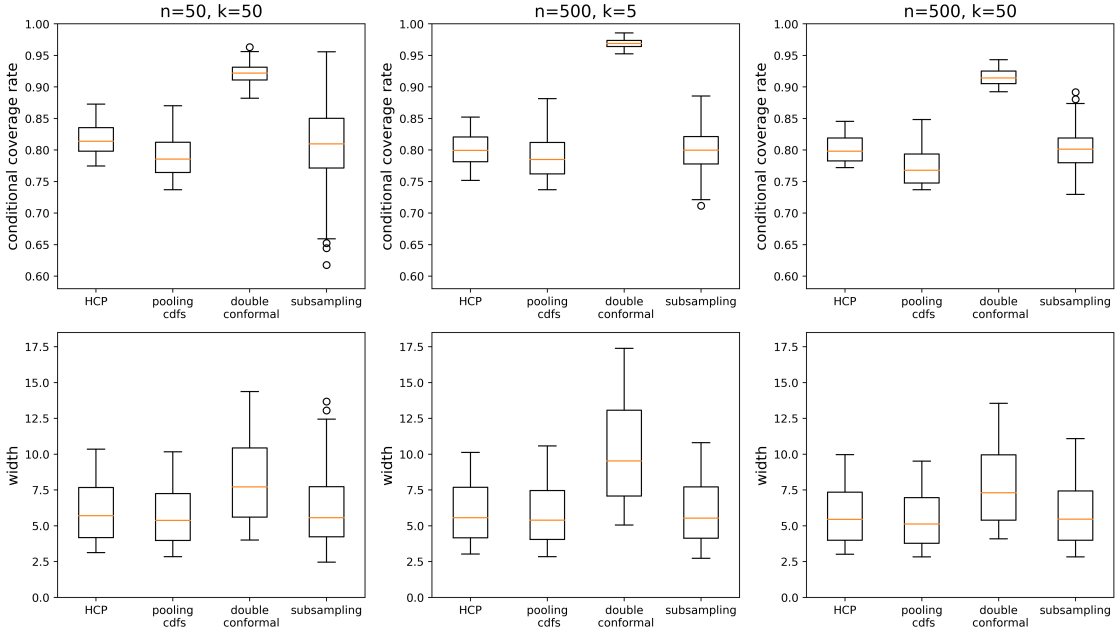

Figure 4.1: Conditional coverage rates and widths of hierarchical conformal prediction(HCP), pooling CDFs, double conformal, and subsampling.

The results for double conformal, pooling CDFs, and subsampling coincide with the discussion by [Dunn et al., 2022]. Double conformal tends to provide over-conservative predic-

tion sets, and subsampling suffers from higher variance due to reduced sample size(which is more severe in the first setting where n is small and k is large). Pooling CDFs works relatively well but can undercover for small n, as it only has an asymptotic coverage guarantee. Hierarchical conformal prediction is free from these issues and tends to provide non-conservative prediction sets in all three cases.

4.3 Distribution-free prediction with repeated measurements

We now look further into the setting where we have i.i.d data with repeated measurements, which is a special case of data with hierarchical structure. We assume that we have data $\{(X_i, \tilde{Y}_i) : i = 1, 2, \cdots, n\} \subset \mathcal{X} \times \mathcal{Y}$ ($\mathcal{Y} = \mathbb{R} \cup \mathbb{R}^2 \cup \mathbb{R}^3 \cup \cdots$) which is generated by

$$
\begin{aligned}
&X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} P_X \\
&K_1, K_2, \cdots, K_n \overset{\text{iid}}{\sim} P_K,\ (K_i)_{1 \le i \le n} \perp\!\!\!\perp (X_i)_{1 \le i \le n} \\
&\tilde{Y}_i = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,K_i}) | X_i, K_i \overset{\text{iid}}{\sim} P_{Y|X}.
\end{aligned}
$$

Compared to the general hierarchical data setting, we have a stronger assumption that the repeat number K_i is independent of the data—we will discuss later how this assumption can be relaxed. This setting can be thought of as a hierarchical data setting with $Z_i = (X_i, \tilde{Y}_i)$. The task is to provide inference for Y_{n+1} given a new input X_{n+1}, where Y_{n+1} is a draw from $P_{Y|X=X_{n+1}}$.

4.3.1 Marginal coverage guarantee via hierarchical conformal prediction

Let $s : \mathcal{X} \times \mathbb{R} \to \mathbb{R}^+$ be a score function fitted on separate training data, and define $S_{i,j} = s(X_i, Y_{i,j})$ for each $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, K_i$. For example, we can construct a mean estimator $\hat{\mu}$ and consider score $s(x, y) = |y - \hat{\mu}(x)|$. For any discrete

distribution $\mathcal{D} = \sum_{i \in I} p_i \delta_{x_i}$ and $0 < \beta < 1$, where $I \subset \mathbb{N}$ and $\{x_i : i \in I\} \subset \mathbb{R} \cup \{\infty\}$, define

$$Q_\beta(\mathcal{D}) = \min \left\{ x \in \mathbb{R} \cup \{\infty\} : \sum_{i \in I : x_i \leq x} p_i \geq \beta \right\}.$$

Applying the hierarchical conformal prediction from Theorem 5, we have the following prediction set which provides the marginal coverage guarantee.

Theorem 6 (HCP for repeated measurements setting)**.** *Let*

$$\widehat{C}_n(x) = \left\{ y : s(x,y) \leq Q_{1-\alpha} \left(\sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{S_{i,j}} + \frac{1}{n+1} \cdot \delta_\infty \right) \right\}. \tag{4.5}$$

Then

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

where the probability is taken with respect to $\{(X_i, K_i, \tilde{Y}_i)\}_{1 \leq i \leq n+1} \overset{\text{iid}}{\sim} P$ *and* $Y_{n+1} | X_{n+1}, \{(X_i, \tilde{Y}_i)\}_{1 \leq i \leq n} \sim P_{Y|X=X_{n+1}}$*. Moreover, if all* $S_{i,j}$*'s are distinct almost surely,*

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right\} \leq 1 - \alpha + \frac{2}{n+1}.$$

Remark 2. *The prediction set* (4.5) *is still valid if we only assume hierarchical exchangeability instead of independence, since it's based on the hierarchical conformal prediction.*

Together with the results in Appendix 4.5.1, our results show that we can have a valid distribution-free prediction set with data with repeated measurements, in terms of marginal coverage guarantee. This is of course not surprising since we are assuming that we have more observations compared to the standard setting. Given that we have such more information, can we expect a stronger inference that achieves beyond the marginal coverage guarantee?

4.3.2 Toward inference with conditional coverage guarantees

In this section, we discuss possible targets of distribution-free inference beyond the marginal coverage guarantee, by making use of the repeated measurements. In particular, we study the possibility of having a prediction set with a useful conditional coverage guarantee, which is known to be hard to achieve in the standard setting.

Controlling conditional miscoverage

Consider the miscoverage rate conditional on all the observations we have—training data $\{X_i, \tilde{Y}_i\}_{1\le i\le n}$ and the new observation X_{n+1}. Specifically, define

$$\alpha_n(x) = \mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \;\middle|\; \{X_i, \tilde{Y}_i\}_{1\le i\le n}, X_{n+1} = x\right\},$$

where the probability is taken with respect to $(X_1, \tilde{Y}_1), \cdots, (X_n, \tilde{Y}_n) \overset{\text{iid}}{\sim} P_{X,\tilde{Y}}$ and $Y_{n+1}|X_{n+1} \sim P_{Y|X=X_{n+1}}$ (independently from $(X_i, \tilde{Y}_i)_{1\le i\le n}$). The marginal coverage guarantee

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\right\} \le \alpha$$

can alternatively be expressed as

$$\mathbb{E}\left[\alpha_n(X_{n+1})\right] \le \alpha, \tag{4.6}$$

by the law of iterated expectation.

For the goal of controlling conditional miscoverage, an ideal target would be

$$\alpha_n(X_{n+1}) \le \alpha \text{ almost surely.} \tag{4.7}$$

This condition will ensure that whatever values of observations we have, the resulting predic-

tion set provides a good coverage for any value of new input X_{n+1}. However, guarantee (4.7) is not a realistic target in the distribution-free setting—for example, if P_X is continuous so that with probability 1, X_{n+1} is not equal to any of $X_1, \cdots, X_n$, the only prediction set that can satisfy (4.7) would be $\mathbb{R}$, which is of course not useful.

As an alternative/practical target, we consider the following stronger guarantee, which we denote as *second-moment coverage guarantee*.

$$\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \leq \alpha^2. \tag{4.8}$$

Note that this is a condition in-between the marginal coverage guarantee (4.6) and the ideal condition (4.7), in the sense that (4.7) implies (4.8) and (4.8) implies (4.6). Another way of understanding the guarantee (4.8) as a mechanism for controlling conditional miscoverage is to look at the tail probability. For a constant $c > 1$, the marginal coverage guarantee leads to

$$\mathbb{P}\left\{\alpha_n(X_{n+1}) > c\alpha\right\} \leq \frac{\mathbb{E}\left[\alpha_n(X_{n+1})\right]}{c\alpha} = \frac{1}{c},$$

while the stronger guarantee (4.8) provides

$$\mathbb{P}\left\{\alpha_n(X_{n+1}) > c\alpha\right\} = \mathbb{P}\left\{\alpha_n(X_{n+1})^2 > c^2\alpha^2\right\} \leq \frac{\mathbb{E}\left[\alpha_n(X_{n+1})^2\right]}{c^2\alpha^2} = \frac{1}{c^2}$$

from Markov's inequality. Hence, we can expect more uniformly small conditional miscoverage rate from the second-moment coverage guarantee.

Distribution-free prediction with second-moment coverage guarantee

Now we discuss how the stronger guarantee (4.8) can be achieved in the distribution-free sense. As in Section 4.2, we introduce an extension of split conformal method here, and discuss application of other methods in the Appendix.

Define $S_{i,j}$ as in the previous section. Theorem 7 provides a split conformal-based pre-

diction set that satisfies the second-moment coverage guarantee.

Theorem 7 (HCP^2 for repeated measurements)**.** *Let*

$$\widehat{C}_n(x) = \left\{ y : s(x,y) \leq Q_{1-\alpha^2} \left(\sum_{\substack{i \leq n \\ K_i \geq 2}} \sum_{j_1 < j_2} \frac{1}{(N_{\geq 2}+1) \cdot \binom{K_i}{2}} \cdot \delta_{\min\{S_{i,j_1}, S_{i,j_2}\}} + \frac{1}{N_{\geq 2}+1} \cdot \delta_\infty \right) \right\}, \tag{4.9}$$

where

$$N_{\geq 2} = \sum_{i=1}^{n} \mathbb{1}\left\{K_i \geq 2\right\}.$$

Then it holds that

$$\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \leq \alpha^2,$$

where the expectation is taken with respect to $\{(X_i, K_i, \tilde{Y}_i)\}_{1 \leq i \leq n} \overset{\text{iid}}{\sim} P$ *and* $X_{n+1} | \{(X_i, \tilde{Y}_i)\}_{1 \leq i \leq n} \sim P_X$*. Moreover, if* $S_{i,j}$*'s are all distinct almost surely, the following lower bound holds.*

$$\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \geq \alpha^2 - \frac{2}{(n+1)p_K},$$

where $p_K = \mathbb{P}\left\{K \geq 2\right\}$*.*

The underlying idea of the prediction interval in Theorem 7 is to make use of the exchangeability of X_i's as well as the exchageability of (S_{i,j_1}, S_{i,j_2}) pairs to apply the idea of conformal prediction and obtain a bound for $\min\{S_{n+1,1}, S_{n+1,2}\}$.

