

THE UNIVERSITY OF CHICAGO

ASSUMPTION-LEAN APPROACHES TO MODERN STATISTICAL INFERENCE

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
ROHAN HORE

CHICAGO, ILLINOIS

JUNE 2025

To my parents,
to Debashish Ghosh and Debdarshan Roy,
and to all the teachers who have shaped me
in countless ways.

TABLE OF CONTENTS

LIST OF FIGURES	viii
LIST OF TABLES	xi
ACKNOWLEDGMENTS	xii
ABSTRACT	xiv
1 INTRODUCTION	1
1.1 Distribution-free predictive inference	1
1.1.1 Conformal prediction: a general framework to attain marginal coverage	2
1.1.2 Beyond exchangeability: towards stronger coverage guarantees	5
1.2 Distribution-free tests for conditional independence	6
1.2.1 Permutation test: a general framework to test marginal independence	7
1.2.2 Adapting permutation strategies for conditional independence testing	8
2 DISTRIBUTION-FREE PREDICTIVE INFERENCE FOR PARTIALLY-OBSERVED DATA	9
2.1 Introduction	9
2.1.1 Distribution-free lower prediction bound	11
2.1.2 Initial approach: inference on the censored survival time	12
2.1.3 Filtering out the observations with low censoring time	13
2.1.4 Refinement: a covariate-adaptive cutoff	14
2.2 Theory of existing approaches	17
2.2.1 Covering the censored survival time via CP	17
2.2.2 Using fixed threshold for filtering out low censoring times	19
2.3 Conformalized survival analysis with adaptive cutoffs	20
2.3.1 Our procedure	20
2.3.2 A generalized procedure	22
2.3.3 Theoretical guarantee: a double robustness result	25
2.4 Simulations	28
2.4.1 Synthetic setup	29
2.4.2 Simulation results	30
2.5 Real data application	33
2.6 Discussion	36
3 DISTRIBUTION-FREE CONDITIONAL PREDICTIVE INFERENCE	38
3.1 Introduction: test-conditional coverage	39
3.1.1 Hardness of test-conditional coverage	39

3.1.2	Relaxed notions of conditional coverage	40
3.1.3	Defining the goals of relaxed conditional coverage	42
3.2	Conformal prediction with local weights: existing approaches	44
3.2.1	An initial approach: baseLCP	45
3.2.2	Calibrating to recover marginal coverage: callCP	47
3.2.3	Another look at baseLCP	49
3.3	Randomly-localized conformal prediction	50
3.3.1	The role of a synthetic sample	52
3.4	Theoretical guarantees	53
3.4.1	Marginal coverage and a key property	54
3.4.2	Test-conditional coverage	57
3.4.3	Coverage against arbitrary covariate shift	61
3.5	Experiments	68
3.5.1	Simulations	69
3.5.2	Real data example: predicting age of abalones	77
3.6	Discussion and extensions	80
3.6.1	Related work	81
3.6.2	Is it possible to derandomize RLCP?	85
3.6.3	Reinterpreting callCP as an instance of full conformal prediction	89
4	ASSUMPTION-LEAN TESTS FOR CONDITIONAL INDEPENDENCE	92
4.1	Introduction	92
4.1.1	Hardness of conditional independence testing	93
4.1.2	Isotonic conditional independence	95
4.2	PairSwap-ICI-test: a general approach	97
4.3	Type I error control	99
4.4	Designing a powerful PairSwap-ICI test	100
4.4.1	Specifying the kernel ψ	101
4.4.2	Oracle strategies for choosing the matching and weights	102
4.4.3	Heuristic strategies for choosing the matching and weights	105
4.5	Power analysis	110
4.5.1	Empirical isotonic SNR drives power	111
4.5.2	Oracle isotonic SNR characterizes hardness	113
4.5.3	Near-optimal power guarantees without knowledge of μ_n	115
4.6	Simulations	118
4.6.1	Extent of conservatism under H_0^{ICI}	119
4.6.2	Power under alternatives	119
4.7	Experiment on real data: risk factors for diabetes	122
4.8	Discussion	127
5	ASSUMPTION-LEAN INFERENCE: BROADER APPLICATIONS	129
5.1	Beyond this thesis: assumption-lean confidence sets	130
5.2	Leveraging smoothness: random localization for assumption-lean confidence sets	132

5.3	Leveraging shape constraints: monotonicity for assumption-lean confidence sets	133
5.4	Conclusion	135
REFERENCES		136
A	APPENDICES FOR CHAPTER 2	151
A.1	Proofs	151
A.1.1	Proof of Theorem 3	151
A.1.2	Proof of Theorem 4	156
B	APPENDICES FOR CHAPTER 3	162
B.1	Proofs and few additional results	162
B.1.1	Completing the proof of Theorem 3.6	162
B.1.2	Training-conditional coverage	164
B.1.3	Asymptotic test-conditional coverage	169
B.1.4	Proving the connection between calLCP and full conformal prediction	174
B.1.5	A connection between m -RLCP and baseLCP	175
B.2	Details for smoothing to avoid overcoverage	176
B.2.1	Smoothing for weighted conformal prediction (WCP)	177
B.2.2	From smoothed WCP to smoothed RLCP	181
B.2.3	Smoothed versions of baseLCP and calLCP	183
B.2.4	A note on implementation	185
B.3	Additional experiments and implementation details	187
B.3.1	Quantitative analysis of the effect of randomization on RLCP predic- tion intervals	187
B.3.2	Experiments on m -RLCP	189
B.3.3	Additional experiments on local coverage in univariate settings	190
B.3.4	Additional experiments on local coverage in multivariate setting . . .	193
C	APPENDICES FOR CHAPTER 4	196
C.1	Proof of validity results of PairSwap-ICI test	196
C.1.1	Proof of Theorem 4.2	196
C.1.2	Proof of Theorem 4.3	199
C.2	Proof of the results from Section 4.5	202
C.2.1	A general result on power of PairSwap-ICI test	203
C.2.2	Power of oracle and plug-in matching strategies	212
C.2.3	Power of neighbor matching (proof of Theorem 4.10)	217
C.2.4	Power of cross-bin matching (proof of Theorem 4.11)	220
C.2.5	Proof of propositions, corollaries and lemmas from Section 4.5	225
C.2.6	An oracle matching: isotonic median matching	228
C.2.7	Proof of lemmas from Appendices C.2.1—C.2.4	239
C.3	Auxiliary lemmas	252
D	APPENDICES FOR CHAPTER 5	267

D.1	Proofs	267
D.1.1	Proof of Theorem 5.1	267
D.1.2	Proof of Theorem 5.2	269

LIST OF FIGURES

2.1	Left: an illustration of the training sample for one trial of the experiment. Middle: boxplot of the coverage rate; the red dashed line corresponds to the target coverage rate $1 - \alpha = 90\%$. Right: boxplot of the LPBs. The results are from 100 independent trials.	16
2.2	Illustration of the censoring time C and the uncensored survival time T in settings 1–4 as functions of the univariate covariate X . Both C and T are on a log scale.	30
2.3	Empirical coverage (top) and average LPBs (bottom) of all the candidate methods under settings 1–4, where X is univariate. The boxplot shows results from 100 independent draw of datasets. The dashed red line corresponds to the target coverage level $1 - \alpha = 90\%$	31
2.4	Boxplots of empirical coverage (left) and average LPBs (right) in the multivariate experimental settings. The details are otherwise the same as in Figure 2.3 . . .	31
2.5	Running time of all candidate methods in setting 3.	33
2.6	(a) Active time in three weeks for $i = 1, \dots, 10$; (b) histogram of censored survival times for all users.	34
2.7	Left: lower bound β_{lo} of the empirical coverage rate; middle: upper bound β_{hi} of the empirical coverage rate; right: average LPBs.	35
3.1	Marginal coverage of each method in two univariate simulation settings, across different bandwidths. Results are shown for 50 independent trials. See Section 3.5.1 for details.	70
3.2	Top panels: the average prediction intervals produced by each method (at each feature value x , the endpoints of the prediction interval are averaged over 50 independent trials), compared to the oracle prediction interval. Bottom panels: The corresponding local coverage of each method, averaged over 50 independent trials. See Section 3.5.1 for details.	71
3.3	Predictive coverage conditional on $X_{n+1} \in B$, for three choices of the set B (namely, B_{in} , B_{out} , and $B_{in} \cup B_{out} = \mathcal{X}$). Results for each method with fixed bandwidth $h = 1.5$ are plotted with respect to dimension d of the feature space. Results are averaged over 50 independent trials. See Section 3.5.1 for details. . .	74
3.4	Predictive coverage conditional on $X_{n+1} \in B$, for three choices of the set B (namely, B_{in} , B_{out} , and $B_{in} \cup B_{out} = \mathcal{X}$). Results for each method, averaged over 50 independent trials, are plotted with respect to dimension d of the feature space. Each method is run with its own bandwidth choice, so that a constant 50 effective sample size is maintained across dimensions. See Section 3.5.1 for more experimental details.	76
3.5	Marginal coverage (averaged over the test set) and prediction interval width (median over the test set), for each of the three methods across different bandwidths and for different choices of the pretrained base predictor $\hat{f}$, for the real data experiment. Results are averaged over 50 random splits of the data into pretraining, calibration, and test sets. See Section 3.5.2 for details.	79

3.6	Comparison of local coverage of the three methods, conditioning on X_{sex} (left) and on X_{length} (right). See Section 3.5.2 for details.	81
3.7	A general workflow of existing conformal approaches in literature to attain test-conditional coverage	82
3.8	A visualization of the variability of the RLCP prediction band for a fixed data set, showing the 5th and 95th percentile of 300 draws of constructing RLCP using the same simulated data set. We also compare to the oracle prediction interval and to the calLCP prediction interval. See Section 3.6.2 for details.	87
4.1	The workflow of PairSwap -ICI testing procedure	101
4.2	Demonstration of the cross-bin matching scheme described in Algorithm 3. The dashed lines indicate the median within each bin.	109
4.3	Simulation results illustrating Type I error control under the null H_0^{ICI} for two forms of the conditional mean $\mathbb{E}[X \mid Z]$. Each subplot shows the rejection probability of the PairSwap -ICI test on a data set of size 1000, averaged over 10^4 simulation trials, as a function of the strength of stochastic monotonicity γ	120
4.4	Simulation results demonstrating the power of the PairSwap -ICI for two alternatives at level $\alpha = 0.1$. Each subplot shows the rejection probability, averaged over 10^4 simulation trials, as a function of the sample size n . Columns correspond to different relationships between Y and Z . In each setting, X follows a Gaussian linear model with mean $\beta_n Y + \gamma Z$, where $\beta_n = n^{-1/3}$ and $\gamma = 0.5$ (above) or $\gamma = 2$ (below).	121
4.5	A scatter plot (jittered for better visibility) of Age and Diabetes Incidence along with the fitted logistic regression model to demonstrate the stochastic monotonicity.	124
4.6	Scatter plots of X (jittered for better visibility) and other feature variables along with the fitted logistic regression models to demonstrate the dependence among these variables and Diabetes Incidence	125
B.1	Quantification of the effect of randomization on RLCP via the estimated deviation $D(h)$, for the univariate simulations (Section 3.5.1) and the real data example (Section 3.5.2). See Section B.3.1 for details.	188
B.2	Marginal coverage of m -RLCP against increasing dimension for different choices of m	189
B.3	Additional results for Setting 1. Top panels: the average prediction intervals produced by each method (at each feature value x , the endpoints of the prediction interval are averaged over 50 independent trials), compared to the oracle prediction interval. Bottom panels: The corresponding local coverage of each method, averaged over 50 independent trials. See Section 3.5.1 and Appendix B.3.3 for details.	190

B.4 Additional results for Setting 2. Top panels: the average prediction intervals produced by each method (at each feature value x , the endpoints of the prediction interval are averaged over 50 independent trials), compared to the oracle prediction interval. Bottom panels: The corresponding local coverage of each method, averaged over 50 independent trials. See Section 3.5.1 and Appendix B.3.3 for details. 191

B.5 Local coverage of RLCP and the method from Gibbs et al. [2023] in setting 1 for $d = 1$ 193

B.6 Local coverage of calLCP and RLCP prediction intervals across different dimensions. 195

LIST OF TABLES

2.1	Parameters used in the six experimental settings: $P_X = \text{Unif}([0, 4]^p)$, $\mathbb{P}\{T \mid X\} = \exp(\mathcal{N}(\mu(X), \sigma^2(X)))$	30
4.1	p -values, averaged over random 3,000 subsamples along with the standard errors (within brackets) for the different tests from different experiments, as outlined in Section 4.7. The p -values significant at the 0.05 level are marked in bold.	126

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Rina Foygel Barber, for her unwavering support, guidance, and mentorship throughout my doctoral journey. Her intellectual sharpness, deep passion for the subject, and the excitement she brings to every new result have made research both inspiring and deeply fulfilling. What I admire most is her ability to immerse herself in intricate theory while seamlessly stepping back to view problems from multiple angles—balancing depth with practicality in a way that makes every discussion insightful. I am incredibly grateful for the countless discussions, insights, and encouragement she has provided.

I am also deeply grateful to Professor Richard J. Samworth, whose guidance and support since hosting my visit to the University of Cambridge have had a lasting impact. His deep knowledge, keen attention to detail, and effortless recall of the statistical literature have been truly inspiring. Beyond research, his advice on communication and professional growth has been invaluable, and I greatly appreciate his mentorship.

My heartfelt thanks go to my committee members, Rebecca Willett and Cong Ma, for their invaluable feedback and support. Their thoughtful insights have helped shape my research and pushed me to think more critically throughout this journey.

A special thanks to my collaborators, Yu Gui, Zhimei Ren, and Jake Soloff, for the stimulating discussions and shared enthusiasm for research. Working with them has been an enriching experience, filled with insightful debates, creative problem-solving, and a shared curiosity that made even the toughest challenges enjoyable.

I would also like to express my sincere gratitude to Professor Abhik Ghosh from ISI Kolkata, whose encouragement during the difficult times of COVID played a crucial role in my decision to pursue a PhD. His belief in me and words of motivation gave me the confidence to take

this path, and for that, I will always be grateful.

I am incredibly fortunate to have had the support of my colleagues and friends in the Department of Statistics at the University of Chicago. Conversations—whether about research, a new idea, or simply the world at large—have been both stimulating and enjoyable. Learning about others’ work, even outside my own focus, and exchanging thoughts on everything from statistics to philosophy have made this journey all the more rewarding. I am also grateful to the department’s faculty and staff for fostering such a warm and welcoming environment.

Finally, I want to thank my family and friends, especially my parents, for their endless encouragement and belief in me. Despite being miles away, their love and support—through every call, message, and moment of reassurance—have given me the strength to push forward. Even from afar, they have been my loudest cheerleaders, celebrating every small victory and lifting me up through the toughest moments. I could not have done this without them.

ABSTRACT

In the modern era of machine learning, where increasingly complex models drive decision-making, statistical inference becomes challenging. Assumption-lean inference minimizes reliance on distributional assumptions, ensuring robustness across diverse settings. This thesis explores two key aspects—predictive inference via conformal prediction and conditional independence testing—each providing practical, interpretable solutions to real-world challenges.

The first part investigates distribution-free predictive inference. Conformal prediction (CP) constructs prediction intervals with distribution-free guarantees, but its validity relies on exchangeability and only ensures marginal coverage, leading to failures in non-i.i.d. settings, undercoverage for certain subpopulations, or discrepancies when test and training distributions differ. We address these limitations in two steps. First, in survival analysis, where right-censoring restricts observation of true survival times, we propose a novel lower prediction bound using a data-adaptive filter that excludes low censoring times, ensuring valid marginal coverage under mild conditions. Next, we tackle local coverage, aiming for prediction intervals that retain validity for each feature configuration. While exact local coverage is theoretically impossible, we introduce randomly-localized conformal prediction (RLCP), a framework that provides interpretable guarantees for relaxed notions of local validity, such as approximate coverage under smooth covariate shifts or within sufficiently large subpopulations.

The second part develops assumption-lean tests for conditional independence. Given the inherent difficulties of testing conditional independence in a distribution-free manner, we propose a test for the conditional independence of X and Y given Z , under the assumption that X is stochastically increasing in Z . Our procedure, **PairSwap-ICI**, offers significant flexibility while ensuring finite-sample Type I error control and achieving high power against a broad range of alternatives that deviate sufficiently from the null.

CHAPTER 1

INTRODUCTION

In today’s landscape of statistical inference and machine learning, the increasing complexity of models and algorithms necessitates robust inference strategies that remain valid across diverse data domains. Traditional statistical methods often rely on rigid parametric assumptions or depend on asymptotic approximations in nonparametric settings, thereby limiting their practical applicability. Consequently, developing new approaches that embrace the modern goal of assumption-lean inference is more critical than ever. In this thesis, we concentrate on two key pillars of statistical inference—predictive inference and conditional independence testing—within a distribution-free or assumption-lean framework, aiming to create reliable and interpretable methodologies that overcome the limitations of classical approaches.

1.1 Distribution-free predictive inference

Consider a training dataset $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ and a test point (X_{n+1}, Y_{n+1}) , where each data point (X_i, Y_i) consists of a (possibly high-dimensional) feature $X_i \in \mathcal{X}$ and a response $Y_i \in \mathcal{Y}$. The problem of predictive inference is as follows: if we observe training data $\{(X_i, Y_i)\}_{i \in [n]}$ (where $[n]$ denotes $\{1, \dots, n\}$), and are given a new feature vector X_{n+1} for a new test point, we want to construct a prediction interval $\hat{C}_n(X_{n+1})$ —a subset of $\mathcal{Y}$ that is likely to contain, or “cover”, the true value of the unseen test response Y_{n+1} . The framework of *distribution-free prediction* aims to provide this type of guarantee with no assumptions on the distribution P of the data: that is, the construction $\hat{C}_n$ should satisfy

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} \geq 1 - \alpha, \quad (1.1)$$

for some target coverage level $1 - \alpha$, for data drawn i.i.d. from *any* distribution P . This type of distribution-free guarantee is achieved by the conformal prediction (CP) method [Vovk et al., 2005] and related methodologies, which we describe below in detail.

1.1.1 Conformal prediction: a general framework to attain marginal coverage

Conformal inference [Vovk et al., 2005] is an extremely popular framework for calibrating prediction uncertainty, and is used to build prediction intervals satisfying distribution-free marginal validity as in (1.1). A common variant of this methodology is the *split conformal prediction* procedure [Papadopoulos et al., 2002], which we briefly define now. Assume we have two datasets available: $\{(X_i^{\text{pre}}, Y_i^{\text{pre}})\}_{i=1, \dots, n_{\text{pre}}}$, a dataset used for pretraining, and $\{(X_i, Y_i)\}_{i=1, \dots, n}$, the holdout set which will be used for calibration.¹

- **Step 1: defining a conformal score:** We fit a model on $\{(X_i^{\text{pre}}, Y_i^{\text{pre}})\}_{i=1, \dots, n_{\text{pre}}}$, and obtain a *conformity score function* $s(x, y)$ that returns a large value if the point (x, y) appears unusual (does not “conform”) relative to the trends observed in the data. For example, if we run a regression procedure to train a model $\hat{f}(x)$ (which predicts Y given $X = x$), we might choose the commonly used *residual score* $s(x, y) = |y - \hat{f}(x)|$, which returns a large value if the value of y is not near the predicted value $\hat{f}(x)$.
- **Step 2: computing a threshold for the test score:** The dataset $\{(X_i, Y_i)\}_{i=1, \dots, n}$, which contains n many data points that were not used for training $\hat{f}$, is then used for calibrating a threshold for the score function by defining $\hat{q}$ as the $(1 - \alpha)(1 + 1/n)$ quantile of the holdout scores $\{s(X_i, Y_i)\}_{i=1, \dots, n}$, or equivalently the $(1 - \alpha)$ quantile

1. In practice, of course, we would simply partition a single available dataset into two disjoint subsets—but it will be convenient here to use the notation of two separate datasets.

of the list of all holdout scores $\{s(X_i, Y_i)\}_{i=1, \dots, n}$ with an appended $+\infty$ score, i.e.,

$$\hat{q} = \text{Quantile}_{1-\alpha} \left(\sum_{i=1}^n \frac{1}{n+1} \delta_{s(X_i, Y_i)} + \frac{1}{n+1} \delta_{+\infty} \right),$$

where δ_s denotes the point mass at s .

- **Step 3: constructing the prediction interval:** Finally, we return the prediction interval²

$$\hat{C}_n(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}\}. \quad (1.2)$$

The split CP intervals are computationally efficient and satisfy a distribution-free coverage guarantee as in (1.1).

Theorem 1.1. *Suppose $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are exchangeable. For any score function s , and any $\alpha \in (0, 1)$, the split conformal prediction interval satisfies*

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

An extension: weighted conformal prediction

The validity of split CP intervals relies strongly on the assumption of exchangeability. But often, this exchangeability is not satisfied in real data settings. One interesting example is the setting of covariate shift. More formally, suppose that the test point is drawn from a distribution $\tilde{P} = \tilde{P}_X \times \tilde{P}_{Y|X}$, while training data points are drawn from $P = P_X \times P_{Y|X}$. Here P_X denotes the feature distribution, i.e., the marginal distribution of X , while $P_{Y|X}$ is the conditional distribution of $Y \mid X$, and similarly for $\tilde{P}_X$ and $\tilde{P}_{Y|X}$. The covariate

2. We use the terminology “prediction interval” throughout the thesis as this is most familiar in statistical settings where often we have a real-valued response (i.e., $\mathcal{Y} = \mathbb{R}$), but it might be more correct to say “prediction set”, since even when $\mathcal{Y} = \mathbb{R}$ the resulting set may not necessarily consist of a single interval.

shift setting assumes that the conditional distribution of $Y|X$ is shared across the two, with $P_{Y|X} = \tilde{P}_{Y|X}$; the difference then lies in the marginal distributions of the features (or covariates) X , i.e., P_X versus $\tilde{P}_X$. In this scenario, we seek a coverage guarantee that would satisfy

$$\mathbb{P}_{P^n \times \tilde{P}} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} \geq 1 - \alpha, \quad (1.3)$$

where the probability is taken with respect to training data $(X_1, Y_1), \dots, (X_n, Y_n)$ sampled i.i.d. from P , and test point (X_{n+1}, Y_{n+1}) sampled from $\tilde{P}$.

The Weighted Conformal Prediction (WCP) method of Tibshirani et al. [2019] uses a weighted quantile, with weights reflecting the shift from P_X to $\tilde{P}_X$, to define $\hat{q}$ in a way that achieves this desired coverage guarantee. More specifically, suppose $\tilde{P}_X \ll P_X$ and we follow the steps in the aforementioned algorithm of split CP except that the threshold $\hat{q}$ in (1.2) is replaced with a *weighted* quantile,

$$\hat{q} = \text{Quantile}_{1-\alpha} \left(\sum_{i=1}^n w_i \delta_{s(X_i, Y_i)} + w_{n+1} \delta_{+\infty} \right),$$

where weights $w_i = \frac{d\tilde{P}_X}{dP_X}(X_i)$ are chosen to correct for the covariate shift; for example, if P_X has density f_X and $\tilde{P}_X$ has density $\tilde{f}_X$, we choose $w_i \propto \tilde{f}_X(X_i)/f_X(X_i)$.

Theorem 1.2. *Suppose $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{iid}{\sim} P_X \times P_{Y|X}$, and independently, $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times \tilde{P}_{Y|X}$. For any score function s and any $\alpha \in (0, 1)$, the weighted conformal prediction interval, as defined above, satisfies*

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

Thus, WCP is a notable variant of conformal inference tailored for scenarios with distribution shift. However, its validity relies on the assumption of a fixed and *known* covariate shift—a

condition that is often not met in practical applications.

1.1.2 Beyond exchangeability: towards stronger coverage guarantees

The challenges and objectives of predictive inference in real-world settings—such as dealing with non-exchangeability or providing stronger coverage guarantees—demand robust and adaptable variants of conformal prediction. This thesis tackles several of these challenges by proposing novel methods that enhance the utility and reliability of conformal inference in complex, data-driven environments.

In Chapter 2, we consider the setting of survival analysis, where we want to build prediction intervals for the true survival time T at a test feature vector X but the challenge is: we only get to observe T if it occurs before the censoring time C , i.e., only $\min\{T, C\}$ and the features X are available in training data—a scenario that existing conformal inference methods do not address. To overcome this, we propose a novel approach for constructing a distribution-free lower prediction bound (LPB) by filtering out observations with low censoring times using a data-adaptive cutoff. We further show that under a double robustness assumption—where either the censoring mechanism or the true quantiles of $T \mid X$ are well estimated, our method achieves approximate marginal coverage.

Next, in Chapter 3, we turn to the problem of building distribution-free prediction intervals that are valid for any given feature value—a guarantee that the usual split conformal prediction intervals fail to satisfy. This form of coverage guarantee is especially important to practitioners because marginal coverage alone does not ensure valid coverage across all subpopulations, potentially leading to severe undercoverage in certain cases. We introduce a novel framework called randomly localized conformal prediction, which not only maintains marginal coverage but also provides approximate local coverage and robustness against

certain covariate shifts.

1.2 Distribution-free tests for conditional independence

In the second part of the thesis, we will focus on the problem of conditional independence testing. Formally, we consider the following setting. Suppose that $(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n) \stackrel{iid}{\sim} P$ for some unknown distribution P on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. We consider the null hypothesis of conditional independence,

$$H_0^{\text{CI}} : X \perp\!\!\!\perp Y \mid Z.$$

Our goal is to construct a test function $\phi : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \rightarrow [0, 1]$ (potentially randomized)³ so that the type I error can be controlled at any target level α , i.e.,

$$\sup_{P \in H_0^{\text{CI}}} \mathbb{E}_P \left[\phi \left(\{X_i, Y_i, Z_i\}_{i \in [n]} \right) \right] \leq \alpha,$$

where we abuse notation to write $P \in H_0^{\text{CI}}$ to denote all distributions P that satisfy the null hypothesis H_0^{CI} . Moreover, we expect this test to have non-trivial power against some interesting alternatives, i.e., we expect that for some $\mathcal{Q} \subseteq (H_0^{\text{CI}})^c$ and $Q \in \mathcal{Q}$,

$$\mathbb{E}_Q \left[\phi \left(\{X_i, Y_i, Z_i\}_{i \in [n]} \right) \right] \gg \alpha.$$

The difficulty of the testing problem varies depending on the choice of the conditional variable Z . At one extreme, when $Z \equiv 1$, i.e., when the conditioning is trivial, then the null hypothesis boils down to the hypothesis of marginal independence $H_0^{\text{MI}} : X \perp\!\!\!\perp Y$. The permutation

3. Traditionally, we think of a hypothesis test ϕ as a map from data to a decision, i.e., $\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \{0, 1\}$. Why, then, do we define tests ϕ as mapping to the space $[0, 1]$? This is because, in some settings, we may want to consider randomized tests—for instance, in the notation above where ϕ maps to $[0, 1]$, an outcome $\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = 0.75$ represents that, given the data, our randomized test rejects the null with probability 0.75. Of course, a nonrandomized test is simply a special case, obtained by restricting the output of ϕ to lie in $\{0, 1\}$.

test is a well-known simple framework for testing marginal independence, which we briefly review next.

1.2.1 Permutation test: a general framework to test marginal independence

Let $\mathbf{X}$ denote $(X_i)_{i=1}^n$ and $\mathbf{Y}$ denote $(Y_i)_{i=1}^n$. We can reframe the problem of testing marginal independence as testing whether the entries of $\mathbf{X}$ are independent and identically distributed given $\mathbf{Y}$. With this formulation in place, permutation tests look for violations of exchangeability of $\mathbf{X}$ given $\mathbf{Y}$.

The general approach proceeds as follows: based on $\mathbf{Y}$, the analyst chooses any statistic $T : \mathcal{X}^n \rightarrow \mathbb{R}$, with larger values of $T(\mathbf{X})$ indicating evidence against the independence null.⁴ Write $\mathcal{S}_n$ for the set of permutations of $[n] := \{1, \dots, n\}$, and for $\sigma \in \mathcal{S}_n$, let $T_\sigma = T(\mathbf{X}^\sigma)$ denote the value of the statistic when the entries of $\mathbf{X}$ are permuted according to σ —that is, $\mathbf{X}^\sigma = (X_{\sigma(1)}, \dots, X_{\sigma(n)})$. Finally, define a p -value

$$p = \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \mathbb{1}\{T_\sigma \geq T\}. \quad (1.4)$$

Under the null of marginal independence, p follows a super-uniform distribution:

Theorem 1.3. *Let $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{iid}{\sim} P$ for some distribution P . Suppose P satisfy the assumption of marginal independence, i.e., H_0^{MI} . Then, for any $\alpha \in [0, 1]$ the p -value defined in (1.4) satisfies*

$$\mathbb{P}\{p \leq \alpha\} \leq \alpha.$$

In other words, this construction produces a valid p -value for *any* choice of test statistic T ,

4. We emphasize that $T(\mathbf{X})$ is allowed to depend on both $\mathbf{X}$ and $\mathbf{Y}$ —for instance, we may define the function as $T(\mathbf{x}) = |\sum_{i=1}^n x_i Y_i|$, though we suppress the dependence on $\mathbf{Y}$ in our notation.

and the statistic T can be tailored to have power against certain specific alternatives.

This strategy has been successfully employed to construct independence tests via nearest neighbour distances and mutual information [Berrett and Samworth, 2019], moment methods [Kim et al., 2022a], kernels [Pfister et al., 2018], the Hilbert–Schmidt independence criterion [Albert et al., 2022] and U -statistics [Berrett and Samworth, 2021, Berrett et al., 2021].

1.2.2 Adapting permutation strategies for conditional independence testing

The permutation test can naturally be extended to assess the hypothesis of conditional independence, H_0^{CI} . In particular, when Z is categorical, a straightforward extension is to group observations by their Z values and permute within each group. For example, if $Z \in \{0, 1\}$, we could draw a permutation π that permutes the X_i s within $\{i : Z_i = 0\}$ and separately within $\{i : Z_i = 1\}$. While this cannot be directly applied to continuous Z , we can bin Z -values into a categorical variable $\tilde{Z}$ to instead test the relaxed null $H_0 : X \perp\!\!\!\perp Y \mid \tilde{Z}$. Under suitable smoothness assumptions, this yields an approximate valid test for the null H_0^{CI} , as investigated in Kim et al. [2022b].

However, smoothness assumptions are typically untestable and often restrictive. To address this, in Chapter 4 we relax the conditional independence null by assuming that X is stochastically increasing in Z . This shape constraint is often more naturally justified in real-world settings, making it a practical approach for assumption-lean inference. Drawing inspiration from permutation tests, we introduce the **PairSwap**-ICI test, which swaps X values between pairs of Z values. Under the shape constraint, we expect a higher X value to typically accompany a higher Z values; significant deviations from this provide evidence against the null hypothesis. Our method controls Type I error without further model assumptions and demonstrates robust power across a broad range of alternatives.

CHAPTER 2

DISTRIBUTION-FREE PREDICTIVE INFERENCE FOR PARTIALLY-OBSERVED DATA

In this chapter, we consider the problem of constructing distribution-free prediction intervals in a non-i.i.d. setting—specifically, in survival analysis, where survival times are only partially observed due to right-censoring [Leung et al., 1997]. For example, T may be the survival time of a patient (measured as time since diagnosis), which may be censored at time C , marking the endpoint of the study monitoring the patient.

To formally describe the setting of survival analysis, let $X \in \mathcal{X}$ denote the covariate vector, $T \in \mathbb{R}_{\geq 0}$ the survival time, and $C \in \mathbb{R}_{\geq 0}$ the censoring time. Under censoring, the survival time T is observed only if it occurs before the censoring time C . In other words, while the features X and the censoring time C are both observed, the survival time is observed only indirectly, via the censored survival time as $\tilde{T} = \min(T, C)$ (which may not be equal to T). Shortly, we will describe how to construct distribution-free prediction intervals for $T \mid X_{n+1}$ at a new test feature X_{n+1} , drawn from the same population, given access only to the dataset $\left\{ (X_i, C_i, \tilde{T}_i) \right\}_{i=1}^n$.

This chapter presents work co-authored with *Yu Gui, Zhimei Ren, and Rina Foygel Barber*, which has been published in *Biometrika* under the title ‘Conformalized Survival Analysis with Adaptive Cut-Offs’ [Gui et al., 2024].

2.1 Introduction

Survival analysis lies in the core of many important questions in clinical trials [Fleming and Lin, 2000, Singh and Mukhopadhyay, 2011], ecology [Muenchow, 1986], and other applied

fields. In particular, one important problem is that of studying the behavior of survival time T , and how it relates to other features of the data, which we denote by a potentially high-dimensional feature vector X . Modeling the association between X and T can in turn play a crucial role in enabling more useful and reliable policy making. However, the major challenge is that these survival times are only partially observed, which makes the statistical analysis quite non-routine.

One of many goals of survival analysis is to infer the survival function—the probability of survival beyond a given time, given the censored data. The Kaplan-Meier curve [Kaplan and Meier, 1958] can produce such inferences for sub-population with a particular covariate structure while making no assumption on the distribution of survival times but it requires sufficiently many events in each subgroup [Kalbfleisch and Prentice, 2011]. This is no longer the case in the modern era of big data, where with the ever-increasing ability to collect and store data we may have access to a large number of (potentially continuous) covariates.

Over the years, many tools have been developed to cope with such high dimensionality, offering estimation of the conditional survival function. One popular example in that line is the Cox model which posits a proportional hazard model: an unspecified non-parametric baseline is modified via a parametric model describing how the hazard varies in response to explanatory covariates [Cox, 1972, Breslow, 1975]. Other popular parametric approaches include the accelerated failure time (AFT) model [Cox, 1972, Wei, 1992] and the proportional odds model [Murphy et al., 1997, Harrell Jr, 2015]. More recently, we have witnessed more complex survival analysis methods that are based on machine learning/deep learning [Faraggi and Simon, 1995, Tibshirani, 1997, Gui and Li, 2005, Katzman et al., 2016, Lao et al., 2017, Wang et al., 2019, Li and Bradic, 2020]. Despite the success of these methods in many areas, it remains largely unclear how to provide reliable uncertainty quantification for these methods. This is mainly because they posit model assumptions that are hard to verify

and/or the algorithms themselves are too complicated to be analyzed. For these reasons, it is desirable to find a more assumption-lean or distribution-free approach towards reliable inference in survival analysis.

The recent work of Candès et al. [2021] proposes such an approach (which we will describe in detail below). As the target of inference, they propose computing a $100(1 - \alpha)\%$ lower prediction bound (LPB) for the survival time of a patient/unit, where α is a pre-specified level; it means that the patient/unit is expected to survive beyond this predicted time with at least $100(1 - \alpha)\%$ probability. The LPB is used to provide a summary of what we can infer about the individual's survival time given available data, and it is important to note that the LPB can be low when either the true survival time is low or there is not enough information for us to get an informative lower bound; in other words, insufficient data should not lead to an invalid claim, but instead may lead to a less informative output.

2.1.1 Distribution-free lower prediction bound

We now give the formal definition of a marginally calibrated LPB. Throughout, for a joint distribution P on (X, C, T) , we will write P_X , $P_{(X,T)}$, $P_{(X,\tilde{T})}$, etc, to denote the corresponding marginal distributions, and $P_{C|X}$, $P_{T|X}$, $P_{\tilde{T}|X}$, etc, to denote the corresponding conditional distributions.

Definition 2.1 (Marginally calibrated LPB). *Let $(X_i, C_i, T_i) \stackrel{\text{iid}}{\sim} P$ for data points $i = 1, \dots, n$, and let $\hat{L}$ be a function of the observed data $\mathcal{D} = \{(X_i, C_i, \tilde{T}_i) : 1 \leq i \leq n\}$, where $\tilde{T}_i = \min(T_i, C_i)$ is the censored survival time. Then we say that $\hat{L}$ is a marginally calibrated LPB at level $1 - \alpha$ if it satisfies*

$$\mathbb{P}_{(X,T) \sim P} \left\{ T \geq \hat{L}(X) \right\} \geq 1 - \alpha, \quad (2.1)$$

where this probability is taken with respect to both the available data $\mathcal{D}$ and a new data point $(X, T) \sim P_{(X, T)}$.

The marginally calibrated LPB provides guarantee in an average sense—that is, over all the possible draws of the data, the coverage of the LPBs is guaranteed. However, in practical settings, we may be more concerned about the coverage guarantee we can obtain given the data at hand. There, the probably approximately correct (PAC)-type LPB defined below can be more informative [see also Vovk, 2012, Bates et al., 2021, Angelopoulos et al., 2021, Jin et al., 2021].

Definition 2.2 (PAC-type LPB). *Under the same notation as Definition 2.1, we say that $\hat{L}$ is a PAC-type LPB at level α with tolerance δ if, with probability at least $1 - \delta$ over the draw of $\mathcal{D}$,*

$$\mathbb{P}_{(X, T) \sim P} \left\{ T \geq \hat{L}(X) \mid \mathcal{D} \right\} \geq 1 - \alpha$$

where the probability is now taken with respect to a new data point $(X, T) \sim P_{(X, T)}$.

Throughout, we adopt the *conditionally independent censoring* assumption.

Assumption A1 (Conditionally independent censoring). *The joint distribution P of (X, C, T) satisfies $C \perp\!\!\!\perp T \mid X$.*

This assumption is standard in the survival analysis literature, in order to ensure identifiability of the problem [see e.g., Kalbfleisch and Prentice, 2011].

2.1.2 Initial approach: inference on the censored survival time

As discussed by Candès et al. [2021], since the censored survival time $\tilde{T}$ cannot be larger than T by definition, any valid lower bound on $\tilde{T}$ is trivially a lower bound on T . In other words, if an estimated lower bound $\hat{L}$ satisfies $\mathbb{P}_{(X, C, T) \sim P} \left\{ \tilde{T} \geq \hat{L}(X) \right\} \geq 1 - \alpha$, then

trivially Definition 2.1 is satisfied and so $\hat{L}$ is a marginally calibrated LPB. (Similarly, if $\mathbb{P}\left\{\tilde{T} \geq \hat{L}(X) \mid \mathcal{D}\right\} \geq 1 - \alpha$ with probability at least $1 - \delta$, then by Definition 2.2 $\hat{L}$ is a PAC-type LPB.) Since the censored survival time $\tilde{T}$ can be observed in the dataset at hand (and so $\hat{L}$ can be constructed to satisfy this property), this provides a mechanism for providing a valid LPB.

However, if the censoring time C is often substantially smaller than T , then a valid lower bound on $\tilde{T}$ may be extremely conservative as a lower bound on T itself, thus reducing the utility of the constructed LPB. This suggests that such an approach may not be optimal for most applications. On the other hand, Candès et al. [2021] prove that, in the absence of any assumptions on the distribution P on (X, C, T) , it is impossible to improve on this type of approach—specifically, their result [Candès et al., 2021, Theorem 1] proves that, for any construction $\hat{L}$ that satisfies Definition 2.1 universally over all distributions P , $\hat{L}$ must also satisfy $\mathbb{P}\left\{\tilde{T} \geq \hat{L}(X)\right\} \geq 1 - \alpha$. This motivates their introduction of an additional assumption, as we describe next.