Remark 3. *As in Theorem 6, the assumptions can be weakened. It is sufficient to assume that* $\{(X_i, K_i, \tilde{Y}_i) : i = 1, 2, \cdots, n\}$ *are exchangeable. However, it is necessary to assume that* $Y_{i,1}, \cdots, Y_{i,K_i}$ *are i.i.d given* X_i *and* K_i*, for each* i*, to have the bound for the squared conditional miscoverage rate.*

Remark 4. *It is possible to have similar results also in the case that* K *depends on* X*. Specifically,*

1. *For HCP, Theorem 6 holds also in the case that K depends on X. HCP^2 in Theorem 7 satisfies $\mathbb{E}_{X,\geq 2}\left[\alpha_n(X_{n+1})^2\right] \leq \alpha^2$, where $P_{X,\geq 2}$ denotes the conditional distribution of X given $K \geq 2$.*

2. *If $P_{K|X}$ is known, then we can apply weighted conformal prediction ([Tibshirani et al., 2019]) to get coverage under P_X rather than under $P_{X,\geq 2}$.*

3. *The guarantee $\mathbb{E}_{X,\geq 2}\left[\alpha_n(X_{n+1})^2\right] \leq \alpha^2$ implies*

$$\begin{aligned}
&\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \\
&= \mathbb{P}\left\{K_{n+1} \geq 2\right\} \cdot \mathbb{E}\left[\alpha_n(X_{n+1})^2 \;\middle|\; K_{n+1} \geq 2\right] \\
&\qquad + \mathbb{P}\left\{K_{n+1} = 1\right\} \cdot \mathbb{E}\left[\alpha_n(X_{n+1})^2 \;\middle|\; K_{n+1} = 1\right] \\
&\leq \alpha^2 + \mathbb{P}\left\{K_{n+1} = 1\right\} \cdot (1 - \alpha^2)
\end{aligned}$$

   Therefore, for a sufficiently large repeat probability (i.e., small $\mathbb{P}\left\{K_{n+1} = 1\right\}$), we can obtain a useful bound for $\mathbb{E}\left[\alpha_n(X_{n+1})^2\right]$.

4.3.3 Examples

We look into examples with popular choices of score function. Suppose we fit a mean function $\hat{\mu}$ and use the residual score $s(x, y) = |y - \hat{\mu}(x)|$. HCP with this score provides the following prediction set.

$$\widehat{C}_n(x) = \hat{\mu(x)} \pm Q_{1-\alpha}\left(\sum_{i=1}^{n}\sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{R_{i,j}} + \frac{1}{n+1} \cdot \delta_\infty\right), \tag{4.10}$$

where $R_{i,j} = |Y_{i,j} - \hat{\mu}(X_i)|$. HCP^2 provides

$$\widehat{C}_n(x) = \hat{\mu}(x) \pm Q_{1-\alpha^2}\left(\sum_{\substack{i\le n\\ K_i\ge 2}}\sum_{j_1<j_2} \frac{1}{(N_{\ge 2}+1)\cdot\binom{K_i}{2}}\cdot\delta_{\min\{R_{i,j_1},R_{i,j_2}\}} + \frac{1}{N_{\ge 2}+1}\cdot\delta_\infty\right). \tag{4.11}$$

The widths of both prediction sets are determined by the quantiles and do not depend on the new input X_{n+1}. In cases where we desire prediction set that the width differs by the value of X_{n+1}, we can consider alternative score functions. For example, in the setting where we have sufficiently large number of repeats, we might prefer to construct an estimator $\hat{\sigma}$ of the conditional variance $\mathrm{var}(Y|X)$ together with the conditional mean estimator $\hat{\mu}$, and choose to use the following rescaled residual instead (see. e.g., [Lei et al., 2018] for more discussions on the use of this nonconformity score).

$$S_{i,j} = s(X_i, Y_{i,j}), \text{ where } s(x,y) = \frac{|y-\hat{\mu}(x)|}{\hat{\sigma}(x)}.$$

Applying HCP and HCP^2 with the standardized residuals, we have

$$\widehat{C}_n(x) = \hat{\mu}(x) \pm \hat{\sigma}(x)\cdot Q_{1-\alpha}\left(\sum_{i=1}^{n}\sum_{j=1}^{K_i}\frac{1}{(n+1)K_i}\cdot\delta_{S_{i,j}} + \frac{1}{n+1}\cdot\delta_\infty\right) \tag{4.12}$$

and

$$\widehat{C}_n(x) = \hat{\mu}(x) \pm \hat{\sigma}(x)\cdot Q_{1-\alpha^2}\left(\sum_{\substack{i\le n\\ K_i\ge 2}}\sum_{j_1<j_2} \frac{1}{(N_{\ge 2}+1)\cdot\binom{K_i}{2}}\cdot\delta_{\min\{S_{i,j_1},S_{i,j_2}\}} + \frac{1}{N_{\ge 2}+1}\cdot\delta_\infty\right). \tag{4.13}$$

The widths of the prediction sets (4.12) and (4.13) depend on the input x, enabling us to avoid situations where we have unnecessarily wide prediction sets even for the values of X_{n+1} where Y has small conditional variance $\mathrm{Var}(Y \mid X = X_{n+1})$. However, a good variance estimator (in addition to the mean estimator) would also be necessary for these

prediction sets to be useful.

Remark 5. *If we use a constant estimator $\hat{\sigma} \equiv c$ for some fixed $c > 0$, then the prediction sets* (4.12) *and* (4.13) *are equivalent to* (4.10) *and* (4.11)*, respectively.*

4.3.4 Additional remarks

The prediction sets in Theorem 6 and 7 apply split conformal prediction, and the same idea can be applied to have full conformal-based prediction sets. For example, as a full conformal version of (4.5), we can construct

$$\tilde{C}_n(X_{n+1}) = \left\{ y \in \mathcal{Y} : s^y(X_{n+1}, y) \leq Q_{1-\alpha} \left(\sum_{i=1}^{n+1} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{s^y(X_i, Y_{i,j})} \right) \right\},$$

where s^y denotes the score fitted on $(X_1, \tilde{Y}_1), \cdots, (X_n, \tilde{Y}_n), (X_{n+1}, y)$, and show that $P_1 \tilde{C}_n(X_{n+1})$ is a valid $(1-\alpha)$-prediction set(where $P_1 : \mathcal{Y} \to \mathbb{R}$ is a projection to the first component). As in the relation between the full conformal and the split conformal in the standard setting, we have from the above prediction set the advantage of using more data, with the disadvantage of having a larger computational cost. However, in our case, the disadvantage can easily outweigh the advantage as we need to repeat the computations for the elements in $\mathcal{Y} = \mathbb{R} \cup \mathbb{R}^2 \cup \mathbb{R}^3 \cup \cdots$. For practical purpose, we limit our discussion to the split conformal-based prediction sets in this work.

Another possible extension of the idea used in this section is to aim for an even stronger guarantee

$$\mathbb{E}\left[\alpha_n(X_{n+1})^l\right] \leq \alpha^l \tag{4.14}$$

for $l \geq 3$. This guarantee with a larger l will work as a better proxy of the ideal target (4.7), but is followed by several limitations:

1. To apply the idea of Theorem 7 to achieve (4.14), we can use $N_{\geq l} = \sum_{i=1}^{n} \mathbb{1}\{K_i \geq l\}$

points only.

2. Even if we have a large enough $N_{\geq l}$, we would have the $(1-\alpha^l)$-quantile in place of the $(1-\alpha^2)$ quantile in (4.9) (the distribution inside will have masses on the minimums of sets of l residuals), and it is likely to be infinity unless the sample size is extremely large.

4.3.5 Simulations

We demonstrate a few simulation results to illustrate the performance of the prediction sets from the procedures we proposed. The data is generated as follows:

$$
\begin{aligned}
&X_i \sim \text{Unif}([0,5]) \\
&K_i \equiv 2, \tilde{Y}_i = (Y_{i,1}, Y_{i,2}) | X_i = x \overset{\text{iid}}{\sim} \mathcal{N}(\mu(x), \sigma(x)) \\
&\mu(x) = 1 + x + 0.1 \cdot x^2.
\end{aligned}
$$

For the conditional variance $\sigma(x)$, we look into two settings:

$$
\begin{aligned}
&\text{Setting 1} : \sigma(x) \equiv 2 \\
&\text{Setting 2} : \sigma(x) = \mathbb{1}\{x < 3\} + (1 + 4(x-3)^4) \cdot \mathbb{1}\{3 \leq x < 4\} + 5 \cdot \mathbb{1}\{x \geq 4\}
\end{aligned}
$$

Setting 2 reflects the case where we have different difficulty levels of prediction, while setting 1 represents the homogeneous difficulty case. Figure 4.2 illustrates this.

We use training size $n = 500$ and level $\alpha = 0.2$. We repeat generating the training data and the prediction set 500 times so that we have 1000 samples of $\alpha_n(X_{n+1})$'s. To have the estimate $\hat{\mu}$, we generate a separate training data of size 500 and fit linear regression, and for $\hat{\sigma}$ we compute the sample standard deviation for each individual and fit kernel regression with box kernel $K_h(x) = \frac{1}{2h} \cdot \mathbb{1}\{|x| < h\}$ with bandwidth $h = 0.5$.

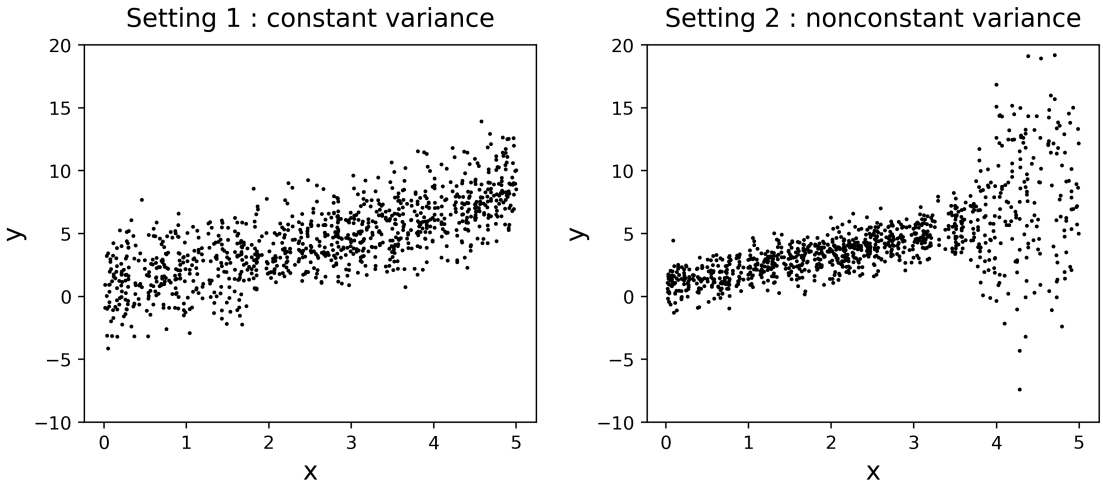

Figure 4.2: Scatter plot of datasets from setting 1(constant variance case) and setting 2(nonconstant variance case).

The first row of Figure 4.3 shows the distributions of $\alpha_n(X_{n+1})$ and the length of $\widehat{C}_n(X_{n+1})$ for HCP and HCP^2 with score $s(x, y) = |y - \hat{\mu}(x)|$ ((4.10) and (4.11)), in setting 1. In this case the conditional coverage rates and the lengths of the two prediction intervals have similar distributions, suggesting that the second-moment coverage guarantee does not lead to an overly conservative prediction set when it is possible to attain small conditional coverage rates with a short width. The second row shows the result for setting 2. The marginal coverage guarantee provides a narrower prediction interval, but with the cost of 'giving up' on the coverage for some values of the feature. The second-moment coverage guarantee leads to a wide prediction interval but instead achieves a small conditional miscoverage for most feature values.

These results suggest that the marginal coverage guarantee and the second-moment coverage guarantee return similar prediction sets in the 'easy' case, while they make different choices in the 'hard' case where it is difficult to achieve both the short length of the interval and a good control of conditional coverage. In the tradeoff between short prediction interval and uniformly small conditional miscoverage, the marginal coverage guarantee prioritizes the

former while the second-moment coverage guarantee prioritizes the latter.