2.1.3 Filtering out the observations with low censoring time

As described above, constructing an LPB on the censored survival time $\tilde{T}$ may be too conservative in applications where the censoring time C is frequently low, leading to censored times $\tilde{T}$ that are far smaller than the true target of inference T . Candès et al. [2021]’s approach is to avoid this issue by discarding any training data points where C is very low—specifically, for a constant cutoff c_0 , they subset the data $\mathcal{D}$ to keep only data points $(X_i, C_i, \tilde{T}_i)$ for which $C_i \geq c_0$. After this filtering step, any lower bound $\hat{L}$ on the remaining censored survival time $\tilde{T}$ is no longer necessarily overly conservative, since the condition $C \geq c_0$ (with a well-chosen c_0) ensures that $\tilde{T}$ is less likely to be far smaller than T . Thus, we can proceed by constructing an LBP $\hat{L}$ that is a lower bound on $\tilde{T}$, in this new training sample.

Of course, we must then be careful about biasing the results because of this cutoff. In particular, since the event $C \geq c_0$ may be highly dependent on the covariates X , the remaining data is drawn from a distribution that is different from the target distribution P . To be more precise, writing $P^{\geq c_0}$ to denote the distribution of a data point $(X, C, T) \sim P$ given the event $C \geq c_0$, we see that the remaining data consists of samples from $P^{\geq c_0}$ while the inference goal is to provide coverage under the original distribution P . In other words, we would like to ensure that the marginal coverage bound (2.1) holds, but calibrating $\hat{L}(\cdot)$ naively on the remaining data would instead only ensure that $\mathbb{P}_{(X,T) \sim P^{\geq c_0}} \{T \geq \hat{L}(X)\} \geq 1 - \alpha$, or equivalently, $\mathbb{P}_{(X,T,C) \sim P} \{T \geq \hat{L}(X) \mid C \geq c_0\} \geq 1 - \alpha$.

To account for this shift in the distribution, Candès et al. [2021] utilize the method of *conformal prediction under covariate shift* [Tibshirani et al., 2019], which builds on the well-known conformal prediction framework for distribution-free predictive inference [Vovk et al., 2005]. To do so, they additionally assume that we have exact or approximate knowledge of the dependence of censoring time C on the covariates X —that is, knowledge of $P_{C|X}$, or more specifically, $\mathbb{P}\{C \geq c_0 \mid X\}$. With this additional information, we can reweight the remaining data points to correct for the change in distribution—essentially, similarly to inverse propensity score weighting, weights $1/\mathbb{P}\{C \geq c_0 \mid X\}$ can account for the difference between the target distribution P and its filtered version $P^{\geq c_0}$. (Of course, the best value of c_0 will depend on the data distribution, and in practice can be chosen on a training set.)

2.1.4 Refinement: a covariate-adaptive cutoff

In the method described above, how should the cutoff c_0 be chosen? The choice of c_0 presents a tradeoff: if c_0 is chosen to be too small, then the inequality $\tilde{T} \leq T$ might be quite loose, and the constructed LPB $\hat{L}$ might still be very conservative even after filtering the data with the cutoff. On the other hand, if c_0 is chosen to be too large, then $\mathbb{P}\{C \geq c_0 \mid X\}$ may be quite small (at least, for many values of X), leading to a low effective sample size, large

weights $1/\mathbb{P}\{C \geq c_0 \mid X\}$ on these data points, and highly unstable behavior. In fact, it is not always possible to find a constant c_0 that yields good LPBs, especially in cases when the censoring time varies substantially with respect to the covariates X —selecting a large value of c_0 could cause highly unstable LPBs in areas where censoring times are low, whereas selecting a small value of c_0 leads to conservative LPBs in areas where censoring times are actually high. To be more specific, think of a simple example where $X \sim \text{Unif}([0, 1])$ and $C = a\mathbb{1}\left\{X \geq \frac{1}{2}\right\} + b\mathbb{1}\left\{X < \frac{1}{2}\right\}$ with $a \gg b$; choosing c_0 to be greater than b requires dropping half of the data and leads to increased variability; instead, selecting a $c_0 \leq b$ yields very conservative LPBs for $X \geq \frac{1}{2}$.

From the above discussion, we can see that it may be beneficial to allow c_0 to depend on X . That is, if $\mathbb{P}\{C \geq c_0 \mid X\}$ is extremely small then we may need to instead choose a lower value of c_0 to avoid high variance, but if $\mathbb{P}\{C \geq c_0 \mid X\}$ is close to 1 then we can afford to increase the value of c_0 , thus avoiding an overly conservative LPB. To illustrate the benefits of this more flexible approach, we show a small simulated example.

We consider a univariate-covariate case, where T and C depend on X via different models (the details are to be given in Section 2.4). The left panel of Figure 2.1 visualizes (one realization of) the censoring time and survival time as functions of the covariate. In this example, units with larger values of X tend to have lower censoring times ($P_{C|X} = \text{Exp}(0.25 + (6+x)/100)$), and thus we should choose a lower value of c_0 to avoid high variance (i.e., to avoid overly large weights $1/\mathbb{P}\{C \geq c_0 \mid X\}$; units with smaller values of X , on the other hand, tend to have larger values of C and so we can afford to increase the value of c_0 , leading to a less conservative LPB.

From this model, $n = 2,000$ independent samples are generated. We compare the baseline method introduced in Section 2.1.2 (referred to as DFT-baseline, where DFT is short for “distribution-free (LPB) for T ”), Candès et al. [2021]’s fixed cutoff method (referred to as

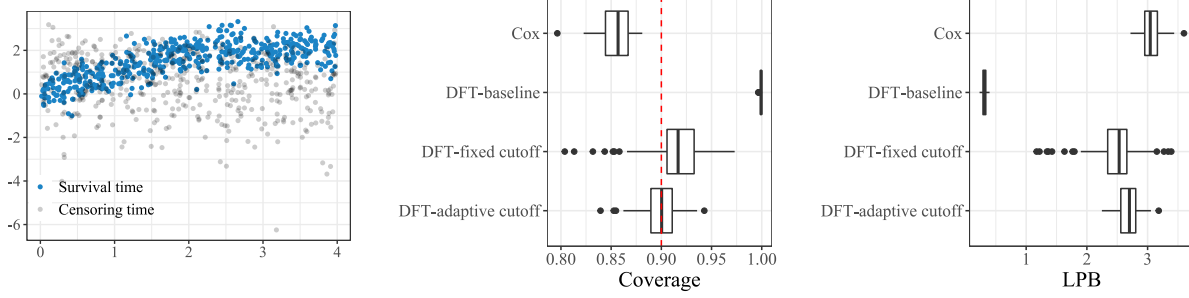


Figure 2.1: Left: an illustration of the training sample for one trial of the experiment. Middle: boxplot of the coverage rate; the red dashed line corresponds to the target coverage rate $1 - \alpha = 90\%$. Right: boxplot of the LPBs. The results are from 100 independent trials.

DFT-fixed cutoff), our new adaptive cutoff method (referred to as DFT-adaptive cutoff and to be defined shortly), and the Cox parametric model. The generated LPBs are then evaluated with an independent dataset of 5,000 test samples, and we display the coverage rate and the resulting LPB in the middle and right panels of Figure 2.1, respectively, with results gathered from 100 independent trials. The parametric method fails to cover the true survival time with desired probability; the baseline method and, to a lesser extent, the fixed cutoff method are conservative in this setting, returning a low (i.e., less informative) LPB. On the other hand, our adaptive cutoff method is able to avoid under- or over-coverage; it achieves essentially the target coverage rate and returns a higher (i.e., more precise) LPB.

Organisation of the chapter The remainder of the chapter is organized as follows. First, in Section 2.2, we will give a more detailed background on the existing methods described briefly above, providing a LPB on the censored survival time $\tilde{T}$ (as in Section 2.1.2) or by using a fixed cutoff c_0 (as in Section 2.1.3). In Section 2.3, we propose our new method, which offers a data-driven and covariate-adaptive method for determining a cutoff, alleviating the limitations of the fixed c_0 approach. We give theoretical results showing that our resulting LBPs satisfy a doubly robustness property—i.e., if either the distribution of $C \mid X$ or the quantile of $T \mid X$ is estimated accurately, our procedure achieves a PAC-type guarantee. The benefits of our proposed approach are demonstrated on both simulated and real data

in Sections 2.4 and 2.5, and we conclude with a short discussion in Section 2.6. Proofs are deferred to the Appendix A.

2.2 Theory of existing approaches

2.2.1 Covering the censored survival time via CP

As described in Section 2.1.2, it is possible to provide an LPB on T with no further assumptions by simply finding a lower bound on the censored survival time $\tilde{T} \leq T$. To do so, one approach is to use the Conformalized Quantile Regression (CQR) framework of Romano et al. [2019b]. To begin, we first partition the available n data points into two data sets¹, a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$ —for instance, into two sets of size $n/2$. Using the training set, we fit a quantile regression: $x \mapsto \hat{q}_\alpha(x)$, which estimates the conditional α -quantile of T given X . This may be done using an arbitrary algorithm, for instance, linear regression or random forests. If this quantile regression were fitted accurately, then we could simply use $\hat{q}_\alpha(X)$ as a LPB for T —if indeed this is the α -quantile of $T \mid X$, then $T \geq \hat{q}_\alpha(X)$ holds with probability $1 - \alpha$, as desired. However, due to potential issues of overfitting, model misspecification, etc, we cannot rely on this being the case, and so the calibration set is then used to correct for any errors in the initial model fitting stage. For each $i \in \mathcal{I}_2$ (the calibration points), define a score $S_i = \hat{q}_\alpha(X_i) - \tilde{T}_i$, and then define the LPB as

$$\hat{L}_{\text{baseline}}(X) = \hat{q}_\alpha(X) - Q_{1-\alpha} \left(\sum_{i \in \mathcal{I}_2} \frac{1}{1 + |\mathcal{I}_2|} \cdot \delta_{S_i} + \frac{1}{1 + |\mathcal{I}_2|} \cdot \delta_{+\infty} \right),$$

where $Q_{1-\alpha}(\cdot)$ denotes the $(1 - \alpha)$ -quantile of a distribution and where δ_v is the point mass at v . The intuition here is that the $Q_{1-\alpha}(\dots)$ term adds a correction to the original fitted

1. In Section 1.1.1, these splits were referred to as the pre-training set and holdout set. Here, we deviate from that notation because in this chapter we evaluate coverage based on how well we train our scores, whereas earlier, we assumed the scores were pre-learned on a separate dataset and provided for conformal calibration.

model to ensure that $\hat{L}_{\text{baseline}}(X)$ has the right coverage level on the calibration set drawn i.i.d. from P , and will thus have the right coverage level on a future draw (X, T) from $P_{(X, T)}$ as well.

Note that the resulting value $\hat{L}_{\text{baseline}}(X)$ may be higher (less conservative) or lower (more conservative) than the initial fitted model $\hat{q}_\alpha(X)$, depending on whether the original fitted model $\hat{q}_\alpha$ is over- or under-covering on the calibration set. In practice, it is likely that we will have undercoverage of the original fitted model (see, e.g., the simulation results in Romano et al. [2019b], Lei and Candès [2020], Candès et al. [2021] and Figure 2.6), leading to a quantile $Q_{1-\alpha}(\dots)$ that is positive, and a LPB $\hat{L}(X)$ that is lower (more conservative) than the original fitted model.

The following result proves that this is a marginally calibrated LPB:

Theorem 2.1 (Adapted from Theorem 1 of Romano et al. [2019b]). *Suppose $(X_i, C_i, T_i) \stackrel{\text{iid}}{\sim} P$. Then $\hat{L}_{\text{baseline}}(X)$ is a marginally calibrated LPB at level $1 - \alpha$, and moreover, satisfies*

$$\mathbb{P}_{(X, T, C) \sim P}(\tilde{T} \geq \hat{L}_{\text{baseline}}(X)) \geq 1 - \alpha.$$

Since $\tilde{T}$ may be often much smaller than T if the censoring is severe, this result indicates that such an LPB may be quite conservative as a lower bound for T . This conservativeness is however inescapable without further assumptions—Candès et al. [2021, Theorem 1] establish that, under mild conditions, for any marginally calibrated LPB $\hat{L}$ for the (uncensored) survival time T which is valid universally over all distributions P on the data, $\hat{L}$ must also be an LPB for $\tilde{T}$ whenever $P_{(C, T)}$ is either discrete or continuous.

2.2.2 Using fixed threshold for filtering out low censoring times

Next we give details for Candès et al. [2021]’s proposed method, which uses a fixed threshold c_0 to avoid an overly conservative LPB. As mentioned above, their work shows that, without further assumptions, it is not possible to improve on the LPB for $\tilde{T}$; therefore, they make the additional assumption that the conditional distribution $P_{C|X}$ is known (or is estimated accurately).

As for CQR, their method begins by partitioning the data into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$, and uses the training set to fit a quantile regression,² $x \mapsto \hat{q}_\alpha(x)$, for the conditional α -quantile of T given X . The cutoff c_0 for the censoring time may also be chosen as a function of the training data. Furthermore, define $\hat{w}(x)$ to be an estimate of $1/\mathbb{P}\{C \geq c_0 \mid X = x\}$ (or, approximately proportional to this quantity), also fitted on the training data.

Next, on the calibration set, we use c_0 to filter the data and define $\mathcal{I}'_2 = \{i \in \mathcal{I}_2 : C_i \geq c_0\}$. For all these remaining calibration points, note that $\tilde{T}_i \wedge c_0 = T_i \wedge c_0$ (that is, $T_i \wedge c_0$ is observed). We then calculate scores $S_i = \hat{q}_\alpha(X_i) - T_i \wedge c_0$ for all $i \in \mathcal{I}'_2$, and return the LPB

$$\hat{L}_{\text{fixed-cutoff}}(X) = \hat{q}_\alpha(X) - Q_{1-\alpha} \left(\frac{\sum_{i \in \mathcal{I}'_2} \hat{w}(X_i) \cdot \delta_{S_i} + \hat{w}(X) \cdot \delta_{+\infty}}{\sum_{i \in \mathcal{I}'_2} \hat{w}(X_i) + \hat{w}(X)} \right).$$

The intuition here is that the calibration set $\mathcal{I}'_2$ consists of data points drawn from the shifted distribution $P^{\geq c_0}$, and the likelihood ratio between the target distribution P and this distribution $P^{\geq c_0}$ is $\mathbb{P}\{C \geq c_0\} / \mathbb{P}\{C \geq c_0 \mid X\}$; since $\hat{w}(X)$ is an estimate of the likelihood ratio (up to constants), reweighting the calibration data points with weights $\hat{w}(X_i)$ ensures coverage with respect to the actual target distribution P .

2. While their proposed method is defined via a more general construction, here we focus on a single version that is most relevant for comparison to our own methods.

Building on the framework of conformal prediction with covariate shift [Tibshirani et al., 2019], Candès et al. [2021]’s result proves that this construction yields a valid LPB.

Theorem 2.2 (Proposition 1 of Candès et al. [2021]). *Suppose $(X_i, C_i, T_i) \stackrel{\text{iid}}{\sim} P$, and suppose $\widehat{w}(x) = 1/\mathbb{P}\{C \geq c_0 \mid X = x\}$, i.e., this probability was fitted exactly. Then $\widehat{L}_{\text{fixed-cutoff}}(X)$ is a marginally calibrated LPB for $T \wedge c_0$, and therefore also for T .*

Moreover, Candès et al. [2021, Theorem 2] establish a double robustness result: if either $\widehat{w}(x)$ was fitted accurately (i.e., is a good approximation of $1/\mathbb{P}\{C \geq c_0 \mid X = x\}$ or the quantile regression was fitted accurately (i.e., $\widehat{q}_\alpha(x)$ is a good approximation of the α -quantile of $T \mid X$), then $\widehat{L}_{\text{fixed-cutoff}}$ approximately satisfies the criterion for a marginally calibrated LPB.

2.3 Conformalized survival analysis with adaptive cutoffs

2.3.1 Our procedure

As before, we first partition the data into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$. On the training set, we fit a *family* of estimated quantile regression functions, $(x, a) \mapsto \widehat{q}_a(x)$, mapping x to the estimated a -quantile of the conditional distribution of T given $X = x$, for all $a \in [0, 1]$. We assume that, for any x , $a \mapsto \widehat{q}_a(x)$ is nondecreasing.³ (In contrast, for the existing methods defined in Section 2.2, this regression is run only at a single value of a .)

Next, we need to use the calibration set in order to choose a value of a , for which returning $\widehat{q}_a(X)$ is a valid LPB—that is, we need to find a value a such that $\mathbb{P}\{T < \widehat{q}_a(X)\} = \alpha$ (where we implicitly treat the fitted quantile function $\widehat{q}_a$ as fixed, and take the probability over $(X, T) \sim P$). Note that, if the original regression were estimated perfectly, then we

3. If our estimators $\widehat{q}_a$ are computed independently for each a and this constraint is violated, monotonicity can easily be restored via sorting the outputs—see e.g., Koenker [1994]. $\widehat{q}_a$ is defined to be $-\infty$ at $a = 0$.

would expect to return $a = \alpha$, i.e., the estimated α -quantile $\hat{q}_\alpha(X)$ would already be a valid LPB. In practice, as discussed earlier, we expect to see overfitting in most real-data settings and thus we expect to return $a < \alpha$. To choose a appropriately, we could consider solving for a in the following expression:

$$\begin{aligned}
\alpha &= \mathbb{P}\{T < \hat{q}_a(X)\} = \mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \mid X\}] \\
&= \mathbb{E}\left[\mathbb{P}\{T < \hat{q}_a(X) \mid X\} \cdot \frac{\mathbb{P}\{\hat{q}_a(X) \leq C \mid X\}}{\mathbb{P}\{\hat{q}_a(X) \leq C \mid X\}}\right] \\
&\approx \mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \mid X\} \cdot \mathbb{P}\{\hat{q}_a(X) \leq C \mid X\} \cdot \hat{w}_a(X)] \\
&= \mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \leq C \mid X\} \cdot \hat{w}_a(X)] \\
&= \mathbb{E}[\mathbb{1}\{T < \hat{q}_a(X) \leq C\} \cdot \hat{w}_a(X)], \tag{2.2}
\end{aligned}$$

where now $\hat{w}_a(x)$ is chosen to be (approximately) equal to $1/\mathbb{P}\{C \geq \hat{q}_a(X) \mid X = x\}$, and where the next-to-last step holds by Assumption A1. If instead we only assume that our estimate $\hat{w}_a(x)$ is *proportional* to $1/\mathbb{P}\{C \geq \hat{q}_a(X) \mid X = x\}$, then we want to solve for a in the equation

$$\begin{aligned}
\alpha &= \mathbb{P}\{T < \hat{q}_a(X)\} \approx \frac{\mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \mid X\} \cdot \mathbb{P}\{\hat{q}_a(X) \leq C \mid X\} \cdot \hat{w}_a(X)]}{\mathbb{E}[\mathbb{P}\{\hat{q}_a(X) \leq C \mid X\} \cdot \hat{w}_a(X)]} \\
&= \frac{\mathbb{E}[\mathbb{1}\{T < \hat{q}_a(X) \leq C\} \cdot \hat{w}_a(X)]}{\mathbb{E}[\mathbb{1}\{\hat{q}_a(X) \leq C\} \cdot \hat{w}_a(X)]}. \tag{2.3}
\end{aligned}$$

Now we note that the events $\mathbb{1}\{T_i < \hat{q}_a(X_i)\}$ cannot be observed on the calibration set (since we only observe the censored survival time, $\tilde{T}_i$). On the other hand, the filtered events $\mathbb{1}\{T_i < \hat{q}_a(X_i) \leq C_i\}$ *can* be observed (since if $C_i \geq \hat{q}_a(X_i)$, $\mathbb{1}\{T_i < \hat{q}_a(X_i)\} = \mathbb{1}\{\tilde{T}_i < \hat{q}_a(X_i)\}$). Therefore, the calibration set can indeed be used to find a value a so that the equation (2.2) or (2.3) is (approximately) satisfied.

Now we formally describe how to select a using the calibration set. For each value a , we

estimate the miscoverage rate $\mathbb{P}\{T < \hat{q}_a(X)\}$ as follows:

$$\hat{\alpha}(a) = \frac{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i) \cdot \mathbb{1}\{T_i < \hat{q}_a(X_i) \leq C_i\}}{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i) \cdot \mathbb{1}\{\hat{q}_a(X_i) \leq C_i\}}.$$

This empirical quantity estimates $\alpha^*(a) = \mathbb{P}\{T < \hat{q}_a(X)\}$. Since our aim is to find a value of a sufficiently small so that $\alpha^*(a) \leq \alpha$, we will instead search for a satisfying $\hat{\alpha}(a) \leq \alpha$. However, while $\alpha^*(a)$ is monotone in a (since $\hat{q}_a(x)$ is monotone in a), this property may not hold for the estimator $\hat{\alpha}(a)$; we therefore define $\hat{a} = \sup\{a \in [0, 1] : \sup_{a' \leq a} \hat{\alpha}(a') \leq \alpha\}$. Finally, we output the LPB $\hat{L}(X) := \hat{q}_{\hat{a}}(X)$.

Below, we will give a double robustness result proving that this choice of $\hat{L}$ is (approximately) a marginally calibrated LPB, as long as *either* the weights $\hat{w}_a$ *or* the quantiles $\hat{q}_a$ are fitted accurately; furthermore, when the quantiles are fitted accurately, the LPBs are (approximately) conditionally valid. Before giving our theoretical results, we first present a more general form of this procedure.

2.3.2 A generalized procedure

The procedure described above tends to perform well in settings where $\hat{q}_a(x)$ (for relevant values of a) is not too large—so that $\mathbb{P}\{C \geq \hat{q}_a(x) \mid X = x\}$ is not close to zero and the weights $\hat{w}_a(X_i)$ on the calibration points are not too large. In other settings, however, the procedure may be somewhat unstable. Specifically, in scenarios where C is often much smaller than T (as in the example given in Section 2.1.4 above), we might have a very small probability $\mathbb{P}\{C \geq \hat{q}_a(X) \mid X = x\}$; this is problematic since the inverse weight $\hat{w}(x)$ will then be extremely large. To alleviate this, we now generalize the procedure sketched above to allow for a more stable and robust method. Define a family of functions for $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $(x, a) \mapsto \hat{f}_a(x)$, which are fitted on the training set, such that for each fixed x this map is nondecreasing in a (note that the estimated quantiles, $\hat{q}_a(x)$, are simply a special

case). Our aim is now to use the calibration set in order to choose a so that $\widehat{L}(X) = \widehat{f}_a(X)$ offers a calibrated LPB. With the same rationale as before, we wish to find a to satisfy

$$\begin{aligned}\alpha = \mathbb{P}\left\{T < \widehat{f}_a(X)\right\} &\approx \frac{\mathbb{E}\left[\mathbb{P}\left\{T < \widehat{f}_a(X) \mid X\right\} \cdot \mathbb{P}\left\{\widehat{f}_a(X) \leq C \mid X\right\} \cdot \widehat{w}_a(X)\right]}{\mathbb{E}\left[\mathbb{P}\left\{\widehat{f}_a(X) \leq C \mid X\right\} \cdot \widehat{w}_a(X)\right]} \\ &= \frac{\mathbb{E}\left[\mathbb{1}\left\{T < \widehat{f}_a(X) \leq C\right\} \cdot \widehat{w}_a(X)\right]}{\mathbb{E}\left[\mathbb{1}\left\{\widehat{f}_a(X) \leq C\right\} \cdot \widehat{w}_a(X)\right]},\end{aligned}\tag{2.4}$$

where $\widehat{w}_a$ is again fitted on the training data but is now chosen to be (approximately) proportional to $1/\mathbb{P}\left\{C \geq \widehat{f}_a(X) \mid X = x\right\}$.

From this point on, we proceed exactly as before, but with $\widehat{f}_a$ in place of $\widehat{q}_a$ —we define

$$\widehat{\alpha}(a) = \frac{\sum_{i \in \mathcal{I}_2} \widehat{w}_a(X_i) \cdot \mathbb{1}\left\{T_i < \widehat{f}_a(X_i) \leq C_i\right\}}{\sum_{i \in \mathcal{I}_2} \widehat{w}_a(X_i) \cdot \mathbb{1}\left\{\widehat{f}_a(X_i) \leq C_i\right\}},$$

which estimates $\alpha^*(a) = \mathbb{P}\left\{T < \widehat{f}_a(X)\right\}$. As before, we compute

$$\widehat{a} = \sup\left\{a \in [0, 1] : \sup_{a' \leq a} \widehat{\alpha}(a') \leq \alpha\right\},\tag{2.5}$$

and return the LPB $\widehat{L}(X) := \widehat{f}_{\widehat{a}}(X)$.

Choosing the family of bounds. In this more general procedure, how should the family $\widehat{f}_a(x)$ be chosen? The LPB will be approximately valid regardless of our choice, but the utility of the method will depend strongly on choosing a reasonable family of functions. We consider two goals when choosing the family:

- We would like to closely approximate the “oracle” LPB, $L(X) = q_\alpha(X)$, where $q_\alpha(x)$ is the true α -quantile of T given $X = x$. As a result, for a such that $\widehat{q}_a(x)$ is close to

$q_a(x)$, we would like to have $\hat{f}_a(x) \approx \hat{q}_a(x)$, our estimated quantiles for T given X .

- On the other hand, we would like for the weights $\hat{w}_a(x)$ to not be too large, or equivalently, for $\mathbb{P}\left\{C \geq \hat{f}_a(X) \mid X = x\right\}$ to not be too small for any a . Consequently, we might want to require $\hat{f}_a(x) \leq \hat{q}_{1-\beta}^C(x)$, where $\hat{q}_{1-\beta}^C(x)$ estimates the conditional quantile of C given $X = x$, and we choose some constant value β .

To balance between these two goals, we propose selecting $\hat{f}_a(x) = \min\left\{\hat{q}_a(x), \hat{q}_{1-\beta}^C(x)\right\}$. As for the choice of β , we use $\beta = 1/\log|\mathcal{I}_2|$ in our implementation such that $\hat{w}_a(X_i) \leq \log|\mathcal{I}_2| = o(\sqrt{|\mathcal{I}_2|})$. In the simulations, we will compare this choice against the “canonical” version of the method with $\hat{f}_a(x) = \hat{q}_a(x)$, to see how this new choice adds stability to the method.

Implementation details. Next we describe how the threshold $\hat{a}$ in (2.5) can be computed efficiently in practice. We note that $\sup_{a' \leq a} \hat{\alpha}(a')$ is a non-decreasing piecewise constant function in a , with no more than $2n$ knots—values of a at which the indicators $\mathbb{1}\left\{T_i < \hat{f}_a(X_i) \leq C_i\right\}$ or $\mathbb{1}\left\{\hat{f}_a(X_i) \leq C_i\right\}$ change signs. Denote $\bar{a}_i = \sup_{a \in [0,1]} \left\{\hat{f}_a(X_i) \leq \tilde{T}_i\right\}$, $\tilde{a}_i = \sup_{a \in [0,1]} \left\{\hat{f}_a(X_i) \leq C_i\right\}$ and $\mathcal{A}_1 = \{\bar{a}_i : i = 1, \dots, n\}$, $\mathcal{A}_2 = \{\tilde{a}_i : i = 1, \dots, n\}$. Then by definition, the breakpoints of the piecewise constant map $a \mapsto \hat{\alpha}(a)$ must all lie in $\mathcal{A}_1 \cup \mathcal{A}_2$. In the implementation, in order to obtain $\hat{a}$, we only need to search through the finite grids

$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \{0\}. \quad (2.6)$$

A complete description of the general procedure can be found in Algorithm 1.

Computational complexity. The computational cost of our proposed procedure can be decomposed into the cost of model fitting on $\mathcal{I}_1$ and that of finding $\hat{a}$ on $\mathcal{I}_2$. The cost of the first stage heavily depends on the type of models chosen by the user. For the second stage, we

Algorithm 1 Conformalized survival analysis with adaptive cutoffs

Input: Level α ; data $\mathcal{D} = (X_i, \tilde{T}_i, C_i)_{i \in [n]}$.

Procedure:

1. Split the data into two folds: the training fold $\mathcal{I}_1$ and the calibration fold $\mathcal{I}_2$.
2. Using $\mathcal{I}_1$ as input, apply any algorithm to fit the candidate LPBs $\{\hat{f}_a(\cdot)\}_{a \in [0,1]}$.
3. Using $\mathcal{I}_1$ as input, apply any algorithm to construct estimates $\hat{w}_a(x)$ of $\mathbb{P}\{C \geq \hat{f}_a(x) \mid X = x\}$.
4. Determine $\mathcal{A}$ according to (2.6).
5. **For** a in $\mathcal{A}$ **do:** Compute the estimated miscoverage rate

$$\hat{\alpha}(a) = \frac{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i) \cdot \mathbb{1}\{T_i < \hat{f}_a(X_i) \leq C_i\}}{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i) \cdot \mathbb{1}\{\hat{f}_a(X_i) \leq C_i\}}. \quad (2.7)$$

6. Compute the threshold: $\hat{a} = \sup\{a \in \mathcal{A} : \sup_{a' \leq a, a' \in \mathcal{A}} \hat{\alpha}(a') \leq \alpha\}$.

Return: The calibrated LPB: $\hat{L}(\cdot) = \hat{f}_{\hat{a}}(\cdot)$.

first need to find the set of “knots” $\mathcal{A}$ defined in (2.6). For each $i \in \mathcal{I}_2$, finding $\bar{a}_i$ (resp. $\tilde{a}_i$) requires finding the supremum over a such that $\hat{f}_a(X_i) \leq \tilde{T}_i$ (resp. $\hat{f}_a(X_i) \leq C_i$). Since $\hat{f}_a(X_i)$ is nondecreasing in a , finding $\bar{a}_i$ or $\tilde{a}_i$ can be done efficiently via binary search. More specifically, given a tolerance level ϵ , we can obtain an ϵ -accurate solution within $O(\log(1/\epsilon))$ runs. Repeating the above for all $i \in \mathcal{I}_2$ requires $O(|\mathcal{I}_2| \cdot \log(1/\epsilon))$ runs. Finally, evaluating $\hat{\alpha}(a)$ for $a \in \mathcal{A}$ and finding $\hat{a}$ requires $O(|\mathcal{I}_2|)$ runs. Overall, the computational complexity of the second stage is of the order $O(|\mathcal{I}_2| \cdot (1 + \log(1/\epsilon)))$.

2.3.3 Theoretical guarantee: a double robustness result

In this section, we establish the theoretical guarantees for the LPBs produced by Algorithm 1. In particular, we show that the LPBs enjoy a double-robustness property in the following sense: the LPBs are approximately marginally calibrated if *either* the censoring mechanism *or* the conditional quantile of survival times can be estimated well; when the latter is true, the LPBs are furthermore approximately conditionally calibrated.

Given the class of functions $\{\hat{f}_a(\cdot)\}_{a \in [0,1]}$, we define the oracle weights

$$w_a(x) = \left(\mathbb{P} \left\{ C \geq \hat{f}_a(X) \mid \mathcal{I}_1, X = x \right\} \right)^{-1},$$

(here we condition on $\hat{f}_a$, i.e., the function is treated as fixed), and the following oracle quantity for any $\beta \in [0, 1]$:

$$a(\beta) = \sup \left\{ a \in [0, 1] : \mathbb{P} \left\{ T < \hat{f}_a(X) \mid \mathcal{I}_1 \right\} \leq \beta \right\}.$$

Theorem 2.3 and 2.4 develop the coverage guarantee for the LPBs.

Theorem 2.3. *Fix any $\delta, \alpha \in (0, 1)$. Assume that $\hat{f}_a(x)$ is continuous in a , and that for any $a \in [0, 1]$, there exists some constant $\hat{\gamma}_a > 0$ such that $\hat{w}_a(x) \leq \hat{\gamma}_a$ for P_X -almost all x . Then with probability at least $1 - \delta$ over the draw of $\mathcal{D}$, the LPB produced by Algorithm 1 satisfies*

$$\begin{aligned} \mathbb{P}_{(X,T) \sim P} \left\{ T \geq \hat{L}(X) \mid \mathcal{D} \right\} &\geq 1 - \alpha \\ &- \sup_{a \in [0,1]} \left(\mathbb{E} \left[\left| \frac{\hat{w}_a(X)}{w_a(X) \hat{\pi}_a} - 1 \right| \mid \mathcal{I}_1 \right] + \sqrt{\frac{1 + \frac{\hat{\gamma}_a^2}{\hat{\pi}_a^2} + \max(1, \frac{\hat{\gamma}_a}{\hat{\pi}_a} - 1)^2}{|\mathcal{I}_2|} \cdot \log \left(\frac{1}{\delta} \right)} \right), \end{aligned}$$

where the probability is taken with respect to a new data point $(X, T) \sim P_{(X,T)}$, and where we define $\hat{\pi}_a = \mathbb{E}_{X \sim P_X} \left[\frac{\hat{w}_a(X)}{w_a(X)} \mid \mathcal{I}_1 \right]$ for any $a \in [0, 1]$.

The proof of Theorem 2.3 is deferred to Appendix A. In other words, if the estimates $\hat{w}_a$ are accurate approximations of w_a (up to rescaling by a constant), then we have $\hat{w}_a(X)/w_a(X)\hat{\pi}_a \approx 1$, and approximate coverage is guaranteed.

Next, we show that we also achieve approximate coverage when $T \mid X$ can be accurately

modeled.

Theorem 2.4. *Fix any $\delta, \alpha \in (0, 1)$. Assume the same conditions as Theorem 2.3, and assume further that the conditional distribution of $T \mid X$ is continuous, with its conditional density upper bounded by a constant $B > 0$, and that there exists a constant $r > 0$ such that*

1. $\sup_{\xi \in [a(\alpha), a(\alpha+r)+\psi]} w_\xi(x) \leq \gamma$ and $\sup_{\xi \in [a(\alpha), a(\alpha+r)+\psi]} \hat{w}_\xi(x) \leq \hat{\gamma}$ for some constants $\psi, \gamma, \hat{\gamma} > 0$;
2. $\sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in \mathcal{X}} \left\{ \max(B, 1) \cdot |\hat{f}_{a(\beta)}(x) - q_\beta(x)| \right\} + \hat{\gamma} \gamma \sqrt{\frac{\log(1/\delta)}{|\mathcal{I}_2|}} \leq r$, where $q_\beta(x)$ is the β -quantile of T conditional on $X = x$.

Then with probability at least $1 - \delta$ over the draw of $\mathcal{D}$, the LPB produced by Algorithm 1 satisfies that for P_X -almost all x ,

$$\begin{aligned} \mathbb{P}_{(X,T) \sim P} \left\{ T \geq \hat{L}(x) \mid \mathcal{D}, X = x \right\} &\geq 1 - \alpha \\ &- \sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in \mathcal{X}} \left\{ 2B \cdot |\hat{f}_{a(\beta)}(x) - q_\beta(x)| \right\} - \hat{\gamma} \gamma \sqrt{\frac{1}{|\mathcal{I}_2|} \cdot \log\left(\frac{1}{\delta}\right)}. \end{aligned}$$

The proof of Theorem 2.4 is deferred to the Appendix A, where we in fact prove a more general version. The implication of Theorem 2.4 is that if $T \mid X$ can be modeled well, the conditional miscoverage rate will be small (which also implies that the marginal coverage rate will be small).

Remark 1. *The assumption on the continuity $\hat{f}_a(x)$ in a and the boundedness on the estimated weights can simply be satisfied by choosing the appropriate class of functions in the training stage (i.e., the fitting procedure using $\mathcal{I}_1$). The additional assumption (a) requires the oracle weights $w_\beta(x)$ is bounded as least in a neighborhood of $a(\alpha)$; (b) is satisfied when $T \mid X$ is estimated uniformly well in a neighborhood of α and when $|\mathcal{I}_2|$ is sufficiently large.*

2.4 Simulations

We set up six synthetic experiments, and under each setting we generate $N = 100$ i.i.d. datasets.⁴

Each dataset consists of the training set $\mathcal{I}_1$, calibration set $\mathcal{I}_2$, and the test set $\mathcal{I}_3$, where $|\mathcal{I}_1| = 1,000$, $|\mathcal{I}_2| = 1,000$ and $|\mathcal{I}_3| = 5,000$. For all experiments, the target level is $1 - \alpha = 90\%$. In these experiments, we implement our proposed method with two families of bounds:

- DFT-adaptive-T: the candidate LPB is given by $\hat{f}_a(x) = \hat{q}_a(x)$, where $\hat{q}_a(x)$ is the estimated a -th conditional quantile of T given $X = x$.
- DFT-adaptive-CT: the candidate LPB is given by $\hat{f}_a(x) = \min \{ \hat{q}_a(x), \hat{q}_{1-\log(1/|\mathcal{I}_2|)}^C(x) \}$, where $\hat{q}_a(x)$ is as before and $\hat{q}_b^C(x)$ is the estimated b -th conditional quantile of C given $X = x$.

We also obtain LPBs based on parametric models and other distribution-free methods:

- Cox: LPBs generated by the estimated Cox model that is implemented as in Therneau [2020]
- RandomForest: LPBs returned by the censored quantile regression forest [Li and Bradic, 2020, Athey et al., 2019]; the implementation is based on Li and Bradic [2020].
- DFT-baseline: The distribution-free LPBs obtained by applying conformal quantile regression [Romano et al., 2019b] to generating bounds for $\tilde{T}$.
- DFT-fixed: conformalized LPB with a fixed thresholded c_0 ; the implemental details are as suggested by Candès et al. [2021].

4. The code for reproducing all numerical results from the simulation and the real data analysis can be found at https://github.com/zhimeir/adaptive_conformal_survival_paper.

For all conformalized method, the base algorithm for fitting conditional quantile of $T \mid X$ is the Cox model, and a Gaussian process model is fitted to approximate $C \mid X$ (this is implemented by the `GauPro` R-package [?]). For each dataset, we compute the following two quantities with the test set:

$$\text{Empirical coverage} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \mathbb{1} \left\{ \widehat{L}(X_i) < T_i \right\}, \quad \text{Average LPB} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \widehat{L}(X_i).$$

An ideal method would have empirical coverage $\approx 1 - \alpha$, and average LPB as low as possible. We shall demonstrate boxplots of the empirical coverage and average LPB resulting from the 100 datasets.

2.4.1 Synthetic setup

We consider six data generating models, where settings 1–4 concern univariate X and settings 5–6 multivariate X . For all settings, the marginal distribution of the covariates is given by $P_X = \text{Unif}([0, 4]^p)$; conditional on X , we generate T and C via distributions $\log T \mid X \sim \mathcal{N}(\mu(X), \sigma^2(X))$ and $C \mid X \sim P_{C|X}$.

In settings 1 and 2, $p = 1$ and $C \mid X \sim \text{Exp}(0.1)$ —the censoring mechanism is completely exogenous; in settings 3 and 4, $p = 1$ and we allow C to depend on X . In particular, setting 3 corresponds to the example shown in the introduction. Settings 5 and 6 consider multivariate X , where $p = 10$; in setting 5, $\sigma(x) = 1$, and in setting 6 $\sigma(x)$ depends on X . Table 2.1 summarizes the parameters used in the six settings.