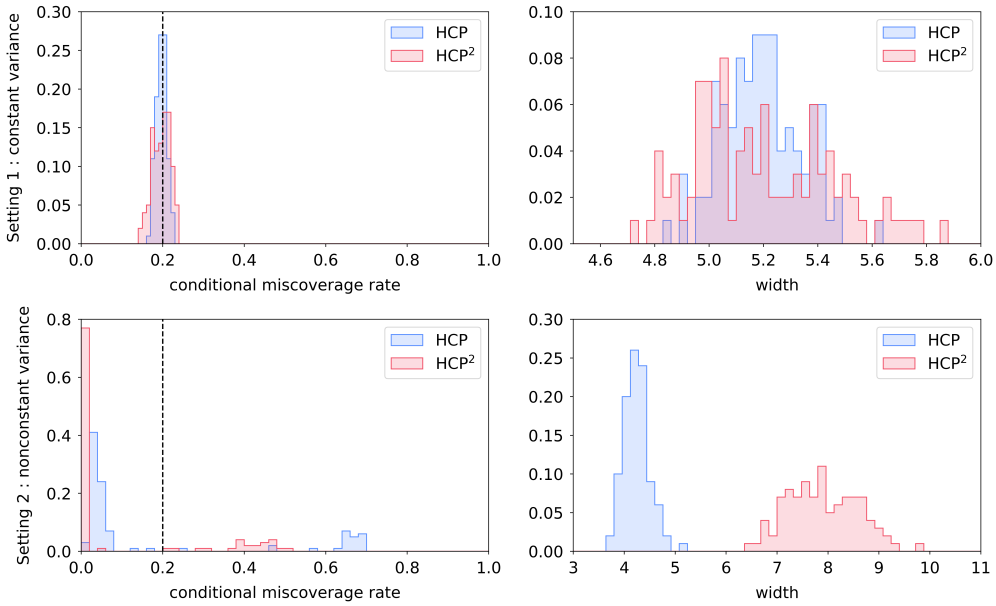

Figure 4.3: Conditional miscoverage rates and widths of HCP and HCP^2 constructed via score $s(x, y) = |y - \hat{\mu}(x)|$ ((4.10) and (4.11)).

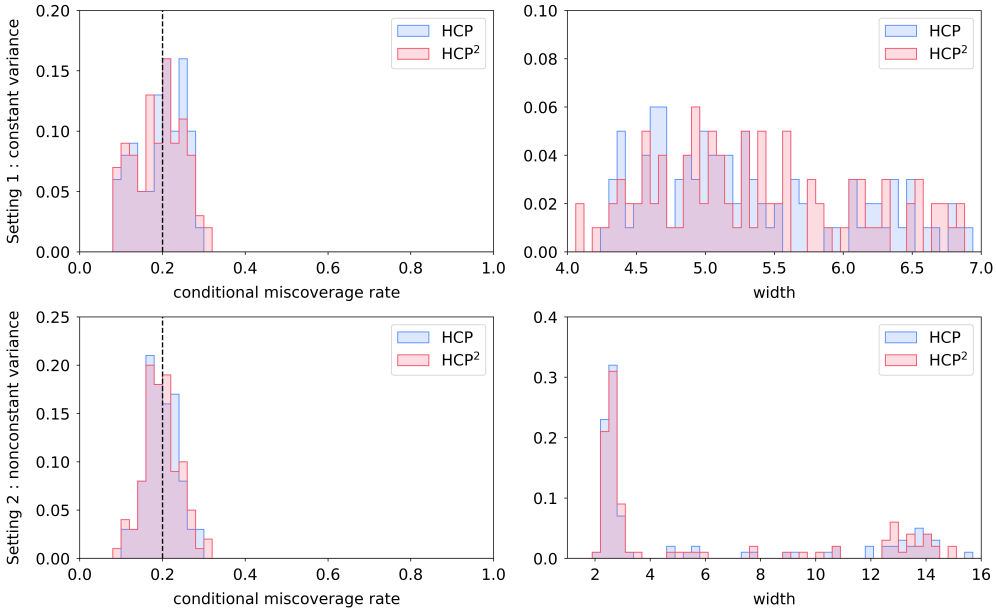

Figure 4.4: Conditional miscoverage rates and widths of HCP and HCP^2 constructed via score $s(x, y) = |y - \hat{\mu}(x)|/\hat{\sigma}(x)$ ((4.12) and (4.13)).

Next, we investigate the performance of the prediction sets from the score $s(x, y) = |y - \hat{\mu}(x)|/\hat{\sigma}(x)$ ((4.12) and (4.13)). The two guarantees now provide similar prediction sets in both settings (Figure 4.4). Note that the prediction sets (4.10) and (4.11) can be thought of as special cases of (4.12) and (4.13), with $\hat{\sigma} \equiv 1$ (or equivalently, $\hat{\sigma} \equiv c$ for any $c > 0$). Therefore, one possible interpretation for these results would be: if the choice of the rescaled residual as score is good and the model is accurate so that we have good estimators $\hat{\mu}$ and $\hat{\sigma}$, then it is possible to achieve the second-moment coverage guarantee with a non-conservative prediction set (Figure 4.3 and 4.4), whereas there is a severe tradeoff between the conservativeness and the conditional miscoverage control in the case that the quality of estimators is poor. We can extend these observations to leverage the conservativeness level of HCP^2 prediction set to evaluate the model and estimators (in terms of how good they are overall across the input space).

4.4 Discussion

In this chapter, we looked into the problem of distribution-free inference in the setting where we have a data with hierarchical structure and proposed hierarchical conformal prediction, and then discussed the repeated measurements setting with a focus on the control of conditional miscoverage. The empirical results support that the target guarantee we propose in this work can work as a good conditional miscoverage controller, and that if we have a good estimate of the conditional variance then we can obtain a prediction set which is as short as the one from marginal coverage guarantee while having a guarantee for the conditional coverage rates.

Many open questions are remaining. In the setting where we can determine the number of repeats K_i in the data collection stage, what would be the optimal strategy—in terms of the distribution of K_i, ratio of repeats and the total number of data, etc.? In our simulations where $K \equiv 2$ was used, we saw that the prediction set with the stronger guarantee can work

as a good conditional miscoverage controller. Therefore, we might prefer to set K_i's small and instead have a larger number of individuals in our sample. However, having a larger repeat number has advantages in terms of the accuracy of the conditional variance estimator. What would be the best choice in this tradeoff? Another important tradeoff is in the target guarantee. Between the marginal coverage guarantee and the strict guarantee (4.7), our choice in this work was (4.8), which bounds the expected squared conditional miscoverage. Would it be an optimal choice also in the case where we have a large sample size or large number of repeats?

Similarly to Lee and Barber [2021]'s work, our results show that having even a small amount of repeats can significantly expand the realm of what distribution-free inference can do. This raises a more general question: what reasonable additions can we make to the setting so that a more useful inference is possible in a distribution-free manner? We aim to explore more of these types of problems in our future works.

4.5 Appendix

4.5.1 Application of other distribution-free methods

In section 4.2 and 4.3, we introduced methods that apply split conformal prediction to construct prediction sets with the marginal coverage guarantee and the stronger guarantee. Similar ideas can be applied to derive extensions of other methods—here we discuss extensions of full conformal [Vovk et al., 2005] and jackknife+ [Barber et al., 2021b] methods which enable inference without loss in data size.

Hierarchical full conformal prediction

The following theorem provides a full conformal-based prediction set which provides a marginal coverage guarantee in the hierarchical data setting.

Theorem 8. *Given dataset* $\{\tilde{Z}_i : i = 1, 2, \cdots, n\}$ *with hierarchical exchangeability and a symmetric algorithm* $\mathcal{A} : (\mathcal{Z} \cup \mathcal{Z}^2 \cup \cdots)^{n+1} \to \mathbb{R}^{n+1}$*, define* $\widehat{C}_n = P_1 \tilde{C}_n$ *where* P_1 *denotes the projection to the first component and*

$$\tilde{C}_n = \left\{ z \in \mathcal{Z} \cup \mathcal{Z}^2 \cup \cdots : s^z_{n+1,1} \leq Q_{1-\alpha} \left(\sum_{i=1}^{n+1} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{s^z_{i,j}} \right) \right\},$$

where $(\tilde{s}^z_1, \tilde{s}^z_2, \cdots, \tilde{s}^z_n, \tilde{s}^z_{n+1}) = \mathcal{A}(\tilde{Z}_1, \tilde{Z}_2, \cdots, \tilde{Z}_n, z)$ *and* $\tilde{s}^z_i = (s^z_{i,1}, s^z_{i,2}, \cdots, s^z_{i,K_i})$ *for each* $1 \leq i \leq n+1$.

Then it holds that

$$\mathbb{P}\left\{ Z_{n+1} \in \widehat{C}_n \right\} \geq 1 - \alpha.$$

We omit the proof as it is almost the same as the proof of Theorem 5. Since $\mathcal{Z} \cup \mathcal{Z}^2 \cup \cdots$ is a space of vectors of arbitrary lengths, the prediction set in Theorem 8 has a heavier computational cost than the full conformal method in standard settings, which already is known to have a high computational cost. Note also that the prediction set contains z's with a condition that depends only on the first component of z. It is therefore unlikely to be useful in practice, and we introduce the extension here just to show that it is theoretically possible to have an extension of full conformal prediction.

In the repeated measurements setting, the extension of full conformal for the stronger guarantee is also possible in a similar way, but we omit the result as it is again impractical.

Hierarchical jackknife+

Next, we introduce methods based on jackknife+ [Barber et al., 2021b], which provides a valid distribution-free prediction set without data splitting and with a relatively low computational cost. Here we restrict our discussion to the repeated measurements setting with hierarchical i.i.d structure and construct prediction sets with the marginal coverage guarantee and the second-moment coverage guarantee.

Given dataset $\{(X_i, \tilde{Y}_i)\}_{1 \le i \le n}$ with repeated measurements, consider leave-one-out residuals

$$R_{ij}^{\mathrm{LOO}} = |Y_{ij} - \hat{\mu}_{-i}(X_i)|, i = 1, 2, \cdots, n, j = 1, 2, \cdots, K_i,$$

where $\hat{\mu}_{-i} = \mathcal{A}((X_1, \tilde{Y}_1), \cdots, (X_{i-1}, \tilde{Y}_{i-1}), (X_{i+1}, \tilde{Y}_{i+1}), \cdots, (X_n, \tilde{Y}_n))$ denotes the estimator from a symmetric procedure $\mathcal{A}$ and the training data with $(X_i, \tilde{Y}_i)$ excluded. For any distribution $\mathcal{D}$, define

$$Q_\alpha^-(\mathcal{D}) = \sup\{t : \mathbb{P}_{X \sim \mathcal{D}}\{X < t\} < \alpha\}$$

and

$$Q_\alpha^+(\mathcal{D}) = \inf\{t : \mathbb{P}_{X \sim \mathcal{D}}\{X > t\} < \alpha\}.$$

We allow the input $\mathcal{D}$ to be a conditional distribution, e.g., $P_{X|Z}$ for some random variables X and Z, in which case $Q_\alpha^-(\mathcal{D})$ and $Q_\alpha^+(\mathcal{D})$ are random. The following prediction set provides the marginal coverage guarantee in the repeated measurements setting.

Theorem 9. *Let*

$$\widehat{C}_n(x) = \left[Q_\alpha^- \left(\sum_{i=1}^n \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{\hat{\mu}_{-i}(x) - R_{ij}^{LOO}} + \frac{1}{n+1} \cdot \delta_{-\infty} \right), \right.$$
$$\left. Q_\alpha^+ \left(\sum_{i=1}^n \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{\hat{\mu}_{-i}(x) + R_{ij}^{LOO}} + \frac{1}{n+1} \cdot \delta_{\infty} \right) \right].$$

Then

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right\} \ge 1 - 2\alpha,$$

where the probability is taken with respect to $\{(X_i, K_i, \tilde{Y}_i)\}_{1 \le i \le n+1} \overset{\text{iid}}{\sim} P$ *and* $Y_{n+1} | X_{n+1}, \{(X_i, \tilde{Y}_i)\}_{1 \le i \le n} \sim P_{Y|X=X_{n+1}}$.

Note that as in the original jackknife+ methods for the standard setting, we have $1 - 2\alpha$ as the lower bound. Similarly, we can construct a jackknife+-based prediction set with a

bound on the second moment of $\alpha_n(X_{n+1})$.