Figure 2.2 shows the scatterplots of the survival time T and censoring time C against the univariate covariate X in univariate experimental settings. Settings 2–4 are more challenging than setting 1, as they all have scenarios where there are roughly two sub-populations: the sub-population with smaller values of X has relatively higher censoring time, leading to a low

Setting	p	$\mu(x)$	$\sigma(x)$	$P_{C X}$
1	1	$0.632x$	2	$\text{Exp}(0.1)$
2	1	$3 \cdot \mathbb{1}\{x > 2\} + x \cdot \mathbb{1}\{x \leq 2\}$	0.5	$\text{Exp}(0.1)$
3	1	$2 \cdot \mathbb{1}\{x > 2\} + x \cdot \mathbb{1}\{x \leq 2\}$	0.5	$\text{Exp}(0.25 + \frac{6+x}{100})$
4	1	$3 \cdot \mathbb{1}\{x > 2\} + 1.5x \cdot \mathbb{1}\{x \leq 2\}$	0.5	$\text{lognormal}(2 + \frac{(2-x)}{50}, 0.5)$
5	10	$0.126(x_1 + \sqrt{x_3 x_5}) + 1$	1	$\text{Exp}(\frac{x_{10}}{10} + \frac{1}{20})$
6	10	$0.126(x_1 + \sqrt{x_3 x_5}) + 1$	$\frac{x_2+2}{4}$	$\text{Exp}(\frac{x_{10}}{10} + \frac{1}{20})$

Table 2.1: Parameters used in the six experimental settings: $P_X = \text{Unif}([0, 4]^p)$, $\mathbb{P}\{T \mid X\} = \exp(\mathcal{N}(\mu(X), \sigma^2(X)))$.

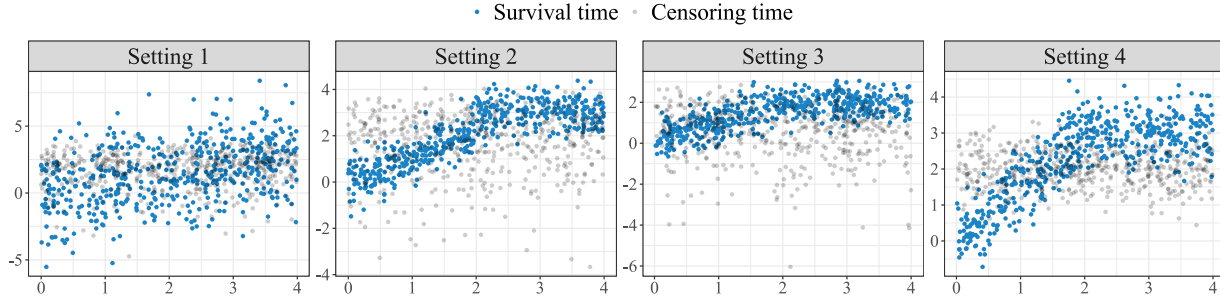


Figure 2.2: Illustration of the censoring time C and the uncensored survival time T in settings 1–4 as functions of the univariate covariate X . Both C and T are on a log scale.

censoring zone while the sub-population with larger values of X has comparatively higher survival time, hence a very high censoring zone. In settings 5–6, there is a similar challenge: the distribution $P_{C|X}$ depends on X_{10} and thus we will have low censoring times for certain values of X and higher censoring times for others.

2.4.2 Simulation results

Figure 2.3 plots the empirical coverage and average LPBs of all candidate methods under the univariate settings. First we consider the methods without distribution-free type guarantees. As we see in the figures, Random Forest exhibits undercoverage in all four settings. Even in setting 1, a simple case where the error is homogeneous and C does not depend on X , RandomForest does not return valid LPBs. Cox shows undercoverage as well, except for the simple regime of setting 1, and the miscoverage gaps are even larger than those of

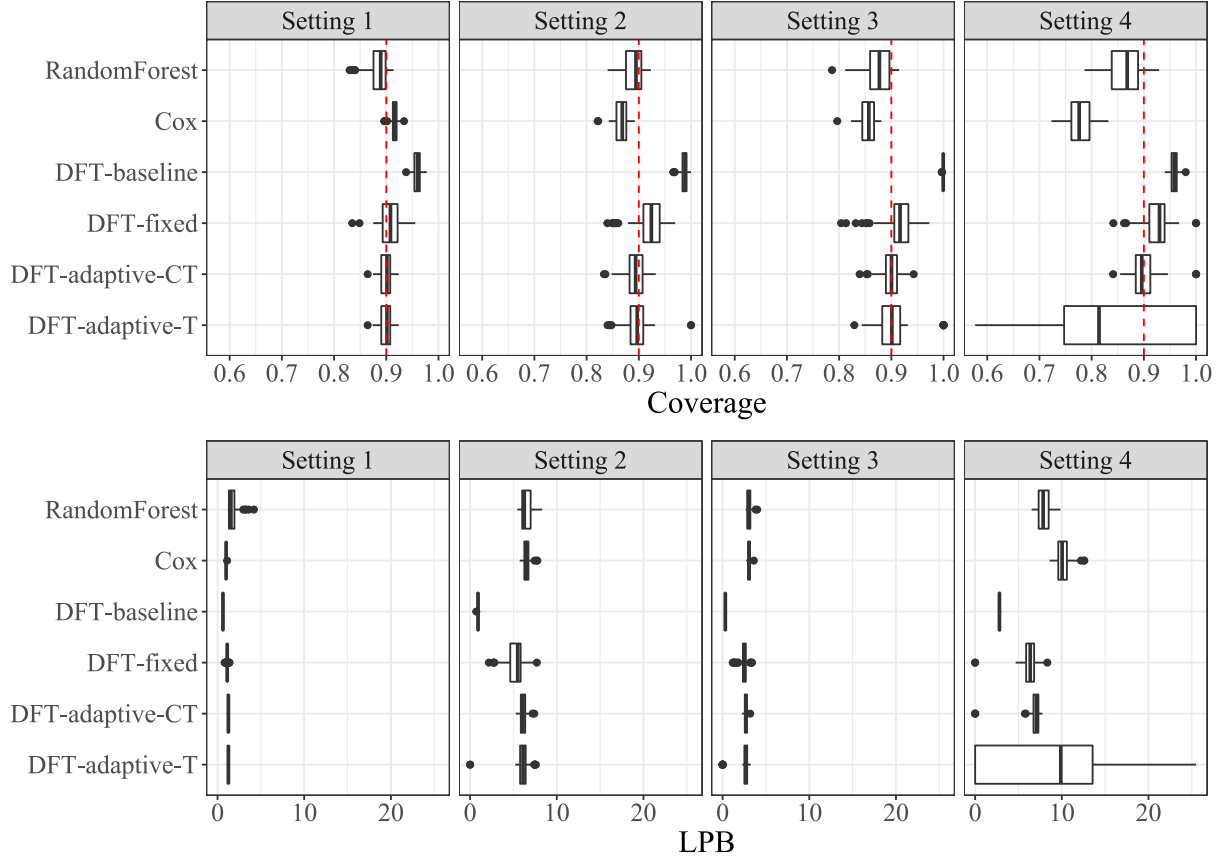


Figure 2.3: Empirical coverage (top) and average LPBs (bottom) of all the candidate methods under settings 1–4, where X is univariate. The boxplot shows results from 100 independent draw of datasets. The dashed red line corresponds to the target coverage level $1 - \alpha = 90\%$.

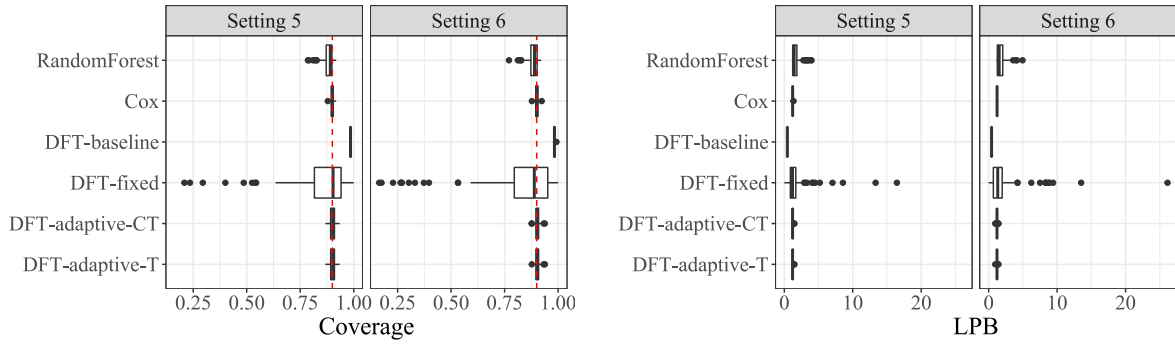


Figure 2.4: Boxplots of empirical coverage (left) and average LPBs (right) in the multivariate experimental settings. The details are otherwise the same as in Figure 2.3

RandomForest in settings 2, 3, and 4.

Next we consider the DFT methods. The LPBs returned by DFT-baseline are very conservative in all four settings due to the censoring issue, as we expected (recall from our discussion in Section 2.1.2 that this method covers T by covering the censored time $\tilde{T}$). DFT-fixed LPBs are more conservative than our proposed DFT-adaptive-CT LPBs, especially in settings 2-4 where the relationship between C and T changes drastically in different subpopulations. Finally, we can see the canonical version of our method, DFT-adaptive-T, exhibits high variability in settings 3 and 4, verifying our statement on stability in Section 2.3.2 and highlighting the potential advantages of DFT-adaptive-CT.

Next, Figure 2.4 demonstrates the results under the multivariate settings. Here, we observe that RandomForest shows slight undercoverage under both settings; Cox performs well and does not undercover, but does not offer a distribution-free coverage guarantee. Turning to the DFT methods, DFT-baseline is again very conservative, and DFT-fixed LPBs exhibit high variability. Both of our proposed methods, DFT-adaptive-T and DFT-adaptive-CT are able achieve exact coverage, and the variability is much lower than that of DFT-fixed LPBs.

Finally, we show in Figure 2.5 the running time of all the candidate methods under setting 3 (the example in Section 2.1.4) with different choices of n . We can see that DFT-fixed, DFT-adaptive-T, and DFT-adaptive-CT are more computationally expensive than the other methods; the running time of DFT-fixed and DFT-adaptive-CT is comparable, while that of DFT-adaptive-T is somewhat shorter.

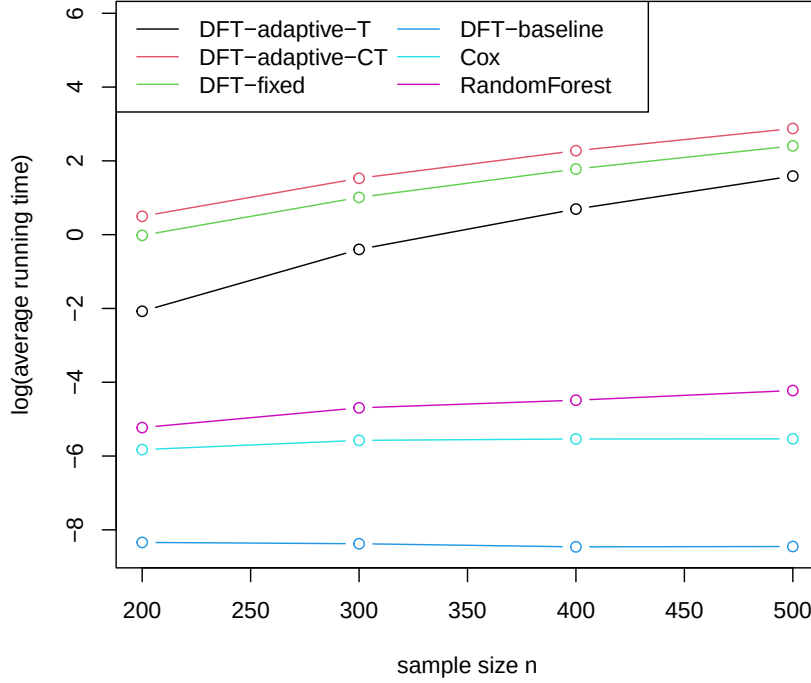


Figure 2.5: Running time of all candidate methods in setting 3.

2.5 Real data application

In this section, we apply our proposed method to predicting users' active time on a mobile app with a publicly available dataset.⁵ This dataset records the time stamps of pings for a cohort of 2,476 users in a shared window of three weeks, where a ping represents a login activity or a received message. As is shown in Figure 2.6(a), a user's pings gathered during an active day form a line segment, whose length is proportional to the span of active time during that day (the time span is standardized so that the total time window is mapped to the interval $[0, 21]$ to represent the total number of days). The number of line segments and the length of line segments vary for different users, reflecting different types of user behavior. For each user, the time is recorded from the user's first active day, i.e., if the first active

5. The data is downloaded from <https://www.kaggle.com/datasets/bhuvanchennoju/mobile-usage-time-prediction?select=pings.csv>.

time for a particular user is 1.5, then the sequence of active time for this user is shifted by $\lfloor 1.5 \rfloor = 1$ and is censored at time $C = 21 - 1 = 20$.

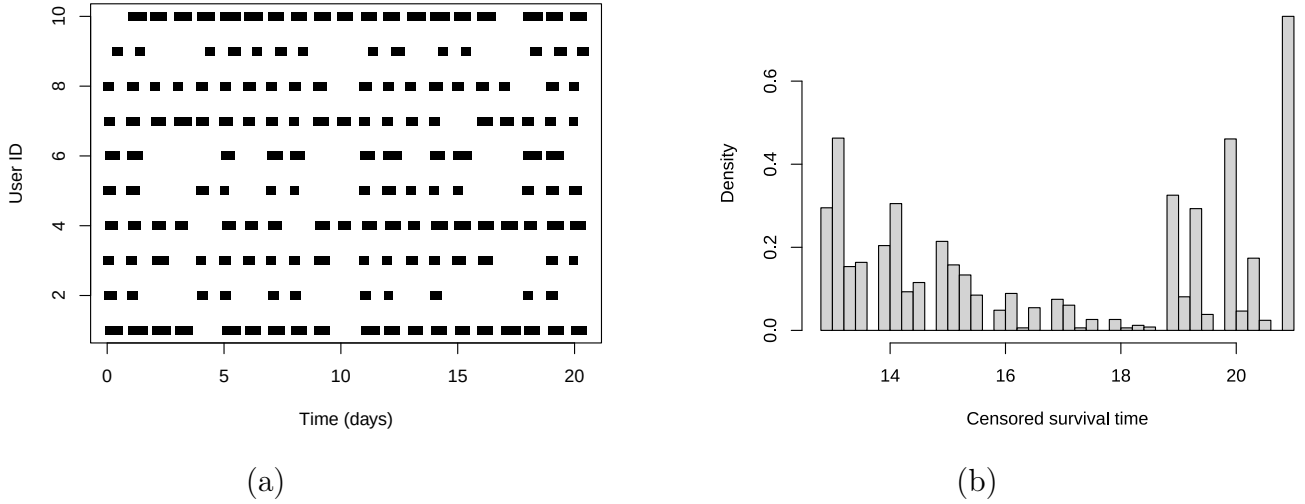


Figure 2.6: (a) Active time in three weeks for $i = 1, \dots, 10$; (b) histogram of censored survival times for all users.

With this dataset, we focus on predicting *the beginning of a user's 14th active day*. In practice, the prediction lower bounds can be informative if, for instance, the mobile app wishes to launch a promotion, offering a discount for in-app purchases at the beginning of a user's 14th active day. For a user who is active for less than 14 days within the time window, the survival time is therefore censored and only $\tilde{T} = \min(C, T)$ can be observed. Figure 2.6(b) is the histogram of the censored survival time. Besides the time stamps, there are three covariates in this dataset related to users' characteristics: X_1 (gender), X_2 (age), and X_3 (number of children).

To implement the method, we begin by choosing $|\mathcal{I}_1| = 500$ data points as the training set, and keep this set fixed throughout. Among the remaining data points, for 50 independent random trials, we sample $|\mathcal{I}_2| = 500$ data points as the calibration set and another $|\mathcal{I}_3| = 500$ as the test set, uniformly without replacement. All the methods are applied with the target

level $1 - \alpha$ at 90%. Since the true survival time for censored data points are not available, we instead empirically evaluate the upper and lower bounds of the coverage rate: we compute $\beta_{\text{lo}} := \mathbb{P}\{\tilde{T} \geq \hat{L}(X)\} \leq \mathbb{P}\{T \geq \hat{L}(X)\}$, and also $\beta_{\text{hi}} := 1 - \mathbb{P}\{\tilde{T} < \hat{L}(X), T \leq C\} \geq \mathbb{P}\{T \geq \hat{L}(X)\}$, so that, by construction, β_{lo} is an underestimate of our target coverage rate, and β_{hi} is an overestimate.

The upper and lower bounds for methods in comparison are reported in Figure 2.7. Since

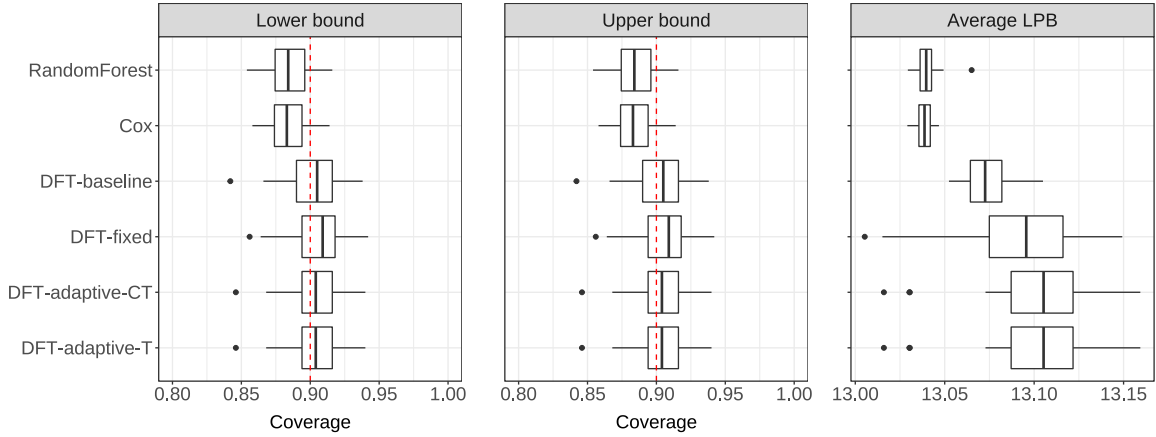


Figure 2.7: Left: lower bound β_{lo} of the empirical coverage rate; middle: upper bound β_{hi} of the empirical coverage rate; right: average LPBs.

this setting has a low censoring rate (19.7%), the difference between DFT-adaptive-T and DFT-adaptive-CT is negligible. Both DFT-adaptive-T and DFT-adaptive-CT attain nearly exact coverage at 90%, while Cox and Random Forests have coverage below the target level. In comparison, although DFT-baseline has only slightly inflated coverage, the average LPB is lower than our methods and is thus less accurate in practice; in the meantime, DFT-fixed has coverage rate slightly higher than 90% and shows larger variance than our methods with adaptive cutoffs.

2.6 Discussion

This chapter offers a data-adaptive tool for conformalized survival analysis. By using covariate-dependent cutoff event to subset the data, i.e., considering data satisfying $C \geq f(X)$ for an appropriately chosen function f , our method enables higher power in a broader range of scenarios, where the distribution of $C \mid X$ can vary highly with X without creating overly conservative bounds, to improve on earlier work using a fixed cutoff, $C \geq c_0$ [Candès et al., 2021].

As in Candès et al. [2021], this work has been primarily focusing on the Type I censoring, where the censoring time for each individual is assumed observable; this is typically the case when the censoring time is the termination of a study. Another common type of censoring time is the loss-to-follow-up censoring. When the event is death, the loss-to-follow-up censoring time is not observed for patients who did not survive, and our method no longer applies. As discussed in Candès et al. [2021], our method can however provide informative LPBs beyond the setting of Type I censoring: when we have both the end-of-study censoring time C_{end} and the loss-to-follow-up censoring time C_{loss} , the censored survival time is then given by $\tilde{T} = T \wedge C_{\text{end}} \wedge C_{\text{loss}}$. Under the assumption that $(T, C_{\text{loss}}) \perp\!\!\!\perp C_{\text{end}} \mid X$, we can treat $T' := T \wedge C_{\text{loss}}$ as the true survival time and apply our procedure, producing an LPB on T' . We can thus alleviate the conservativeness caused by C_{end} , especially in studies with short duration.

We close the chapter by a discussion on extensions and interesting directions for future work. First, the theoretical guarantees shown in this work focus on constructing the PAC-type LPB. It can also be of interest to see if one can derive marginal guarantees for the proposed method, where the weighted conformal inference technique is not applicable. (Recent work by Angelopoulos et al. [2022] on a related problem suggest tools for converting a PAC-type bound to a finite-sample bound in expectation, and may be applicable to the survival

analysis setting as well.) Second, as with many double-robustness type results, our theoretical guarantees rely on high accuracy of our estimate of *either* the conditional distribution of $C \mid X$ or of $T \mid X$, but it may be possible to establish a better bound where moderately accurate estimates of *both* distributions contribute multiplicatively to a single unifying bound; this may be more relevant to practical settings, where we might expect moderate accuracy for each estimation problem. Finally, as discussed earlier, the cutoff introduces a variance-bias tradeoff—with a large cutoff, the observed survival time is closer to the true survival time but the effective sample size is reduced, and vice versa. It is interesting to quantitatively characterize this phenomenon, and derive an optimal choice of candidate LPBs based on this characterization.

CHAPTER 3

DISTRIBUTION-FREE CONDITIONAL PREDICTIVE INFERENCE

In this chapter, we address the problem of constructing distribution-free prediction intervals that remain *valid conditional on any feature value*. Earlier, in Section 1.1.1, we introduced the split-conformal prediction algorithm, an efficient method for constructing distribution-free prediction intervals for Y_{n+1} at a test feature X_{n+1} , given training data $(X_1, Y_1), \dots, (X_n, Y_n)$.

While the marginal coverage guarantee from Theorem 1.1 avoids strong distributional assumptions, it has limitations. The coverage probability is averaged over both the training data (the data points $(X_1, Y_1), \dots, (X_n, Y_n)$ used to fit the prediction interval $\hat{C}_n$) and the test point (X_{n+1}, Y_{n+1}) under the assumption that the latter follows the same distribution as the training sample. In applications like medicine i.e., predicting a patient’s future outcome, this means a nominal 90% coverage guarantee holds *on average* over new patients sampled from the same population but does not ensure coverage within specific subgroups (e.g., patients over age 50) or if the test population differs from the training sample (e.g., due to shifts in age distribution).

In this chapter, we construct a distribution-free prediction interval that addresses these shortcomings. The work presented here is co-authored with *Rina Foygel Barber*, which has been published in *Journal of the Royal Statistical Society Series B* under the title ‘Conformal prediction with local weights: randomization enables robust guarantees’ [Hore and Barber, 2024].

3.1 Introduction: test-conditional coverage

As motivated above, we seek assurance that the prediction interval $\hat{C}_n(X_{n+1})$ contains the test response Y_{n+1} with fairly high probability at all feature values X_{n+1} (i.e., among all patient demographics), rather than only on average. In other words, we would ideally like to be able to achieve the *test-conditional coverage* guarantee (sometimes also referred to as *local coverage*):

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid X_{n+1} \right\} \geq 1 - \alpha \text{ for all } P, \text{ almost surely over } X_{n+1}. \quad (3.1)$$

Compared to the marginal coverage guarantee (1.1), this stronger guarantee would address the issues of uneven coverage over different patient demographics, and of differences between the feature distribution within training sample versus the general population.

3.1.1 Hardness of test-conditional coverage

Unfortunately, in the setting of a nonatomic feature distribution (i.e., the distribution of X has no point masses, meaning, no value x has $\mathbb{P} \{X = x\} > 0$), this guarantee is known to be impossible, aside from trivial solutions such as returning a meaningless prediction interval $\hat{C}_n(X_{n+1}) \equiv \mathcal{Y}$.

Theorem 3.1 (adapted from [Vovk, 2012, Lei and Wasserman, 2014]). *Let $\mathcal{Y} = \mathbb{R}$, and suppose a prediction set $\hat{C}_n$ satisfies (3.1). Then, for any distribution P and for P_X -almost every x_0 , we have that if x_0 is not an atom of P_X , then*

$$\mathbb{E} \left[\text{leb} \left(\hat{C}_n(x_0) \right) \right] = \infty,$$

where leb denotes the Lebesgue measure supported on $\mathbb{R}$.

This means is that any conditionally valid prediction set will be extremely large, and non-informative.

Consequently, we are interested in finding a meaningful compromise between test-conditional coverage (3.1), which cannot be achieved in a distribution-free regime (aside from settings where X is discrete), and marginal coverage (1.1), which can be achieved distribution-free but is too weak for many practical applications. In particular, we might ask for coverage to hold when we condition, not on a specific value of X_{n+1} , but instead in a weaker sense, e.g., for any X_{n+1} within some neighbourhood—for instance, 90% coverage may not be guaranteed for patients who are exactly 25.75 years old, but perhaps could be guaranteed to hold on average over any 5-year age range.

3.1.2 Relaxed notions of conditional coverage

Since exact test-conditional coverage is unattainable in a distribution-free setting, we explore two alternative relaxations: (i) ensuring valid coverage over structured subsets of the feature space and (ii) maintaining coverage under controlled covariate shifts. We formalize these approaches below.

Set-conditional coverage

Let's consider a collection $\mathcal{B}$ of subsets of feature space, $B \subseteq \mathcal{X}$, such that we seek valid coverage guarantees if the test feature lies in B , for each $B \in \mathcal{B}$ (for instance, each B might be a subset of patients defined by a particular 5-year age range). This can be expressed as the following desired guarantee:

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid X_{n+1} \in B \right\} \geq 1 - \alpha \quad \text{for all } P \text{ and for all } B \in \mathcal{B}. \quad (3.2)$$

Different choices of the collection of subsets $\mathcal{B}$ represent different coverage goals for $\hat{C}_n(X_{n+1})$. For example, if we take $\mathcal{B} = \{\mathcal{X}\}$, then the coverage guarantee (3.2) is simply the marginal coverage guarantee (1.1). At the other extreme, if $\mathcal{B} = \{\{x\} : x \in \mathcal{X}\}$ contains all the singleton sets, then (3.2) becomes equivalent to the (impossible) pointwise test-conditional coverage guarantee (3.1) (and more broadly, the hardness result in Barber et al. [2021] states that it is impossible to achieve such a guarantee whenever the Vapnik–Chervonenkis (VC)-dimension of $\mathcal{B}$ is too large). Therefore, to capture a notion of coverage that lies in between the two, we are interested in collections $\mathcal{B}$ that lie somewhere in between these two extremes—for instance, if $\mathcal{X} = \mathbb{R}^d$, we might consider all open balls of a certain radius.

If our goal is to achieve test-conditional coverage relative to a *single* known subset $B \subseteq \mathcal{X}$, we can simply run split conformal prediction using only holdout points i with $X_i \in B$ to calibrate $\hat{q}$. But, ensuring valid coverage across multiple subsets (potentially infinite many) remains a non-trivial challenge.

Coverage with respect to covariate shifts.

A different relaxation of pointwise test-conditional coverage (3.1) frames the question as a covariate shift problem. If pointwise test-conditional coverage (3.1) were to hold, then this would imply that

$$\mathbb{P}_{P^n \times \tilde{P}} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} \geq 1 - \alpha \quad \text{for all } P, \tilde{P} \text{ with } P_{Y|X} = \tilde{P}_{Y|X}, \quad (3.3)$$

where the probability is taken with respect to training data $(X_1, Y_1), \dots, (X_n, Y_n)$ sampled i.i.d. from P , and test point (X_{n+1}, Y_{n+1}) sampled from $\tilde{P}$. This aim is again impossible (indeed, by taking $\tilde{P}_X$ to be a point mass at some point x , we see that the guarantee (3.3) would imply pointwise test-conditional coverage (3.1), which is impossible to attain). Thus,

we again consider a relaxed formulation:

$$\mathbb{P}_{P^n \times \tilde{P}} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} \geq 1 - \alpha \quad \text{for all } P, \tilde{P} \text{ with } P_{Y|X} = \tilde{P}_{Y|X} \text{ and } \tilde{P}_X \in \mathcal{P}, \quad (3.4)$$

for some class of allowed covariate shifts $\mathcal{P}$ —e.g., we might constrain $\tilde{P}_X$ to be near P_X in some metric, or, we might constrain $\tilde{P}_X$ to be sufficiently smooth relative to P_X in some sense.

As before, if our goal is to achieve coverage under covariate for a single known shifted distribution $\tilde{P}_X$, we can resort to the weighted conformal method in 1.1.1, but ensuring coverage across multiple covariate shifts remains an open challenge..

3.1.3 Defining the goals of relaxed conditional coverage

Building on our discussion of relaxed coverage notions, we now ask: how can we achieve coverage conditional on $X_{n+1} \in B$ for *any* B in a broad class of sets $\mathcal{B}$ (e.g., all balls of a certain radius), or ensure coverage under covariate shift for *any* $\tilde{P}_X$ within an allowed class of shifted distributions?

For these more challenging problems, a range of methods have been proposed in the literature, including work by Lei and Wasserman [2014], Cauchois et al. [2020a], Izbicki et al. [2022], Guan [2023], Gibbs et al. [2023]. Some of these methods offer theoretical results aiming towards relaxed guarantees similar to (3.2) (for test-conditional coverage) or to (3.4) (for covariate shift). We defer a more detailed discussion of this broad field to Section 3.6.1, where we will be able to compare more closely to our own results.

Most of these methods approximately attain conditional coverage empirically; however, combining sharp empirical performance with interpretable theoretical guarantees remains challenging. In our work, we aim to unify these questions, and bridge the gap between theoretical

validity and empirical performance for these problems. Our aim is to provide a $\hat{C}_n$ that simultaneously satisfies all of the following aims:

- (A1) **Marginal coverage:** $\hat{C}_n$ provably achieves exact finite-sample distribution-free marginal coverage, i.e., $\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} \geq 1 - \alpha$ for data drawn i.i.d. from any distribution P .
- (A2) **Coverage within subpopulations:** $\hat{C}_n$ provably achieves a meaningful relaxed test-conditional coverage guarantee, with $\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid X_{n+1} \in B \right\} \gtrsim 1 - \alpha$ holding over all sets B satisfying some regularity conditions that we will detail later on.
- (A3) **Coverage under covariate shifts:** $\hat{C}_n$ provably achieves approximate coverage guarantees for covariate shift problems, with $\mathbb{P}_{Pn \times \tilde{P}} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} \gtrsim 1 - \alpha$ holding over all P and all covariate-shifted distributions $\tilde{P} = \tilde{P}_X \times P_{Y|X}$, where the shift from P_X to $\tilde{P}_X$ satisfies some regularity conditions that we will detail later on.

Of course, all of these aims could trivially be met by always returning the uninformative interval $\hat{C}_n(X_{n+1}) \equiv \mathcal{Y}$ —so our goal is to provide a $\hat{C}_n$ that satisfies aims (A1)–(A3) without being too conservative, i.e., producing an interval that is as narrow as possible. We will see that randomly localized conformal prediction (RLCP) provides a tool that can allow us to achieve these aims without being overly conservative.

Organization of the chapter. The remainder of the chapter is organized as follows. We will begin by reviewing the literature on framework for locally weighted conformal prediction methods in Section 3.2, giving details on some existing methods and their shortcomings in order to gain motivation for our proposed method RLCP. Then in Section 3.3, we formally introduce RLCP, and in Section 3.4, we present theoretical results towards establishing that RLCP satisfies the desired properties (A1), (A2), (A3). In Section 3.5, we demonstrate the performance of RLCP (and compare it with these other locally weighted CP methods

described in Section 3.2) on both simulated and real data. Finally, we conclude with a discussion in Section 3.6, where we compare our method with existing literature and examine the role that randomization plays in our method. Some proofs and some details of implementing the methods are deferred to the Appendix B.

3.2 Conformal prediction with local weights: existing approaches

In this section, we will review the generalized framework of locally weighted conformal prediction from Guan [2023] where two locally adaptive conformal methods have been introduced. Our terminology here will differ somewhat from the terminology used in that reference, to help us introduce the new perspectives and frameworks in this chapter later on.

As for the split conformal method defined in Section 1.1.1, we begin by pretraining a score function $s = s(x, y)$, where large values indicate that (x, y) does not agree or “conform” with the trends observed in the pretraining set $\{(X_i^{\text{pre}}, Y_i^{\text{pre}})\}_{i=1, \dots, n_{\text{pre}}}$. Unlike the split conformal method, however, under a localized framework we will compute a local threshold for the score function, to produce a prediction interval of the form

$$\hat{C}_n(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}(X_{n+1})\}, \quad (3.5)$$

where $\hat{q}(x)$ may now vary with x , to reflect local differences in the distribution of the score. For example, for the regression score $s(x, y) = |y - \hat{f}(x)|$, if the fitted model $\hat{f}$ appears to be a poor fit at (or near) some particular x , we would want $\hat{q}(x)$ to be larger at (or near) this value x to ensure local predictive coverage.

The remaining question, then, is how to compute this local threshold $\hat{q}(x)$. Guan [2023]’s localized conformal prediction (LCP) framework addresses this question, as we will now describe.

3.2.1 An initial approach: baseLCP

The threshold $\hat{q}$ constructed by the split conformal method may not lead to local coverage at (or near) some particular x , because the distribution of the scores $s(X_i, Y_i)$ over the holdout set $i = 1, \dots, n$ may be quite different from the distribution of the score $s(X_{n+1}, Y_{n+1})$ if we condition on $X_{n+1} = x$ (or $X_{n+1} \approx x$). To address this, a natural approach is to compute a quantile $\hat{q}(x)$ that places more weight on holdout points X_i that lie near x . This method is discussed by Guan [2023], and formalized as follows. We first choose some *localization kernel* given by a function

$$H : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0},$$

for instance, for $\mathcal{X} = \mathbb{R}^d$, common choices include a Gaussian kernel (with bandwidth h) given by

$$H(x, x') = \exp \left\{ -\|x - x'\|_2^2 / 2h^2 \right\},$$

or a box kernel (again with bandwidth h), given by

$$H(x, x') = \mathbb{1} \{ \|x - x'\|_2 \leq h \}.$$

The local quantile is then computed as

$$\hat{q}_{1-\alpha}(X_{n+1}) = \text{Quantile}_{1-\alpha} \left(\sum_{i=1}^n w_i \delta_{s(X_i, Y_i)} + w_{n+1} \delta_{+\infty} \right), \quad (3.6)$$

where δ_s is the point mass at s , and where the weights are given by

$$w_i = \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})}, \quad i \in [n+1].$$

(Note that $\hat{q}_{1-\alpha}(X_{n+1})$ depends on the training data $\{(X_i, Y_i)\}_{i \in [n]}$ as well as the test point X_{n+1} , but this dependence is not made explicit in the notation.)

Finally, the resulting prediction interval is defined as in (3.5), i.e.,

$$\hat{C}_n^{\text{baseLCP}}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}_{1-\alpha}(X_{n+1})\}.$$

Observe that, if we choose the trivial localization kernel $H(x, x') \equiv 1$ (i.e., no localization), then we would have $\hat{q}_{1-\alpha}(x)$ constant over x , and we recover the original split conformal method (1.2).

For the Gaussian kernel and box kernel described above (as well as other natural choice of H), the choice of bandwidth h controls the extent of localization. At a high level, for larger values of h the method behaves similarly to split conformal, while for smaller h the method becomes more localized, and can offer better empirical local coverage. However, if we choose h too small, performance will suffer due to low effective sample size—the quantile $\hat{q}(x)$ may even be estimated as $+\infty$. We will return later on to the question of choosing a bandwidth h that appropriately addresses this tradeoff.

This approach is very intuitive: to (approximately) achieve local coverage at x , we simply give more weight to holdout points for which $X_i \approx x$. Interestingly, however, this method may fail in practice: even marginal coverage may fail; empirically, the marginal coverage may be either higher or lower than $1 - \alpha$, and Guan [2023, Propositions 2.4, 2.5] show theoretically that marginal coverage may be arbitrarily high or arbitrarily low for this method. Consequently, we will refer to this method as the “base” version of localized conformal prediction, or “baseLCP” for short, to indicate that it forms a baseline that other methods will build upon. To correct for this issue in the marginal coverage level, Guan [2023] proposes an adjusted version of the method to address the problems of baseLCP, which we will present next.

3.2.2 Calibrating to recover marginal coverage: calLCP

Now we describe Guan [2023]’s proposed method—this method is referred to as localized conformal prediction in that paper but we will refer to it as “calibrated” localized conformal prediction, or calLCP for short, to distinguish it from baseLCP and from our own method to be defined later on. The idea is to recover marginal coverage by changing the definition of the local threshold $\hat{q}(x)$.

Recall that baseLCP uses the threshold $\hat{q}_{1-\alpha}(X_{n+1})$ for the score $s(X_{n+1}, y)$, to determine whether this value y belongs in the prediction interval—but as we have discussed above, this might lead to actual coverage being below or above the target level $1 - \alpha$. The idea of calLCP is to replace the level $1 - \alpha$ in the quantile with a different value, to recalibrate to the target coverage level.

Specifically, Guan [2023] defines the prediction interval as

$$\hat{C}_n^{\text{calLCP}}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}_{1-\tilde{\alpha}(X_{n+1}, y)}(X_{n+1}) \right\},$$

for

$$\hat{q}_{1-\tilde{\alpha}(X_{n+1}, y)}(X_{n+1}) = \text{Quantile}_{1-\tilde{\alpha}(X_{n+1}, y)} \left(\sum_{i=1}^n w_i \delta_{s(X_i, Y_i)} + w_{n+1} \delta_{s(X_{n+1}, y)} \right),$$

where the adjusted quantile value $1 - \tilde{\alpha}(X_{n+1}, y)$ is chosen ensure marginal coverage, as follows. To define this value, we introduce notation

$$Y_i^y = \begin{cases} Y_i, & i \in [n], \\ y, & i = n + 1, \end{cases}$$

i.e., $(Y_1^y, \dots, Y_{n+1}^y)$ is equal to the vector containing the n training data response values

$Y_1, \dots, Y_n$, augmented with the value y in the last argument. Then $\tilde{\alpha}(X_{n+1}, y)$ is defined as

$$\max \left\{ a \in \Gamma(w) : \sum_{i=1}^{n+1} \mathbb{1} \left\{ s(X_i, Y_i^y) \leq \text{Quantile}_{1-a} \left(\sum_{j=1}^{n+1} w_{i,j} \delta_{s(X_i, Y_i^y)} \right) \right\} \geq (1 - \alpha)(n + 1) \right\}, \quad (3.7)$$

where, overloading notation, we let

$$w_{i,j} = \frac{H(X_j, X_i)}{\sum_{k=1}^{n+1} H(X_k, X_i)}, \quad i, j \in [n + 1]$$

(and note that $w_{n+1,i}$ coincides with the weight w_i defined above), and where

$$\Gamma(w) = \left\{ \sum_{j \in I} w_{i,j} : i \in [n + 1], \quad I \subseteq [n + 1] \right\} \quad (3.8)$$

is the set of possible sums of weights that might occur in the calculation of the discrete CDF inside the quantile.