Theorem 10. *Let*

$$\widehat{C}_n(x) = \left[Q^-_{\alpha^2}\left(\sum_{\substack{1\le i\le n \\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{(N_{\ge 2}+1)\binom{K_i}{2}} \cdot \delta_{\hat{\mu}_{-i}(x)-\min\{R^{LOO}_{ij_1}, R^{LOO}_{ij_2}\}} + \frac{1}{N_{\ge 2}+1}\cdot\delta_{-\infty} \middle| N_{\ge 2}\right), \right.$$
$$\left. Q^+_{\alpha^2}\left(\sum_{\substack{1\le i\le n \\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{(N_{\ge 2}+1)\binom{K_i}{2}} \cdot \delta_{\hat{\mu}_{-i}(x)+\min\{R^{LOO}_{ij_1}, R^{LOO}_{ij_2}\}} + \frac{1}{N_{\ge 2}+1}\cdot\delta_{\infty} \middle| N_{\ge 2}\right) \right].$$

Then

$$\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \le 4\alpha^2.$$

Remark 6. *Similarly to the split conformal-based methods, the results hold also in the case that K and X are dependent, as long as we replace the result in Theorem 10 by $\mathbb{E}_{P_{X,\ge 2}}\left[\alpha_n(X_{n+1})^2\right] \le 4\alpha^2$, where $P_{X,\ge 2}$ is defined in Remark 4.*

4.5.2 Extension—inference for regression with repeated measurements

We look into the regression problem in the repeated measurements setting (which is defined in Section 4.3), where the task is to provide inference for conditional mean of Y_{n+1} given X_{n+1}. Suppose Y is bounded, and for simplicity assume $Y \in [0,1]$. Given a dataset $\{X_i, \tilde{Y}_i\}_{1\le i\le n}$ with repeated label observations, we investigate the possibility of a useful distribution-free inference on the conditional mean $\mathbb{E}\left[Y_{n+1} \mid X_{n+1}\right]$ at a new input X_{n+1}.

To make this concrete, let us write $\mu_P(x)$ to denote the conditional mean $\mathbb{E}_{(X,Y)\sim P}\left[Y \mid X = x\right]$. Our task is to construct a set $\widehat{C}_n(X_{n+1})$ such that

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \ge 1-\alpha \tag{4.15}$$

holds under any distribution P.

In the standard setting where we have dataset $\{(X_i, Y_i)\}$ without repeated measurements, prior works have shown impossibility results for the above goal in the case P_X is nonatomic, while in the case X is discrete, there's a possibility of having a useful distribution-free confidence set [Barber, 2020, Lee and Barber, 2021]. In the setting with repeated measurements, it turns out that it is possible to have a meaningful confidence interval also for nonatomic X.

We construct an algorithm that provides a confidence set that satisfies the above guarantee and at the same time has a vanishing length, following the idea of Lee and Barber [2021]. First, we prepare an estimate $\mu : \mathcal{X} \to [0, 1]$ of μ_p—for example, we can split the training data and use one split for the construction of the estimate. Given an estimate $\mu(\cdot)$ and the training data $\{X_i, \tilde{Y}_i\}_{1 \le i \le n}$, we compute

$$Z = \frac{1}{N_{\ge 2}} \cdot \sum_{i : K_i \ge 2} \left((\bar{Y}_i - \mu(X_i))^2 - \frac{S_i^2}{K_i} \right)$$

in the case $N_{\ge 2} \ge 1$, where

$$N_{\ge 2} = \sum_{i=1}^{n} \mathbb{1}\{K_i \ge 2\}$$

denotes the number of individuals with repeated label observations, and

$$\bar{Y}_i = \frac{1}{K_i} \sum_{j=1}^{K_i} Y_{i,j}, \; S_i^2 = \frac{1}{K_i - 1} \sum_{i=1}^{K_i} (Y_{i,j} - \bar{Y}_i)^2$$

denote the sample mean and the sample variance of $\{Y_{ij} : 1 \le j \le K_i\}$. Next, for a fixed $\delta \in (0, \alpha)$, let

$$\hat{\Delta} = \begin{cases} Z_+ + \frac{5}{4} \sqrt{\frac{1}{2N_{\ge 2}} \cdot \log \frac{1}{\delta}}, & N_{\ge 2} \ge 1 \\ \infty, & N_{\ge 2} = 0, \end{cases}$$

where $Z_+ = \max\{Z, 0\}$. We then define

$$\widehat{C}_n(x) = \left[\max\left\{0, \mu(x) - \sqrt{(\alpha-\delta)^{-1}\hat{\Delta}}\right\}, \min\left\{1, \mu(x) + \sqrt{(\alpha-\delta)^{-1}\hat{\Delta}}\right\}\right]. \tag{4.16}$$

The following theorem proves that the confidence interval (4.16) is a valid distribution-free confidence interval for the conditional mean, and that it can have a vanishing length.

Theorem 11. *The confidence interval constructed in* (4.16) *satisfies coverage guarantee* (4.15)*. Moreover, it holds that*

$$\mathbb{E}\left[\mathrm{leb}(\widehat{C}_n(X_{n+1})) \;\middle|\; N_{\geq 2} \geq 1\right] \leq c \cdot \left[\sqrt{err_\mu} + n^{-\frac{1}{4}}\right],$$

where $err_\mu = (\mathbb{E}_{X\sim P_X}\left[(\mu_P(X) - \mu(X))^2\right])^{1/2}$ *and* $c > 0$ *depends only on* α, δ*, and* $p_K = \mathbb{P}\{K \geq 2\}$*.*

4.5.3 Proof of Theorem 5

The proof follows the idea of Tibshirani et al. [2019]. Let $S_{i,j} = s(Z_{i,j})$ be the score of $Z_{i,j}$. Note that it is sufficient to prove

$$\mathbb{P}\left\{Z_{n+1,1} \in \widehat{C}_n\right\} \geq 1 - \alpha,$$

where we assume $\tilde{Z}_{n+1} = (Z_{n+1,1}, \cdots, Z_{n+1,K_{n+1}})$ is exchangeable with $\tilde{Z}_1, \cdots, \tilde{Z}_n$. Assume there are no ties among $S_{i,j}$'s almost surely, and let E_s be the event that $\{S_1, S_2, \cdots, S_{n+1}\} = \{s_1, s_2, \cdots, s_{n+1}\}$ where $s_i = \{s_{i,1}, \cdots, s_{i,k_i}\} \subset \mathbb{R}$ for each $1 \leq i \leq n+1$ and we write $s = \{s_1, \cdots, s_{n+1}\}$. Then

$$\mathbb{P}\left\{S_{n+1,1} = s_{i,j} \mid E_s\right\} = \mathbb{P}\left\{S_{n+1,1} = s_{i,j} \mid S_i = s_i\right\} \cdot \mathbb{P}\left\{S_i = s_i \mid E_s\right\} = \frac{1}{k_i} \cdot \frac{1}{n+1},$$

where the last equality follows from the exchangeability of $(S_{n+1,1}, \cdots, S_{n+1,K_{n+1}})$ and the exchangeability of $(S_1, \cdots, S_{n+1})$. Therefore,

$$S_{n+1,1}|E_s \sim \sum_{i=1}^{n+1} \sum_{j=1}^{k_i} \frac{1}{(n+1)k_i} \cdot \delta_{s_{i,j}}.$$

This result holds also in the case we have ties: Let $A_s = \{s_{i,j} : 1 \le i \le n+1, 1 \le j \le k_i\}$ be the set of $s_{i,j}$ values and define $I_s = \{(i,j) : s_{i,j} = s\}$ for each $s \in A_s$. Note that $\{I_s : s \in A_s\}$ is a partition of $\{(i,j) : 1 \le i \le n+1, 1 \le j \le k_i\}$. Then we have

$$\mathbb{P}\left\{S_{n+1,1} = s \;\middle|\; E_s\right\} = \sum_{(i,j)\in I_s} \frac{1}{k_i} \cdot \frac{1}{n+1}$$

from the two layes of exchangeability, and thus

$$S_{n+1,1}|E_s \sim \sum_{s\in A_s} \sum_{(i,j)\in I_s} \frac{1}{(n+1)k_i} \cdot \delta_s \equiv \sum_{i=1}^{n+1} \sum_{j=1}^{k_i} \frac{1}{(n+1)k_i} \cdot \delta_{s_{i,j}}.$$

Now let $\mathcal{D}^s$ denote the above distribution. By definition of $Q_{1-\alpha}$, we have

$$1 - \alpha \le \mathbb{P}\left\{S_{n+1} \le Q_{1-\alpha}(\mathcal{D}_s) \mid E_s\right\}.$$

We next define distribution $\tilde{\mathcal{D}}^s$ by

$$\sum_{i=1}^{n} \sum_{j=1}^{k_i} \frac{1}{(n+1)k_i} \cdot \delta_{s_{i,j}} + \frac{1}{n+1} \cdot \delta_\infty.$$

Since $Q_{1-\alpha}(\mathcal{D}_s) \le Q_{1-\alpha}(\tilde{\mathcal{D}}_s)$ clearly, it holds that

$$1 - \alpha \le \mathbb{P}\left\{S_{n+1} \le Q_{1-\alpha}(\tilde{\mathcal{D}}_s) \;\middle|\; E_s\right\}.$$

Next, under the assumption that s_{ij}'s have no tie, the two distributions $\mathcal{D}_s$ and $\tilde{\mathcal{D}}_s$ differ only at $s_{n+1,1}, \cdots, s_{n+1,k_{n+1}}$ of $\mathcal{D}_s$ and $+\infty$ of $\tilde{\mathcal{D}}_s$ which have total mass $1/(n+1)$. Hence, for any $t \in \mathbb{R}$, it holds that

$$\mathbb{P}_{S\sim\mathcal{D}_s}\{S < t\} \leq \mathbb{P}_{S\sim\tilde{\mathcal{D}}_s}\{S < t\} + \frac{1}{n+1}.$$

From this inequality and the definition of $Q_{1-\alpha}$, we have

$$\mathbb{P}_{S\sim\mathcal{D}_s}\left\{S < Q_{1-\alpha}(\tilde{\mathcal{D}}_s)\right\} \leq \mathbb{P}_{S\sim\tilde{\mathcal{D}}_s}\left\{S < Q_{1-\alpha}(\tilde{\mathcal{D}}_s)\right\} + \frac{1}{n+1} < 1-\alpha+\frac{1}{n+1},$$

which implies

$$Q_{1-\alpha+\frac{1}{n+1}}(\mathcal{D}_s) \geq Q_{1-\alpha}(\tilde{\mathcal{D}}_s).$$

It follows that

$$\begin{aligned}\mathbb{P}\left\{S_{n+1,1} \leq Q_{1-\alpha}(\tilde{\mathcal{D}}^s) \;\middle|\; E_s\right\} &\leq \mathbb{P}\left\{S_{n+1,1} \leq Q_{1-\alpha+\frac{1}{n+1}}(\mathcal{D}^s) \;\middle|\; E_s\right\} \\ &\leq 1-\alpha+\frac{1}{n+1}+\mathbb{P}\left\{S_{n+1,1} = Q_{1-\alpha}(\tilde{\mathcal{D}}^s) \;\middle|\; E_s\right\} \leq 1-\alpha+\frac{2}{n+1}.\end{aligned}$$

From the above results, we have

$$\begin{aligned}1-\alpha \leq \mathbb{P}\left\{S_{n+1,1} \leq Q_{1-\alpha}\left(\sum_{i=1}^{n}\sum_{j=1}^{K_i}\frac{1}{(n+1)K_i}\cdot\delta_{S_{i,j}}+\frac{1}{n+1}\delta_\infty\right)\right\} \\ = \mathbb{E}\left[\mathbb{P}\left\{S_{n+1,1} \leq Q_{1-\alpha}(\tilde{\mathcal{D}}^s) \;\middle|\; E_s\right\}\right] \leq 1-\alpha+\frac{2}{n+1},\end{aligned}$$

where the right inequality holds only under the assumption that $S_{i,j}$'s have no tie almost surely. Therefore,

$$1-\alpha \leq \mathbb{P}\left\{Z_{n+1} \in \widehat{C}_n\right\} = \mathbb{P}\left\{Z_{n+1,1} \in \widehat{C}_n\right\} \leq 1-\alpha+\frac{2}{n+1}.$$

4.5.4 Proof of Theorem 6

The desired inequality follows directly by applying Theorem 5 with $Z_{i,j} = (X_i, Y_{i,j})$ and $S_{i,j} = s(X_i, Y_{i,j})$.

4.5.5 Proof of Theorem 7

Again, assume there is no tie almost surely and define E_s to be the event that $\{S_1, S_2, \cdots, S_{n+1}\} = \{s_1, s_2, \cdots, s_{n+1}\}$ where $s_i = \{s_{i,1}, \cdots, s_{i,k_i}\} \subset \mathbb{R}$ for each $1 \leq i \leq n+1$.

Then for any i with $k_i \geq 2$ and $1 \leq j_1 \neq j_2 \leq k_i$,

$$\begin{aligned}
&\mathbb{P}\left\{S_{n+1,1} = s_{i,j_1}, S_{n+1,2} = s_{i,j_2} \;\middle|\; E_s \cap \{K_{n+1} \geq 2\}\right\} \\
=&\mathbb{P}\left\{S_{n+1,1} = s_{i,j_1}, S_{n+1,2} = s_{i,j_2} \;\middle|\; S_{n+1} = s_i\right\} \cdot \mathbb{P}\left\{S_{n+1} = s_i \mid E_s \cap \{K_{n+1} \geq 2\}\right\} \\
=&\frac{1}{k_i(k_i-1)} \cdot \frac{1}{\sum_{i=1}^{n+1} \mathbb{1}\left\{k_i \geq 2\right\}}.
\end{aligned}$$

Therefore,

$$(S_{n+1,1}, S_{n+1,2}) | E_s \cap \{K_{n+1} \geq 2\} \sim \sum_{i:k_i \geq 2} \sum_{1 \leq j_1 \neq j_2 \leq k_i} \frac{1}{\sum_{i=1}^{n+1} \mathbb{1}\left\{k_i \geq 2\right\} \cdot k_i(k_i-1)} \cdot \delta_{s_{i,j_1}, s_{i,j_2}}.$$

Again, this holds also in the case there are ties, and it can be proved by the same logic in the proof of Theorem 5. Hence,

$$\begin{aligned}
&\mathbb{P}\Bigg\{ \min\{S_{n+1,1}, S_{n+1,2}\} \leq Q_{1-\alpha^2}\left(\sum_{i:K_i \geq 2} \sum_{1 \leq j_1 < j_2 \leq K_i} \frac{1}{(N_{\geq 2}+1) \cdot \binom{K_i}{2}} \cdot \delta_{\min\{S_{i,j_1}, S_{i,j_2}\}} \right) \\
&\qquad\qquad \Bigg| E_s \cap \{K_{n+1} \geq 2\} \Bigg\} \\
=&\mathbb{P}\Bigg\{ \min\{S_{n+1,1}, S_{n+1,2}\}
\end{aligned}$$

$$\leq Q_{1-\alpha^2}\left(\sum_{i:K_i\geq 2}\sum_{1\leq j_1<j_2\leq K_i}\frac{1}{(\sum_{i=1}^{n+1}\mathbb{1}\{K_i\geq 2\})\cdot\binom{K_i}{2}}\cdot\delta_{\min\{S_{i,j_1},S_{i,j_2}\}}\right)\Bigg|E_s\cap\{K_{n+1}\geq 2\}\Bigg\}$$

$$=\mathbb{P}\Bigg\{\min\{S_{n+1,1},S_{n+1,2}\}\leq Q_{1-\alpha^2}\left(\sum_{i:k_i\geq 2}\sum_{1\leq j_1<j_2\leq k_i}\frac{1}{(\sum_{i=1}^{n+1}\mathbb{1}\{k_i\geq 2\})\cdot\binom{k_i}{2}}\cdot\delta_{\min\{s_{i,j_1},s_{i,j_2}\}}\right)$$

$$\Bigg|E_s\cap\{K_{n+1}\geq 2\}\Bigg\}$$

$$=\mathbb{P}\Bigg\{\min\{S_{n+1,1},S_{n+1,2}\}\leq Q_{1-\alpha^2}\left(\sum_{i:k_i\geq 2}\sum_{j_1\neq j_2}\frac{1}{(\sum_{i=1}^{n+1}\mathbb{1}\{k_i\geq 2\})\cdot k_i(k_i-1)}\cdot\delta_{\min\{s_{i,j_1},s_{i,j_2}\}}\right)$$

$$\Bigg|E_s\cap\{K_{n+1}\geq 2\}\Bigg\}$$

$$\geq 1-\alpha^2$$

holds for any E_s with $\sum_{i=1}^{n+1}\mathbb{1}\{k_i\geq 2\}\geq 1$, where the last equality holds since $\min\{s_{i,j_1},s_{i,j_2}\}=\min\{s_{i,j_2},s_{i,j_1}\}$. It follows that

$$\mathbb{P}\left\{\min\{S_{n+1,1},S_{n+1,2}\}\leq q_{1-\alpha^2}\;\middle|\;E_s\cap\{K_{n+1}\geq 2\}\right\}\geq 1-\alpha^2,$$

where

$$q_{1-\alpha^2}=Q_{1-\alpha^2}\left(\sum_{i\leq n:K_i\geq 2}\sum_{1\leq j_1<j_2\leq K_i}\frac{1}{(N_{\geq 2}+1)\cdot\binom{K_i}{2}}\cdot\delta_{\min\{S_{i,j_1},S_{i,j_2}\}}+\frac{1}{N_{\geq 2}+1}\cdot\delta_{\infty}\right).$$

Note that $q_{1-\alpha^2}=\infty$ if $N_{\geq 2}=0$ so that the above inequality holds also for E_s with no repeats ($k_i=1\,\forall i$). From marginalization it follows that

$$\mathbb{P}\left\{\min\{S_{n+1,1},S_{n+1,2}\}\leq q_{1-\alpha^2}\;\middle|\;K_{n+1}\geq 2\right\}\geq 1-\alpha^2,$$

and hence

$$\begin{aligned}
\alpha^2 &\geq \mathbb{P}\left\{\min\{S_{n+1,1}, S_{n+1,2}\} > q_{1-\alpha^2} \;\middle|\; K_{n+1} \geq 2\right\} \\
&= \mathbb{E}\left[\mathbb{P}\left\{\min\{S_{n+1,1}, S_{n+1,2}\} > q_{1-\alpha^2} \;\middle|\; \mathcal{D}_n, X_{n+1}, \{K_{n+1} \geq 2\}\right\} \;\middle|\; K_{n+1} \geq 2\right] \\
&= \mathbb{E}\left[\mathbb{P}\left\{\min\{S_{n+1,1}, S_{n+1,2}\} > q_{1-\alpha^2} \;\middle|\; \mathcal{D}_n, X_{n+1}\right\} \;\middle|\; K_{n+1} \geq 2\right] \\
&= \mathbb{E}\left[\mathbb{P}\left\{S_{n+1,1} > q_{1-\alpha^2} \;\middle|\; \mathcal{D}_n, X_{n+1}\right\}^2 \;\middle|\; K_{n+1} \geq 2\right] \\
&= \mathbb{E}\left[\alpha_n(X_{n+1})^2 \;\middle|\; K_{n+1} \geq 2\right] \\
&= \mathbb{E}\left[\alpha_n(X_{n+1})^2\right],
\end{aligned}$$

where $\mathcal{D}_n$ denotes $\{(X_i, \tilde{Y}_i)\}_{1 \leq i \leq n}$ and the last step holds since X and K are independent. Note that without the independence assumption, we have $\mathbb{E}_{P_{X,\geq 2}}\left[\alpha_n(X_{n+1})^2\right]$ in the last step, which proves the statement in Remark 4.

The lower bound follows from the observation that if there's no tie among $S_{i,j}$'s, then any tie of $\min\{S_{i,j_1}, S_{i,j_2}\}$'s should share i, and that for a specific tie, the number of (i, j_1, j_2)'s with the same $\min\{S_{i,j_1}, S_{i,j_2}\}$ value cannot exceed $K_i - 1$. By a similar argument to the proof of Theorem 5, we have

$$\begin{aligned}
&\mathbb{P}\left\{\min\{S_{n+1,1}, S_{n+1,2}\} \leq q_{1-\alpha^2} \;\middle|\; E_s \cap \{K_{n+1} \geq 2\}\right\} \\
&\qquad\qquad \leq 1 - \alpha^2 + \frac{2(k_i - 1)}{(N_{\geq 2} + 1) \cdot \binom{k_i}{2}} \leq 1 - \alpha^2 + \frac{2}{N_{\geq 2} + 1}.
\end{aligned}$$

Applying marginalization and similar steps as the upper bound, it follows that

$$\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \geq \alpha^2 - \mathbb{E}\left[\frac{2}{N_{\geq 2} + 1}\right].$$

Note that $N_{\geq 2} \sim \text{Binomial}(n, p_K)$, so that

$$\begin{aligned}
\mathbb{E}\left[\frac{2}{N_{\geq 2}+1}\right] &= \sum_{l=0}^{n} \frac{2}{l+1} \cdot \binom{n}{l} \cdot p_K^l (1-p_K)^{n-l} \\
&= \frac{2}{(n+1)p_K} \cdot \sum_{l=0}^{n} \binom{n+1}{l+1} \cdot p_K^{l+1} (1-p_K)^{n-l} \\
&\leq \frac{2}{(n+1)p_K} \cdot \sum_{l=-1}^{n} \binom{n+1}{l+1} \cdot p_K^{l+1} (1-p_K)^{n-l} \\
&= \frac{2}{(n+1)p_K}.
\end{aligned}$$

This proves the lower bound.

4.5.6 Proof of Theorem 9

The proof applies the idea of the proof for the original jackknife+ [Barber et al., 2021b]. Similarly to the proof for the split conformal-based methods, we assume we have $\tilde{Y}_{n+1}$ which is exchangeable with $\tilde{Y}_1, \cdots, \tilde{Y}_n$, and then look into the coverage for $Y_{n+1,1}$. For each $1 \leq i \neq i' \leq n+1$, let $\tilde{\mu}_{-(i,i')} = \mathcal{A}((X_l, \tilde{Y}_l)_{l \in [n+1] \setminus \{i,i'\}})$ be the estimator from the expanded dataset with $(X_{n+1}, \tilde{Y}_{n+1})$ after excluding the i-th and i'-th point, and define $R_{(i,j),i'}$ by

$$R_{(i,j),i'} = \begin{cases} |Y_{ij} - \tilde{\mu}_{-(i,i')}(X_i)|, & i \neq i' \\ +\infty, & i = i' \end{cases}$$

for $1 \leq j \leq K_i$. Next, define

$$A_{(i,j),(i',j')} = \mathbb{1}\left\{R_{(i,j),i'} > R_{(i',j'),i}\right\},$$

and

$$A_{(i,j),\bullet} = \sum_{i'=1}^{n+1} \sum_{j'=1}^{K_{i'}} \frac{1}{(n+1)K_{i'}} \cdot A_{(i,j),(i',j')}.$$

Write A to denote the array of all $A_{(i,j),(i',j')}$'s:

$$A = \{A_{(i,j),(i',j')} : 1 \leq i, i' \leq n+1, 1 \leq j \leq K_i, 1 \leq j' \leq K_{i'}\}$$

Finally, define the set $S(A)$ by

$$S(A) = \{(i,j) : A_{(i,j),\bullet} \geq 1 - \alpha\},$$

and let

$$s(A) = \sum_{(i,j) \in S(A)} \frac{1}{(n+1)K_i}.$$

We first show that $s(A) \leq 2\alpha$ holds. Define

$$t(A) = \sum_{(i,j) \in S(A)} \sum_{(i',j') \in S(A)} \frac{1}{(n+1)K_i} \cdot \frac{1}{(n+1)K_{i'}} \cdot A_{(i,j),(i',j')}.$$