Examining this definition, it appears that implementing callLCP requires computing a recalibrated threshold $\tilde{\alpha}(X_{n+1}, y)$ for every value $y \in \mathcal{Y}$ (i.e., a similar computational challenge to that faced by full conformal prediction [Vovk et al., 2005]). However, one can bypasses this issue of computational complexity via an efficient implementation proposed in Guan [2023, Section 3.2].

With this recalibration step, Guan [2023, Theorem 2.1] establishes that marginal coverage at level $1 - \alpha$ holds for the callLCP method (unlike for baseLCP). Empirically, the method also shows improvement in local coverage, but there are no known distribution-free theoretical results to guarantee this (however, under additional smoothness assumptions, a conditional coverage result is established in Guan [2023, Theorem 5.2]). Interestingly, the marginal coverage result can be recovered via a new insight: we will show in Section 3.6.3 that Guan

[2023]’s calLCP method is, in fact, an instance of full conformal prediction—but implemented with a modified score function, not with the original pretrained score function $s = s(x, y)$.

3.2.3 Another look at baseLCP: achieving a different coverage guarantee

In this section, we will talk about an alternative formulation of baseLCP in order to get a better understanding of the properties of this method, and why marginal coverage can fail. (This in turn will help motivate the construction of RLCP, which we will present next.)

Consider a new feature variable $\tilde{X}_{n+1}$, generated as $\tilde{X}_{n+1}|X_{n+1} \sim P_X \circ H(\cdot, X_{n+1})$, where for any function $g : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ with $0 < \mathbb{E}_{P_X} [g(X)] < \infty$, we define $P_X \circ g$ as the distribution P_X reweighted by g , i.e.,

$$(P_X \circ g)(A) = \frac{\int_A g(x) \, dP_X(x)}{\int_{\mathcal{X}} g(x) \, dP_X(x)} \text{ for all } A \subseteq \mathcal{X}. \quad (3.9)$$

If the localization kernel H has a reasonably small bandwidth, then the reweighted distribution $P_X \circ H(\cdot, X_{n+1})$ places most of its mass near X_{n+1} . As a result, we can interpret $\tilde{X}_{n+1}$ as a sort of “synthetic sample” designed to be similar to the test feature X_{n+1} —in a setting where each data point is a patient, we are generating new patient data $\tilde{X}_{n+1}$ for a patient that is *similar* (in feature space) to the test patient for whom we want to perform prediction.

Now we return to the baseLCP method. How does constructing the “synthetic sample” $\tilde{X}_{n+1}$ relate to this procedure? It turns out that baseLCP is actually related to running weighted conformal prediction (introduced in Section 1.1.1) on the covariate shift problem defined by

taking $\tilde{P}_X = P_X \circ H(\cdot, X_{n+1})$. Specifically, if we define

$$\hat{C}_n^{\text{WCP}}(X_{n+1}, \tilde{X}_{n+1}) = \left\{ y \in \mathcal{Y} : s(\tilde{X}_{n+1}, y) \leq \hat{q}_{1-\alpha}(X_{n+1}) \right\},$$

then this prediction interval will have probability $\geq 1 - \alpha$ of covering a new response value $\tilde{Y}_{n+1}$, drawn from the conditional distribution $P_{Y|X}$ at $X = \tilde{X}_{n+1}$ (as established by Guan [2023, Theorem 5.3]). An equivalent way of thinking about this is that the threshold $\hat{q}_{1-\alpha}(X_{n+1})$ offers a $1 - \alpha$ probability of bounding the score for a synthetic test point, $s(\tilde{X}_{n+1}, \tilde{Y}_{n+1})$ —but not necessarily the original score $s(X_{n+1}, Y_{n+1})$ itself for the actual test point of interest.

This interpretation leads us towards the RLCP method. In baseLCP, using the kernel $H(\cdot, X_{n+1})$ centered at the true test point X_{n+1} , leads to a threshold $\hat{q}_{1-\alpha}(X_{n+1})$ that provides coverage for a synthetic sample $\tilde{X}_{n+1}$ —that is, for data points that are *similar* to X_{n+1} , but not necessarily for X_{n+1} itself. For RLCP, then, we will instead use a kernel that is *not* centered at X_{n+1} , but will aim to achieve coverage for X_{n+1} itself.

3.3 Randomly-localized conformal prediction

From our discussion above, we saw that baseLCP constructs the threshold $\hat{q}_{1-\alpha}(X_{n+1})$ with weights defined by a kernel centered at X_{n+1} , which then leads to coverage for a “synthetic sample” $\tilde{X}_{n+1}$ that is sampled *near* X_{n+1} . To define randomly-localized conformal prediction (RLCP), we will reverse this construction: we will define the threshold using weights centered at a random $\tilde{X}_{n+1}$ sampled near X_{n+1} , and will see later on that this leads to marginal coverage (and also approximate local coverage) at X_{n+1} .

We now define the method. We again begin with a function $H : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$, with an added assumption that, for each x , $H(x, \cdot)$ defines a density with respect to some base

measure ν on $\mathcal{X}$:

$$\int_{\mathcal{X}} H(x, x') \, d\nu(x') = 1 \text{ for all } x \in \mathcal{X}. \quad (3.10)$$

Returning to the earlier examples of the Gaussian kernel and the box kernel in the setting $\mathcal{X} = \mathbb{R}^d$, here we can use the same kernels up to a normalizing constant (and take ν to be Lebesgue measure): for the Gaussian kernel with bandwidth h , we take

$$H(x, x') = \frac{1}{(2\pi h^2)^{d/2}} \exp\left(-\frac{\|x - x'\|_2^2}{2h^2}\right), \quad (3.11)$$

while for the box kernel with bandwidth h , we take

$$H(x, x') = \frac{1}{V_d h^d} \mathbb{1}\{\|x - x'\|_2 \leq h\}, \quad (3.12)$$

where V_d is the volume of the unit ball in $\mathbb{R}^d$. Although both of these choices for H happen to be symmetric in the two arguments, we do not assume this condition.

With this function selected, we now define the RLCP method: given the pretrained score $s = s(x, y)$, the holdout data $(X_1, Y_1), \dots, (X_n, Y_n)$, and the test point X_{n+1} ,

Sample $\tilde{X}_{n+1}$ from density $H(X_{n+1}, \cdot)$,

$$\text{then return } \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})\}, \quad (3.13)$$

where analogously to before, the threshold $\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$ is computed as

$$\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^n \tilde{w}_i \delta_{s(X_i, Y_i)} + \tilde{w}_{n+1} \delta_{+\infty}\right),$$

for weights

$$\tilde{w}_i = \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})}, \quad i \in [n+1].$$

(As before, $\hat{q}(X_{n+1}, \tilde{X}_{n+1})$ depends also on the training data $\{(X_i, Y_i)\}_{i \in [n]}$, but this dependence is not explicit in the notation.)

The only difference between RLCP and baseLCP, then, is that the weights are given by the $\tilde{w}_i$'s, computed with the kernel H centered at $\tilde{X}_{n+1}$, rather than the original w_i 's, which were computed with H centered at X_{n+1} .

3.3.1 The role of a synthetic sample

Before proceeding, we pause to discuss $\tilde{X}_{n+1}$ a bit more. First, observe that the randomness of $\tilde{X}_{n+1}$ means that the resulting prediction interval is no longer a deterministic function of the data—it is randomized because the threshold $\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$ is computed with the kernel H centered at a random point $\tilde{X}_{n+1}$ (hence the term “randomly-localized” in the name of the method). To compare again with baseLCP,

- For baseLCP, the weights used for computing the threshold $\hat{q}_{1-\alpha}(X_{n+1})$ are calculated by centering the kernel H at X_{n+1} , and can lead to coverage for a “synthetic sample” $\tilde{X}_{n+1}$, drawn from $P_X \circ H(\cdot, X_{n+1})$. In other words, we can think of the original test point X_{n+1} as a prototype: coverage will hold relative to the reweighted feature distribution $P_X \circ H(\cdot, X_{n+1})$, which generates new “synthetic samples” $\tilde{X}_{n+1}$ which are potentially very similar to this prototype—but coverage may fail for X_{n+1} itself, i.e., marginal coverage may fail to hold, as observed by Guan [2023].
- For RLCP, the weights used for computing the threshold $\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$ are calculated by centering the kernel H at a random point $\tilde{X}_{n+1}$ that was drawn near X_{n+1} , and (as we will see below) will lead to coverage for X_{n+1} itself. We can now think of $\tilde{X}_{n+1}$ as a “synthetic prototype”, and conditionally, the test sample X_{n+1} is likely to lie near to this prototype $\tilde{X}_{n+1}$. More concretely, we will see that the conditional distribution of X_{n+1} given $\tilde{X}_{n+1}$ is equal to the reweighted feature distribution

$P_X \circ H(\cdot, \tilde{X}_{n+1})$, which (for a strongly localizing kernel H) places most of its mass near the “synthetic prototype” $\tilde{X}_{n+1}$.

We observe that, for RLCP, the “synthetic prototype” $\tilde{X}_{n+1}$ may *not* be a plausible data value—for example, the feature distribution P_X might be supported on integer values, but we might still choose to use a Gaussian kernel for H , leading to noninteger values of $\tilde{X}_{n+1}$ when it is sampled from density $H(X_{n+1}, \cdot)$. In contrast, for baseLCP, $\tilde{X}_{n+1}$ is drawn from $P_X \circ H(\cdot, X_{n+1})$, whose support is equal to (or is a subset of) the support of P_X , and so we think of $\tilde{X}_{n+1}$ as a “synthetic sample”—a new data point (e.g., a new patient) from the same population, that was drawn in such a way that it is similar to X_{n+1} .

3.4 Theoretical guarantees

In this section, we present theoretical coverage guarantees for randomly-localized conformal prediction (RLCP). We will see that these guarantees verify that RLCP achieves the aims outlined earlier in **(A1)**–**(A3)**.

We will begin in Section 3.4.1 by establishing a key property of the RLCP method in Proposition 3.2 that underlies in all its coverage guarantees. With this result in place, we then present the marginal coverage guarantee (aim **(A1)**). Approximate test-conditional coverage (aim **(A2)**) and robustness to covariate shift (aim **(A3)**) will be addressed in Sections 3.4.2 and 3.4.3, respectively, where we will also explore the relationship between these two problems. RLCP also enjoys a training-conditional coverage guarantee, but since this result is of a different flavor, we defer it to Appendix B.1.2.

3.4.1 Marginal coverage and a key property

We begin by defining some notation. In the field of distribution-free prediction, test-conditional coverage is often quantified by studying the conditional miscoverage rate

$$\alpha(X_{n+1}) = \mathbb{P} \left\{ Y_{n+1} \notin \hat{C}_n(X_{n+1}) \mid X_{n+1} \right\},$$

where $\hat{C}_n$ is the prediction interval produced by the procedure of interest. In other words, $\alpha(X_{n+1})$ is the miscoverage rate conditional on X_{n+1} , but marginalizing over the random draw of the training data (and the draw of Y_{n+1}). To relate this random variable to our earlier terminology and notation, we can observe that the marginal coverage guarantee (1.1) is then equivalent to requiring $\mathbb{E}[\alpha(X_{n+1})] \leq \alpha$, while the pointwise test-conditional coverage guarantee (3.1) would require $\alpha(X_{n+1}) \leq \alpha$ almost surely.

Now we extend this idea to RLCP. For our method, after marginalizing over the randomness in the training data, we have an additional remaining random variable $\tilde{X}_{n+1}$ in addition to the test feature X_{n+1} . We then define

$$\alpha(X_{n+1}, \tilde{X}_{n+1}) = \mathbb{P} \left\{ Y_{n+1} \notin \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1}, \tilde{X}_{n+1} \right\}.$$

With this notation in place, the following result establishes a key property of the RLCP method. For this proposition, and all subsequent results, we assume that the RLCP method is implemented with a localization kernel H that satisfies the condition (3.10), i.e., $H(x, \cdot)$ is a density (with respect to some fixed base measure) for each x .

Proposition 3.2 (Key property of RLCP). *Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$ for any*

distribution P . Then the RLCP method defined in (3.13) satisfies

$$\mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \right] \leq \alpha \text{ almost surely.}$$

This new quantity $\alpha(X_{n+1}, \tilde{X}_{n+1})$ can again be related to marginal coverage by taking an expected value, and thus Proposition 3.2 immediately yields a marginal coverage guarantee.

Theorem 3.3 (Marginal coverage guarantee). *The RLCP method defined in (3.13) satisfies the marginal coverage property (1.1), i.e., for any distribution P , if $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$, then*

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha.$$

Proof of Theorem 3.3. Using the tower law, we calculate

$$\begin{aligned} \mathbb{P} \left\{ Y_{n+1} \notin \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} &= \mathbb{E} \left[\mathbb{P} \left\{ Y_{n+1} \notin \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1}, \tilde{X}_{n+1} \right\} \right] \\ &= \mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \right] = \mathbb{E} \left[\mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \right] \right] \leq \mathbb{E}[\alpha] = \alpha, \end{aligned}$$

where the inequality applies Proposition 3.2. $\square$

In other words, this theorem follows immediately from Proposition 3.2 by observing that the new notation offers an equivalent formulation of marginal coverage (1.1),

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha \iff \mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \right] \leq \alpha.$$

We remark that we can similarly obtain an equivalent formulation of pointwise test-conditional coverage (3.1) since the test-conditional coverage $\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \right\}$ can be equivalently expressed as $\mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \right]$. However, this last equivalence does not lend itself to an immediate application of Proposition 3.2, since we do not have a

bound on $\mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \right]$ (because the proposition conditions on $\tilde{X}_{n+1}$ rather than on X_{n+1}). Instead, we will need to take a different approach for the test-conditional coverage problem, as we will see in the next section. Before proceeding, however, we first prove the proposition, to examine how the construction of RLCP (and specifically, the manner in which $\tilde{X}_{n+1}$ is drawn) allows for this key property to hold.

Proof of Proposition 3.2. First, we compute the distribution of the data conditional on $\tilde{X}_{n+1}$. Since the data points are i.i.d., and so the training data is independent of $\tilde{X}_{n+1}$ by construction, the training data has the same distribution as before, $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$. For the test point, however, the distribution has changed: by construction, the joint distribution of $(X_{n+1}, Y_{n+1}, \tilde{X}_{n+1})$ is given by

$$\begin{cases} X_{n+1} \sim P_X, \\ Y_{n+1} \mid X_{n+1} \sim P_{Y|X}(\cdot \mid X_{n+1}), \\ \tilde{X}_{n+1} \mid (X_{n+1}, Y_{n+1}) \sim H(X_{n+1}, \cdot). \end{cases}$$

A direct calculation then yields

$$(X_{n+1}, Y_{n+1}) \mid \tilde{X}_{n+1} \sim (P_X \circ H(\cdot, \tilde{X}_{n+1})) \times P_{Y|X}. \quad (3.14)$$

We can see, then, that this is an instance of a covariate shift problem: the distribution of $X_{n+1} \mid \tilde{X}_{n+1}$ is no longer equal to P_X , but the conditional distribution of Y_{n+1} remains the same as for the training data, i.e., $P_{Y|X}$.

Now we return to the RLCP method. The quantile $\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$ is calculated using a reweighted empirical distribution of the scores. Examining the definition of these weights $\tilde{w}_i$ (defined in Section 3.3), and recalling the weighted conformal prediction method introduced in Section 1.1.1, we can see that, conditional on $\tilde{X}_{n+1}$, the RLCP procedure is equivalent to

running weighted conformal prediction (WCP) for this covariate shift problem—i.e., when the shifted covariate distribution is given by $\tilde{P}_X = P_X \circ H(\cdot, \tilde{X}_{n+1})$. The desired coverage result then follows as a direct application of Theorem 1.2, which guarantees coverage at level $1 - \alpha$ when weighted conformal prediction is applied to a setting with a (known) covariate shift. $\square$

3.4.2 Test-conditional coverage

Next, we shift our attention to relaxed test-conditional coverage guarantees of the form outlined in Aim (A2). Our goal is to establish that, for a broad class of “nice” sets $B \subseteq \mathcal{X}$, the RLCP prediction interval offers conditional coverage at level $\gtrsim 1 - \alpha$. To state this result, we first need to introduce some more notation. First, we write $P_{(X, \tilde{X})}$ to denote the distribution on $(X, \tilde{X}) \in \mathcal{X} \times \mathcal{X}$ defined by sampling

$$X \sim P_X, \tilde{X} \mid X \sim H(X, \cdot).$$

Next, for a set $B \subseteq \mathcal{X}$ that we will condition on, for any $r > 0$ define

$$\text{bd}_r(B) = \left\{ x \in B : \inf_{x' \in B^c} \|x - x'\| \leq r \right\},$$

i.e., any $x \in B$ that lies within distance r of the complement, B^c .¹ We can think of $\text{bd}_r(B)$ as an inflation of the boundary of the set B .

Theorem 3.4 (Test-conditional coverage guarantee). *The RLCP method defined in (3.13) satisfies the following property: for any distribution P , if $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$,*

1. Implicitly, we are assuming that $\mathcal{X}$ is a metric space, and that any $B \subseteq \mathcal{X}$ we condition on is a measurable set. The norm $\|\cdot\|$ may be any norm on $\mathcal{X}$.

then for every $B \subseteq \mathcal{X}$, it holds that

$$\begin{aligned} \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \in B \right\} \\ \geq 1 - \alpha - \frac{\inf_{\epsilon > 0} \left\{ P_{(X, \tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} + P_X(\text{bd}_{2\epsilon}(B)) \right\}}{P_X(B)}. \end{aligned}$$

Let us pause to unpack this result and examine the terms appearing in the bound. We would like to determine for what sets B , and for what choice of the localization kernel H , the excess error bound

$$\frac{P_{(X, \tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} + P_X(\text{bd}_{2\epsilon}(B))}{P_X(B)}$$

is likely to be small (for some ϵ). First, in the denominator, the term $P_X(B)$ is the mass placed by the feature distribution P_X on the set B , and hence can be thought of as an proxy for effective sample size. A low value of the denominator means that our dataset likely contains very few data points (X_i, Y_i) with $X_i \in B$, and hence ensuring coverage conditional on this set is not feasible.

Now we turn to the numerator. For the first term, $P_{(X, \tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \}$, this term is small as long as ϵ is chosen to be sufficiently large relative to the bandwidth of the localization kernel H —for example, when H is a box kernel, this term is zero if we choose $\epsilon = h$. The second term in the numerator essentially measures regularity properties of the set B . For instance, if P_X has a bounded density, then $P_X(\text{bd}_{2\epsilon}(B))$ scales with the volume of the set $\text{bd}_{2\epsilon}(B)$, which is a 2ϵ -inflated boundary of the set B and will have relatively low volume for many classes of “nice” sets B —as we will see next for a specific example, namely, by taking B to be a ball in the ℓ_2 norm.

Example: coverage conditional on a ball

To make the results of Theorem 3.4 more concrete, let us see what these results imply for a specific setting: assuming $\mathcal{X} = \mathbb{R}^d$, we will take $B = \mathbb{B}(x_0, r) = \{x \in \mathbb{R}^d : \|x - x_0\|_2 \leq r\}$, the ball of radius r centered at some point x_0 . We will also take H to be the normalized box kernel with bandwidth h , $H(x, x') = \frac{1}{V_d h^d} \mathbb{1}_{\{\|x - x'\|_2 \leq h\}}$, with $2h \leq r$. Applying Theorem 3.4 to this setting (with $\varepsilon = h$) yields the guarantee

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \in \mathbb{B}(x_0; r) \right\} \geq 1 - \alpha - \frac{P_X(\mathbb{B}(x_0; r) \setminus \mathbb{B}(x_0; r - 2h))}{P_X(\mathbb{B}(x_0; r))},$$

since we have $\text{bd}_{2\epsilon}(B) = \mathbb{B}(x_0; r) \setminus \mathbb{B}(x_0; r - 2h)$ for our choice of set B .

In particular, if P_X is a continuous distribution, then the ratio appearing in the coverage guarantee will be small—if we assume P_X has density f_X with respect to Lebesgue measure, for instance, then

$$\begin{aligned} \frac{P_X(\mathbb{B}(x_0; r) \setminus \mathbb{B}(x_0; r - 2h))}{P_X(\mathbb{B}(x_0; r))} &\leq \frac{\sup_{x \in \mathbb{B}(x_0; r)} f_X(x) \cdot \text{Leb}(\mathbb{B}(x_0; r) \setminus \mathbb{B}(x_0; r - 2h))}{\inf_{x \in \mathbb{B}(x_0; r)} f_X(x) \cdot \text{Leb}(\mathbb{B}(x_0; r))} \\ &= \frac{\sup_{x \in \mathbb{B}(x_0; r)} f_X(x)}{\inf_{x \in \mathbb{B}(x_0; r)} f_X(x)} \cdot \left[1 - \left(1 - \frac{2h}{r} \right)^d \right]. \end{aligned}$$

An asymptotic result. While the goal of our work is to provide finite-sample theoretical results, for intuition it is useful to consider how this guarantee behaves asymptotically. Suppose that the distribution P is fixed (in particular, dimension d is taken to be constant), and P_X has density f_X as above, but we take sample size $n \rightarrow \infty$. At each n , we will run RLCP with a box kernel of bandwidth h_n , and will examine coverage with respect to a ball of radius $r_n \geq 2h_n$. Rewriting the above calculation into this new asymptotic notation, we

have

$$\begin{aligned} \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \in \mathbb{B}(x_0; r_n) \right\} \\ \geq 1 - \underbrace{\alpha - \frac{\sup_{x \in \mathbb{B}(x_0; r_n)} f_X(x)}{\inf_{x \in \mathbb{B}(x_0; r_n)} f_X(x)}}_{\text{Term 1}} \cdot \underbrace{\left[1 - \left(1 - \frac{2h_n}{r_n} \right)^d \right]}_{\text{Term 2}}. \end{aligned}$$

Term 1 has to do with the regularity of the distribution P_X . In particular, if $f_X(x)$ is continuous and positive at $x = x_0$, then $r_n \rightarrow 0$ implies that Term 1 converges to 1 as $n \rightarrow \infty$. Term 2 has to do with the set B that we are conditioning on—we need our localization (controlled by h_n) to be sufficiently strong for ensuring coverage conditional on $X_{n+1} \in B$ (and this event’s probability is controlled by r_n). In particular, if we choose h_n, r_n such that $h_n/r_n \rightarrow 0$, then Term 2 is vanishing as $n \rightarrow \infty$. To summarize, we have seen that in this setting, the following asymptotic result applies:

Corollary 3.5. *Under the notation and assumptions defined above, if $r_n \rightarrow 0$ and $h_n/r_n \rightarrow 0$, then*

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \in \mathbb{B}(x_0; r_n) \right\} \geq 1 - \alpha - o(1).$$

We remark that under some additional smoothness assumptions, we can also derive asymptotic test conditional coverage of the form (3.1) under this setting—that is, pointwise coverage, which conditions on $X_{n+1} = x_0$, rather than $X_{n+1} \approx x_0$ as in the result above. We defer the theorem and its proof to the Appendix B.1.3.

The role of the bandwidth

The coverage guarantee in Theorem 3.4 is valid irrespective of how we choose the kernel H —in particular, we are free to choose any bandwidth h . However, as mentioned earlier

in Section 3.2.1, we need to choose the bandwidth carefully in order for the guarantee, and the prediction interval itself, to be meaningful. If h is too large, then the method is not very localized—and in particular, the numerator in the excess error term appearing in Theorem 3.4 might then be arbitrarily large (since we would need to choose a very large value of ε in order for $P_{(X, \tilde{X})}\{\|X - \tilde{X}\| > \varepsilon\}$ to be small). If instead we choose h to be too small, on the other hand, the excess error term in Theorem 3.4 will generally be very small, but this comes at the cost of an overly wide (perhaps even infinitely wide) prediction interval.

To make this more concrete, suppose H is the box kernel of bandwidth h . When constructing the RLCP prediction set, we need a large effective sample size—that is, a large number of training points $i \in [n]$ such that $\|X_i - \tilde{X}_{n+1}\| \leq h$. At the extreme, if *none* of the training points $i \in [n]$ satisfy this bound, then the quantile $\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$ will be $+\infty$.

Now consider an asymptotic regime (as discussed above, for coverage conditional on a ball)—suppose that at sample size n , we take H to be a box kernel with bandwidth h_n , where our features lie in $\mathcal{X} = \mathbb{R}^d$. Under regularity conditions on P_X , the mass of the ball $\mathbb{B}(\tilde{X}_{n+1}, h_n)$ will typically scale as $n(h_n)^d$ —i.e., this is the effective sample size after localization. Hence, in an asymptotic setting, to ensure that RLCP produces meaningful intervals we should choose the bandwidth h_n such that $h_n \rightarrow 0$ but $h_n \cdot n^{1/d} \rightarrow \infty$.

3.4.3 Coverage against arbitrary covariate shift

As have been stated before, test-conditional coverage (or approximations of test-conditional coverage, as in Theorem 3.4) can also be viewed as a special case of coverage against covariate shift settings. To consider a general setting, recall that, for any reweighting function $g : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$, with $\mathbb{E}_{P_X}[g(X)]$ positive and finite, the reweighted covariate distribution $P_X \circ g$ is defined as in (3.9). We will consider a setting where training data is drawn from $P =$

$P_X \times P_{Y|X}$, while the test point distribution (i.e., the target distribution for which we would like to have predictive coverage) is $(P_X \circ g) \times P_{Y|X}$. For example, in a setting where each data point represents a patient and the feature space $\mathcal{X}$ includes information about age, g might be a monotone increasing function of age, indicating that older patients comprise a *larger* fraction of the general population as compared to the training sample; equivalently, older patients are *underrepresented* in the training sample.

Our goal, as described earlier in Aim **(A3)**, is to provide a predictive interval that offers an approximate coverage guarantee with respect to a large class of “nice” reweighting functions g . Leveraging our new notation, we would like to prove a bound of the form

$$\mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \right] \lesssim \alpha,$$

or equivalently,

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \gtrsim 1 - \alpha,$$

where expectation and probability are taken with respect to the following data distribution: the training points $(X_i, Y_i) \stackrel{\text{iid}}{\sim} P$, the test point $(X_{n+1}, Y_{n+1}) \sim (P_X \circ g) \times P_{Y|X}$, and then by construction of the RLCP method, also $\tilde{X}_{n+1} \mid X_{n+1} \sim H(X_{n+1}, \cdot)$.

In this section, we will work towards a general result of this type for a broad class of reweighting functions $g : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$. We will then see that the test-conditional coverage result in Theorem 3.4 can be derived as a special case.

To present the result, we need to define a parameter that captures the variability of the reweighting function g : for any $r > 0$ and any $A \subseteq \mathcal{X}$, define

$$L_{g,r,A} = \sup_{\substack{x, x' \in A \\ \|x - x'\| \leq r}} \frac{|g(x) - g(x')|}{r}.$$

We can essentially interpret $L_{g,r,A}$ as a relaxed version of the Lipschitz constant of g on the set A . In particular, if g is L_g -Lipschitz, then $L_{g,r,A} \leq L_g$ for all r and all A . We will also write $\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)| = \sup_{x \in \mathcal{X}} g(x)$.

We are now ready to state our main result.

Theorem 3.6 (Robustness to covariate shift). *The RLCP method defined in (3.13) satisfies the following property: for any distribution P , and every $g : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ with $P_X \circ g$ well-defined (i.e., with $0 < \mathbb{E}_{P_X} [g(X)] < \infty$), it holds that*

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha - \frac{\inf_{A \subseteq \mathcal{X}, \epsilon > 0} \left\{ \epsilon \cdot L_{g,2\epsilon,A} + \|g\|_\infty \cdot P_{(X,\tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} + \|g\|_\infty \cdot P_X(A^c) \right\}}{\mathbb{E}_{P_X} [g(X)]},$$

where the probability on the left-hand side is taken with respect to $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$ and $(X_{n+1}, Y_{n+1}) \sim (P_X \circ g) \times P_{Y|X}$, and $\tilde{X}_{n+1} \mid X_{n+1} \sim H(X_{n+1}, \cdot)$ as defined by the RLCP method (3.13).

In general, computing the infimum in this upper bound, over all $A \subseteq \mathcal{X}$ and $\epsilon > 0$, might be highly nontrivial. However, in some special cases, we can choose A strategically to get an interpretable upper bound on the miscoverage. These special cases include:

- **Lipschitz functions:** if g is L_g -Lipschitz, then by choosing $A = \mathcal{X}$, the result of the theorem yields

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha - \frac{\inf_{\epsilon > 0} \left\{ \epsilon \cdot L_g + \|g\|_\infty \cdot P_{(X,\tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} \right\}}{\mathbb{E}_{P_X} [g(X)]}. \quad (3.15)$$

- **Indicator functions:** suppose $g(x) = \mathbb{1}\{x \in B\}$ is the indicator function for some set $B \subseteq \mathcal{X}$. In this case, we can see that requiring coverage with respect to covariate shift (with $P_X \circ g$ as the shifted distribution) is exactly equivalent to requiring coverage conditional on $X_{n+1} \in B$. Consequently, by considering indicator functions, this general result leads to an immediate proof of Theorem 3.4.

Proof of Theorem 3.4. For any $B \subseteq \mathcal{X}$, define $g(x) = \mathbb{1}\{x \in B\}$. Define the set $A = (\text{bd}_{2\epsilon}(B))^c \subseteq \mathcal{X}$. Observe that for any $x, x' \in A$, if $\|x - x'\| \leq 2\epsilon$, we cannot have $x \in B$ and $x' \in B^c$ (because this would imply $x \in \text{bd}_{2\epsilon}(B)$, which is a contradiction); we must either have $x, x' \in B$ or $x, x' \in B^c$, and in particular this means $g(x) = g(x')$. Thus, we have $L_{g, 2\epsilon, A} = 0$, and so Theorem 3.6 yields

$$\mathbb{P}\left\{Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \in B\right\} \geq 1 - \alpha - \frac{\inf_{\epsilon > 0} \left\{ \|g\|_\infty \cdot P_{(X, \tilde{X})} \{\|X - \tilde{X}\| > \epsilon\} + \|g\|_\infty \cdot P_X(\text{bd}_{2\epsilon}(B)) \right\}}{\mathbb{E}_{P_X}[g(X)]}.$$

Finally, by definition of g , we calculate $\|g\|_\infty = 1$ and $\mathbb{E}_{P_X}[g(X)] = P_X(B)$, which yields the desired bound. $\square$

More generally, we can interpret the various terms in Theorem 3.6's excess error bound as follows. To make the first two terms in the numerator small, we need to ensure that the localization kernel H is indeed "local": there needs to be some small value $\epsilon > 0$ for which $\mathbb{P}\{\|X - \tilde{X}\| > \epsilon\}$ is low. For the first and third term in the numerator to both be small, we need to choose a function g that, while perhaps not globally Lipschitz, is nonetheless (approximately) Lipschitz over the bulk of the distribution of features X . Finally, the denominator simply serves to make the result scale-invariant: if we consider a reweighting function g' defined as $g'(x) = c \cdot g(x)$ for some positive constant c , the two shifted distributions $P_X \circ g$ and $P_X \circ g'$ are equal, so our upper bound needs to be invariant to this type of rescaling

in order to be meaningful.

A closer look at the covariate shift guarantee: proof sketch of

Theorem 3.6

To develop our intuition for the construction of the RLCP method, and how the construction of $\tilde{X}$ plays a role, here we will take a closer look at some of the key ideas underlying the proof of Theorem 3.6.

In the covariate shift setting, by definition, the joint distribution of the data along with the random point $\tilde{X}_{n+1}$ is given by

$$\begin{cases} \{(X_i, Y_i)\}_{i \in [n]} \stackrel{\text{iid}}{\sim} P, \\ (X_{n+1}, Y_{n+1}) \sim (P_X \circ g) \times P_{Y|X}, \\ \tilde{X}_{n+1} \mid X_{n+1} \sim H(X_{n+1}, \cdot). \end{cases}$$

Equivalently, we can write this joint distribution as

$$\begin{cases} \{(X_i, Y_i)\}_{i \in [n]} \stackrel{\text{iid}}{\sim} P, \\ \tilde{X}_{n+1} \sim \tilde{P}_{\tilde{X}}, \\ (X_{n+1}, Y_{n+1}) \mid \tilde{X}_{n+1} \sim (P_X \circ g \circ H(\cdot, \tilde{X}_{n+1})) \times P_{Y|X}, \end{cases} \quad (3.16)$$

where $\tilde{P}_{\tilde{X}}$ is the marginal distribution of $\tilde{X}_{n+1}$ under the covariate shift model: this distribution can be characterized by the marginal distribution of $\tilde{X}$ when we first draw $X \sim P_X \circ g$, and then draw $\tilde{X} \sim H(X, \cdot)$. In order to establish approximate coverage under covariate shift, then, we need to show that

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \gtrsim 1 - \alpha \text{ with respect to the joint distribution (3.16).}$$

Now we will consider an approximation to this joint distribution of the data, which no longer depends on the unknown reweighting function g . Specifically, compare to the following joint distribution:

$$\begin{cases} \{(X_i, Y_i)\}_{i \in [n]} \stackrel{\text{iid}}{\sim} P, \\ \tilde{X}_{n+1} \sim \tilde{P}_{\tilde{X}}, \\ (X_{n+1}, Y_{n+1}) \mid \tilde{X}_{n+1} \sim (P_X \circ H(\cdot, \tilde{X}_{n+1})) \times P_{Y \mid X}. \end{cases} \quad (3.17)$$

We observe the only difference between this distribution, and the joint distribution that appears in (3.16), lies in the conditional distribution of $(X_{n+1}, Y_{n+1}) \mid \tilde{X}_{n+1}$. Will marginal coverage hold with respect to this new distribution? In fact, in this new joint distribution (3.17), the conditional distribution of the training data and test point $\{(X_i, Y_i)\}_{i \in [n+1]}$ conditional on $\tilde{X}_{n+1}$, is exactly the same as that calculated in the proof of Proposition 3.2 for the i.i.d. data setting (without covariate shift). If we examine the result of Proposition 3.2, we see that it establishes that RLCP has $\geq 1 - \alpha$ coverage conditional on $\tilde{X}_{n+1}$,

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \right\} \geq 1 - \alpha,$$

with respect to drawing $\{(X_i, Y_i)\}_{i \in [n]} \stackrel{\text{iid}}{\sim} P$ and $(X_{n+1}, Y_{n+1}) \mid \tilde{X}_{n+1} \sim (P_X \circ H(\cdot, \tilde{X}_{n+1})) \times P_{Y \mid X}$, as in (3.17). In particular, averaging over a random draw $\tilde{X}_{n+1} \sim \tilde{P}_{\tilde{X}}$ it holds that

$$\mathbb{E}_{\tilde{X}_{n+1} \sim \tilde{P}_{\tilde{X}}} \left[\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \right\} \right] \geq 1 - \alpha$$

or equivalently,

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha \text{ with respect to the joint distribution (3.17).}$$

Therefore, we can see that, for the covariate shift setting (as in (3.16)), we have the following

lower bound on coverage:

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha - d_{\text{TV}} \left(\begin{array}{cc} \text{the joint distribution} & \text{the joint distribution} \\ \text{of the data} & \text{of the data} \\ \text{appearing in (3.16)} & \text{appearing in (3.17)} \end{array} \right).$$

Examining the difference between the two joint distributions derived in (3.16) and in (3.17), we can simplify this result to

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha - \mathbb{E}_{\tilde{P}_{\tilde{X}}} \left[d_{\text{TV}} \left(P_X \circ H(\cdot, \tilde{X}), P_X \circ g \circ H(\cdot, \tilde{X}) \right) \right].$$

From this point on, then, our only task is to bound this total variation distance. (In fact, we will explain in Appendix B.2.2 that this total variation distance can also be used to bound the extent to which RLCP can *overcover*, i.e., to ensure that the method is not too conservative.)

Now we can ask, for what functions g might we expect this total variation distance to be small? If H is a localizing kernel with a fairly small bandwidth, then the reweighted distribution $P_X \circ H(\cdot, \tilde{X})$ is likely to place nearly all of its mass in a small neighborhood around $\tilde{X}$. Consequently, if g is Lipschitz (or some similar condition), then g will be approximately constant on the bulk of the distribution $P_X \circ H(\cdot, \tilde{X})$, and thus reweighting additionally by g (i.e., computing the new distribution $P_X \circ g \circ H(\cdot, \tilde{X})$) will only minimally change this distribution. The localization is very much necessary to this argument: if we ran conformal prediction without localization (i.e., CP rather than one of the LCP methods), following the same arguments would lead us to a term $d_{\text{TV}}(P_X, P_X \circ g)$, which is likely to be large even if g is Lipschitz (aside from the trivial case where g is a constant function, i.e., $P_X \circ g = P_X$).

To build a more precise understanding of this intuition, later on in Appendix B.1.1, we will establish the bound

$$\mathbb{E}_{\tilde{P}_{\tilde{X}}} \left[d_{\text{TV}} \left(P_X \circ H(\cdot, \tilde{X}), P_X \circ g \circ H(\cdot, \tilde{X}) \right) \right] \leq \frac{\mathbb{E}_{P_{\tilde{X}}} \left[\sqrt{\frac{1}{2} \text{Var}_{P_{X|\tilde{X}}} (g(X) | \tilde{X})} \right]}{\mathbb{E}_{P_X} [g(X)]}. \quad (3.18)$$

We can see that the conditional variance in the numerator is likely to be small when H is strongly localized (i.e., when the conditional distribution of $X | \tilde{X}$ is concentrated in a small neighborhood), while g is Lipschitz, so that $g(X)$ has low variability conditional on $\tilde{X}$. The remaining details, including proving the bound (3.18) and also establishing the upper bound in the theorem, are deferred to Appendix B.1.1.

3.5 Experiments

In our experiments², we will examine the empirical performance of our method, RLCP, and will compare its performance to the two locally weighted CP methods described in Section 3.2—baseLCP and calLCP [Guan, 2023].

Before we describe the experiments, we pause to comment on a technicality. For calLCP and RLCP, the marginal coverage guarantees ensure that coverage will be *at least* $1 - \alpha$; in some settings, the methods may be conservative, with coverage higher than the target level. (Of course, as discussed earlier, baseLCP does not give a marginal coverage guarantee.) The possibility of marginal overcoverage means that comparing the methods meaningfully is challenging: if one method achieves better local coverage, perhaps this might be because this method is more conservative in terms of its marginal coverage. With this in mind, in order to allow for a more meaningful comparison, in our experiments we implement *smoothed* versions of each of the three methods, which (for calLCP and RLCP) ensures that marginal

2. Code for reproducing all experiments can be found at <https://github.com/rohanhore/RLCP>.

coverage is exactly $1 - \alpha$, and leads to a more fair comparison. Details on the implementation of the smoothed versions of each of the methods can be found in Appendix B.2.