Then we have

$$\begin{aligned} 2 \cdot t(A) &= \sum_{(i,j) \in S(A)} \sum_{(i',j') \in S(A)} \frac{1}{(n+1)K_i} \cdot \frac{1}{(n+1)K_{i'}} \cdot (A_{(i,j),(i',j')} + A_{(i',j'),(i,j)}) \\ &= \sum_{(i,j) \in S(A)} \sum_{(i',j') \in S(A)} \frac{1}{(n+1)K_i} \cdot \frac{1}{(n+1)K_{i'}} \cdot \mathbb{1}\left\{R_{(i,j),i'} \neq R_{(i',j'),i}\right\} \\ &= s(A)^2 - u(A), \end{aligned}$$

where

$$u(A) = \sum_{(i,j) \in S(A)} \sum_{(i',j') \in S(A)} \frac{1}{(n+1)K_i} \cdot \frac{1}{(n+1)K_{i'}} \cdot \mathbb{1}\left\{R_{(i,j),i'} = R_{(i',j'),i}\right\}.$$

It also holds that

$$
\begin{aligned}
t(A) &= \sum_{(i,j)\in S(A)} \frac{1}{(n+1)K_i} \cdot \sum_{(i',j')\in S(A)} \frac{1}{(n+1)K_{i'}} \cdot A_{(i',j'),(i,j)} \\
&= \sum_{(i,j)\in S(A)} \frac{1}{(n+1)K_i} \cdot \sum_{(i',j')\in S(A)} \frac{1}{(n+1)K_{i'}} \cdot \\
&\qquad\qquad \left[1 - A_{(i,j),(i',j')} - \mathbb{1}\left\{R_{(i,j),i'} = R_{(i',j'),i}\right\}\right] \\
&= \sum_{(i,j)\in S(A)} \frac{1}{(n+1)K_i} \cdot \sum_{(i',j')\in S(A)} \frac{1}{(n+1)K_{i'}} \cdot \left[1 - A_{(i,j),(i',j')}\right] - u(A) \\
&\leq \sum_{(i,j)\in S(A)} \frac{1}{(n+1)K_i} \cdot \sum_{i'=1}^{n+1} \sum_{j'=1}^{K_{i'}} \frac{1}{(n+1)K_{i'}} \cdot \left[1 - A_{(i,j),(i',j')}\right] - u(A) \\
&= \sum_{(i,j)\in S(A)} \frac{1}{(n+1)K_i} \cdot \left[1 - A_{(i,j),\bullet}\right] - u(A) \\
&\leq \sum_{(i,j)\in S(A)} \frac{1}{(n+1)K_i} \cdot \alpha - u(A) \\
&= \alpha \cdot s(A) - u(A).
\end{aligned}
$$

Putting all together, we have

$$
\begin{aligned}
s(A)^2 &= 2 \cdot t(A) + u(A) \\
&\leq 2\alpha \cdot s(A) - 2u(A) + u(A) \\
&\leq 2\alpha \cdot s(A),
\end{aligned}
$$

which implies $s(A) \leq 2\alpha$.

Next, by the exchangeability of $(Y_{n+1,j})_{1\leq j\leq K_i}$ and the exchangeability of $(X_i, \tilde{Y}_i)_{1\leq i\leq n+1}$, we have

$$
\mathbb{P}\left\{(n+1,1) \in S(A) \mid K_{n+1} = k\right\} = \mathbb{P}\left\{(n+1,j) \in S(A) \mid K_{n+1} = k\right\}
$$

for all $j = 1, 2, \cdots, k$, where k is any positive integer, and

$$\mathbb{E}\left[\frac{1}{K_{n+1}} \cdot \sum_{j=1}^{K_{n+1}} \mathbb{1}\left\{(n+1, j) \in S(A)\right\}\right] = \mathbb{E}\left[\frac{1}{K_i} \cdot \sum_{j=1}^{K_i} \mathbb{1}\left\{(i, j) \in S(A)\right\}\right]$$

for all $i = 1, 2, \cdots, n$. It follows that

$$\begin{aligned}
\mathbb{P}\left\{(n+1, 1) \in S(A)\right\} &= \mathbb{E}\left[\mathbb{P}\left\{(n+1, 1) \in S(A) \mid K_{n+1}\right\}\right] \\
&= \mathbb{E}\left[\frac{1}{K_{n+1}} \cdot \sum_{j=1}^{K_{n+1}} \mathbb{P}\left\{(n+1, j) \in S(A) \mid K_{n+1}\right\}\right] \\
&= \mathbb{E}\left[\frac{1}{K_{n+1}} \cdot \sum_{j=1}^{K_{n+1}} \mathbb{E}\left[\mathbb{1}\left\{(n+1, j) \in S(A)\right\} \mid K_{n+1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{K_{n+1}} \cdot \sum_{j=1}^{K_{n+1}} \mathbb{1}\left\{(n+1, j) \in S(A)\right\} \,\middle|\, K_{n+1}\right]\right] \\
&= \mathbb{E}\left[\frac{1}{K_{n+1}} \cdot \sum_{j=1}^{K_{n+1}} \mathbb{1}\left\{(n+1, j) \in S(A)\right\}\right] \\
&= \mathbb{E}\left[\frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{K_i} \cdot \sum_{j=1}^{K_i} \mathbb{1}\left\{(i, j) \in S(A)\right\}\right] \\
&= \mathbb{E}\left[\sum_{(i,j) \in S(A)} \frac{1}{(n+1)K_i}\right] \\
&= \mathbb{E}\left[s(A)\right] \\
&\leq 2\alpha,
\end{aligned}$$

where the last inequality holds since we have shown that $s(A) \leq 2\alpha$ holds deterministically.

Now suppose $Y_{n+1,1} \notin \widehat{C}_n(X_{n+1})$. This implies

$$Y_{n+1,1} < Q_\alpha^- \left(\sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{\hat{\mu}_{-i}(X_{n+1}) - R_{ij}^{\mathrm{LOO}}} + \frac{1}{n+1} \cdot \delta_{-\infty}\right),$$

or

$$Y_{n+1,1} > Q_\alpha^+ \left(\sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \delta_{\hat{\mu}_{-i}(X_{n+1}) + R_{ij}^{\mathrm{LOO}}} + \frac{1}{n+1} \cdot \delta_\infty \right).$$

From the definition of Q_α^- and Q_α^+, we then have

$$\sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \mathbb{1}\left\{ \hat{\mu}_{-i}(X_{n+1}) - R_{ij}^{\mathrm{LOO}} < Y_{n+1,1} \right\} + \frac{1}{n+1} < \alpha,$$

or

$$\sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \mathbb{1}\left\{ \hat{\mu}_{-i}(X_{n+1}) + R_{ij}^{\mathrm{LOO}} > Y_{n+1,1} \right\} + \frac{1}{n+1} < \alpha.$$

In either case, we have

$$\begin{aligned}
1 - \alpha &\leq \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \mathbb{1}\left\{ Y_{n+1,1} \notin \hat{\mu}_{-i}(X_{n+1}) \pm R_{ij}^{\mathrm{LOO}} \right\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \mathbb{1}\left\{ |Y_{n+1,1} - \hat{\mu}_{-i}(X_{n+1})| > |Y_{ij} - \hat{\mu}_{-i}(X_i)| \right\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot \mathbb{1}\left\{ R_{(n+1,1),(i,j)} > R_{(i,j),(n+1,1)} \right\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot A_{(n+1,1),(i,j)} \\
&= \sum_{i=1}^{n+1} \sum_{j=1}^{K_i} \frac{1}{(n+1)K_i} \cdot A_{(n+1,1),(i,j)} \\
&= A_{(n+1,1),\bullet\cdot}
\end{aligned}$$

Hence, we have shown that $Y_{n+1,1} \notin \widehat{C}_n(X_{n+1})$ implies $(n+1,1) \in S(A)$.

Therefore,

$$\mathbb{P}\left\{ Y_{n+1,1} \notin \widehat{C}_n(X_{n+1}) \right\} \leq \mathbb{P}\left\{ (n+1,1) \in S(A) \right\} \leq 2\alpha.$$

4.5.7 Proof of Theorem 10

For $1 \le i \ne i' \le n+1$ with $K_i, K_{i'} \ge 2$, $1 \le j_1 < j_2 \le K_i$, and $1 \le j_1' < j_2' \le K_{i'}$, define $\tilde{\mu}_{-(i,i')}$ as in the proof of Theorem 9, and then define

$$R_{(i,j_1,j_2),i'} = \begin{cases} \min\{|Y_{i,j_1} - \tilde{\mu}_{-(i,i')}(X_i)|, |Y_{i,j_2} - \tilde{\mu}_{-(i,i')}(X_i)|\}, & i \ne i' \\ +\infty, & i = i' \end{cases}$$

and

$$A_{(i,j_1,j_2),(i',j_1',j_2')} = \mathbb{1}\left\{R_{(i,j_1,j_2),i'} > R_{(i',j_1',j_2'),i}\right\}.$$

Next, let

$$S(A) = \{(i,j_1,j_2) : A_{(i,j_1,j_2),\bullet} \ge 1 - 2\alpha^2, 1 \le i \le n+1, K_i \ge 2, 1 \le j_1 < j_2 \le K_i\}$$

where

$$A_{(i,j_1,j_2),\bullet} = \sum_{\substack{1 \le i' \le n+1 \\ K_{i'} \ge 2}} \sum_{1 \le j_1' < j_2' \le K_{i'}} \frac{1}{\tilde{N}_{\ge 2}\binom{K_{i'}}{2}} \cdot A_{(i,j_1,j_2),(i',j_1',j_2')}$$

and

$$\tilde{N}_{\ge 2} = \sum_{i=1}^{n+1} \mathbb{1}\{K_i \ge 2\}.$$

Let

$$\begin{aligned} s(A) &= \sum_{(i,j_1,j_2) \in S(A)} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}}, \\ t(A) &= \sum_{(i,j_1,j_2) \in S(A)} \sum_{(i',j_1',j_2') \in S(A)} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}} \cdot \frac{1}{\tilde{N}_{\ge 2}\binom{K_{i'}}{2}} \cdot A_{(i,j_1,j_2),(i',j_1',j_2')}. \end{aligned}$$

It holds that

$$
\begin{aligned}
&2t(A)\\
&= \sum_{(i,j_1,j_2)\in S(A)} \sum_{(i',j_1',j_2')\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \frac{1}{\tilde{N}_{\geq 2}\binom{K_{i'}}{2}} \cdot \left(A_{(i,j_1,j_2),(i',j_1',j_2')} + A_{(i',j_1',j_2'),(i,j_1,j_2)}\right)\\
&= \sum_{(i,j_1,j_2)\in S(A)} \sum_{(i',j_1',j_2')\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \frac{1}{\tilde{N}_{\geq 2}\binom{K_{i'}}{2}} \cdot \mathbb{1}\left\{R_{(i,j_1,j_2),i'} \neq R_{(i',j_1',j_2'),i}\right\}\\
&= s(A)^2 - u(A),
\end{aligned}
$$

where

$$
u(A) = \sum_{(i,j_1,j_2)\in S(A)} \sum_{(i',j_1',j_2')\in S(A)} \frac{1}{(\tilde{N}_{\geq 2})^2\binom{K_i}{2}\binom{K_{i'}}{2}} \cdot \mathbb{1}\left\{R_{(i,j_1,j_2),i'} = R_{(i',j_1',j_2'),i}\right\}.
$$

It also holds that

$$
\begin{aligned}
t(A) &= \sum_{(i,j_1,j_2)\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \left[\sum_{(i',j_1',j_2')\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_{i'}}{2}} \cdot A_{(i',j_1',j_2'),(i,j_1,j_2)}\right]\\
&= \sum_{(i,j_1,j_2)\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \left[\sum_{(i',j_1',j_2')\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_{i'}}{2}} \cdot \left(1 - A_{(i,j_1,j_2),(i',j_1',j_2')}\right.\right.\\
&\qquad\qquad \left.\left. - \mathbb{1}\left\{R_{(i,j_1,j_2),i'} = R_{(i',j_1',j_2'),i}\right\}\right)\right]\\
&= \sum_{(i,j_1,j_2)\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \left[\sum_{(i',j_1',j_2')\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_{i'}}{2}} \cdot \left(1 - A_{(i,j_1,j_2),(i',j_1',j_2')}\right)\right] - u(A)\\
&\leq \sum_{(i,j_1,j_2)\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \left[\sum_{i'=1}^{n+1} \sum_{j_1'<j_2'} \frac{1}{\tilde{N}_{\geq 2}\binom{K_{i'}}{2}} \cdot \left(1 - A_{(i,j_1,j_2),(i',j_1',j_2')}\right)\right] - u(A)\\
&= \sum_{(i,j_1,j_2)\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot \left[1 - A_{(i,j_1,j_2),\bullet}\right] - u(A)
\end{aligned}
$$

$$\leq \sum_{(i,j_1,j_2)\in S(A)} \frac{1}{\tilde{N}_{\geq 2}\binom{K_i}{2}} \cdot 2\alpha^2 - u(A)$$
$$= 2\alpha^2 \cdot s(A) - u(A).$$