3.5.1 Simulations

For all our simulation experiments, our aim will be to achieve 90% coverage, i.e., we choose $\alpha = 0.1$. The feature space is given by $\mathcal{X} = \mathbb{R}^d$, and all experiments are run with localization kernel H given by the Gaussian kernel, $H(x, x') = \frac{1}{(2\pi h^2)^{d/2}} \exp\left(-\frac{\|x - x'\|_2^2}{2h^2}\right)$.

Univariate setting

We begin with experiments in a univariate setting, i.e., $d = 1$. We consider the following two data generating distributions:

- **Setting 1:** $X \sim \mathcal{N}(0, 1)$, $Y|X \sim \mathcal{N}\left(\frac{X}{2}, |\sin(X)|\right)$
- **Setting 2:** $X \sim \mathcal{N}(0, 1)$, $Y|X \sim \mathcal{N}\left(\frac{X}{2}, \frac{4}{3}\phi\left(\frac{2X}{3}\right)\right)$, where $\phi(\cdot)$ is the density of a standard Gaussian.

The two settings share the same feature distribution, $P_X = \mathcal{N}(0, 1)$. The difference lies in the conditional distribution of the response, $P_{Y|X}$. In Setting 2, the response has more variance for values of X near the center of the distribution. In contrast, for Setting 1, the response has more variance for values of X lying away from the mean. These different trends can be seen in Figure 3.2.

For both settings, our score function is given by $s(x, y) = |y - \hat{f}(x)|$, where $\hat{f}$ is a pretrained predictive model (trained using linear regression, using a separate dataset of 2000 data points). For each setting, we implement the methods using the localization kernel H at five different bandwidths, $h \in \{0.1, 0.2, 0.4, 0.8, 1.6\}$; the lowest value represents a highly local H , while the highest value leads to a kernel H that is essentially flat over the bulk of the

feature distribution. The methods are run with sample size $n = 2000$, and evaluated on 2000 test points. The entire experiment is repeated for 50 independent trials.

Results: marginal coverage. Figure 3.1 displays the marginal coverage obtained by each of the three methods on the 2000 test points, plotted over the 50 independent trials of the experiment. As expected from the theoretical guarantees, both RLCP and callLCP achieve the target coverage level $1 - \alpha = 90\%$. On the other hand, depending on the bandwidth choice, baseLCP can have marginal coverage smaller or larger than $1 - \alpha$.

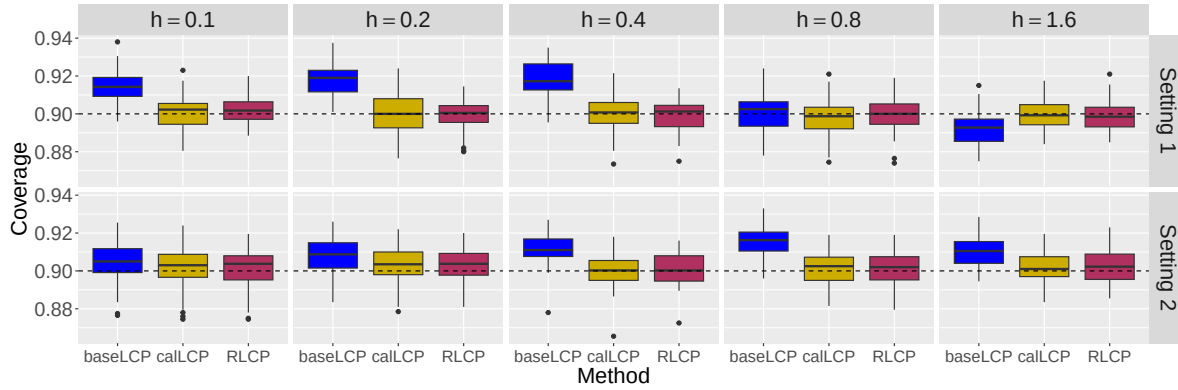


Figure 3.1: Marginal coverage of each method in two univariate simulation settings, across different bandwidths. Results are shown for 50 independent trials. See Section 3.5.1 for details.

Results: local adaptivity. One of the main goals addressed by callLCP in Guan [2023]’s work, and by RLCP in our work, is the goal of (approximate) test-conditional coverage, as in Aim (A2). In order to study whether these methods are successfully addressing this aim empirically, we should look beyond marginal coverage, and examine whether the methods appear to provide local coverage in our experiments.

In the top panels of Figure 3.2, we display the average values of the prediction interval returned by each method (that is, for each feature value $x \in \mathbb{R}$, we compute the mean lower endpoint, and mean upper endpoint, of the prediction interval over the 50 trials of the

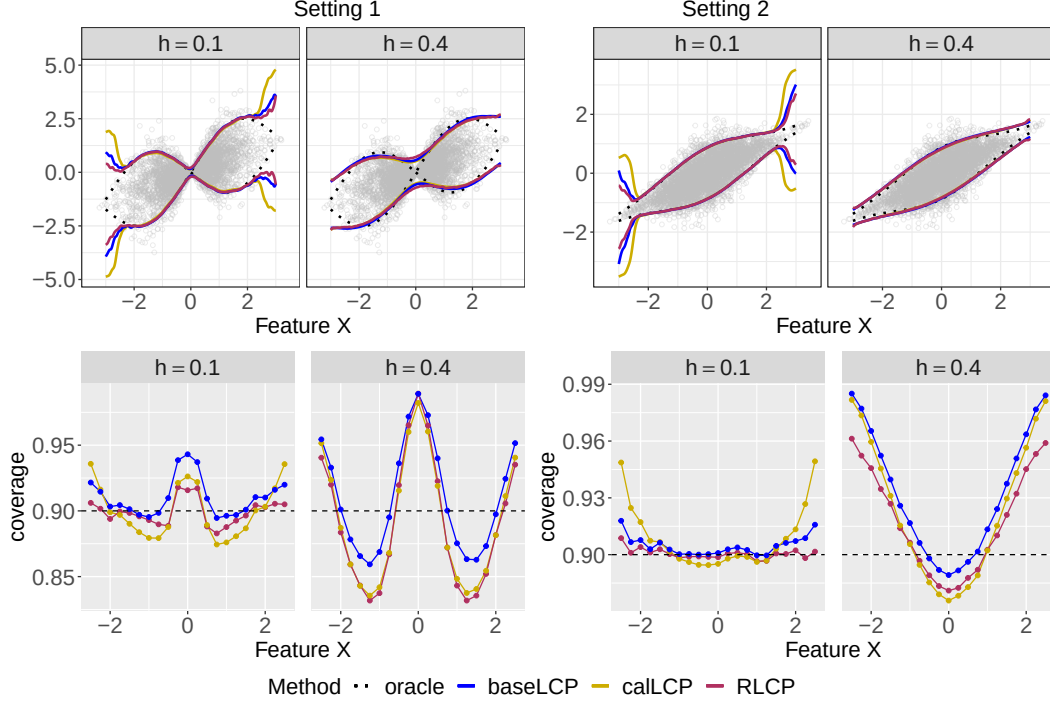


Figure 3.2: Top panels: the average prediction intervals produced by each method (at each feature value x , the endpoints of the prediction interval are averaged over 50 independent trials), compared to the oracle prediction interval. Bottom panels: The corresponding local coverage of each method, averaged over 50 independent trials. See Section 3.5.1 for details.

experiment). We compare the results from the three methods against the “oracle” prediction interval, computed with knowledge of the true distribution $P_{Y|X}$ —i.e., the interval defined by the 5th and 95th percentiles of the conditional distribution of Y given $X = x$, at each value x . In the bottom panels, we plot the local coverage of our prediction intervals, computed as the empirical coverage in balls of radius 0.4 around a feature point, averaged over 50 independent trials, and over the test observations lying within the ball. For brevity, we only present result for two bandwidths here; additional plots are shown in Appendix B.3.3.

For a smaller bandwidth h , the strong localization effect of the kernel H ensures that the prediction intervals, for each of the three methods, adapt to the local variance of $P_{Y|X}$. For a larger bandwidth, however, this desirable performance is lost; the localization effect diminishes (to be specific, the weighted quantile calculations in each method are being calculated

with weights that are approximately constant), hence all three methods return approximately constant-width intervals that are not locally adaptive. Even for small h , we can see that the intervals appear conservative in the tails of the X distribution, likely due to low effective sample size in these regions—however, this effect is more pronounced for calLCP, while RLCP offers more constant local coverage at the smaller h .

Additional results. In Appendix B.3.3, in addition to showing plots for all values of the bandwidth h (since Figure 3.2 only shows results for $h = 0.1, 0.4$), we also compare to a related method from the recent literature—the work of Gibbs et al. [2023], which we will describe below in Section 3.6.1.

Multivariate setting

Next, we will study the performance of these methods in a multivariate setting, again with the goal of comparing their local coverage properties. To do this in dimension $d > 1$, since we cannot plot the prediction intervals at each value $x \in \mathbb{R}^d$, we will instead examine the coverage properties of the methods on meaningful subsets of the feature space $\mathcal{X} = \mathbb{R}^d$.

We begin by generalizing the univariate setting: for a dimension $d \geq 1$, we consider the data distribution

$$X \sim \mathcal{N}_d(0, \mathbf{I}_d), \quad Y|X \sim \mathcal{N}\left(\sum_{i=1}^d \frac{X_i}{2}, \sum_{i=1}^d |\sin(X_i)|\right).$$

Note that in dimension $d = 1$, this coincides with Setting 1 from our univariate simulation in Section 3.5.1. The details for implementing baseLCP, calLCP, and RLCP are the same as for the univariate setting: we again use score function $s(x, y) = |y - \hat{f}(x)|$ for a pretrained (via linear regression) $\hat{f}$, and again use a Gaussian kernel H , and sample size $n = 2000$.

We will now examine the local coverage properties of each method in several different ways.

Results: coverage conditional on a set (with constant bandwidth h). To examine the test-conditional properties of these methods in high dimensions, we will focus on a specific task: coverage conditional on $X \in B$ for specific sets B . Concretely, we will define

$$B_{\text{in}} = \{x \in \mathbb{R}^d : \|x\|_2 \leq \tau_d\}, \quad B_{\text{out}} = \{x \in \mathbb{R}^d : \|x\|_2 > \tau_d\}$$

as the set of points x lying inside and outside a ball of radius τ_d . Here τ_d^2 is chosen as the median of the χ_d^2 distribution, which ensures that $P_X(B_{\text{in}}) = P_X(B_{\text{out}}) = 0.5$.

We will calculate local measures of predictive coverage for each of the three methods, across dimension $d \in \{1, 5, 10, 15, \dots, 50\}$. For each method and each dimension d , we compute an empirical estimate of the conditional coverage probability³

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid X_{n+1} \in B \right\},$$

for three different choices of the set B : $B = B_{\text{in}}$, $B = B_{\text{out}}$, and $B = B_{\text{in}} \cup B_{\text{out}} = \mathcal{X}$ (i.e., this last choice corresponds to marginal rather than conditional coverage). Intuitively, a method with strong local coverage properties should show approximately equal coverage on the inner and outer sets, B_{in} and B_{out} .

We start with exploring the empirical local coverage of the conformal methods with a fixed bandwidth $h = 1.5$ for all the methods across all dimensions (we will address the question of letting h vary with dimension d shortly). The results are shown in Figure 3.3. RLCP performs significantly better than both baseLCP and callCP in this setting. Aside from the smallest dimensions d , RLCP provides approximately equal coverage on both B_{in} and B_{out} , with both conditional coverage levels approximately equal to the marginal coverage

3. For discussing the comparison between the three methods, we will denote the prediction interval as $\hat{C}_n(X_{n+1})$ for each method, even though for RLCP, the prediction interval is also a function of the random variable $\tilde{X}_{n+1}$.

level $1 - \alpha$. In contrast, while calLCP maintains exact marginal coverage (as guaranteed by the theory), it exhibits severely uneven coverages on the two sets B_{in} and B_{out} . Finally, baseLCP shows extremely inconsistent performance across varying d , and can show marginal over- or undercoverage at various dimensions, but interestingly, the coverage level conditional on B_{in} or on B_{out} are more similar to each other than for calLCP, in this setting.

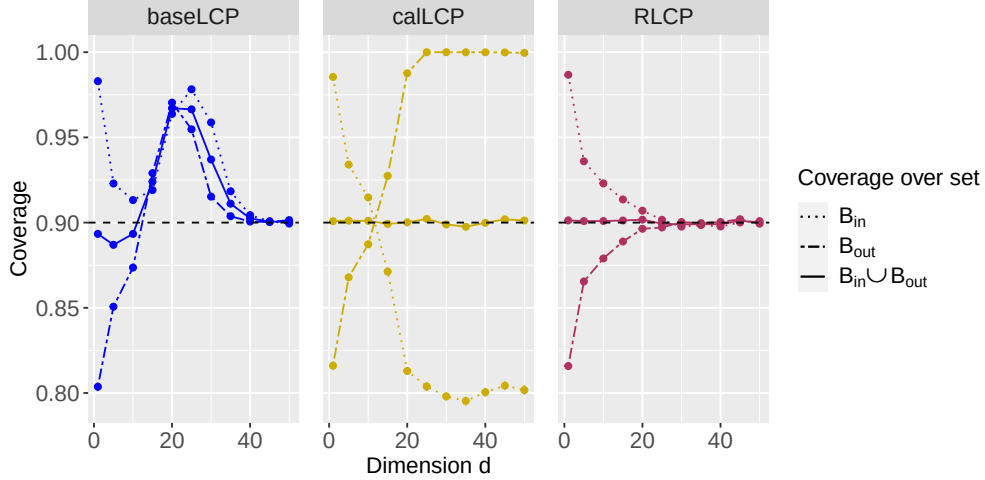


Figure 3.3: Predictive coverage conditional on $X_{n+1} \in B$, for three choices of the set B (namely, B_{in} , B_{out} , and $B_{\text{in}} \cup B_{\text{out}} = \mathcal{X}$). Results for each method with fixed bandwidth $h = 1.5$ are plotted with respect to dimension d of the feature space. Results are averaged over 50 independent trials. See Section 3.5.1 for details.

Results: coverage conditional on a set (with bandwidth h varying with dimension d). The trends in performance observed above may be heavily influenced by the choice of the bandwidth h —and in particular, due to the curse of dimensionality, for a fixed value h that does not change with dimension d , the amount of mass captured by P_X within a radius- h neighborhood around any given x will decrease exponentially rapidly as d grows. In other words, using a fixed h as d increases, will lead to a sharply decreasing effective sample size for calculating the weighted quantiles; in a sense, the method is “too localized” for higher d . Therefore, for a more meaningful comparison across a range of dimensions d , we should choose the bandwidth h in a dimension-adaptive way, to try to maintain a constant notion

of effective sample size.

Given feature values $X_1, \dots, X_{n+1}$, we recall the weights

$$w_i = \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})}.$$

These weights appear in the calculation of the weighted quantile for both baseLCP and calLCP. To understand how the w_i 's relate to effective sample size, let's consider the simple case where H is a box kernel and each X is likely to pick k many neighbors out of a sample of size n with this kernel: then for each i , we have (approximately) k many w_i 's with value (approximately) $1/k$, and the rest zero. In this setting, for each i we have $\sum_i w_i^2 \approx 1/k$. Equivalently, we can write

$$k \cdot \sum_i H(X_i, X_{n+1})^2 \approx \left(\sum_i H(X_i, X_{n+1}) \right)^2.$$

Now, the terms in this expression are random, but for large n we can approximate this calculation by

$$k \cdot n \mathbb{E} \left[H(X, X')^2 \right] \approx \mathbb{E} \left[\left(n \mathbb{E} \left[H(X, X') \mid X \right] \right)^2 \right],$$

where the expected values are taken with respect to $X, X' \stackrel{\text{iid}}{\sim} P_X$. This then motivates the following definition of effective sample size:

$$n_{\text{eff}}(h) = n \cdot \frac{\mathbb{E} \left[\mathbb{E} \left[H(X, X')^2 \mid X \right] \right]}{\mathbb{E} \left[H(X, X')^2 \right]}. \quad (3.19)$$

For baseLCP and calLCP, which each construct $\hat{C}_n(X_{n+1})$ with quantile weights w_i , we use a bandwidth h that satisfies $n_{\text{eff}}(h) = 50$ (where $n_{\text{eff}}(h)$ is computed as in (3.19), after using the pretraining set to compute estimates of the necessary expected values). For RLCP, on

the other hand, the quantile calculation that defines the prediction interval is constructed using weights $\tilde{w}_i$, which is defined by calculating kernel values $H(X_i, \tilde{X}_{n+1})$ rather than $H(X_i, X_{n+1})$. Consequently, for this method, we define

$$\tilde{n}_{\text{eff}}(h) = n \cdot \frac{\mathbb{E} \left[\mathbb{E} \left[H(X, \tilde{X}') \mid X \right]^2 \right]}{\mathbb{E} \left[H(X, \tilde{X}')^2 \right]},$$

where now the expected value is taken with respect to drawing $X, X' \stackrel{\text{iid}}{\sim} P_X$ and then sampling $\tilde{X}' \sim H(X', \cdot)$. We then implement RLCP with a bandwidth h chosen to satisfy $\tilde{n}_{\text{eff}}(h) = 50$ (where again, the expected values defining our notion of effective sample size, are estimated empirically on the pretraining set). Consequently, the various methods are being run with (potentially) different bandwidths, to ensure that they each have the same effective sample size—that is, to ensure that we are comparing these methods at approximately the same level of localization.

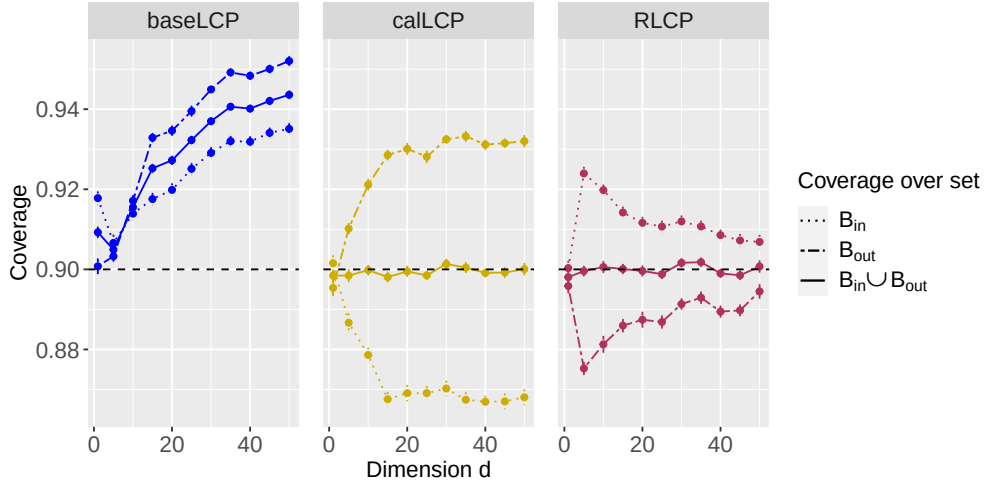


Figure 3.4: Predictive coverage conditional on $X_{n+1} \in B$, for three choices of the set B (namely, B_{in} , B_{out} , and $B_{\text{in}} \cup B_{\text{out}} = \mathcal{X}$). Results for each method, averaged over 50 independent trials, are plotted with respect to dimension d of the feature space. Each method is run with its own bandwidth choice, so that a constant 50 effective sample size is maintained across dimensions. See Section 3.5.1 for more experimental details.

The results, which are displayed in Figure 3.4, show that callCP has more disparity in the

coverage conditional on B_{in} versus on B_{out} , as compared to RLCP: aside from the lowest values of dimension d , the coverage of RLCP on these subsets are around 1% off from the target level, compared to around 3% for calLCP. For both of these methods, as expected from the theory, the marginal coverage is exactly $1 - \alpha$. (For baseLCP, the marginal coverage is quite conservative except at the smallest values of d , and hence the conditional coverage levels are high as well.)

Additional results. The results shown above consider conditional coverage relative to only two specific sets B —namely, B_{in} and B_{out} . For a more complete picture of the performance of the methods, in Appendix B.3.4, we consider a more general collection of sets B , and compare coverage of all these localized conformal methods conditioned on those sets. We will see a similar pattern of performance—divergence of coverage of calLCP from the target level is significantly higher compared to that of RLCP, for higher dimensions d .

3.5.2 Real data example: predicting age of abalones

Next, we will compare the performance of the local CP methods on a real dataset on abalones (a species of mollusk) [Nash et al., 1994].⁴ The objective is to predict the age of abalones based on their physical measurements. Since the economic value of abalone is positively correlated with its age, but determining age precisely requires a costly procedure, accurately predicting age from physical characteristics alone becomes both a valuable and practical problem. This dataset has 4177 data points, where the response **rings** determines the age of the abalone, and the features consist of the abalone’s sex (categorized as Male, Female, or Infant), length (the longest shell measurement), diameter (perpendicular to length), height, whole weight, shucked weight (i.e., weight of the meat inside), viscera weight (i.e., gut weight after bleeding) and shell weight (the weight after drying).

4. The data for this experiment was obtained from <https://archive.ics.uci.edu/dataset/1/abalone>.

For our experiment, we will use $Y = \text{ring}$ as the response, and X_{sex} , X_{length} , X_{diameter} , X_{height} and $X_{\text{whole_weight}}$ as the features. After splitting the data equally into a pretraining set, a calibration set, and a test set,⁵ we train three different base predictors $\hat{f}$ on the pretraining set, using a linear regression, a 2-layer neural network, or a random forest. For each $\hat{f}$ we define the corresponding pretrained score function $s(x, y) = |y - \hat{f}(x)|$. Next, we will define the localization kernel H to place high weight for pairs of patients that have the same sex, and similar values for length, diameter, height and whole weight: fixing a choice of bandwidths h_{length} , h_{diameter} , h_{height} and $h_{\text{whole_weight}}$, we define

$$H(x, x') = \frac{1}{2h_{\text{sex}} \cdot 2h_{\text{diameter}} \cdot 2h_{\text{height}} \cdot 2h_{\text{whole_weight}}} \cdot \mathbb{1} \left\{ \begin{aligned} &x_{\text{sex}} = x'_{\text{sex}}, \\ &|x_{\text{length}} - x'_{\text{length}}| \leq h_{\text{length}}, |x_{\text{diameter}} - x'_{\text{diameter}}| \leq h_{\text{diameter}}, \\ &|x_{\text{height}} - x'_{\text{height}}| \leq h_{\text{height}}, |x_{\text{whole_weight}} - x'_{\text{whole_weight}}| \leq h_{\text{whole_weight}} \end{aligned} \right\}.$$

In our implementation, for simplicity we will take $h_{\text{length}} = h_{\text{diameter}} = h_{\text{height}} = h_{\text{whole_weight}} = h$ to always equal the same value, chosen from $h \in \{0.05, 0.1, \dots, 0.25, 0.3\}$. This satisfies the condition (3.10), i.e., that $H(x, \cdot)$ defines a density with respect to an appropriate base measure: namely, sampling from $H(x, \cdot)$ consists of sampling each of length, diameter, height and whole_weight measurement uniformly from a $2h$ length bin around the corresponding test data measurement (e.g. length is an uniform random sample from $[x_{\text{length}} - h, x_{\text{length}} + h)$, while copying the value of x_{sex} .

Marginal coverage and prediction interval width. In Figure 3.5, for each of the three localized CP methods, we present the marginal coverage and the median width of the

5. Here we depart from the terminology of the earlier sections in this chapter—the term *calibration set* refers to the data $(X_1, Y_1), \dots, (X_n, Y_n)$ on which we implement the methods baseLCP, calLCP, and RLCP, using a predefined score function s that is based on a pretrained predictor $\hat{f}$. This same set was previously called the *training set* throughout the chapter, but here we refer to it as the calibration set to distinguish it from the pretraining set used for fitting $\hat{f}$.

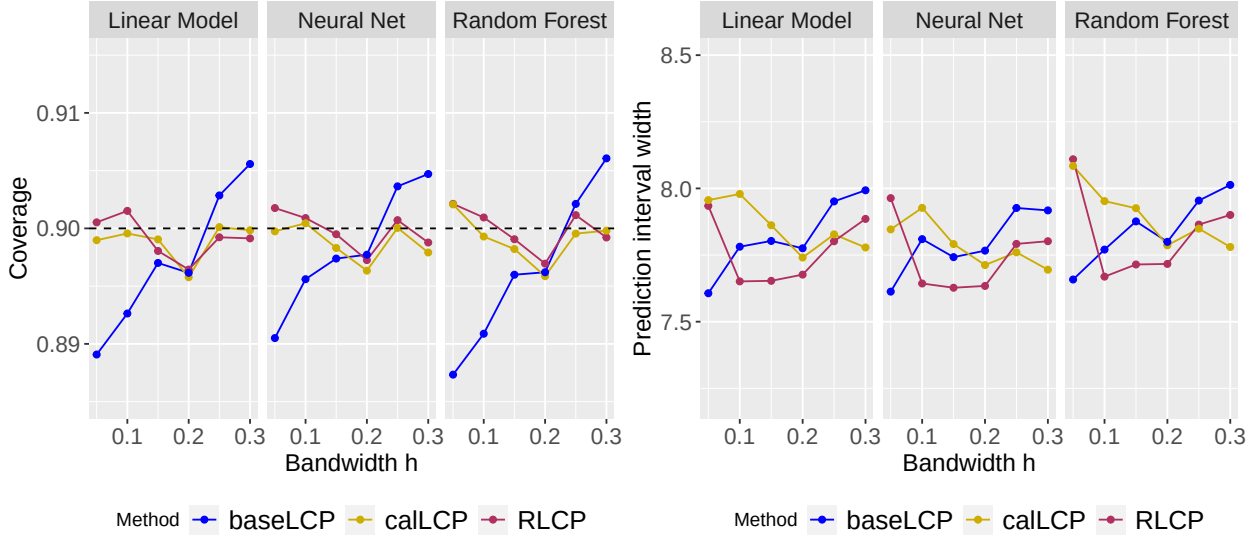


Figure 3.5: Marginal coverage (averaged over the test set) and prediction interval width (median over the test set), for each of the three methods across different bandwidths and for different choices of the pretrained base predictor $\hat{f}$, for the real data experiment. Results are averaged over 50 random splits of the data into pretraining, calibration, and test sets. See Section 3.5.2 for details.

prediction intervals on the test set, averaged over 50 random splits of the data. While RLCP and callLCP attains exact target level of marginal coverage, coverage of baseLCP prediction intervals are lower and higher respectively for smaller and larger values of h . We further see that median length of prediction intervals are slightly narrower for RLCP prediction intervals for most bandwidth choices and for the various base algorithm choices. We will see next that RLCP offers better empirical conditional coverage at lower bandwidth values, which may justify the need for the prediction intervals with slightly larger width at the lowest values of h .

Conditional coverage. In this real data setting, there are many practical questions we can ask regarding test-conditional coverage. In particular, since each data point corresponds to an individual abalone, test-conditional coverage can be interpreted as asking whether the predictive intervals are equally valid across different subgroups of abalones (e.g., smaller vs larger abalones). Here we aim to gain some insight into these questions by examining the

empirical coverage of each method conditional on certain physical measurement for abalone. Ideally, we would expect approximately constant conditional coverage across different subgroups of abalones. In the left panel of Figure 3.6, we show coverage conditional on sex, i.e., conditional on the value of the three levels of covariate X_{sex} . The bar plot shows the value of $\mathbb{P}\left\{Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid (X_{n+1})_{\text{sex}} = x\right\}$ for each $x \in \{\text{Male}, \text{Female}, \text{Infant}\}$. While RLCP shows approximately equal coverage levels for the three groups across all three base predictors $\hat{f}$, for calLCP the performance is quite variable, with extremely disparate coverage levels for some choices of the base algorithm.

In the right panel of Figure 3.6, we ask the same question with respect to X_{length} . Since this is a continuous covariate, we cannot compute conditional coverage exactly, but instead estimate it with a sliding window whose width is chosen to have 5% mass: that is, we are plotting $\mathbb{P}\left\{Y_{n+1} \in \hat{C}_n(X_{n+1}) \mid (X_{n+1})_{\text{length}} \in [x - \Delta_0, x + \Delta_1)\right\}$ where, for each x , Δ_0 and Δ_1 are chosen so that 2.5% of the abalone population has length in the range $[x - \Delta_0, x)$, and same for the range $[x, x + \Delta_1)$. Here we again see that RLCP shows better conditional coverage compared to calLCP particularly at lower bandwidth $h = 0.05$ —the estimated conditional coverage for RLCP is approximately constant at the target level of 90% while for calLCP, we see very prominent over- or undercoverage at various values of length.

3.6 Discussion and extensions

In this work, we have introduced a novel variant of the localized conformal prediction framework, randomly-localized conformal prediction (RLCP), which attains distribution-free approximate coverage guarantees for relaxed test-conditional coverage and for smooth covariate shift problems while maintaining exact marginal coverage. As explained in Section 3.4.3, all of these coverage results are consequences of a core total variation distance based argu-

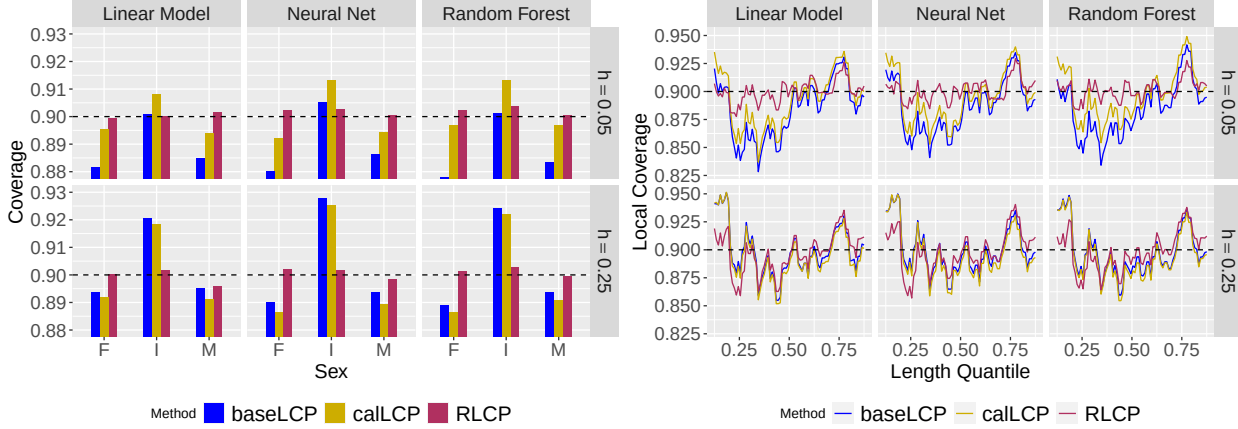


Figure 3.6: Comparison of local coverage of the three methods, conditioning on X_{sex} (left) and on X_{length} (right). See Section 3.5.2 for details.

ment, which could potentially be used to derive other types of coverage guarantees as well. Additionally, RLCP also enjoys a training-conditional coverage guarantee. Our numerical simulations and data experiments advocate the use of RLCP over existing locally-weighted conformal prediction algorithms, due to its better performance at maintaining conditional coverage across a range of settings.

In the remainder of this section, we will first compare RLCP with existing literature and methodology in closer detail, and will then explore the role of the randomization in RLCP. We will also consider an alternative formulation of calLCP, relating it to full conformal prediction.

3.6.1 Related work

As we have described earlier, it is well known that achieving distribution-free pointwise test-conditional coverage is impossible to do (aside from trivial solutions such as $\hat{C}_n(X_{n+1}) \equiv \mathcal{Y}$) [Vovk, 2012, Lei and Wasserman, 2014]; even natural relaxations of the test-conditional coverage property also suffer from fundamental hardness results in the distribution-free setting [Barber et al., 2021]. With these hardness results in mind, any progress on conditional

coverage for conformal prediction methods must necessarily make compromises, whether by aiming for a weaker theoretical guarantee, by placing some assumptions on the problem, or by targeting empirical rather than theoretical performance.

Almost all of these existing approaches in the conformal literature, including our proposal, can be thought of as a specific instance of the following three-step general workflow:

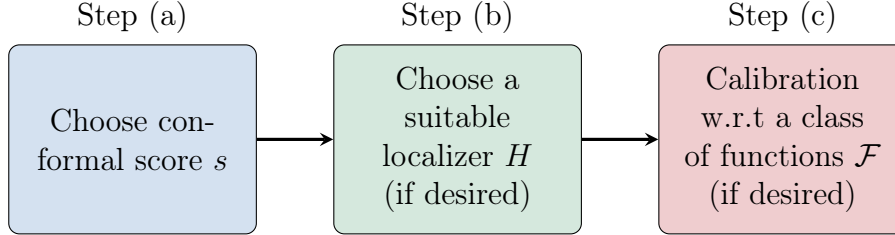


Figure 3.7: A general workflow of existing conformal approaches in literature to attain test-conditional coverage

We now explain these steps, and highlight how some existing methods can be interpreted in terms of this three-step workflow.

Step (a): choosing a score s . The split conformal method, defined in Section 1.1.1, can be implemented with any choice of the score function s (as in Step (a)). For example, we might choose the residual score function, $s(x, y) = |y - \hat{f}(x)|$ (for a prefitted regression function $\hat{f}$), as discussed earlier. A popular alternative is the conformalized quantile regression (CQR) method [Romano et al., 2019b], which uses the score function $s(x, y) \max\{\hat{Q}_\alpha(x) - y, y - \hat{Q}_{1-\alpha}(x)\}$, where $\hat{Q}$ denotes a pretrained estimated quantile.

In general, a score function s that captures the underlying heterogeneity of $P_{Y|X}$ (such as CQR) will tend to exhibit good local coverage empirically. For additional examples of score functions of this type, see the works of Lei et al. [2018], Györfi and Walk [2019], Chernozhukov et al. [2021], Sesia and Romano [2021], Deutschmann et al. [2023], Feldman et al. [2021]; see also Gupta et al. [2022], Sesia and Candès [2020] for some empirical comparison of local

coverage properties of various choices of s . Some of these methods offer (asymptotic) local coverage guarantees under additional distributional assumptions, e.g., Sesia and Romano [2021], Györfi and Walk [2019]. However, split conformal—with *any* choice of s —does not have any distribution-free guarantees for approximate local coverage.

Step (b): choosing a localizer H . Next, we turn to Step (b)—choosing a localizer H . This means that when constructing the prediction interval, instead of using a single threshold $\hat{q}$ for *any* value of the test point X_{n+1} , we instead compute a localized threshold via a weighted quantile, with weights determined by the localizing kernel H —exactly as in both calLCP and RLCP. In particular, both calLCP and RLCP require *both* choosing a score function s and then choosing a localizer H , in order to implement an instance of the method (e.g., we might combine the CQR score with a box kernel localizer). As we have seen in this work, for RLCP, adding Step (b) into our workflow allows for distribution-free theoretical guarantees that approximate test-conditional coverage, as in Theorem 3.4.

Another type of approach that can be viewed through the lens of Step (b), is the approach of partitioning $\mathcal{X}$ into bins, and then computing $\hat{q}$ separately within each bin; methods of this type have been explored by, e.g., Lei and Wasserman [2014], Izbicki et al. [2022], LeRoy and Zhao [2021], and can be related to Step (b) by choosing $H(x, x') = \mathbb{1}\{x, x' \text{ are in the same bin}\}$. Alternatively, H can be chosen as a nearest-neighbor based kernel [Ghosh et al., 2023].

We remark that for the split conformal method (which does not incorporate Step (b)), we can equivalently view this method as choosing the *trivial* localizer H given by $H(x, x') \equiv 1$, since computing an unweighted quantile (to define the threshold $\hat{q}$ for the split conformal interval) is equivalent to computing a weighted quantile with constant weights $w_1 = \dots = w_{n+1}$.

Step (c): choosing a function class $\mathcal{F}$. Finally, we turn to the third step: calibration with respect to a class of functions $\mathcal{F}$. This step does not appear in our RLCP method, but plays a key role in some of the related methods in the literature. Specifically, in the work of Gibbs et al. [2023], calibration with respect to $\mathcal{F}$ means that prediction intervals are required to satisfy⁶

$$\mathbb{E} \left[f(X_{n+1}) \cdot \left(\mathbb{1} \left\{ Y_{n+1} \in \hat{C}_n(X_{n+1}) \right\} - (1 - \alpha) \right) \right] \geq 0 \text{ for all } f \in \mathcal{F}.$$

If we take $\mathcal{F}$ to be the set of all measurable functions, this would be equivalent to a test-conditional coverage guarantee as in (3.1)—but as discussed before, is impossible to achieve distribution-free. Gibbs et al. [2023] therefore implement their method with more restricted classes of functions $\mathcal{F}$, for instance linear functions or functions from a RKHS, to achieve an approximate notion of local coverage; their method is implemented via solving a quantile regression problem over the class $\mathcal{F}$. As another example, we can choose $\mathcal{F} = \{ \mathbb{1} \{ \cdot \in G \} : G \in \mathcal{G} \}$ for a collection $\mathcal{G}$ of subsets of $\mathcal{X}$; this amounts to requiring group-conditional coverage for all groups from the collection $\mathcal{G}$, as has been studied in Jung et al. [2022].

At the other extreme, a marginal coverage guarantee can be interpreted as requiring calibration with respect to the trivial class of functions, $\mathcal{F} = \{ \mathbf{1} \}$ (where $\mathbf{1}$ is the constant function $x \mapsto 1$); thus methods such as split conformal, calLCP, or RLCP, which do not incorporate Step (c) into their workflow, can be equivalently interpreted as applying Step (c) with this trivial choice $\mathcal{F} = \{ \mathbf{1} \}$.

6. In the work of Gibbs et al. [2023], functions $f \in \mathcal{F}$ are not constrained to be nonnegative, and so this requirement is written as $= 0$, not ≥ 0 (of course, obtaining coverage at level exactly $1 - \alpha$ requires using a smoothed version of the method). Here we use a formulation more comparable to the notation of this chapter.

Implications for local coverage and covariate shift. To summarize, we have now seen that many existing methods (and our proposed RLCP method) can be compared under this common high-level framework. What does this tell us in terms of their local coverage properties (and coverage under covariate shift)? As mentioned above, methods that only use Step (a) (and therefore can be viewed as choosing the trivial localizer $H(x, x') \equiv 1$ for Step (b), and the trivial function class $\mathcal{F} = \{1\}$ for Step (c)) do not achieve distribution-free theoretical guarantees of (approximate) local coverage, although of course the choice of score s in split conformal can lead to substantial empirical improvements in local coverage in practice.

On the other hand, methods that use *either* localization at Step (b) (such as our RLCP procedure) or a nontrivial function class at Step (c) (such as the method of Gibbs et al. [2023]) can offer meaningful approximate local coverage theory in addition to strong performance empirically. Using localization (Step (b)) or calibration with respect to a function class (Step (c)) can therefore be viewed as complementary options towards achieving the same type of aim—we will further investigate this claim empirically by comparing the local coverage of RLCP and the method from Gibbs et al. [2023] in some of our simulation settings in Appendix B.3.3. An interesting open question is whether stronger guarantees and/or better empirical conditional coverage might be obtained by developing methods that incorporate both Step (b) and Step (c) simultaneously.

3.6.2 Is it possible to derandomize RLCP?

In our discussion in Section 3.6.1, we showed that a range of methods from the literature can all be viewed through a three-step framework for constructing localized versions of conformal prediction—including our proposed method, RLCP. One way that RLCP differs from the other methods discussed, however, is the fact that randomization is inherent in its construction. In this section, we take a closer look at the question of randomization: is

the randomized construction of RLCP necessary for ensuring its favorable theoretical and empirical properties, and does it cause substantial variability in the output in practice?

We will begin with the latter question, by examining empirically whether the random draw of $\tilde{X}_{n+1}$ in the construction of RLCP causes substantial randomness in the output. To explore this, we return to the univariate simulations from Section 3.5.1. Whereas before we plotted the *average* lower and upper endpoints of the prediction intervals across 100 independent trials of the experiment (in Figure 3.2), now we generate a *single* dataset and compute both the callCP and RLCP prediction intervals, but for RLCP, we repeat the random draw of $\tilde{X}_{n+1}$ 100 times to examine the randomness in the construction. Figure 3.8 shows the results: the shaded area visualizes the 5th and 95th percentile of the lower and upper endpoints of the RLCP prediction band. We can see that, on average, callCP and RLCP are giving similar results here (as was already shown in Figure 3.2), but the variability of RLCP is substantial in some settings (e.g., particularly for Setting 1 at the lower bandwidth values, where the localization is strong). We also examine the randomness of RLCP in a more quantitative way in Appendix B.3.1; we will see that, in the worst case, the RLCP prediction interval can vary by around 20–25% in width relative to the average output.

Hence, depending on the underlying setting, this variability of RLCP can be substantial, and may be a substantial limitation of this method in practice. On the other hand, the randomized construction is at the heart of the ability of RLCP to obtain stronger and cleaner theoretical guarantees relative to existing methods (as we have discussed in Sections 3.3.1 and 3.4.3), as well as its favorable empirical performance in terms of local coverage.

With this tradeoff in mind, this naturally suggests the following question: is it possible to “derandomize” the RLCP method, perhaps by averaging or aggregating across multiple runs of RLCP on a given data set, to retain its strong theoretical and empirical properties but reduce the issue of variability due to randomization?

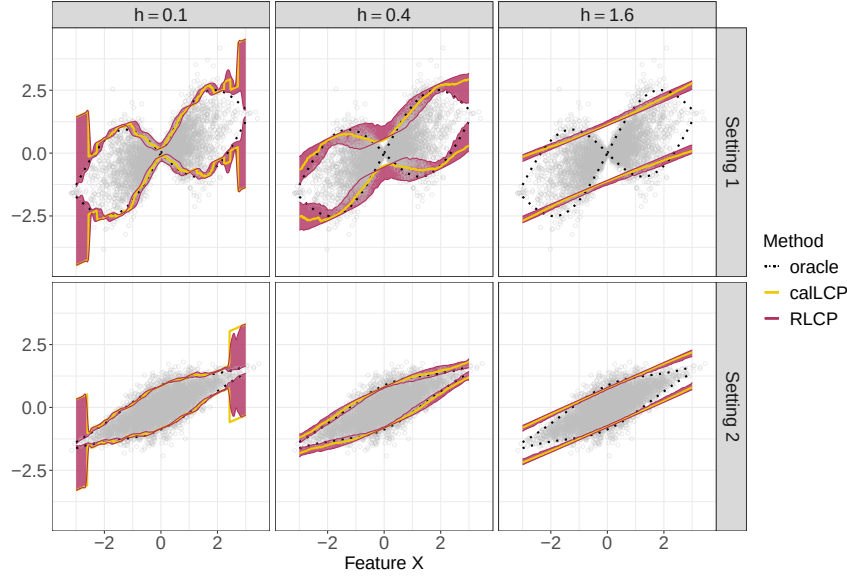


Figure 3.8: A visualization of the variability of the RLCP prediction band for a fixed data set, showing the 5th and 95th percentile of 300 draws of constructing RLCP using the same simulated data set. We also compare to the oracle prediction interval and to the callLCP prediction interval. See Section 3.6.2 for details.

Derandomizing RLCP via m -randomly-localized CP (m -RLCP)

We will now propose one possible approach towards derandomizing RLCP, by averaging over m many runs in a particular way. We will see that this construction is very natural, but unfortunately its empirical performance is substantially worse than RLCP itself. Through this exploration, we hope to inspire continued research on this question to determine whether the randomness of RLCP is somehow critical to its strong performance, or whether a different derandomization scheme may instead offer better results.

Before we begin, we first reformulate RLCP through the language of p-values. For all methods within the conformal prediction framework, the prediction interval can be constructed by considering, for each possible value y , the hypothesis test that asks whether we can reject the hypothesis that $Y_{n+1} = y$ (in which case y is excluded from the prediction interval), or cannot reject it (in which case y is included); the p-value from this hypothesis test can be viewed as an equivalent formulation of the conformal prediction methodology. For the case

of RLCP specifically, given the training set $\{(X_i, Y_i)\}_{i \in [n]}$, define a *randomized* p-value

$$p(y; X_{n+1}, \tilde{X}_{n+1}) = \sum_{i=1}^n \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \mathbb{1} \{s(X_i, Y_i) \geq s(X_{n+1}, y)\} + \frac{H(X_{n+1}, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})}. \quad (3.20)$$

As for other conformal methods, the prediction interval for RLCP can equivalently be constructed as

$$\hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) = \{y \in \mathcal{Y} : p(y; X_{n+1}, \tilde{X}_{n+1}) > \alpha\}.$$

We are now ready to derandomize. Firstly, given the test feature we generate m copies of $\tilde{X}_{n+1}$, i.e. $\tilde{X}_{n+1}^{[m]} = (\tilde{X}_{n+1}^1, \tilde{X}_{n+1}^2, \dots, \tilde{X}_{n+1}^m)$, with each value drawn i.i.d. from $H(X_{n+1}, \cdot)$.

For each $y \in \mathcal{Y}$, we then calculate the corresponding m p-values,

$$p(y; X_{n+1}, \tilde{X}_{n+1}^1), \dots, p(y; X_{n+1}, \tilde{X}_{n+1}^m)$$

as in (3.20) above, and then average these p-values,

$$\bar{p}(y; X_{n+1}, \tilde{X}_{n+1}^{[m]}) = \frac{p(y; X_{n+1}, \tilde{X}_{n+1}^1) + \dots + p(y; X_{n+1}, \tilde{X}_{n+1}^m)}{m}. \quad (3.21)$$

Finally, we return the prediction interval

$$\hat{C}_n^{m\text{-RLCP}}(X_{n+1}, \tilde{X}_{n+1}^{[m]}) = \{Y_{n+1} : \bar{p}(y; X_{n+1}, \tilde{X}_{n+1}^{[m]}) > \alpha\}.$$

We will refer to this method as m -randomly-localized conformal prediction (m -RLCP).

Since we are averaging the p-value over m random draws of $\tilde{X}_{n+1}$, for large m , the m -RLCP prediction interval is essentially a deterministic function of the data (i.e., of the training data

$\{(X_i, Y_i)\}_{i \in [n]}$ along with the test feature X_{n+1} .

It is well known that the average of valid p-values is itself, up to a factor of 2, a valid p-value [Rüschendorf, 1982, Vovk and Wang, 2020], i.e., we have

$$\mathbb{P} \left\{ \bar{p}(Y_{n+1}; X_{n+1}, \tilde{X}_{n+1}^{[m]}) \leq \alpha \right\} \leq 2\alpha.$$

This theoretical guarantee suggests that the m -RLCP prediction interval may be somewhat anticonservative, i.e., coverage is only guaranteed at level $\geq 1 - 2\alpha$. Although this result only addresses the potential under-coverage, in practice we observe that m -RLCP can in fact severely *overcover*, i.e., may be extremely conservative (we show an empirical comparison of RLCP in m -RLCP later on in Appendix B.3.2). In fact, a heuristic argument explains that m -RLCP behaves similarly to baseLCP, but with a different localizer kernel; our empirical findings on m -RLCP thus agree with the fact that baseLCP can often be conservative in practice, as we have seen empirically in the earlier experiments. We provide more details on the connection between m -RLCP and baseLCP in Appendix B.1.5.

To summarize, this intuitive approach towards derandomizing RLCP shows substantial drawbacks in terms of empirical performance (as well as weaker theory as compared to RLCP). Thus, finding a better way to aggregate across multiple runs of RLCP, to achieve a less random method without compromising on theoretical properties or empirical performance, remains an important open question.

3.6.3 Reinterpreting callCP as an instance of full conformal prediction

To conclude our discussion of related work, we will now examine Guan [2023]’s callCP method (as we recall, simply referred to as “LCP” in that work) through a new lens. Specifi-

cally, we make the observation that, with a properly chosen score function, calLCP is actually equivalent to running full conformal prediction [Vovk et al., 2005].

In this section, we will use the following formulation of full conformal prediction. We are given a function $T : (\mathcal{X} \times \mathcal{Y})^{n+1} \rightarrow \mathbb{R}^{n+1}$, which maps datasets of size $n + 1$ to a vector of scores (i.e., the i th score corresponds to the i th data point, with large scores indicating that the data point appears unusual relative to the other data in the dataset). The map T is required to commute with permutations (i.e., swapping data points i and j results in swapping the i th and j th scores returned by T). The full conformal prediction interval (or more generally, prediction set) is then defined as

$$\hat{C}_n^{\text{fullCP}}(X_{n+1}) = \{y \in \mathcal{Y} : T_{n+1}^y \leq \text{Quantile}_{1-\alpha}(T_1^y, \dots, T_n^y, T_{n+1}^y)\}, \quad (3.22)$$

where for any $y \in \mathcal{Y}$, the scores are defined as

$$T_i^y = \left(T((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)) \right)_i, \quad i = 1, \dots, n+1,$$

i.e., the scores of the $n+1$ data points, if we assume that the test point is equal to (X_{n+1}, y) . Full conformal prediction satisfies the marginal coverage guarantee for i.i.d. (or exchangeable) data [Vovk et al., 2005].

Now we make the correspondence between calLCP and full CP precise. After pretraining the original score function $s = s(x, y)$ on a separate data set, we define the map T as follows: for any dataset $(x_1, y_1), \dots, (x_{n+1}, y_{n+1})$, the new score for the i th data point is given by

$$\left(T((x_1, y_1), \dots, (x_{n+1}, y_{n+1})) \right)_i = \sum_{j=1}^{n+1} w_{i,j} \mathbb{1} \{s(x_j, y_j) < s(x_i, y_i)\},$$

where

$$w_{i,j} = \frac{H(x_j, x_i)}{\sum_{k=1}^{n+1} H(x_k, x_i)}.$$

in Appendix B.1.4, we will prove that this is an equivalent representation of the callCP method:

Proposition 3.7. *The prediction interval $\hat{C}_n^{\text{callCP}}(X_{n+1})$ returned by callCP, is always equal to the prediction interval $\hat{C}_n^{\text{fullCP}}(X_{n+1})$ returned by full conformal prediction, when implemented with the score map T defined as above.*

Consequently, the marginal coverage guarantee for callCP follows as a corollary: since full CP satisfies marginal coverage, callCP must therefore inherit this property, since it is an instance of full CP. On the other hand, establishing conditional coverage type properties is quite challenging for callCP (indeed, this is often the case for full CP in general), while the randomized construction of RLCP opens the door to the local theoretical guarantees that we have established above.

CHAPTER 4

ASSUMPTION-LEAN TESTS FOR CONDITIONAL INDEPENDENCE

In this chapter, we examine the problem of conditional independence (CI) testing. We first discuss the challenges of constructing an assumption-free test for the CI null hypothesis and then explore how incorporating a shape constraint enables a simple powerful test. In the broader context of testing challenging hypotheses, smoothness assumptions can help mitigate the difficulties of distribution-free inference; however, these assumptions are typically untestable. Instead, in certain settings, a shape constraint is often more easily justified and serves as a viable alternative to smoothness assumptions. In this chapter, we examine this behavior in the context of CI testing to construct a valid test which also attains meaningful power against some interesting alternatives.

This chapter presents work co-authored with *Jake A. Soloff, Rina Foygel Barber and Richard J. Samworth* is available on [ArXiv:2501.06133](https://arxiv.org/abs/2501.06133) under the title ‘Testing conditional independence under isotonicity’ [Hore et al., 2025].

4.1 Introduction

The statistical problem of testing conditional independence has been extensively studied for decades, with foundational contributions spanning multiple disciplines. Among the earliest developments, R.A. Fisher introduced the notion of the partial correlation coefficient [Fisher et al., 1924], which provides a valid test for CI under the assumption of a multivariate Gaussian distribution for (X, Y, Z) . CI testing also plays a central role in the analysis of contingency tables [Agresti, 2012]. In such discrete setting, a widely used example is the Cochran–Mantel–Haenszel test, which accounts for stratification when assessing associations

between categorical variables.

Beyond these classical approaches, the importance of conditional independence has been evident across various scientific domains. Early epidemiological work, such as Cornfield et al. [1959]’s study on smoking and lung cancer, underscored the necessity of adjusting for confounders to distinguish causal relationships from spurious correlations. Moreover, conditional independence plays a crucial role in many statistical applications, including the development of graphical models [Koller and Friedman, 2009]. In modern statistical inference, variable selection [Williamson et al., 2021, Dai et al., 2022]—aimed at identifying the most predictive subset of features for a given response variable—remains a critical component of practice. Collectively, these examples illustrate that conditional independence testing has been a fundamental challenge in statistical theory and practice for over a century, shaping advancements in fields as diverse as social sciences, genetics, and machine learning.

Despite its rich history, testing the conditional independence null remains fundamentally challenging in the absence of additional structural assumptions—precisely the challenge that assumption-lean inference seeks to address.

4.1.1 Hardness of conditional independence testing

Next, we formally present the core impossibility result in CI testing and discuss existing approaches that address this challenge. Suppose $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$ and $\mathcal{Z} = \mathbb{R}^{d_3}$, where $d_1 + d_2 + d_3 = d$. Let $\mathcal{P}_{\mathbb{R}^d}$ be the set of all distributions for $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ supported on $\mathbb{R}^d$, $\mathcal{P}_{\mathbb{R}^d, \text{Cont}} \subset \mathcal{P}_{\mathbb{R}^d}$ be the subset of all distributions that are absolutely continuous with respect to the d -dimensional Lebesgue measure, and let $\mathcal{P}_{K, \text{Cont}} \subset \mathcal{P}_{\mathbb{R}^d, \text{Cont}}$ be the set of all distributions on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure, and are supported within a L_∞ ball of radius K .

Finally, let $\mathcal{P}_{0, K, \text{Cont}} \subset \mathcal{P}_{K, \text{Cont}}$ denote the set of distributions that satisfy the conditional

independence assumption, i.e., $X \perp\!\!\!\perp Y \mid Z$. The following theorem establishes one of the key impossibility results in the context of conditional independence testing.

Theorem 4.1 (Shah and Peters [2020]). *Fix any $n \in \mathbb{N}$, any $\alpha \in (0, 1)$, and $K \in (0, \infty]$. Consider any test¹ $\psi : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \rightarrow [0, 1]$ that controls Type I error at level α for all distributions in $\mathcal{P}_{0,K,\text{Cont}}$, i.e.,*

$$\sup_{P \in \mathcal{P}_{0,K,\text{Cont}}} \mathbb{E}_P [\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha.$$

Then, for any distribution $Q \in \mathcal{P}_{K,\text{Cont}} \setminus \mathcal{P}_{0,K,\text{Cont}}$, it holds that $\mathbb{E}_Q [\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha$.

In other words, no *universally valid* test for the null hypothesis H_0^{CI} can achieve meaningful power against *any* fixed alternative. Moreover, it also does not help to take K finite—in which case the random variables X, Y and Z are all bounded, so we would have all the averages converge asymptotically to Gaussian limits uniformly over $\mathcal{P}_{0,K,\text{Cont}}$; however, as the result states that this does not allow us to construct a non-trivial test for H_0^{CI} . A similar hardness result regarding CI testing is also established in Kim et al. [2022b].

Existing approaches

Existing approaches to testing conditional independence have sought to guarantee validity (Type I error control) over restricted classes of null distributions that impose one of the following additional structures:

- (a) a parametric model, such as assuming joint Gaussianity of (X, Y, Z) , or a Gaussian

1. Traditionally, we think of a hypothesis test ϕ as a map from data to a decision, i.e., $\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \{0, 1\}$. Why, then, do we define tests ϕ as mapping to the space $[0, 1]$? This is because, in some settings, we may want to consider randomized tests—for instance, in the notation above where ϕ maps to $[0, 1]$, an outcome $\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = 0.75$ represents that, given the data, our randomized test rejects the null with probability 0.75. Of course, a nonrandomized test is simply a special case, obtained by restricting the output of ϕ to lie in $\{0, 1\}$.

linear model for $Y \mid (X, Z)$ (e.g., Kalisch and Bühlmann [2007], Sacha Epskamp and Borsboom [2018], Colombo et al. [2012]);

- (b) a known or well-estimated conditional distribution $P_{X|Z}$ [Candès et al., 2018, Barber et al., 2020, Berrett et al., 2020, Niu et al., 2024]; or
- (c) smoothness of the conditional distribution $P_{X|Z}$ [Shah and Peters, 2020, Lundborg et al., 2022, Kim et al., 2022b, Lundborg et al., 2024], e.g., it belongs to a Hölder class.

4.1.2 Isotonic conditional independence

In contrast to the existing approaches in the literature, we consider a nonparametric framework for testing conditional independence by imposing a shape constraint—namely, a form of stochastic monotonicity—on the conditional distribution of $X \mid Z$. We show that this assumption alleviates the inherent difficulty of conditional independence testing, allowing us to construct a general class of tests, called the `PairSwap`-ICI test, which maintains validity under the relaxed null while achieving meaningful power against a broad range of alternatives.

Informally, our shape constraint imposes the assumption that larger values of Z tend to correspond to larger values of X . This constraint is motivated by various applications, particularly in biomedicine, where factors such as age are often associated with increased risk of certain diseases or conditions. Similar trends also appear in economics, where greater professional experience is linked to higher salaries, and in agriculture, where crop yields tend to increase with optimal rainfall levels. Inspired by these real-world patterns, we formally introduce the following constraint:

Assumption 1 (Monotonicity of the conditional distribution $P_{X|Z}$). *Let $\mathcal{X} \subseteq \mathbb{R}$ and let $\preceq$*

be a partial order on $\mathbb{R}^d Z$. We assume that X is stochastically increasing in Z , meaning that

$$\text{if } z \preceq z' \text{ then } \mathbb{P}_P \{X \geq x \mid Z = z\} \leq \mathbb{P}_P \{X \geq x \mid Z = z'\} \text{ for all } x.$$

This assumption does not fall into any of the categories (a)–(c) above. We will often consider the case where the control variable Z is univariate, i.e., $\mathcal{Z} \subseteq \mathbb{R}$, under the usual total order $\leq$ on $\mathbb{R}$. Nevertheless, our framework also allows for multivariate $Z \in \mathbb{R}^d$, in which case the most common partial order is the coordinatewise order.

Based on this shape constraint, we now consider the relaxed testing problem of the isotonic conditional independence (ICI) null hypothesis

$$H_0^{\text{ICI}} : X \perp\!\!\!\perp Y \mid Z, \text{ and } P_{X|Z} \text{ satisfies Assumption 1.} \quad (4.1)$$

Naturally, this relaxation should only be applied in settings where the monotonicity condition of Assumption 1 is well-motivated, so that a rejection of H_0^{ICI} can reasonably be interpreted as evidence that $X \not\perp\!\!\!\perp Y \mid Z$. However, this should not be seen as a limitation of this approach since we again recall that some additional assumption beyond H_0^{CI} is essential for any valid test to have non-trivial power at any alternative.

The idea behind our framework is to consider carefully-chosen pairs of data points (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j) with $Z_i \preceq Z_j$. Under H_0^{ICI} , we expect $X_i \leq X_j$ more often than not, and a substantial violation of this provides evidence of the influence of Y , i.e. evidence against H_0^{ICI} . In Section 4.2, we introduce the **PairSwap**-ICI procedure, which allows the practitioner the flexibility to choose from a broad family of test statistics, and is calibrated by comparison with the same statistic computed on swapped pairs of X values. In particular, the procedure offers a finite-sample guarantee of Type I error control against the null H_0^{ICI} .

Organization of the chapter In Section 4.2, we define a general class of test algorithms for the CI null, called the **PairSwap-ICI** test, which provides practitioners with the flexibility to choose from a broad family of test statistics and is calibrated by comparing the statistic to its counterpart computed on swapped pairs of X values. In Section 4.3, we formally establish Type I error control for this class of tests. Section 4.4 discusses how practitioners can tailor our approach to achieve high power against specific alternatives of interest. Next, in Section 4.5, we study the power properties of the **PairSwap-ICI** test and characterize its asymptotic power via upper and lower bounds under a natural regression model for X . Finally, in Sections 4.6 and 4.7, we validate our findings through numerical simulations and real-data experiments, concluding with a brief discussion in Section 4.8.

4.2 PairSwap-ICI-test: a general approach

First we present our general procedure, called the **PairSwap-ICI** test, for testing this relaxed null H_0^{ICI} . This builds on the intuition outlined earlier, whereby pairs of data points (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j) with large positive differences $X_i - X_j$ and with $Z_i \preceq Z_j$ provide evidence against the null.

More formally, based on $\mathbf{Y}$ and $\mathbf{Z}$, and without access to $\mathbf{X}$, we first choose:

1. A sequence of ordered pairs

$$(i_1, j_1), \dots, (i_L, j_L)$$

of indices in $[n]$, where all $2L$ entries are distinct. We require the pairs to be ordered in the sense that

$$Z_{i_\ell} \preceq Z_{j_\ell} \tag{4.2}$$

for each $\ell \in [L]$. Such a choice of ordered pairs (with any $L \leq \lfloor n/2 \rfloor$) is referred to as a *matching*, and we denote the set of all possible matchings in $[n]$ satisfying (4.2)

as $\mathcal{M}_n(\mathbf{Z})$.

2. A sequence of functions $\psi_1, \dots, \psi_L$, where each $\psi_\ell : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ satisfies the following *anti-monotonicity* property:

Definition 4.1. *A function $\psi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is said to satisfy anti-monotonicity if*

$$\psi(x + \Delta, x' - \Delta') - \psi(x' - \Delta', x + \Delta) \geq \psi(x, x') - \psi(x', x) \quad (4.3)$$

for all x, x' and all $\Delta, \Delta' \geq 0$.

An example of a class of functions ψ that satisfy anti-monotonicity is $\psi(x, x') = f(x - x')$ for any monotone nondecreasing function f that is also *odd* in the sense that $f(x) = -f(-x)$, for instance, $\psi_\ell(x, x') = x - x'$ or $\psi_\ell(x, x') = \text{sign}(x - x')$.

With these choices in place, our test statistic $T : \mathcal{X}^n \rightarrow \mathbb{R}$ is defined as

$$T(\mathbf{x}) = \sum_{\ell=1}^L \psi_\ell(x_{i_\ell}, x_{j_\ell}). \quad (4.4)$$

Here we suppress dependence on $\mathbf{Y}$ and $\mathbf{Z}$ in the notation $T(\mathbf{x})$, even though this statistic does depend on $\mathbf{Y}$ and $\mathbf{Z}$ through the choices of the matched pairs (i_ℓ, j_ℓ) and functions ψ_ℓ .

Now in order to calibrate the test, we compare the observed test statistic $T = T(\mathbf{X})$ with versions of T where indices within pairs (i_ℓ, j_ℓ) are randomly swapped. Specifically, for $\mathbf{s} = (s_1, \dots, s_L) \in \{\pm 1\}^L$, define $T_{\mathbf{s}} = T(\mathbf{X}^{\mathbf{s}})$, where $\mathbf{X}^{\mathbf{s}}$ is a swapped version of the data vector $\mathbf{X}$, with entries

$$\begin{cases} (X_{i_\ell}^{\mathbf{s}}, X_{j_\ell}^{\mathbf{s}}) = (X_{i_\ell}, X_{j_\ell}) & s_\ell = +1, \\ (X_{i_\ell}^{\mathbf{s}}, X_{j_\ell}^{\mathbf{s}}) = (X_{j_\ell}, X_{i_\ell}) & s_\ell = -1. \end{cases}$$

Thus, $s_\ell = -1$ indicates that the random variables X_{i_ℓ} and X_{j_ℓ} are swapped in $(X_{i_\ell}^{\mathbf{s}}, X_{j_\ell}^{\mathbf{s}})$

compared with (X_{i_ℓ}, X_{j_ℓ}) , while $s_\ell = +1$ indicates no swap. Informally, the two constraints (4.2) and (4.3) ensure that, under the null, each $\psi_\ell(X_{i_\ell}, X_{j_\ell})$ is likely to be no larger than its swapped version, $\psi_\ell(X_{j_\ell}, X_{i_\ell})$ —and thus, the statistic $T = T(\mathbf{X})$ is likely to be no larger than its swapped copies, $T_{\mathbf{s}} = T(\mathbf{X}^{\mathbf{s}})$. If instead $T > T_{\mathbf{s}}$ for many swaps $\mathbf{s}$, this indicates evidence against the null. Indeed, we may define

$$p := \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}\{T_{\mathbf{s}} \geq T\}, \quad (4.5)$$

and in Theorem 4.2 we show that this is a valid p -value.

Our method is quite flexible, in that one may select any pairs (i_ℓ, j_ℓ) subject to the ordering constraint (4.2), and any functions ψ_ℓ satisfying anti-monotonicity (4.3), to construct a valid test. In particular, one can decide on these aspects of the test *after* exploring the data $\mathbf{Y}, \mathbf{Z}$, choosing to include a pair (i_ℓ, j_ℓ) if the data observed in $\mathbf{Y}$ indicates that $X_{i_\ell} > X_{j_\ell}$ would be likely under the alternative. However, the quality of the matches (i_ℓ, j_ℓ) and functions ψ_ℓ affects the power of our test; we discuss effective strategies for designing a statistic T in Section 4.4.

4.3 Type I error control

Firstly, we observe that the general class of PairSwap-ICI test indeed has valid level under the null hypothesis H_0^{ICI} . More formally, the following theorem establishes that the p -value defined in (4.5) is super-uniform under H_0^{ICI} .

Theorem 4.2. *Under H_0^{ICI} , the conditional Type I error of the PairSwap-ICI test satisfies $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \alpha$ for all $\alpha \in [0, 1]$. In particular, the test enjoys marginal error control: $\mathbb{P}\{p \leq \alpha\} \leq \alpha$ for all α .*

The p -value constructed in (4.5) requires computing $T_{\mathbf{s}} = T(\mathbf{X}^{\mathbf{s}})$ for all 2^L values of $\mathbf{s} \in \{-1, 1\}^L$, which may be computationally prohibitive for moderate or large L . In practice, it is common to use a Monte Carlo approximation to the p -value: we sample $\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)} \stackrel{\text{iid}}{\sim} \text{Unif}(\{\pm 1\}^L)$, and then compute

$$\hat{p}_M = \frac{1 + \sum_{m=1}^M \mathbb{1} \left\{ T_{\mathbf{s}^{(m)}} \geq T \right\}}{1 + M}$$

The extra ‘1+’ term appearing in the numerator and denominator is necessary to ensure error control for this Monte Carlo version of our test [Davison and Hinkley, 1997, Phipson and Smyth, 2010]; in particular, this correction ensures that we cannot have $\hat{p}_M = 0$. The following theorem verifies that this version of the test also controls the Type I error.

Theorem 4.3. *Fix any $M \in \mathbb{N}$. Under H_0^{ICI} , it holds that $\mathbb{P} \{ \hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \} \leq \alpha$ for all $\alpha \in [0, 1]$, and consequently, $\mathbb{P} \{ \hat{p}_M \leq \alpha \} \leq \alpha$.*

4.4 Designing a powerful PairSwap-ICI test

In this Subsection, we construct a principled, powerful implementation of our test, by designing concrete choices for the pairs (i_ℓ, j_ℓ) and the functions ψ_ℓ introduced in Subection 4.2. Throughout, we will restrict our attention to test statistics $T(\mathbf{x})$ of the form

$$T(\mathbf{x}) = \sum_{\ell=1}^L w_\ell \psi(x_{i_\ell}, x_{j_\ell}) \tag{4.6}$$

for some $w_1, \dots, w_L \geq 0$. That is, in the original definition of the test statistic (4.4), we take $\psi_\ell(\cdot) = w_\ell \psi(\cdot)$ for some vector of non-negative weights $\mathbf{w} = (w_\ell)_{\ell=1}^L$ and some *fixed* kernel ψ , which is required to satisfy the anti-monotonicity property (4.3).

With this simplification, designing a test statistic now requires specifying the kernel ψ ,

deciding which pairs (i_ℓ, j_ℓ) are matched, and finally, how much weight w_ℓ to assign to each pair, as depicted in this flowchart:

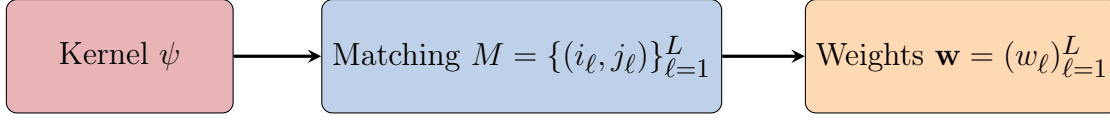


Figure 4.1: The workflow of PairSwap-ICI testing procedure

As guaranteed by Theorem 4.2, our test controls the Type I error for any choice of ψ satisfying (4.3), $M \in \mathcal{M}_n(\mathbf{Z})$, and non-negative weight vector $\mathbf{w}$. However, for the test to be effective, we need to tailor these choices to the specific application of interest. Of course, all of these choices interact with each other: what constitutes a good matching depends on how we choose the weights, and vice versa.

4.4.1 Specifying the kernel ψ

We begin by considering several simple options for the kernel ψ . As a first example, consider $\psi(x, x') = x - x'$. This choice of ψ means that $\psi(X_{i_\ell}, X_{j_\ell})$ is likely to be non-positive under the null (since $Z_{i_\ell} \preceq Z_{j_\ell}$), but under the alternative, may be likely to be positive and large (if the pair (i_ℓ, j_ℓ) is chosen wisely). Alternatively, we may consider nonlinear test statistics to handle a broader range of settings. If X has heavy tails, then the distribution of a linear statistic T can be very sensitive to extreme values. We can ameliorate this sensitivity by using $\psi(x, x') = \text{sign}(x - x')$, or $\psi(x, x') = (-K) \vee (x - x') \wedge K$ for some constant $K > 0$ (i.e., the truncation of $x - x'$ to some bounded interval $[-K, K]$).

Each of the above examples of the kernel ψ satisfies the required anti-monotonicity property (4.3), but actually also satisfies anti-symmetry, which is a strictly stronger property. In fact, we can observe that for any ψ satisfying anti-monotonicity, we can always write

$$\psi(x, x') = \frac{1}{2} \{ \psi(x, x') + \psi(x', x) \} + \frac{1}{2} \{ \psi(x, x') - \psi(x', x) \} =: \psi_{\text{sym}}(x, x') + \psi_{\text{anti}}(x, x'),$$

where by construction, ψ_{sym} is a symmetric function while ψ_{anti} is anti-symmetric. We can observe that ψ_{sym} does not contribute to the p -value—that is, the p -value would be unchanged if we replace ψ with ψ_{anti} (for example, our test with the kernel $\psi(x, x') = x - x'$ coincides with the special case of linear test statistic from Section 4.2). Thus, without loss of generality, from this point on we can assume that we have chosen an anti-symmetric kernel ψ .

4.4.2 Oracle strategies for choosing the matching and weights

We now build intuition for how to choose the matching M and weights $\mathbf{w}$ effectively by sketching the asymptotics of our test, assuming some oracle knowledge (or estimates) of the data distribution; see Subsection 4.5 for a more formal analysis. Let us consider any statistic T of the form (4.6), where (as discussed above) without loss of generality we can take ψ to be an anti-symmetric function. We now wish to choose the weights $\mathbf{w} = (w_\ell)_{\ell \in [L]}$ and matching $M = \{(i_\ell, j_\ell)\}_{\ell \in [L]}$ to maximize the power of our test. Given the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, the reference statistic $T_{\mathbf{s}}$ can be expressed as $T_{\mathbf{s}} = \sum_{\ell=1}^L s_\ell \cdot w_\ell \psi(x_{i_\ell}, x_{j_\ell})$ by construction of the swapping operation, and is therefore a sum of L independent random variables. Under some regularity conditions on the weights $\mathbf{w}$ and the function ψ , a central limit theorem (CLT) approximation gives, for large L , that

$$p \approx \bar{\Phi}(\hat{T}) \quad \text{where} \quad \hat{T} = \hat{T}(\mathbf{w}, M) := \frac{\sum_{\ell=1}^L w_\ell \psi(X_{i_\ell}, X_{j_\ell})}{\sqrt{\sum_{\ell=1}^L w_\ell^2 \psi(X_{i_\ell}, X_{j_\ell})^2}}, \quad (4.7)$$

and where $\bar{\Phi}$ denotes the standard Gaussian survival function, i.e., $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$, where Φ is the standard Gaussian distribution function. The above approximation holds for fixed $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and relies only on the CLT approximation for a weighted sum of L independent signs $s_1, \dots, s_L \in \{\pm 1\}$. We should therefore aim to choose weights that *maximize* the approximate probability of rejection, $\mathbb{P}\{\bar{\Phi}(\hat{T}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$, in order to achieve the best possible

power. Under some conditions this conditional power can be further approximated as

$$\Phi\left(\frac{\sum_{\ell=1}^L w_{\ell} \mathbb{E}\left[\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}\right]}{\sqrt{\sum_{\ell=1}^L w_{\ell}^2 \text{Var}\left(\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}\right)}} - \bar{\Phi}^{-1}(\alpha)\right) \quad (4.8)$$

(see Theorem C.1 in the Appendix for a closer look at this approximation). The following lemma shows how to maximize this approximation over the weights, treating the matching M as fixed.

Lemma 4.4. *Assume that $\text{Var}(\psi_{\ell}(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}) > 0$ for $\ell \in [L]$. Considered as a function of $\mathbf{w} = (w_1, \dots, w_L) \in [0, \infty)^L$, the function in (4.8) is maximized by the choice*

$$w_{\ell}^* = \frac{\max\left\{\mathbb{E}\left[\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}\right], 0\right\}}{\text{Var}(\psi_{\ell}(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z})}, \quad (4.9)$$

for $\ell \in [L]$.

Proof. Writing $u_{\ell} = \mathbb{E}\left[\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}\right]$ and $v_{\ell} = [\text{Var}(\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z})]^{1/2}$ for convenience, for any $w \in [0, \infty)^L$ we calculate

$$\frac{\sum_{\ell=1}^L w_{\ell} u_{\ell}}{\sqrt{\sum_{\ell=1}^L w_{\ell}^2 v_{\ell}^2}} \leq \frac{\sum_{\ell=1}^L w_{\ell} \max\{u_{\ell}, 0\}}{\sqrt{\sum_{\ell=1}^L w_{\ell}^2 v_{\ell}^2}} = \frac{\sum_{\ell=1}^L (w_{\ell} v_{\ell}) \cdot \frac{\max\{u_{\ell}, 0\}}{v_{\ell}}}{\sqrt{\sum_{\ell=1}^L (w_{\ell} v_{\ell})^2}} \leq \left(\sum_{\ell=1}^L \frac{\max\{u_{\ell}, 0\}^2}{v_{\ell}^2}\right)^{1/2},$$

with equality if and only if $w_{\ell} \propto \max\{u_{\ell}, 0\}/v_{\ell}^2$ for $\ell \in [L]$. $\square$

Using a plug-in estimate for the moments. In Lemma 4.4, the oracle weight vector $\mathbf{w}^*$ depends on the conditional expected value and variance, $\mathbb{E}\left[\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}\right]$ and $\text{Var}(\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z})$. For the remainder of this subsection, we will assume that we have access to estimates $\hat{E}_{ij}$ of $\mathbb{E}\left[\psi(X_{i_{\ell}}, X_{j_{\ell}}) \mid \mathbf{Y}, \mathbf{Z}\right]$ and $\hat{V}_{ij} > 0$ of $\text{Var}(\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z})$, for each $i, j \in [n]$, constructed from data that are independent of $\mathbf{X}$. For example, these

might be obtained from a fitted model for the conditional distribution of X given Y, Z based on a separate data set (e.g. via sample splitting). For the linear kernel $\psi(x, x') = x - x'$, this would amount to estimating the first two moments of $X_i \mid Y_i, Z_i$ for $i \in [n]$, whereas if $\psi(x, x') = \text{sign}(x - x')$, then we would need to estimate $\mathbb{P}\{X_i > X_j \mid Y_i, Z_i, Y_j, Z_j\}$ for distinct indices i, j . With these estimates in place, we can seek to maximize the power of our test by choosing weights

$$\hat{w}_\ell = \frac{\max\{\hat{E}_{i_\ell j_\ell}, 0\}}{\hat{V}_{i_\ell j_\ell}} \quad \text{for } \ell \in [L].$$

Choosing the matching. If the oracle weights in (4.9) were known, then the estimator (4.8) of the test's conditional power would be maximized by solving a maximum-weight matching problem, namely that of finding

$$M^* \in \underset{M \in \mathcal{M}_n(\mathbf{Z})}{\operatorname{argmax}} \sum_{(i,j) \in M} (W_{ij}^*)^2 \quad \text{where } W_{ij}^* = \frac{\max\{\mathbb{E}[\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z}], 0\}}{\operatorname{Var}(\psi(X_i, X_j) \mid \mathbf{Y}, \mathbf{Z})}. \quad (4.10)$$

We refer to this version of our procedure as *oracle matching*. In practice, when the moments of the distribution are estimated rather than known, we can instead use the plug-in estimates

$$\hat{M} \in \underset{M \in \mathcal{M}_n(\mathbf{Z})}{\operatorname{argmax}} \sum_{(i,j) \in M} \hat{W}_{ij}^2 \quad \text{where } \hat{W}_{ij} = \frac{\max\{\hat{E}_{ij}, 0\}}{\hat{V}_{ij}}. \quad (4.11)$$

To run **PairSwap-ICI**, we then take weights $w_\ell = \hat{W}_{i_\ell j_\ell}$ for each pair $(i_\ell, j_\ell) \in \hat{M}$. This plug-in matching $\hat{M}$ can be computed efficiently in polynomial time [Edmonds, 1965, Duan and Pettie, 2014] using an algorithm of Gabow [1985].