Therefore, we have

$$s(A)^2 \leq 2t(A) + u(A) \leq 4\alpha^2 \cdot s(A) - 2u(A) + u(A) \leq 4\alpha^2 \cdot s(A),$$

which implies

$$s(A) \leq 4\alpha^2.$$

Next, from the exchangeability of $(X_i, \tilde{Y}_i)_{1\leq i\leq n+1}$ we have

$$\mathbb{P}\left\{(n+1,1,2) \in S(A) \mid K_{n+1} = k\right\} = \mathbb{P}\left\{(n+1,j_1,j_2) \in S(A) \mid K_{n+1} = k\right\},$$

for any $k \geq 2$ and $1 \leq j_1 < j_2 \leq k$. Now let $I_{\geq 2} = \{1 \leq i \leq n+1 : K_i \geq 2\}$ be the set of indices with repeats. Note that $|I_{\geq 2}| = \tilde{N}_{\geq 2}$. By the exchangeability of $(X_i, \tilde{Y}_i)_{1\leq i\leq n+1}$, it holds that

$$\mathbb{E}\left[\frac{1}{\binom{K_{n+1}}{2}} \cdot \sum_{1\leq j_1<j_2\leq K_{n+1}} \mathbb{1}\left\{(n+1,j_1,j_2) \in S(A)\right\} \;\middle|\; I_{\geq 2} = I\right]$$
$$= \mathbb{E}\left[\frac{1}{\binom{K_i}{2}} \cdot \sum_{1\leq j_1<j_2\leq K_i} \mathbb{1}\left\{(i,j_1,j_2) \in S(A)\right\} \;\middle|\; I_{\geq 2} = I\right], \forall\, i \in I$$

for any fixed set $I \subset [n+1]$ that contains $n+1$.

Therefore,

$$\mathbb{P}\left\{(n+1,1,2) \in S(A) \mid K_{n+1} \geq 2\right\}$$
$$= \mathbb{E}\left[\mathbb{P}\left\{(n+1,1,2) \in S(A) \mid K_{n+1}\right\} \mid K_{n+1} \geq 2\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\frac{1}{\binom{K_{n+1}}{2}}\cdot\sum_{1\le j_1<j_2\le K_{n+1}}\mathbb{P}\left\{(n+1,j_1,j_2)\in S(A)\mid K_{n+1}\right\}\;\middle|\;K_{n+1}\ge 2\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\frac{1}{\binom{K_{n+1}}{2}}\cdot\sum_{1\le j_1<j_2\le K_{n+1}}\mathbb{1}\left\{(n+1,j_1,j_2)\in S(A)\right\}\;\middle|\;K_{n+1}\right]\;\middle|\;K_{n+1}\ge 2\right]\\
=&\mathbb{E}\left[\frac{1}{\binom{K_{n+1}}{2}}\cdot\sum_{1\le j_1<j_2\le K_{n+1}}\mathbb{1}\left\{(n+1,j_1,j_2)\in S(A)\right\}\;\middle|\;K_{n+1}\ge 2\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\frac{1}{\binom{K_{n+1}}{2}}\cdot\sum_{1\le j_1<j_2\le K_{n+1}}\mathbb{1}\left\{(n+1,j_1,j_2)\in S(A)\right\}\;\middle|\;I_{\ge 2}\right]\;\middle|\;K_{n+1}\ge 2\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\frac{1}{\tilde{N}_{\ge 2}}\sum_{i\in I_{\ge 2}}\frac{1}{\binom{K_i}{2}}\cdot\sum_{1\le j_1<j_2\le K_i}\mathbb{1}\left\{(i,j_1,j_2)\in S(A)\right\}\;\middle|\;I_{\ge 2}\right]\;\middle|\;K_{n+1}\ge 2\right]\\
=&\mathbb{E}\left[\frac{1}{\tilde{N}_{\ge 2}}\sum_{i\in I_{\ge 2}}\frac{1}{\binom{K_i}{2}}\cdot\sum_{1\le j_1<j_2\le K_i}\mathbb{1}\left\{(i,j_1,j_2)\in S(A)\right\}\;\middle|\;K_{n+1}\ge 2\right]\\
=&\mathbb{E}\left[s(A)\mid K_{n+1}\ge 2\right]\\
\le&4\alpha^2.
\end{aligned}
$$

Now suppose that $K_{n+1}\ge 2$ and that both $Y_{n+1,1}$ and $Y_{n+1,2}$ are not covered by $\widehat{C}_n(X_{n+1})$. Then we have

$$
\begin{aligned}
\alpha^2 >& \sum_{\substack{1\le i\le n\\ K_i\ge 2}}\sum_{1\le j_1<j_2\le K_i}\frac{1}{(N_{\ge 2}+1)\binom{K_i}{2}}\cdot\mathbb{1}\left\{Y_{n+1,1}\in\hat{\mu}_{-i}(X_{n+1})\pm\min\{R^{\mathrm{LOO}}_{ij_1},R^{\mathrm{LOO}}_{ij_2}\}\right\}\\
&\qquad\qquad+\frac{1}{N_{\ge 2}+1}\\
=&\sum_{\substack{1\le i\le n+1\\ K_i\ge 2}}\sum_{1\le j_1<j_2\le K_i}\frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}}\cdot\mathbb{1}\left\{|Y_{n+1,1}-\hat{\mu}_{-i}(X_{n+1})|\le\min\{R^{\mathrm{LOO}}_{ij_1},R^{\mathrm{LOO}}_{ij_2}\}\right\}+\frac{1}{\tilde{N}_{\ge 2}},
\end{aligned}
$$

which implies

$$1-\alpha^2 \le \sum_{\substack{1\le i\le n+1\\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}} \cdot \mathbb{1}\left\{|Y_{n+1,1}-\hat{\mu}_{-i}(X_{n+1})| > \min\{R^{\mathrm{LOO}}_{ij_1}, R^{\mathrm{LOO}}_{ij_2}\}\right\},$$

and the same inequality holds for $Y_{n+1,2}$. Therefore,

$$\begin{aligned}
&2-2\alpha^2 \\
&\le \sum_{\substack{1\le i\le n+1\\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}} \cdot \Bigg[\mathbb{1}\left\{|Y_{n+1,1}-\hat{\mu}_{-i}(X_{n+1})| > \min\{R^{\mathrm{LOO}}_{ij_1}, R^{\mathrm{LOO}}_{ij_2}\}\right\} \\
&\qquad\qquad + \mathbb{1}\left\{|Y_{n+1,2}-\hat{\mu}_{-i}(X_{n+1})| > \min\{R^{\mathrm{LOO}}_{ij_1}, R^{\mathrm{LOO}}_{ij_2}\}\right\}\Bigg] \\
&\le \sum_{\substack{1\le i\le n+1\\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}} \cdot \\
&\quad \left[1+\mathbb{1}\left\{\min\left\{|Y_{n+1,1}-\hat{\mu}_{-i}(X_{n+1})|, |Y_{n+1,2}-\hat{\mu}_{-i}(X_{n+1})|\right\} > \min\{R^{\mathrm{LOO}}_{ij_1}, R^{\mathrm{LOO}}_{ij_2}\}\right\}\right] \\
&\le 1+\sum_{\substack{1\le i\le n+1\\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}} \cdot \mathbb{1}\left\{R_{(n+1,1,2),i} > R_{(i,j_1,j_2),n+1}\right\} \\
&= 1+\sum_{\substack{1\le i\le n+1\\ K_i\ge 2}} \sum_{1\le j_1<j_2\le K_i} \frac{1}{\tilde{N}_{\ge 2}\binom{K_i}{2}} \cdot \left[A_{(n+1,1,2),(i,j_1,j_2)}\right] \\
&= 1+A_{(n+1,1,2),\bullet},
\end{aligned}$$

which implies $(n+1,1,2)\in S(A)$. It follows that

$$\mathbb{P}\left\{Y_{n+1,1}, Y_{n+1,2}\notin \widehat{C}_n(X_{n+1}) \;\middle|\; K_{n+1}\ge 2\right\} \le \mathbb{P}\left\{(n+1,1,2)\in S(A) \mid K_{n+1}\ge 2\right\} \le 4\alpha^2.$$

Note that

$$\mathbb{P}\left\{Y_{n+1,1}, Y_{n+1,2}\notin \widehat{C}_n(X_{n+1}) \;\middle|\; K_{n+1}\ge 2\right\}$$

$$
\begin{aligned}
=&\mathbb{E}\left[\mathbb{P}\left\{Y_{n+1,1}, Y_{n+1,2} \notin \widehat{C}_n(X_{n+1}) \;\middle|\; \mathcal{D}_n, X_{n+1}, K_{n+1} \geq 2\right\} \;\middle|\; K_{n+1} \geq 2\right] \\
=&\mathbb{E}\left[\mathbb{P}\left\{Y_{n+1,1}, Y_{n+1,2} \notin \widehat{C}_n(X_{n+1}) \;\middle|\; \mathcal{D}_n, X_{n+1}\right\} \;\middle|\; K_{n+1} \geq 2\right] \\
=&\mathbb{E}\left[\mathbb{P}\left\{Y_{n+1,1} \notin \widehat{C}_n(X_{n+1}) \;\middle|\; \mathcal{D}_n, X_{n+1}\right\}^2 \;\middle|\; K_{n+1} \geq 2\right] \\
=&\mathbb{E}\left[\alpha_n(X_{n+1})^2 \;\middle|\; K_{n+1} \geq 2\right] \\
=&\mathbb{E}\left[\alpha_n(X_{n+1})^2\right],
\end{aligned}
$$

where the last step holds since K and X are independent. In the case K and X are dependent, we have $\mathbb{E}_{P_{X,\geq 2}}\left[\alpha_n(X_{n+1})^2\right]$ in the last equality and this leads to the result in Remark 6.