4.4.3 Heuristic strategies for choosing the matching and weights

In part 4.4.2 above, the test statistic $T(\mathbf{x})$ was designed with the aim of maximising the power of our test, but its success is predicated on being able to estimate well the first two conditional moments of our data. This may be possible if, for instance, we have been able to fit a model using a separate data set. Naturally, an accurate estimate of the true model can enable a very powerful test (since we have a good approximation of the alternative that we are testing against), but in some settings the approach of part 4.4.2 may not be practical, either because we are not able to reliably estimate the required conditional moments, or because the required computations are too costly (in particular, the matching strategy in (4.11)).

In this section, therefore, we take a completely different approach: we will propose an extremely simple scheme for designing a weight vector $\mathbf{w}$, and two easy strategies for choosing a matching M , that do not require extensive prior knowledge or costly calculations. Of course, this will come at some cost in terms of the resulting power of the test, since we are no longer mimicking an oracle test—but, as we will see in both our theoretical guarantees and our empirical results below, these simple strategies can often attain high power nonetheless.

Throughout this section, we will restrict our attention to the univariate setting where $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \subseteq \mathbb{R}$, and will also choose the kernel $\psi(x, x') = x - x'$ when defining our test statistic as in (4.6). We will work in the setting where we hypothesize that, under the alternative, there is a *positive* association between X and Y even after controlling for Z (of course, if our hypothesis is a negative association, we can follow an analogous strategy). The idea is simple: we will take pairs (i_ℓ, j_ℓ) such that

- $Z_{i_\ell} \leq Z_{j_\ell}$ (as required for validity), but $Z_{i_\ell} \approx Z_{j_\ell}$; and
- $Y_{i_\ell} > Y_{j_\ell}$ (so that, under the alternative, we expect $X_{i_\ell} > X_{j_\ell}$).

A simple weighting scheme

We begin by defining a mechanism for choosing the weights: we will take $w_\ell = (Y_{i_\ell} - Y_{j_\ell})_+$, which is simply the difference in Y values, whenever this difference is positive for the pair (i_ℓ, j_ℓ) . Why is this simple strategy a reasonable choice across broad settings? With this choice of weights, our test statistic $T(\mathbf{X})$ is then given by

$$T(\mathbf{X}) = \sum_{\ell: Y_{i_\ell} > Y_{j_\ell}} (Y_{i_\ell} - Y_{j_\ell}) \cdot (X_{i_\ell} - X_{j_\ell}).$$

Since under the alternative, we expect $X_{i_\ell} > X_{j_\ell}$ whenever $Y_{i_\ell} > Y_{j_\ell}$, this means that the expected value of $T(\mathbf{X})$ is positive under the alternative (but, of course, is non-positive under the null).

As another motivation, consider the partially linear model where

$$\mathbb{E}[X \mid Y = y, Z = z] = \beta^* y + \mu_Z^*(z),$$

and the conditional variance is constant,

$$\text{Var}(X \mid Y = y, Z = z) = \sigma^{*2}.$$

Then, following the oracle strategy of part 4.4.2, as in (4.9) we have oracle weights

$$\begin{aligned} w_\ell^* &= \frac{\max\{\mathbb{E}[X_{i_\ell} - X_{j_\ell} \mid \mathbf{Y}, \mathbf{Z}], 0\}}{\text{Var}(X_{i_\ell} - X_{j_\ell} \mid \mathbf{Y}, \mathbf{Z})} = \frac{\max\{\beta^*(Y_{i_\ell} - Y_{j_\ell}) + (\mu_Z^*(Z_{i_\ell}) - \mu_Z^*(Z_{j_\ell})), 0\}}{2\sigma^{*2}} \\ &\approx \frac{\beta^*}{2\sigma^{*2}} \max\{(Y_{i_\ell} - Y_{j_\ell}), 0\}, \end{aligned}$$

where the last step holds since we have assumed $Z_{i_\ell} \approx Z_{j_\ell}$ in our choice of the pair, and so we can expect $\mu_Z^*(Z_{i_\ell}) \approx \mu_Z^*(Z_{j_\ell})$, at least when μ_Z^* is smooth. But crucially, our test

is invariant to rescaling the weights—that is, choosing weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$ is equivalent to choosing weights $\frac{\beta^*}{2\sigma^*2} \max\{(Y_{i_\ell} - Y_{j_\ell}), 0\}$, and thus is nearly equivalent to the oracle weights.

Two simple matching schemes

Next, we propose two matching strategies that, again, do not require any knowledge or estimate of the model.

neighbor matching. Our first simple matching strategy is to choose pairs that are nearest neighbors in the list of sorted Z values, as formalized in Algorithm 2. This strategy ensures that, for all pairs $(i_\ell, j_\ell) = (\pi(2m - 1), \pi(2m))$ that are included in the matching, we have $Z_{i_\ell} \leq Z_{j_\ell}$ and $Y_{i_\ell} > Y_{j_\ell}$ by construction, and moreover, it likely holds that $Z_{i_\ell} \approx Z_{j_\ell}$ (since we have chosen two consecutive Z values in the sorted list). However, an obvious limitation

Algorithm 2 neighbor matching

Preliminaries: sort Z values, i.e., find a permutation π of $[n]$ such that

$$Z_{\pi(1)} \leq Z_{\pi(2)} \leq \cdots \leq Z_{\pi(n-1)} \leq Z_{\pi(n)}.$$

for $m = 1, \dots, \lfloor n/2 \rfloor$ **do**

If $Y_{\pi(2m-1)} > Y_{\pi(2m)}$, then add the new pair $(\pi(2m - 1), \pi(2m))$ to the matching.

end for

of this naïve matching strategy is that many consecutive pairs $(\pi(2m - 1), \pi(2m))$ in the sorted list can fail to have $Y_{\pi(2m-1)} > Y_{\pi(2m)}$ just by chance. In particular, if the Z values in this pair are approximately equal (as we might expect), then the values $Y_{\pi(2m-1)}, Y_{\pi(2m)}$ are expected to be approximately independent and identically distributed—which means that the event $Y_{\pi(2m-1)} > Y_{\pi(2m)}$ will fail approximately half the time. In other words, we are discarding approximately half of the data: we will expect to have $L \approx n/4$ pairs in this matching (meaning that $2L \approx n/2$ data points have been assigned to a matched pair).

Cross-bin matching. Our next strategy is cross-bin matching, which aims to avoid the inefficiency of neighbor matching by using (nearly) all of the data. To implement this strategy, we will partition the list of sorted Z values into K bins, and will allow a pair of data points to be matched as long as the Z values are in adjacent bins (rather than requiring consecutive Z values, as for neighbor matching).

Algorithm 3 Cross-bin matching

Preliminaries: sort Z values, i.e., find a permutation π of $[n]$ such that

$$Z_{\pi(1)} \leq Z_{\pi(2)} \leq \cdots \leq Z_{\pi(n-1)} \leq Z_{\pi(n)},$$

and define K bins of indices,

$$\begin{aligned} A_1 &= \{\pi(1), \dots, \pi(m)\}, \\ A_2 &= \{\pi(m+1), \dots, \pi(2m)\}, \\ &\vdots \\ A_K &= \{\pi((K-1)m+1), \dots, \pi(Km)\}, \end{aligned}$$

where $m = \lfloor n/K \rfloor$.

for $k = 1, \dots, K$ **do**

 Define $r_{k,1}, \dots, r_{k,m}$ as a permutation of A_k such that $Y_{r_{k,1}} \leq \cdots \leq Y_{r_{k,m}}$.

end for

for $k = 1, \dots, K-1$ **do**

for $s = 1, \dots, \lfloor m/2 \rfloor$ **do**

 If $Y_{r_{k,m+1-s}} > Y_{r_{k+1,s}}$, then add the new pair $(r_{k,m+1-s}, r_{k+1,s})$ to the matching.

end for

end for

To explain this procedure in words:

- First we group the Z values into K bins, with each bin A_k containing $m = \lfloor n/K \rfloor$ consecutive Z values.
- We then attempt to match the largest values of Y in bin k with the smallest values of Y in bin $k+1$ —that is, in the inner “for loop” in Algorithm 3, at step $s = 1$ we are attempting to match the largest Y value in bin k (i.e., $Y_{r_{k,m}}$) with the smallest Y

value in bin $k + 1$ (i.e., $Y_{r_{k+1},1}$), and then at step $s = 2$ we proceed to matching the second-largest and second-smallest, and so on.

This scheme is illustrated in Figure 4.2, and is stated formally in Algorithm 3.

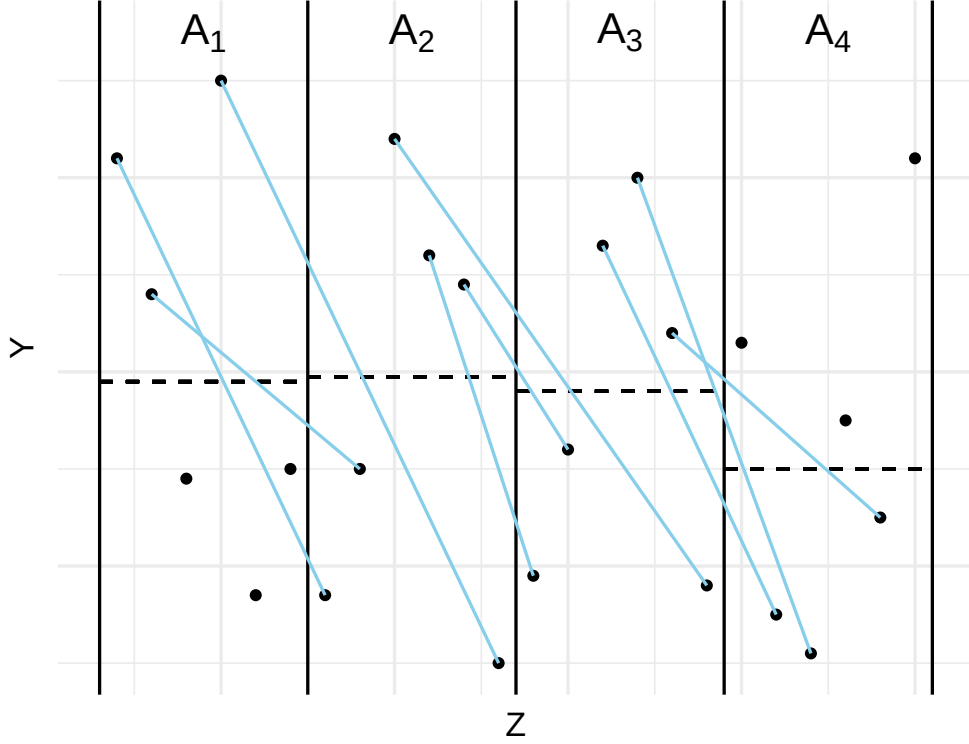


Figure 4.2: Demonstration of the cross-bin matching scheme described in Algorithm 3. The dashed lines indicate the median within each bin.

Why do we expect that this strategy will be more powerful than the simpler neighbor matching strategy? At a high level, while neighbor matching is expected to discard around half of the data, the cross-bin matching strategy can potentially assign nearly all data points to a matched pair. However, there is a potential tradeoff: while the pairs produced by both strategies will likely satisfy $Z_{i_\ell} \approx Z_{j_\ell}$, this approximation will be closer to equality for neighbor matching (where pairs consist of consecutive Z values) than cross-bin matching (where pairs consist of Z values in neighboring bins, i.e., they may be up to $2m$ positions apart in the sorted list). If m is not too large (i.e., the number of bins K is not too small),

though, then we may hope that this difference is negligible. We will examine both methods theoretically in the following section to study these tradeoffs in more detail.

4.5 Power analysis

In this subsection, we study the power of the PairSwap-ICI test under the following general model. We assume that the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (X_i, Y_i, Z_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} P_{X,Y,Z}$ is drawn according to

$$\mathbf{X} = \mu_n(\mathbf{Y}, \mathbf{Z}) + \sigma_n \boldsymbol{\zeta}, \quad (4.12)$$

with $(Y_i, Z_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} P_{Y,Z}$ drawn independently from $(\zeta_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} P_\zeta$. We assume that for any $n \in \mathbb{N}$, $\mu_n : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}$ (applied componentwise) is a measurable function and $\sigma_n > 0$, and also that that P_ζ has mean 0, variance 1, and finite fourth moment. Throughout this subsection, we also assume that the statistic T admits the form in (4.6) with $\psi(x, x') = x - x'$, and we restrict our attention to the setting $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \subseteq \mathbb{R}$; the partial ordering $\preceq$ for Z will simply be the usual ordering $\leq$ on $\mathbb{R}$.

Defining isotonic signal-to-noise ratio (ISNR). Before stating the results on power under this signal plus noise model, we first introduce some notation. We denote by $\mathcal{C}_{\text{ISO}}$ the set of nondecreasing functions on $\mathbb{R}$ and define

$$\text{ISNR}_n = \frac{\inf_{g \in \mathcal{C}_{\text{ISO}}} \mathbb{E}_{P_{Y,Z}^n} [\|\mu_n(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2]}{\sigma_n}, \quad \widehat{\text{ISNR}}_n = \frac{\inf_{g \in \mathcal{C}_{\text{ISO}}} \|\mu_n(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2}{\sigma_n}, \quad (4.13)$$

referring to them as the oracle and empirical *isotonic signal-to-noise ratio* respectively (here $g(\mathbf{Z})$ is again applied componentwise). These quantities capture the extent to which the true model for $X \mid Y, Z$ contains signal that cannot be explained by the isotonic null, H_0^{ICI} . In particular, under H_0^{ICI} , we have $\text{ISNR}_n = \widehat{\text{ISNR}}_n = 0$ (since we can simply take g to be the true regression function, which does not depend on Y), while under the alternative these

quantities might be large. We naturally expect $\widehat{\text{ISNR}}_n \approx \text{ISNR}_n$ when n is large.

Preview of results. To analyze the performance of the **PairSwap**-ICI testing procedure, we study the power conditional on $\mathbf{Y}$ and $\mathbf{Z}$, i.e., $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$. We demonstrate that this power is characterized by the isotonic signal-to-noise ratios ISNR_n and $\widehat{\text{ISNR}}_n$. The organization of this subsection is as follows.

- In part 4.5.1 we study asymptotic upper and lower bounds on the conditional power of oracle matching (as defined in part 4.4.2), and establish that the empirical quantity $\widehat{\text{ISNR}}_n$ governs their behaviour. The same guarantees also hold for the plug-in matching strategy defined in part 4.4.2, as long as we are able to estimate μ_n consistently. Qualitatively, these results show that we can have non-trivial power guarantees against alternatives with large empirical isotonic SNR.
- Next, in part 4.5.2, we show that the converse also holds, i.e., one cannot distinguish the null class from alternatives with small isotonic SNR. More precisely, the conditional power of any valid test is controlled by the oracle quantity ISNR_n . In this sense, the **PairSwap**-ICI test essentially achieves the highest possible power.
- Finally, in part 4.5.3, we specialise our power guarantees to the case of partially linear Gaussian models. We show that even without knowledge of μ_n or an approximation to it, we can achieve near-optimal power with the simpler matching algorithms from part 4.4.3.

4.5.1 Empirical isotonic SNR drives the conditional power of oracle matching

We now study the asymptotic behaviour of the conditional power of the **PairSwap**-ICI procedure with oracle matching. In particular, Theorem 4.5 provides asymptotic upper and

lower bounds on the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$, assuming that the distributions $P_{Y,Z}$ and P_ζ in (4.12) are independent of the sample size n , while the regression function $\mu_n(Y, Z)$ and the noise level σ_n are allowed to depend on n .

Theorem 4.5. *Suppose that $\widehat{\text{ISNR}}_n > 0$ and that $(\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty/\sigma_n) \vee (\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty/\sigma_n)^4 = o_P(\widehat{\text{ISNR}}_n)$. Then, under the model (4.12), the conditional power of the PairSwap-ICI procedure with oracle matching satisfies for $\alpha \in (0, 1)$ that*

$$\Phi\left(\frac{\widehat{\text{ISNR}}_n}{\sqrt{2}} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1) \leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \Phi(\widehat{\text{ISNR}}_n - \bar{\Phi}^{-1}(\alpha)) + o_P(1). \quad (4.14)$$

From Theorem 4.5, we see that the empirical isotonic SNR drives both the upper and lower bound. We also notice that the upper and lower bounds on conditional power match, up to the factor of $\sqrt{2}$ appearing in the expression for the lower bound but not the upper bound. In Theorem 4.11 in part 4.5.3 below, we will see that in a partially linear model setting and under the hypothesis of the symmetry of the distribution $P_{Y|Z}$, it is possible to close this gap.

Remark 2. *The condition $(\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty/\sigma_n) \vee (\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty/\sigma_n)^4 = o_P(\widehat{\text{ISNR}}_n)$ is a natural assumption on the quality of matching. For example, if X is bounded and $\sigma_n = \sigma$ for every $n \in \mathbb{N}$, then $(\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty/\sigma_n) \vee (\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty/\sigma_n)^4 = O(1)$ —and so for this assumption to hold, it is sufficient to have $\widehat{\text{ISNR}}_n \xrightarrow{P} \infty$, i.e., the signal strength diverges to infinity with n .*

Power guarantees for plug-in matching with an estimate of μ_n . Since the oracle matching depends on μ_n , we now shift attention to the more practical setting, where we construct an estimate $\hat{\mu}_n$ of μ_n on one random split of the data, and then implement the PairSwap-ICI test on the remaining data with the plug-in matching $\hat{M}$. While the aforementioned data splitting approach is more accurate and practical, for the sake of simplicity,

in order to state the following result, we assume that we have an independent data set with which we can learn $\hat{\mu}_n$. In this case, Theorem 4.6 below shows that, under suitable consistency assumptions on $\hat{\mu}_n$, we can recover the power guarantees in Theorem 4.5.

Theorem 4.6. *Consider the setting and assumptions of Theorem 4.5. Suppose we use plug-in matching $\hat{M}$ with an estimate $\hat{\mu}_n$ constructed based on independent data, where $\hat{\mu}_n$ satisfies*

$$\frac{\|\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\sigma_n} = o_P(\widehat{\text{ISNR}}_n), \quad \frac{\|\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty^4}{\sigma_n^4} = o_P(\widehat{\text{ISNR}}_n).$$

Then the conditional power of the PairSwap-ICI satisfies (4.14) for $\alpha \in (0, 1)$.

Since the requirements on $\hat{\mu}_n$ in Theorem 4.6 are potentially quite weak, it may be advantageous to employ a data split of unequal sizes when using plug-in matching—in particular, we might choose to use only a small fraction of the available data points to estimate μ_n and compute the weights and matching, and save the rest for the PairSwap-ICI test once these choices are fixed.

4.5.2 Oracle isotonic SNR characterizes hardness of testing the isotonic null

While we have seen that the empirical isotonic SNR governs the asymptotic upper and lower bounds on the power of the PairSwap-ICI test in Theorems 4.5 and 4.6, it turns out that the relationship between the isotonic SNR and power extends beyond any specific choice of test. Here, we establish this connection formally. In particular, we will see that the oracle quantity ISNR_n determines the class of alternatives that are distinguishable from H_0^{ICI} , using *any* valid testing procedure. Towards this goal, we start with a simple total variation calculation to give an upper bound on the power function of any valid test.

Proposition 4.7. Fix $\alpha \in (0, 1)$. Fix any test $\phi : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \rightarrow [0, 1]$ that controls the Type I error at level α , i.e.,

$$\sup_{P \in H_0^{\text{ICI}}} \mathbb{E}_P [\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha.$$

Then for any distribution $P_{X,Y,Z}$,

$$\mathbb{E}_{P_{X,Y,Z}} [\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha + \inf_{Q_{X,Y,Z} \in H_0^{\text{ICI}}} \text{d}_{\text{TV}}(P_{X,Y,Z}^n, Q_{X,Y,Z}^n).$$

This last total variation term is the distance of $P_{X,Y,Z}$ from the null class H_0^{ICI} , meaning the closer P is to null models, the harder it will be to have non-trivial power against P . While in general it is hard to derive exact expressions for the total variation term, under some additional model assumptions on P , we can derive interpretable upper bounds for it. Specifically, consider the model (4.12) where the noise distribution P_ζ has a density function f_ζ satisfying

$$\left\{ \int_{\mathbb{R}} \left(\sqrt{f_\zeta(x)} - \sqrt{f_\zeta(x+t)} \right)^2 dx \right\}^{1/2} \leq L_\zeta |t| \quad \text{for all } t \in \mathbb{R}. \quad (4.15)$$

Note that the left-hand side is equal to $H(P_\zeta, t + P_\zeta)$, where H denotes the Hellinger distance between distributions, and $t + P_\zeta$ denotes the distribution of $t + \zeta$ when $\zeta \sim P_\zeta$ (i.e., a translation of the original distribution P_ζ). In other words, we are assuming that the map $t \mapsto t + P_\zeta$ is Lipschitz with respect to the Hellinger distance. For example, if $P_\zeta = \mathcal{N}(0, 1)$ is the standard Gaussian, then $H^2(P_\zeta, t + P_\zeta) = 2(1 - e^{-t^2/8}) \leq t^2/4$ [Pardo, 2018], and so the assumption (4.15) holds with $L_\zeta = 1/2$.

We now state a result establishing a bound on power for any test, as a function of the isotonic SNR.

Corollary 4.8. *Suppose $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ satisfy (4.12) for some noise distribution P_ζ satisfying the condition (4.15). Then*

$$\inf_{Q_{X,Y,Z} \in H_0^{\text{ICI}}} \text{d}_{\text{TV}} \left(P_{X,Y,Z}^n, Q_{X,Y,Z}^n \right) \leq L_\zeta \cdot \text{ISNR}_n,$$

and consequently, for any test ϕ that controls Type I error at level α ,

$$\mathbb{E}_{P_{X,Y,Z}} [\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha + L_\zeta \cdot \text{ISNR}_n.$$

To summarize, we have seen that under a mild assumption on the noise distribution P_ζ , the oracle isotonic SNR gives a universal upper bound on the power of *any* test of the isotonic null, H_0^{ICI} —no valid test can achieve non-trivial power when ISNR_n is negligible.

4.5.3 Near-optimal power guarantees without knowledge of μ_n

The asymptotic upper and lower bounds on conditional power from Theorems 4.5 and 4.6 require either oracle knowledge of μ_n or an independent estimate of μ_n whose Euclidean norm error is small in probability (as compared to the quantity $\widehat{\text{ISNR}}_n$). If this side information is not available, then a natural question arises: can near-optimal power guarantees be achieved with some of the simpler matching schemes from part 4.4.3? This is indeed possible, under some additional assumptions on the model class (4.12). Specifically, consider the class of partially linear Gaussian models given by

$$(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \text{ satisfy (4.12) with } P_\zeta = \mathcal{N}(0, 1), \mu_n(\mathbf{Y}, \mathbf{Z}) = \mu_0(\mathbf{Z}) + \beta_n \mathbf{Y} \quad (4.16)$$

for some $\mu_0 \in \mathcal{C}_{\text{ISO}}$ (applied componentwise) and some $\beta_n \in (0, \infty)$; we assume also that Y is a mean-zero variable taking values in $[-1, 1]$, with $\mathbb{E}[\text{Var}(Y \mid Z)] > 0$ (i.e., Y is not equal

to a deterministic function of Z). Note that μ_0 is a fixed function, and so the dependence on n of the distribution of (X, Y, Z) is solely through β_n and σ_n . The following lemma relates ISNR_n to a more interpretable quantity, namely the expected conditional variance of Y given Z .

Lemma 4.9. *Under the model class (4.16), provided that $\mathbb{E}|\mu_0(Z)| < \infty$,*

$$\frac{\sqrt{n}\beta_n}{\sigma_n} \cdot (\mathbb{E}[\text{Var}(Y | Z)])^{1/2} (1 + o_P(1)) \leq \text{ISNR}_n \leq \frac{\sqrt{n}\beta_n}{\sigma_n} (\text{Var}(Y))^{1/2} \leq \frac{\sqrt{n}\beta_n}{\sigma_n}.$$

We can draw several conclusions from this result. By Lemma 4.9 and Corollary 4.8, no valid test can achieve asymptotically non-trivial power against an alternative in the model class (4.16) when $\frac{\sqrt{n}\beta_n}{\sigma_n} = o(1)$. On the other hand, if $\frac{\sqrt{n}\beta_n}{\sigma_n} \rightarrow \infty$, then $\text{ISNR}_n \rightarrow \infty$, and the PairSwap-ICI test should have power converging to 1. From this point on, then, we will consider the more interesting regime where

$$n^{-1/2} \lesssim \frac{\beta_n}{\sigma_n} \lesssim 1,$$

to avoid both extremes where our earlier results ensure either trivial power or power tending to 1. We will now develop a more precise characterization of power in this intermediate regime.

Under mild conditions on the model, we also expect that $\widehat{\text{ISNR}}_n = \text{ISNR}_n(1 + o_P(1))$. With this approximation, another consequence of Lemma 4.9 and (4.14) is that the conditional power of oracle matching also satisfies

$$\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \geq \Phi\left(\frac{\sqrt{n}\beta_n \cdot \{\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)]\}^{1/2}}{\sqrt{2}\sigma_n} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1). \quad (4.17)$$

We can now compare the asymptotic conditional power of the PairSwap-ICI test with neighbor matching and cross-bin matching against this lower bound.

Theorem 4.10. *Under the model class (4.16), fix $\alpha \in (0, 1)$, and assume that $\sigma_n \gg \sqrt{\log n/n}$, $1 \gtrsim \beta_n/\sigma_n \gtrsim 1/\sqrt{n}$. Assume also that $\mu_0(Z)$ is a sub-Gaussian random variable. Then the conditional power of neighbor matching (Algorithm 2), implemented with kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$, satisfies*

$$\left| \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} - \Phi\left(\frac{\sqrt{n}\beta_n \cdot \{\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)]\}^{1/2}}{2\sigma_n} - \bar{\Phi}^{-1}(\alpha)\right) \right| = o_P(1).$$

Notably, the conditional power of neighbor matching matches the lower bound from (4.17) up to a factor of $\sqrt{2}$ in the signal-to-noise ratio term. This factor arises due to the inherent inefficiency of neighbor matching: as discussed in part 4.4.3, neighbor matching discards roughly half the data.

We now turn to cross-bin matching. Since we would like paired Z values $Z_{i_\ell} \leq Z_{j_\ell}$ to be approximately equal (in order to avoid an overly conservative test), we further allow the number of bins to depend on sample size n , and write K_n from now on. Consequently, the power of cross-bin matching can then meet the lower bound in (4.17).

Theorem 4.11. *Under the same setting and assumptions as Theorem 4.10, the conditional power of cross-bin matching (Algorithm 3) with K_n bins (where $\sigma_n^{-1}\sqrt{n \log n} \ll K_n \ll n$), implemented with kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$, satisfies*

$$\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \geq \Phi\left(\frac{\sqrt{n}\beta_n \cdot \{\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)]\}^{1/2}}{\sqrt{2}\sigma_n} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1).$$

Further, if the conditional distribution $P_{Y|Z}$ is symmetric about its mean almost surely, then

the conditional power of cross-bin matching satisfies

$$\left| \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} - \Phi\left(\frac{\sqrt{n}\beta_n \cdot \{\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)]\}^{1/2}}{\sigma_n} - \bar{\Phi}^{-1}(\alpha)\right) \right| = o_P(1).$$

In particular, the first part of the result shows that for cross-bin matching, the conditional power asymptotically matches the lower bound in (4.17), without an additional factor of $\sqrt{2}$ as for neighbor matching under the same assumptions. Moreover, with an additional assumption of symmetry of the conditional distributional $P_{Y|Z}$, we can also remove another factor of $\sqrt{2}$. While without the symmetry assumption, the above result provides only an asymptotic lower bound on the conditional power of cross-bin matching, we can characterize the power more exactly; the details are deferred to Theorem C.3 in Appendix C.2.4.

4.6 Simulations

In this section, we evaluate the performance of our method on simulated data², and compare the matching strategies from Section 4.4.3. For simplicity, we focus on the univariate case $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \mathbb{R}$. We will examine two versions of the PairSwap-ICI test:

- neighbor matching (Algorithm 2), with the linear kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$ as discussed in Section 4.4.3;
- Cross-bin matching (Algorithm 3) with $K_n = \lfloor n^{2/3} \rfloor$ bins, again with the linear kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$. This particular choice of K_n is motivated by Theorem 4.11, which suggests choosing the number of bins K_n to satisfy $\sqrt{n \log n} \ll K_n \ll n$, in order to achieve competitive power with neighbor matching.

2. Code for reproducing all experiments can be found at <https://github.com/jake-soloff/PairSwap-ICI-Experiments>.

4.6.1 Extent of conservatism under H_0^{ICI}

Theorem 4.2 establishes valid, finite-sample Type I error control for our method. The purpose of this section is to evaluate the extent of the conservatism of the Type I error under various null distributions. Since our inference relies on the fact that matched pairs (X_{i_ℓ}, X_{j_ℓ}) are stochastically ordered under the null, intuitively this should depend on the strength of the monotonicity in the conditional distribution.

To see how the dependence between X and Z affects the rejection probability, we sample X from an additive noise model

$$X \mid Y, Z \sim \mathcal{N}(\mu(\gamma Z), 1),$$

where Y, Z are independent standard normal random variables. As long as μ is nondecreasing and $\gamma \geq 0$, this joint distribution belongs to the null H_0^{ICI} . The scalar γ controls the strength of the monotonicity of $X \mid Z$. In particular, as $\gamma \searrow 0$ we expect the Type I error $\mathbb{P}\{p \leq \alpha\}$ to approach α .

In our simulations, we consider two functions μ , the identity $\mu(z) = z$ and the standard Gaussian CDF $\mu(z) = \Phi(z)$. Figure 4.3 shows the Type I error as a function of γ for two levels of α . We observe similar results for each α , where the test typically becomes more conservative as γ increases, as expected. Under the null, our test is more conservative for cross-bin matching than for neighbor matching, since the Z values are further apart in cross-bin matching.

4.6.2 Power under alternatives

In Section 4.5.3 we showed theoretically that our heuristic methods—neighbor matching and cross-bin matching—achieve high power (in fact, power that tends to 1) in the partial

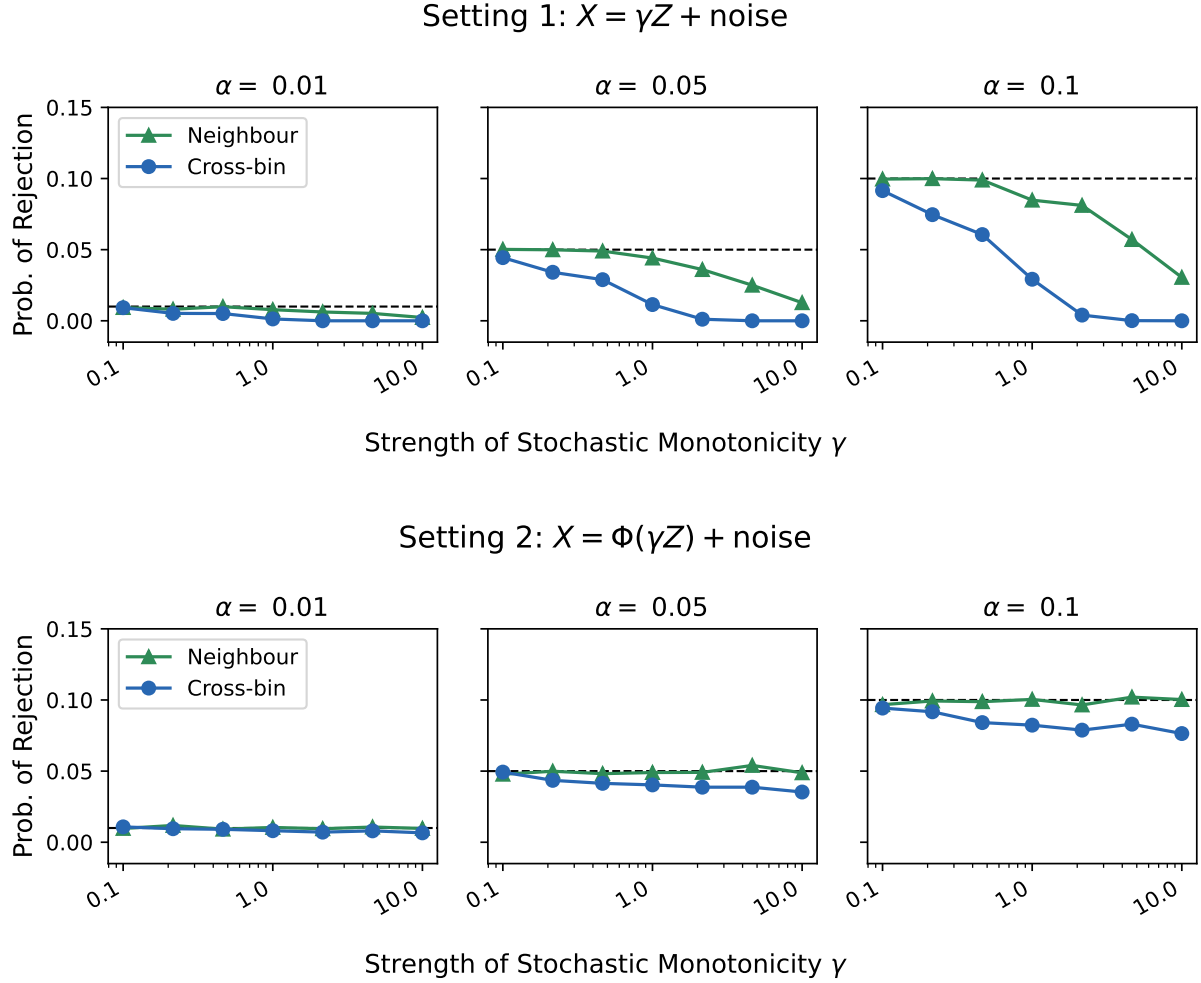


Figure 4.3: Simulation results illustrating Type I error control under the null H_0^{ICI} for two forms of the conditional mean $\mathbb{E}[X | Z]$. Each subplot shows the rejection probability of the PairSwap-ICI test on a data set of size 1000, averaged over 10^4 simulation trials, as a function of the strength of stochastic monotonicity γ .

linear model (4.16) with $\sigma_n = 1$ for all n , provided the signal β_n exceeds the detection threshold $n^{-1/2}$. In this section, we now examine this setting empirically. We sample data from the Gaussian linear model

$$X | Y, Z \sim \mathcal{N}(\beta_n Y + \gamma Z, 1),$$

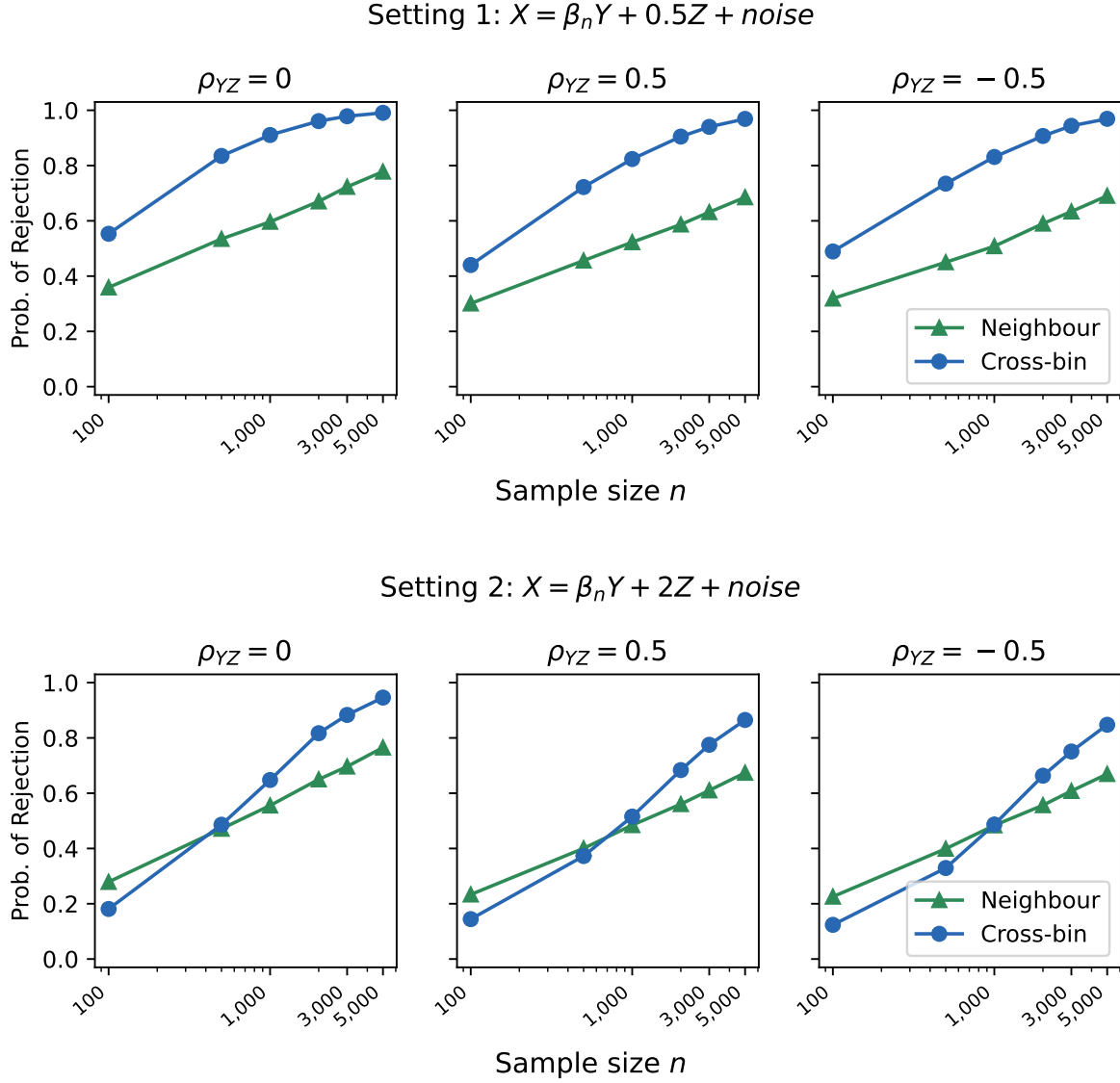


Figure 4.4: Simulation results demonstrating the power of the PairSwap-ICI for two alternatives at level $\alpha = 0.1$. Each subplot shows the rejection probability, averaged over 10^4 simulation trials, as a function of the sample size n . Columns correspond to different relationships between Y and Z . In each setting, X follows a Gaussian linear model with mean $\beta_n Y + \gamma Z$, where $\beta_n = n^{-1/3}$ and $\gamma = 0.5$ (above) or $\gamma = 2$ (below).

with $\beta_n = n^{-1/3}$. The pair (Y, Z) is drawn from a bivariate Gaussian

$$\mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} 1 & \rho_{YZ} \\ \rho_{YZ} & 1 \end{bmatrix}\right).$$

Figure 4.4 shows the (unconditional) power as a function of the sample size n for various choices of γ and ρ_{YZ} . In Setting 1, we set $\gamma = 0.5$, and cross-bin matching uniformly dominates neighbor matching because it allows us to make many more matches of similar quality. On the other hand, in Setting 2 we set $\gamma = 2$, so the strong dependence of X on Z means the quality of a match (i_ℓ, j_ℓ) degrades much more quickly as the gap $Z_{j_\ell} - Z_{i_\ell}$ increases—that is, for cross-bin matching, where Z_{i_ℓ} and Z_{j_ℓ} may be farther apart than for neighbor matching, this gap may lead to conservatism that results in a loss of power. However, with a sufficiently large sample size, cross-bin matching performs at least as well as the neighbor matching. This is because the bin width decreases as n increases, so the quality of the cross-bin matches rivals that of the neighbor matches (with many more matches). The dependence ρ_{YZ} between Y and Z does not have a major impact on the power of these two methods.