Putting everything together, we have

$$
\mathbb{E}\left[\alpha_n(X_{n+1})^2\right] \leq 4\alpha^2.
$$

4.5.8 Proof of Theorem 11

Let $\Delta = \mathbb{E}_{X\sim P_X}\left[(\mu_P(X) - \mu(X))^2\right]$. We first show that

$$
\hat{\Delta} \geq \Delta \;\text{ with probability }\; \geq 1-\delta.
$$

Let $I_{\geq 2} = \{1 \leq i \leq n : K_i \geq 2\}$ be the set of individuals with repeated measurements. Observe that given $I_{\geq 2}$, $\{X_i : i \in I_{\geq 2}\}$ is an i.i.d sample from $P_{X|K\geq 2}$, which is equal to P_X under our assumption that X and K are independent. Next, for each i, we have

$$
\begin{aligned}
\mathbb{E}\left[(\bar{Y}_i - \mu(X_i))^2 - \frac{S_i^2}{K_i} \;\middle|\; i \in I_{\geq 2}\right] &= \mathbb{E}\left[\mathbb{E}\left[(\bar{Y}_i - \mu(X_i))^2 - \frac{S_i^2}{K_i} \;\middle|\; X_i, K_i\right] \;\middle|\; K_i \geq 2\right] \\
&= \mathbb{E}\left[\frac{\sigma_i^2}{K_i} + (\mu_P(X_i) - \mu(X_i))^2 - \frac{\sigma_i^2}{K_i} \;\middle|\; K_i \geq 2\right] \\
&= \mathbb{E}\left[(\mu_P(X_i) - \mu(X_i))^2 \;\middle|\; K_i \geq 2\right]
\end{aligned}
$$

$$= \mathbb{E}\left[(\mu_P(X_i) - \mu(X_i))^2\right]$$
$$= \Delta.$$

Since

$$-\frac{1}{4} \leq -\frac{S_i^2}{K_i} \leq (\bar{Y}_i - \mu(X_i))^2 - \frac{S_i^2}{K_i} \leq (\bar{Y}_i - \mu(X_i))^2 \leq 1$$

for any i, we have from Hoeffding's inequality

$$\mathbb{P}\left\{Z_+ - \Delta < -\epsilon \;\middle|\; I_{\geq 2}\right\} \leq \mathbb{P}\left\{Z - \Delta < -\epsilon \;\middle|\; I_{\geq 2}\right\} \leq \exp\left\{-\frac{2 \cdot N_{\geq 2} \cdot \epsilon^2}{(5/4)^2}\right\},$$

for any $\epsilon > 0$, which implies

$$\mathbb{P}\left\{\hat{\Delta} \geq \Delta \;\middle|\; I_{\geq 2}\right\} \geq 1 - \delta,$$

for any $I_{\geq 2}$ with $|I_{\geq 2}| = N_{\geq 2} \geq 1$. Note that the above inequality holds also for $I_{\geq 2} = \emptyset$, since $\hat{\Delta} = \infty$ in that case. Hence,

$$\mathbb{P}\left\{\hat{\Delta} > \Delta\right\} = \mathbb{E}\left[\mathbb{P}\left\{\hat{\Delta} > \Delta \;\middle|\; I_{\geq 2}\right\}\right] = 1 - \delta.$$

It follows that

$$\begin{aligned}
\mathbb{P}\left\{\mu_P(X_{n+1}) \notin \widehat{C}_n(X_{n+1})\right\} &= \mathbb{P}\left\{|\mu(X_{n+1}) - \mu_P(X_{n+1})| > \sqrt{(\alpha - \delta)^{-1}\hat{\Delta}}\right\} \\
&\leq \mathbb{P}\left\{\hat{\Delta} < \Delta\right\} + \mathbb{P}\left\{|\mu(X_{n+1}) - \mu_P(X_{n+1})| > \sqrt{(\alpha - \delta)^{-1}\Delta}\right\} \\
&\leq \delta + \frac{\mathbb{E}\left[(\mu(X_{n+1}) - \mu_P(X_{n+1}))^2\right]}{(\alpha - \delta)^{-1}\Delta} \\
&= \delta + \alpha - \delta \\
&= \alpha.
\end{aligned}$$

Now we prove the upper bound for the length of $\widehat{C}_n(X_{n+1})$. By construction, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{leb}(\widehat{C}_n(X_{n+1})) \;\middle|\; N_{\geq 2} \geq 1\right] \\
\leq & 2\sqrt{(\alpha-\delta)^{-1}} \cdot \mathbb{E}\left[\sqrt{Z_+ + \frac{5}{4}\sqrt{\frac{1}{2N_{\geq 2}} \cdot \log\frac{1}{\delta}}} \;\middle|\; N_{\geq 2} \geq 1\right] \\
\leq & 2\sqrt{(\alpha-\delta)^{-1}} \cdot \sqrt{\mathbb{E}\left[Z_+ + \frac{5}{4}\sqrt{\frac{1}{2N_{\geq 2}} \cdot \log\frac{1}{\delta}} \;\middle|\; N_{\geq 2} \geq 1\right]} \\
\leq & 2\sqrt{(\alpha-\delta)^{-1}} \cdot \left[\sqrt{\mathbb{E}\left[Z_+ \mid N_{\geq 2} \geq 1\right]} + \sqrt{\frac{5}{4} \cdot \mathbb{E}\left[\sqrt{\frac{1}{2N_{\geq 2}} \cdot \log\frac{1}{\delta}} \;\middle|\; N_{\geq 2} \geq 1\right]}\right] \\
\leq & 2\sqrt{(\alpha-\delta)^{-1}} \cdot \left[\sqrt{\mathbb{E}\left[Z_+ \mid N_{\geq 2} \geq 1\right]} + \left(\mathbb{E}\left[\frac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right] \cdot \log\frac{1}{\delta}\right)^{1/4}\right].
\end{aligned}$$

Following a similar argument to the proof of Theorem 4, we have

$$\begin{aligned}
\mathbb{E}\left[Z_+ \mid N_{\geq 2} \geq 1\right] &= \mathbb{E}\left[Z \mid N_{\geq 2} \geq 1\right] + \mathbb{E}\left[Z_- \mid N_{\geq 2} \geq 1\right] \\
&\leq \mathbb{E}\left[Z \mid N_{\geq 2} \geq 1\right] + \sqrt{\mathbb{E}\left[\frac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right] + 2\mathbb{E}\left[Z \mid N_{\geq 2} \geq 1\right]} \\
&\leq \mathbb{E}\left[Z \mid N_{\geq 2} \geq 1\right] + \sqrt{2\mathbb{E}\left[Z \mid N_{\geq 2} \geq 1\right]} + \sqrt{\mathbb{E}\left[\frac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right]} \\
&= \mathrm{err}_\mu^2 + \sqrt{2} \cdot \mathrm{err}_\mu + \sqrt{\mathbb{E}\left[\frac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right]} \\
&\leq 3\mathrm{err}_\mu + \sqrt{\mathbb{E}\left[\frac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right]}
\end{aligned}$$

since

$$\mathbb{E}\left[Z \mid N_{\geq 2} \geq 1\right] = \mathbb{E}\left[(\mu(X) - \mu_P(X))^2\right] = \mathrm{err}_\mu^2$$

from the previous calculations and $(\mu(x) - \mu_P(x))^2 \leq 1$ for any x.

Next, since $N_{\geq 2} \sim \text{Binomial}(n, p_K)$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right] &\leq \mathbb{E}\left[\frac{2}{N_{\geq 2}+1} \;\middle|\; N_{\geq 2} \geq 1\right] \\
&= 2 \cdot \sum_{l=1}^{n} \frac{1}{l+1} \cdot \binom{n}{l} \cdot p_K^l (1-p_K)^{n-l} \\
&= \frac{2}{(n+1)p_K} \cdot \sum_{l=1}^{n} \binom{n+1}{l+1} \cdot p_K^{l+1} (1-p_K)^{n-l} \\
&\leq \frac{2}{(n+1)p_K} \cdot \sum_{l=-1}^{n} \binom{n+1}{l+1} \cdot p_K^{l+1} (1-p_K)^{n-l} \\
&= \frac{2}{(n+1)p_K}.
\end{aligned}$$

Putting everything together, we have

$$\begin{aligned}
&\mathbb{E}\left[\text{leb}(\widehat{C}_n(X_{n+1})) \;\middle|\; N_{\geq 2} \geq 1\right] \\
&\leq 2\sqrt{(\alpha-\delta)^{-1}} \cdot \left[\sqrt{\mathbb{E}\left[Z_+ \mid N_{\geq 2} \geq 1\right]} + \left(\mathbb{E}\left[\tfrac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right] \cdot \log\frac{1}{\delta}\right)^{1/4}\right] \\
&\leq 2\sqrt{(\alpha-\delta)^{-1}} \cdot \left[\sqrt{3\text{err}_\mu + \sqrt{\mathbb{E}\left[\tfrac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right]}} + \left(\mathbb{E}\left[\tfrac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right] \cdot \log\frac{1}{\delta}\right)^{1/4}\right] \\
&\leq 2\sqrt{(\alpha-\delta)^{-1}} \cdot \left[\sqrt{3\text{err}_\mu} + \left(\mathbb{E}\left[\tfrac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right]\right)^{1/4} + \left(\mathbb{E}\left[\tfrac{1}{N_{\geq 2}} \;\middle|\; N_{\geq 2} \geq 1\right] \cdot \log\frac{1}{\delta}\right)^{1/4}\right] \\
&\leq 2\sqrt{(\alpha-\delta)^{-1}} \cdot \left[\sqrt{3\text{err}_\mu} + \left(\frac{2}{(n+1)p_K}\right)^{1/4} + \left(\frac{2}{(n+1)p_K} \cdot \log\frac{1}{\delta}\right)^{1/4}\right],
\end{aligned}$$

which proves the claim.

CHAPTER 5

DISCUSSION

We studied problems in learning with noisy data and distribution-free inference, and observed that it is possible for the performance of a method to exceed our expectation or to even be counterintuitive.

For the binary classification problem, our work shows that the corruption of labels in the training data can work as a regularization and thus it can be beneficial in some cases. This implies that 'applying no adjustment' can also be considered as a competitive option when analyzing a corrupted data, unless we have additional information about the distribution of data. Our results for the distribution-free regression reveal that there are three regimes in terms of the possibility of useful inference, which include the in-between case where the training data is unlikely to provide information about the distribution of the new point but a meaningful inference is still possible. We provided tools for distribution-free inference in the hierarchical data setting and showed that we can aim for a better control of conditional miscoverage rates by making use of the repeated measurements. Our simulations support that we can observe a significant improvement in conditional coverage control even with small repeat numbers.

These results illustrate that it is often not clear what targets we can learn through statistical procedures, and that the power of a statistical method can exceed what we expect from intuition. In our future works, we hope to explore more such properties in the weak-assumption settings.

REFERENCES

Jayadev Acharya, Clément L Canonne, and Gautam Kamath. A chasm between identity and equivalence testing with conditional queries. *arXiv preprint arXiv:1411.7346*, 2014.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. *arXiv preprint arXiv:1507.05952*, 2015.

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

Rina Foygel Barber. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487–3524, 2020.

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, 2021b.

Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising, 2016.

Víctor Blanco, Alberto Japón, and Justo Puerto. A mathematical programming approach to binary supervised classification with label noise, 2020.

Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 143–158. Springer, 2012.

Jakramate Bootkrajang and Ata Kabán. Learning kernel logistic regression in the presence of class label noise. *Pattern Recognition*, 47(11):3641–3655, 2014.

Timothy I. Cannings, Yingying Fan, and Richard J. Samworth. Classification with imperfect training labels, 2019.

Clément L Canonne, Dana Ron, and Rocco A Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015.

Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International Conference on Machine Learning*, pages 1789–1799. PMLR, 2020.

Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Richard M Dudley, Rimas Norvaiša, and Rimas Norvaiša. *Concrete functional calculus*. Springer, 2011.

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.

B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Benoît Frénay, Ata Kabán, et al. A comprehensive introduction to label noise. In *ESANN*. Citeseer, 2014.

Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *arXiv preprint arXiv:2006.10564*, 2020.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. *arXiv preprint arXiv:2002.09025*, 2020.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36. PMLR, 2014.

Yonghoon Lee and Rina Barber. Distribution-free inference for regression: discrete, continuous, and in between. *Advances in Neural Information Processing Systems*, 34, 2021.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96, 2014.

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.

Jason Z Lin and Jelena Bradic. Learning to combat noisy labels via classification margins. *arXiv preprint arXiv:2102.00751*, 2021.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, Mar 2016. ISSN 2160-9292. doi:10.1109/tpami.2015.2456899.

Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

George G Lorentz, Manfred v Golitschek, and Yuly Makovoz. *Constructive approximation: advanced problems*, volume 304. Springer, 1996.

Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013.

Dhruv Medarametla and Emmanuel J Candès. Distribution-free conditional median inference. *arXiv preprint arXiv:2102.07967*, 2021.

Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent corruption. *arXiv preprint arXiv:1605.00751*, 2016.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, volume 26, pages 1196–1204, 2013.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *International conference on machine learning*, pages 708–717. PMLR, 2016.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

Henry W J Reeve and Ata Kaban. Classification with unknown class-conditional label noise on non-compact feature spaces, 2019a.

Henry W. J. Reeve and Ata Kaban. Fast rates for a knn classifier robust to unknown asymmetric label noise, 2019b.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*, pages 489–511. PMLR, 2013.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 685–694, 2011a.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412. IEEE, 2011b.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

Brendan Van Rooyen, Aditya Krishna Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *arXiv preprint arXiv:1505.07634*, 2015.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 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.