One take-away message from these simulations is that the stronger the stochastic monotonicity is believed to be, the more bins one should use in cross-bin matching, to avoid excessive conservatism. On the other hand, if the extent of stochastic monotonicity is believed to be weak, then one should use fewer bins in cross-bin matching, to allow for a greater number of matches as well as larger weights for the matched pairs, and thereby increase power.

4.7 Experiment on real data: risk factors for diabetes

We now evaluate the performance of our proposed testing procedure on a real data set in three different experimental settings.

We use a data set³ on the incidence of diabetes among the Pima population near Phoenix, Arizona, originally collected by the US National Institute of Diabetes and Digestive and

3. The data for this experiment were obtained from <https://archive.ics.uci.edu/dataset/34/diabetes>. Additional data descriptions can be found in Smith et al. [1988].

Kidney Diseases. The data set comprises 768 observations, and includes information on whether each of the patient has been diagnosed with diabetes according to World Health Organization standards. Additional variables provide data on the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin levels, body mass index (BMI), diabetes pedigree function and age.

It is well-known that the likelihood of developing diabetes increases with age (e.g., the Centers for Disease Control and Prevention⁴ lists advanced age as one of the risk factors for Type 1 and Type 2 diabetes). Therefore, if we choose X to represent the incidence of diabetes and Z as the age of the patient, then we would expect X to exhibit stochastic monotonicity with respect to Z (i.e., we expect that Assumption 1 holds, at least approximately). This is supported by the increasing trend we observe in Figure 4.5. Most of the other variables, such as `BloodPressure`, `BMI`, `Glucose`, and `Pregnancies`, are also considered potential risk factors for diabetes, as visualized in Figure 4.6—but does this association remain after we control for age? In this experiment, we aim to determine whether these variables remain significant risk factors for diabetes, even after controlling for age.

Experiment 1: marginal independence testing: In our first experiment, we consider six variables: `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI`, and aim to assess whether each of them is an individual risk factor for diabetes incidence. Specifically, we test the hypothesis $H_0 : X \perp\!\!\!\perp Y$, where Y represents one of the six variables listed above, while X is `Diabetes` (and since we are testing marginal rather than conditional independence, we do not attempt to control for Z , i.e., `Age`). For this purpose, we will be using the permutation test for independence with $T(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^\top \mathbf{Y}$, as outlined in Section 1.2.1.

4. For more details, see list of diabetes risk factors from the US Centers for Disease Control and Prevention.

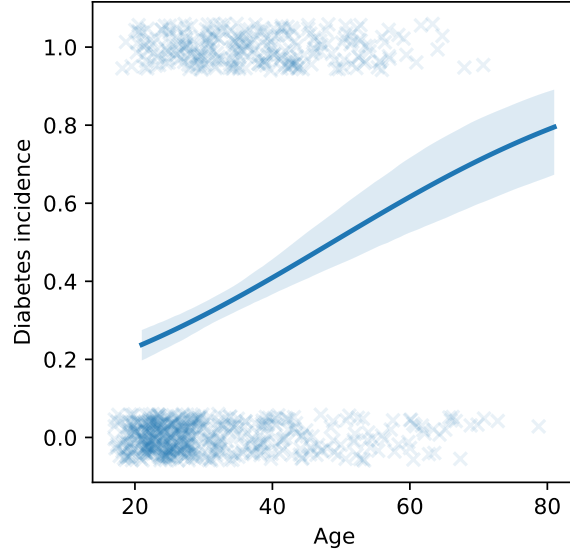


Figure 4.5: A scatter plot (jittered for better visibility) of `Age` and `Diabetes Incidence` along with the fitted logistic regression model to demonstrate the stochastic monotonicity.

Experiment 2: conditional independence testing, after controlling for age: Next, for the same set of six choices for Y , we test the hypothesis $H_0^{\text{ICI}} : X \perp\!\!\!\perp Y \mid Z$, where Z denotes `Age` (and X is `Diabetes` as before). This allows us to identify risk factors for diabetes after controlling for age. As noted earlier, we expect the distribution of $X \mid Z = z$ to be stochastically monotone in z , which supports the application of the `PairSwap-ICI` testing procedure developed in this chapter for this purpose. The `PairSwap-ICI` method is implemented with neighbor matching or with cross-bin matching (with $K = 50$ bins), and our statistic T takes the form in (4.6) with linear kernel $\psi(x, x') = x - x'$ and weights $w_\ell = \max\{Y_{i_\ell} - Y_{j_\ell}, 0\}$.

Experiment 3: conditional independence testing, with synthetic control $\tilde{X}$: Finally, we consider a semi-synthetic experiment where X is replaced by synthetic observations $\tilde{X}$, generated from an estimated model for $P_{X|Z}$ that satisfies stochastic monotonicity. We then test the hypothesis $\tilde{X} \perp\!\!\!\perp Y \mid Z$ for the same choices of Y from Experiment 1. Since

$\tilde{X}$ is generated solely based on Z and obeys stochastic monotonicity, the null hypothesis H_0^{ICI} holds trivially in this synthetic setting. The validity of our procedure should therefore ensure that the p -values generated by PairSwap-ICI are (super-)uniformly distributed.

To generate the synthetic feature $\tilde{X}$, since X is binary, it suffices to fit an isotonic regression to estimate the conditional mean $\mathbb{E}[X | Z]$ and then sample $\tilde{X}$ from the Bernoulli distribution with this fitted conditional mean. Following the theory of Henzi et al. [2021, Theorem 1], this is the best approximation to $P_{X|Z}$ under the *continuous ranked probability score (CRPS)*, while respecting the monotonicity constraint.

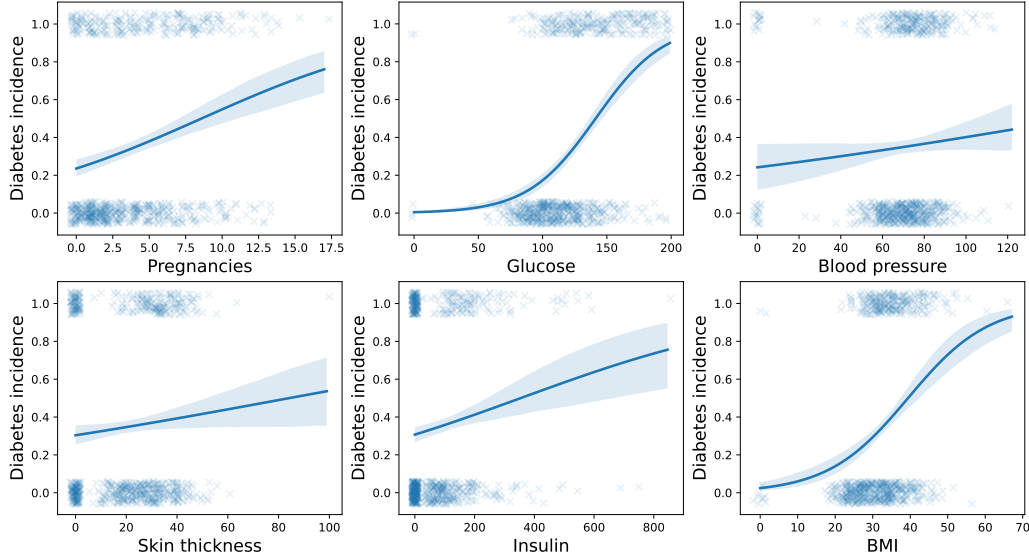


Figure 4.6: Scatter plots of X (jittered for better visibility) and other feature variables along with the fitted logistic regression models to demonstrate the dependence among these variables and **Diabetes Incidence**.

Results: For each experiment, we generate 3,000 random subsamples of the data, each consisting of half the size of the full data set. (For Experiment 3, since constructing a synthetic control requires estimating $P_{X|Z}$, we use the complement of the subsampled data set for this preliminary estimation problem.) We then compute p -values using the permutation

		Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI
Permutation test (testing marginal indep.)		0.001 (0.000)	0.001 (0.000)	0.155(0.003)	0.117(0.002)	0.023 (0.001)	0.001 (0.000)
PairSwap-ICI	neighbor matching	0.433(0.005)	0.002 (0.000)	0.496(0.005)	0.243(0.004)	0.159(0.003)	0.033 (0.001)
	cross-bin matching	0.424(0.004)	0.001 (0.000)	0.499(0.004)	0.123(0.003)	0.224(0.004)	0.001 (0.000)
PairSwap-ICI (w/ synthetic control)	neighbor matching	0.511(0.005)	0.499(0.005)	0.501(0.005)	0.506(0.005)	0.497(0.005)	0.499(0.005)
	cross-bin matching	0.544(0.005)	0.550(0.005)	0.555(0.005)	0.558(0.005)	0.545(0.005)	0.555(0.005)

Table 4.1: p -values, averaged over random 3,000 subsamples along with the standard errors (within brackets) for the different tests from different experiments, as outlined in Section 4.7. The p -values significant at the 0.05 level are marked in bold.

test for marginal independence and the **PairSwap-ICI** test for conditional independence. For all the experiments, p -values are computed on the test set. Finally, we report the average p -values from the 3,000 subsamples, along with the corresponding choice of the Y variable in Table 4.1.

Under the marginal independence test, four of the six variables are identified as having significant association with **Diabetes**, but, once we test conditional independence with the **PairSwap-ICI** test, only two of these associations are identified as significant. Specifically, **Glucose** and **BMI** both are identified as potential risk factors at the 0.05 level of significance by the marginal independence test, and also by the **PairSwap-ICI** test, even after controlling for **Age**. On the other hand, the variables **Pregnancies** and **Insulin** are significant only under the marginal test; this suggests that, after controlling for **Age**, the data does not provide sufficient evidence to support them as risk factors for **Diabetes**.

Finally, we also note that all the averaged p -values from Experiment 3 with synthetic control $\tilde{X}$ are concentrated around 0.5, for each of the choices of Y . Since $(\tilde{X}, Y, Z)$ satisfy H_0^{ICI} , the p -values from Experiment 3 should be roughly uniform (or, if the test is conservative, super-uniform), and thus this behaviour is expected from Theorem 4.2.

4.8 Discussion

In this chapter, we have developed a nonparametric test of conditional independence assuming only stochastic monotonicity of the conditional distribution $P_{X|Z}$. This nonparametric constraint is natural in many applications, and allows us to circumvent the impossibility of assumption-free conditional independence testing [Shah and Peters, 2020]. We have introduced a variety of approaches to constructing a valid test statistic. Our test controls the Type I error in finite samples and has power against an array of alternatives. We close with some connections to the literature, and potential avenues for future work.

- *Optimal power in general settings.* In Theorem 4.11, we were able to eradicate the earlier gap between the upper and lower bounds on the asymptotic conditional power of our test under the assumption that the conditional distribution of $Y | Z$ is symmetric, but it remains an open question whether other tests (or, perhaps, the **PairSwap**-ICI but with a different kernel) may be able to avoid this assumption.
- *Alternative methods for finite-sample Type I error control?* While the stochastic monotonicity assumption facilitates consistent estimation of the conditional distribution $P_{X|Z}$ using isotonic distributional regression [Mösching and Dümbgen, 2020, Henzi et al., 2021], and then applying a conditional independence test which assumes knowledge (or a good estimate) of $P_{X|Z}$ [Berrett et al., 2020, Candès et al., 2018], these tests will only be valid asymptotically. However, under the same setting, the proposed **PairSwap**-ICI test enjoys finite sample Type I error control without any further restrictions on the choice of matching, or weighting. An interesting open question would be to investigate whether these established approaches for CI testing can also be modified to attain validity in finite samples under the monotonicity assumption.
- *Avoiding data splitting.* The oracle matching test derived in Section 4.4 requires mod-

elling the conditional mean and conditional variance of the kernel $\psi(X_i, X_j)$ as a function of Y_i, Y_j, Z_i, Z_j . We proposed to estimate these moments on a hold-out data set. Can we instead perform cross-fitting to improve power and retain finite-sample error control?

- *Connection with knockoffs and conditional randomization tests.* Creating synthetic copies of $\mathbf{X}$ via pairwise swaps resembles other conditional independence testing procedures, such as knockoffs and the conditional randomization test [Candès et al., 2018], and the conditional permutation test [Berrett et al., 2020]. An advantage of the PairSwap-ICI test is that we do not require knowledge of $P_{X|Z}$ apart from its stochastic monotonicity, though the potential price is that the test may be more conservative (i.e., the resulting p -value may be super-uniform), since we are not working under the “sharp null”.
- *Alternative shape constraints.* We view stochastic monotonicity as a natural form of positive dependence for the joint distribution (X, Z) . Are there approaches to test conditional independence under other models of dependence, such as likelihood ratio ordering or total positivity, or under other shape constraints, such as unimodality of $P_{X|Z}$ [Karlin, 1968, Shaked and Shanthikumar, 2007, Mösching and Dümbgen, 2024]?

CHAPTER 5

ASSUMPTION-LEAN INFERENCE: BROADER APPLICATIONS AND INSIGHTS

In this thesis, we tackled the challenge of assumption-lean inference in predictive inference and conditional independence testing. Firstly, we extended the conformal prediction framework to the setting of survival analysis, where right-censoring complicates distribution-free inference for the true survival times. We introduced a covariate-dependent cutoff function $f(X)$ for the censoring time and constructed the prediction interval using only those samples with censoring times larger than $f(X)$. This approach allows the method to adapt to the heterogeneity in the censoring mechanism, leading to improved power over existing assumption-lean methods in the literature. Our results show that if either the censoring mechanism or the conditional quantile of survival time is well estimated, our proposed procedure achieves nearly exact marginal coverage, with the latter also ensuring approximate conditional coverage.

Next, we introduced randomly-localized conformal prediction (RLCP), a novel framework that provides relaxed local coverage guarantees and robustness against smooth covariate shifts. While existing statistical methods typically rely on smoothness assumptions for inference, we take a different approach: instead of imposing smoothness conditions on P , we introduce a synthetic covariate $\tilde{X} \mid X \sim H(X, \cdot)$, which enables predictive inference using the smoothed conditional distribution $P_{Y|\tilde{X}} = \mathbb{E}_{X|\tilde{X}}[P_{Y|X}]$. This approach circumvents the challenges of learning the potentially irregular $P_{Y|X}$ in practice—allowing us to establish results such as Theorem 3.6, where we are able to quantitatively characterize how distributional smoothness impacts performance.

While enforcing smoothness in the space of conditional distributions helps mitigate the

challenges of assumption-lean inference, an alternative and complementary approach is to impose shape constraints on the distributional space. We demonstrate this principle in conditional independence testing through the **PairSwap-ICI** framework, which constructs a powerful test with valid Type I error control under the assumption that the conditional distribution $P_{X|Z=z}$ is stochastically monotone in z . Intuitively, both approaches enhance the effective sample size; whereas smoothness enables local aggregation of information, shape constraints facilitate the aggregation of relative information across any partially ordered space.

These strategies—leveraging smoothness or shape constraints—are not limited to the settings explored in this thesis. Below, we illustrate how these ideas can be further extended to construct distribution-free confidence sets, demonstrating their broader applicability in statistical inference.

5.1 Beyond this thesis: assumption-lean confidence sets

Suppose that $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{iid}{\sim} P$ for some unknown distribution P supported on $\mathcal{X} \times [0, 1]$. We refer to $x \mapsto \mathbb{E}_P[Y \mid X = x]$ as the conditional mean function. Our goal is to construct a distribution-free confidence set (CS) for the conditional mean function, as formally defined below.

Definition 5.1. *We define $\hat{C}_n \equiv \hat{C}_n((X_1, Y_1), \dots, (X_n, Y_n))$ as a distribution-free confidence set for the conditional mean function if, for all distribution P , it holds that*

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

where $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{iid}{\sim} P$.

As observed earlier in the context of predictive inference and conditional independence test-

ing, here as well the conditioning variable determines the difficulty of constructing a confidence set for the corresponding mean function. At one extreme, if $X \equiv 1$ (i.e., trivial conditioning), then our target is the marginal mean of Y , i.e., $\mathbb{E}[Y_{n+1}]$. In this case, we can define the confidence set as

$$\hat{C}_n = \left(\bar{Y}_n \pm \frac{1}{2\sqrt{n\alpha}} \right),$$

where $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ is the empirical mean of $\{Y_i\}_{i \in [n]}$. By Chebyshev's inequality, one can argue that this is a distribution-free valid CS for the marginal mean.

On the other hand, the task becomes significantly more challenging when our target becomes the full conditional mean, $\mathbb{E}[Y_{n+1} \mid X_{n+1}]$. While the length of $\hat{C}_n$ defined above shrinks at a parametric rate of $1/\sqrt{n}$, constructing a distribution-free valid CS for the full conditional mean, whose length diminishes with sample size, is theoretically impossible whenever P_X is non-atomic. For more details, we refer the reader to Barber et al. [2021, Theorem 1]; Barber [2020, Theorem 2].¹ Naturally, the statistical literature has sought a meaningful middle ground between the marginal and full conditional version of the problem.

As observed earlier in the settings of predictive inference and conditional independence testing, both smoothness assumptions and shape constraints play pivotal roles in mitigating these inherent challenges. Building on this idea, we show below how the strategies from Chapter 3 and Chapter 4 introduce a relaxed variant of the problem, enabling a straightforward approach for constructing approximately valid distribution-free confidence set for the conditional mean function.

1. Note that in this referenced paper, $\mathcal{Y} = \{0, 1\}$. To apply their hardness result in our context, we can consider $Y_{\text{bin}} \sim Y$ and extend their conclusion accordingly.

5.2 Leveraging smoothness: random localization for assumption-lean confidence sets

We first illustrate how the workflow of Section 3.3 from Chapter 3 can be adapted to construct a distribution-free confidence set. As in the RLCP framework, we start with generating a synthetic covariate $\tilde{X}_{n+1} \sim H(X_{n+1}, \cdot)$. Instead of constructing a distribution-free confidence set for $\mathbb{E}[Y_{n+1} \mid X_{n+1}]$ as in Definition 5.1, we attempt to do so for $\mathbb{E}[Y_{n+1} \mid \tilde{X}_{n+1}]$. This problem is significantly more tractable since the introduction of the synthetic covariate ensures that

$$\mathbb{E}[Y_{n+1} \mid \tilde{X}_{n+1} = \tilde{x}] = \mathbb{E}_{X_{n+1} \mid \tilde{X}_{n+1} = \tilde{x}}[\mathbb{E}[Y_{n+1} \mid X_{n+1}]],$$

effectively smoothing out the irregularities of the conditional mean function.

For simplicity, suppose $\mathcal{X} = \mathbb{R}^d$, and let H take the form

$$H(x, x'; h) = \frac{1}{h^d} \psi\left(\frac{\|x - x'\|_2}{h}\right), \quad (5.1)$$

for some d -dimensional kernel ψ (e.g., gaussian or box kernel), and a bandwidth parameter $h > 0$. We show that one can construct a distribution-free confidence set (CS) for $\mathbb{E}[Y_{n+1} \mid \tilde{X}_{n+1}]$, which also exhibits diminishing length, although typically at a slower rate than the parametric rate, depending on the choice of h .

Theorem 5.1. *Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{iid}{\sim} P$, and $\tilde{X}_{n+1} \mid X_{n+1} \sim H(X_{n+1}, \cdot; h)$. Suppose that H admits the form given in (5.1) for some bounded kernel $\psi : \mathbb{R}^d \rightarrow [0, B]$. Then the confidence set, given by*

$$\hat{C}_{n,\delta}(\tilde{x}) = \left(\frac{\sum_{i=1}^n Y_i H(X_i, \tilde{x}; h)}{\sum_{i=1}^n H(X_i, \tilde{x}; h)} \pm \frac{4\sqrt{2}B}{h^d \sqrt{n\alpha} \delta} \right)$$

satisfies the following: for any $\alpha, \delta \in (0, 1)$ and $n \geq N_{\delta, h} := \frac{8B^2}{h^{2d}\delta^2\alpha}$, we have

$$\mathbb{P} \left\{ \mathbb{E} \left[Y_{n+1} \mid \tilde{X}_{n+1} \right] \in \hat{C}_{n, \delta}(\tilde{X}_{n+1}) \right\} \geq 1 - \alpha - \mathbb{P}_{\tilde{X}_{n+1}} \left\{ \mathbb{E}_{X \sim P_X} \left[H(X, \tilde{X}_{n+1}; h) \right] < \delta \right\}.$$

The proof of this result is deferred to Appendix D. We note that $\mathbb{E}_{X \sim P_X} [H(X, \tilde{x}; h)]$ represents the density of $\tilde{X}$ at $\tilde{x}$, meaning that the offset term can be made arbitrarily small by choosing δ appropriately. Moreover, we typically expect that $\mathbb{E} [Y_{n+1} \mid \tilde{X}_{n+1}] \approx \mathbb{E} [Y_{n+1} \mid X_{n+1}]$, which allows for a distribution-free confidence set for the latter in *well-behaved* settings (e.g., when the conditional mean function is Lipschitz-smooth).

5.3 Leveraging shape constraints: monotonicity for assumption-lean confidence sets

Next, following the principles of Chapter 4, we demonstrate how a shape constraint can alleviate the difficulty of constructing assumption-lean confidence sets for the conditional mean function. More specifically, suppose $\mathcal{X} = \mathbb{R}^d$, and define the partial order $x \leq y$ for $x, y \in \mathbb{R}^d$ as $x_i \leq y_i$ for all $i \in [d]$. Now, we introduce our shape constraint, which assumes that the conditional mean function $\mathbb{E}[Y \mid X = x]$ is monotonically non-decreasing in x with respect to this partial order. This assumption is natural in many real-world settings, such as economics (where higher income levels often correlate with increased spending), healthcare (where risk factors like age or blood pressure tend to increase disease likelihood).

Under this shape constraint, consider the *one-sided hyperrectangle* $H_{\infty}^{-}(x, h) = \prod_{i=1}^d [x_i - h, x_i]$, where $[a, b]$ denotes the closed interval with endpoints $a, b \in \mathbb{R}$. By construction, we have

$$\mathbb{E}[Y \mid X = x] \geq \mathbb{E}[Y \mid X \in H_{\infty}^{-}(x, h)] \quad \text{for any } x \in \mathbb{R}^d \text{ and any } h > 0. \quad (5.2)$$

This structure enables the construction of a distribution-free lower bound (LB) for the conditional mean function, which we now formally define.

Definition 5.2. We define $\hat{L}_n \equiv \hat{L}_n((X_1, Y_1), \dots, (X_n, Y_n))$ as a distribution-free lower bound for the conditional mean function if, for any distribution P , it satisfies

$$\mathbb{P}_{P^{n+1}} \left\{ \mathbb{E}[Y_{n+1} \mid X_{n+1}] \geq \hat{L}_n(X_{n+1}) \right\} \geq 1 - \alpha,$$

where $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{iid}{\sim} P$.

The inherent difficulty in constructing assumption-lean confidence sets also implies a challenge in obtaining assumption-lean lower bounds. As before, instead of directly constructing a distribution-free LB for $\mathbb{E}[Y_{n+1} \mid X_{n+1}]$, we work with $\mathbb{E}[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h)]$, where the expectation is taken over $(\tilde{X}_{n+1}, \tilde{Y}_{n+1})$, an independent sample drawn from P . This relaxation smooths out irregularities in the conditional mean function, making the problem more tractable. Below, we establish a distribution-free lower bound for this quantity.

Theorem 5.2. Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}), (\tilde{X}_{n+1}, \tilde{Y}_{n+1}) \stackrel{iid}{\sim} P$. Then the lower bound, given by

$$\hat{L}_n(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i \in H_\infty^-(x, h)\}}{\sum_{i=1}^n \mathbb{1}\{X_i \in H_\infty^-(x, h)\}} - \frac{2\sqrt{2}}{h^d \sqrt{n\alpha} \delta}$$

satisfies the following: for any $\alpha, \delta \in (0, 1)$ and any $h > 0$, we have

$$\begin{aligned} \mathbb{P} \left\{ \mathbb{E}[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h)] \geq \hat{L}_n(X_{n+1}) \right\} \\ \geq 1 - \alpha - \mathbb{P}_{X_{n+1}} \left\{ \mathbb{P}_{X \sim P_X} \{X \in H_\infty^-(X_{n+1}, h)\} < \delta h^d \right\}. \end{aligned}$$

The proof of this result is deferred to Appendix D. As in Theorem 5.1, the offset term plays the same role and can be made small by choosing δ suitably in ‘nice’ settings. By (5.2),

it immediately follows that $\hat{L}_n(X_{n+1})$ is also a valid distribution-free lower bound for the conditional mean function.

The role of monotonicity in conditional mean function for assumption-lean inference in regression has also been explored in Müller et al. [2024], where the authors address a related problem of subgroup selection, demonstrating that such shape constraints facilitate valid inference without additional model assumptions.

5.4 Conclusion

We have demonstrated how the principles from Chapter 3 and Chapter 4 can be readily adapted to construct assumption-lean confidence sets for the conditional mean function. Despite their different formulations, both approaches lead to very similar results: the confidence set shrinks at a rate analogous to how the lower bound converges toward the relaxed-conditional mean. Additionally, in both cases, low-mass regions introduce a coverage offset, as seen in Theorems 5.1 and 5.2, dictating the fundamental trade-off associated with the choice of h . Specifically, a larger h introduces more relaxation, smoothing the problem but potentially increasing bias, whereas a smaller h reduces bias but amplifies the challenge of low effective sample size—reflecting the inherent difficulty of assumption-lean inference.

Beyond the specific settings explored in this thesis, these principles can be extended to a broader range of statistical inference tasks. By leveraging smoothness, shape constraints, or other structural assumptions in an assumption-lean framework, we can continue to push the boundaries of distribution-free methods, bridging the gap between theoretical guarantees and practical applicability.

REFERENCES

- Ismail Abbas, Joan Rovira, and Josep Casanovas. Clinical trial optimization: Monte carlo simulation markov model for planning clinical trials recruitment. *Contemporary clinical trials*, 28(3):220–231, 2007.
- Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.
- John Aitchison and Ian Robert Dunsmore. *Statistical prediction analysis*. CUP Archive, 1980.
- Mélanie Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on hsc measures. *The Annals of Statistics*, 50(2):858–879, 2022.
- Paul D Allison. *Event history analysis: Regression for longitudinal event data*. Number 46. Sage, 1984.
- Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.
- Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.
- Vladimir V Anisimov and Valerii V Fedorov. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in medicine*, 26(27):4958–4975, 2007.
- Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.
- Lee Bain. *Statistical analysis of reliability and life-testing models: theory and methods*. Routledge, 2017.
- Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.
- Rina F Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 2019a.
- Rina Foygel Barber. Is distribution-free inference possible for binary regression? 2020.
- Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, 2019b.

- Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431, 2020.
- Rina Foygel Barber, Emmanuel J Candès, 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, 2021.
- Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.
- Richard E Barlow and Frank Proschan. *Statistical Theory of Reliability and Life Testing: Probability Models*, volume 1. Holt, Rinehart and Winston New York, 1975.
- Katharine D Barnard, Louise Dent, and Andrew Cook. A systematic review of models to predict recruitment to multicentre clinical trials. *BMC medical research methodology*, 10(1):1–8, 2010.
- Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6): 1–34, 2021.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001. ISSN 0090-5364,2168-8966. doi:10.1214/aos/1013699998. URL <https://doi.org/10.1214/aos/1013699998>.
- Thomas B Berrett and Richard J Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.
- Thomas B Berrett and Richard J Samworth. Usp: an independence test that improves on pearson’s chi-squared and the g -test. *Proceedings of the Royal Society A*, 477(2256): 20210549, 2021.
- Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(1):175–197, 2020. ISSN 1369-7412,1467-9868.
- Thomas B Berrett, Ioannis Kontoyiannis, and Richard J Samworth. Optimal rates for independence testing via u -statistic permutation tests. *The Annals of Statistics*, 49(5): 2457–2490, 2021.
- Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN

9780199535255. doi:10.1093/acprof:oso/9780199535255.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>.

Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: ‘model- X ’ knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(3):551–577, 2018. ISSN 1369-7412,1467-9868. doi:10.1111/rssb.12265. URL <https://doi.org/10.1111/rssb.12265>.

Emmanuel J Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

Rickey E Carter, Susan C Sonne, and Kathleen T Brady. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC medical research methodology*, 5(1):1–5, 2005.

Rickey Edward Carter. Application of stochastic processes to participant recruitment in clinical trials. *Controlled clinical trials*, 25(5):429–436, 2004.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020a.

Maxime Cauchois, Suyash Gupta, and John Duchi. Knowing what you know: valid confidence sets in multiclass and multilabel prediction. *arXiv preprint arXiv:2004.10181*, 2020b.

Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

David Collett. *Modelling survival data in medical research*. CRC press, 2015.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals*

- of Statistics*, 40(1):294–321, 2012. doi:10.1214/11-AOS940. URL <https://doi.org/10.1214/11-AOS940>.
- Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.
- David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- Ben Dai, Xiaotong Shen, and Wei Pan. Significance tests of feature relevance for a black-box learner. *IEEE transactions on neural networks and learning systems*, 35(2):1898–1911, 2022.
- A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, 1997. ISBN 0-521-57391-2. doi:10.1017/CBO9780511802843. URL <https://doi.org/10.1017/CBO9780511802843>.
- Nicolas Deutschmann, Mattia Rigotti, and Maria Rodriguez Martinez. Adaptive conformal regression with jackknife+ rescaled scores. *arXiv preprint arXiv:2305.19901*, 2023.
- Iván Díaz. Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Statistics in medicine*, 38(15):2735–2748, 2019.
- Lenin Domínguez-Ramírez, Francisca Sosa-Jurado, Guadalupe Díaz-Sampayo, Itzel Solis-Tejeda, Francisco Rodríguez-Pérez, Rosana Pelayo, Gerardo Santos-López, and Paulina Cortes-Hernandez. Age and comorbidities as risk factors for severe covid-19 in mexico, before, during and after massive vaccination. *Vaccines*, 11(11):1676, 2023.
- Ran Duan and Seth Pettie. Linear-time approximation for maximum weight matching. *Journal of the ACM*, 61(1):1–23, 2014.
- Alexander D’Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654, 2021.
- Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.
- B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26, 1979.
- Bradley Efron. Prediction, estimation, and attribution. *International Statistical Review*, 88: S28–S59, 2020.
- Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

- EJ Emanuel, G Persad, R Upshur, B Thome, M Parker, A Glickman, C Zhang, C Boyle, M Smith, and JP Phillips. Fair allocation of scarce medical resources in the time of covid-19. *The New England Journal of Medicine*, 382, 2020.
- Collin Erickson. *GauPro: Gaussian Process Fitting*, 2017. URL <https://CRAN.R-project.org/package=GauPro>. R package version 0.2.2.
- David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.
- Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in Neural Information Processing Systems*, 34: 2060–2071, 2021.
- Ronald Aylmer Fisher et al. 035: The distribution of the partial correlation coefficient. 1924.
- Thomas R Fleming and DY Lin. Survival analysis in clinical trials: past developments and future directions. *Biometrics*, 56(4):971–983, 2000.
- Jerome Friedman and Balasubramanian Narasimhan. *conTree: Contrast Trees and Boosting*, 2020. URL https://jhfh.github.io/conTree_tutorial. R package version 0.2-8.
- Jerome H Friedman. Contrast trees and distribution boosting. *Proceedings of the National Academy of Sciences*, 117(35):21175–21184, 2020.
- Harold N Gabow. A scaling algorithm for weighted matching on general graphs. In *26th Annual Symposium on Foundations of Computer Science (SFCS 1985)*, pages 90–100. IEEE, 1985.
- Byron J Gajewski, Stephen D Simon, and Susan E Carlson. Predicting accrual in clinical trials with bayesian posterior predictive distributions. *Statistics in medicine*, 27(13):2328–2340, 2008.
- Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.
- Seymour Geisser. *Predictive inference*, volume 55. CRC press, 1993.
- Subhankar Ghosh, Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7722–7730, 2023.
- Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

- Jelle J Goeman. L1 penalized estimation in the cox proportional hazards model. *Biometrical journal*, 52(1):70–84, 2010.
- Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.
- Jiang Gui and Hongzhe Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.
- Yu Gui, Rohan Hore, Zhimei Ren, and Rina Foygel Barber. Conformalized survival analysis with adaptive cut-offs. *Biometrika*, 111(2):459–477, 2024.
- Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.
- Laszlo Györfi and Harro Walk. Nearest neighbor based conformal prediction. In *Annales de l’ISUP*, volume 63, pages 173–190, 2019.
- Aric Hagberg, Pieter Swart, and Daniel Schult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
- Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
- Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008. URL <http://www.jstatsoft.org/v27/i05/>.
- Alexander Henzi, Johanna F Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):963–993, 2021.
- Han Hong and Elie Tamer. Inference in censored models with endogenous regressors. *Econometrica*, 71(3):905–932, 2003.
- Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae103, 2024.
- Rohan Hore, Jake A Soloff, Rina Foygel Barber, and Richard J Samworth. Testing conditional independence under isotonicity. *arXiv preprint arXiv:2501.06133*, 2025.
- Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

- Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008.
- Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. CD-split and HPD-split. *Journal of Machine Learning Research*, 23:1–32, 2022.
- Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *arXiv preprint arXiv:2111.12161*, 2021.
- Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.
- John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.
- Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3):613–636, 2007.
- Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.
- Samuel Karlin. *Total Positivity. Vol. I*. Stanford University Press, Stanford, CA, 1968.
- Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *stat*, 1050(2), 2016.
- Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251, 2022a.
- Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022b. ISSN 0090-5364,2168-8966. doi:10.1214/22-aos2233. URL <https://doi.org/10.1214/22-aos2233>.
- John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.
- Roger Koenker. Confidence intervals for regression quantiles. In *Asymptotic statistics*, pages 349–359. Springer, 1994.
- Roger Koenker. *quantreg: Quantile Regression*, 2020. URL <https://CRAN.R-project.org/package=quantreg>. R package version 5.75.
- Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

- Edward L Korn. Censoring distributions as a measure of follow-up in survival analysis. *Statistics in medicine*, 5(3):255–260, 1986.
- Kalimuthu Krishnamoorthy and Thomas Mathew. *Statistical tolerance regions: theory, applications, and computation*, volume 744. John Wiley & Sons, 2009.
- Stephen W Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156, 1979.
- Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):1–8, 2017.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000. ISSN 0090-5364,2168-8966. doi:10.1214/aos/1015957395. URL <https://doi.org/10.1214/aos/1015957395>.
- Jerald F Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011.
- Lucien Marie Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.
- Yonghoon Lee and Rina Barber. Distribution-free inference for regression: discrete, continuous, and in between. *Advances in Neural Information Processing Systems*, 34:7448–7459, 2021.
- E. L. Lehmann. Some concepts of dependence. *Ann. Math. Statist.*, 37:1137–1153, 1966. ISSN 0003-4851. doi:10.1214/aoms/1177699260. URL <https://doi.org/10.1214/aoms/1177699260>.
- Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020. doi:10.3150/19-BEJ1151. URL <https://doi.org/10.3150/19-BEJ1151>.
- 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, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

- Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.
- Benjamin LeRoy and David Zhao. MD-split+: Practical local conformal inference in high dimensions. *arXiv preprint arXiv:2107.03280*, 2021.
- Kwan-Moon Leung, Robert M Elashoff, and Abdelmonem A Afifi. Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104, 1997.
- Alexander Hanbo Li and Jelena Bradic. Censored quantile regression forest. In *International Conference on Artificial Intelligence and Statistics*, pages 2109–2119. PMLR, 2020.
- Anton Rask Lundborg, Rajen D Shah, and Jonas Peters. Conditional independence testing in hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850, 2022.
- Anton Rask Lundborg, Ilmun Kim, Rajen D Shah, and Richard J Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 25:2851–2878, 2024.
- Alexandre Mösching and Lutz Dümbgen. Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14(1):24–49, 2020. ISSN 1935-7524. doi:10.1214/19-EJS1659. URL <https://doi.org/10.1214/19-EJS1659>.
- Alexandre Mösching and Lutz Dümbgen. Estimation of a likelihood ratio ordered family of distributions. *Statistics and Computing*, 34(1):Paper No. 58, 16, 2024. ISSN 0960-3174,1573-1375. doi:10.1007/s11222-023-10370-9. URL <https://doi.org/10.1007/s11222-023-10370-9>.
- Gayle Muenchow. Ecological use of failure time analysis. *Ecology*, 67(1):246–250, 1986.
- Manuel M Müller, Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Isotonic subgroup selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae083, 2024.
- SA Murphy, AJ Rossini, and Aad W van der Vaart. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976, 1997.
- Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn, and Wes B Ford. The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 48:p411, 1994.
- S Edward Nevius, Frank Proschan, and Jayaram Sethuraman. Schur functions in statistics.

- III: A stochastic version of weak majorization, with applications. Technical report, Florida State University Statistics Report M307, 1974.
- Whitney K Newey and James L Powell. Efficient estimation of linear and type i censored regression models under conditional quantile restrictions. *Econometric Theory*, pages 295–317, 1990.
- Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.
- Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5:465–472, 1923/1990. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in *Roczniki Nauk Rolniczych*, Tom X (1923): 1–51 (Annals of Agricultural Sciences).
- Ziang Niu, Abhinav Chakraborty, Oliver Dukes, and Eugene Katsevich. Reconciling model- x and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, 52(3):895–921, 2024.
- Constantinos O’Mahony, Fatima Jichi, Menelaos Pavlou, Lorenzo Monserrat, Aristides Anastakis, Claudio Rapezzi, Elena Biagini, Juan Ramon Gimeno, Giuseppe Limongelli, William J McKenna, et al. A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). *European Heart Journal*, 35(30):2010–2020, 2014.
- Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.
- Leandro Pardo. *Statistical Inference based on Divergence Measures*. Chapman and Hall/CRC, 2018.
- Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=DhP9L8vIyLc>.
- Limin Peng and Yijian Huang. Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482):637–649, 2008.
- Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):5–31, 2018.
- Belinda Phipson and Gordon K. Smyth. Permutation p -values should never be zero: calcu-

- lating exact p -values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Art. 39, 14, 2010. ISSN 1544-6115. doi:10.2202/1544-6115.1585. URL <https://doi.org/10.2202/1544-6115.1585>.
- Stephen Portnoy. Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012, 2003.
- Stephen Portnoy et al. Is ignorance bliss: Fixed vs. random censoring. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková*, pages 215–223. Institute of Mathematical Statistics, 2010.
- James L Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325, 1984.
- James L Powell. Censored regression quantiles. *Journal of econometrics*, 32(1):143–155, 1986.
- Aaditya Ramdas, Rina Foygel Barber, Emmanuel J Candès, and Ryan J Tibshirani. Permutation tests using arbitrary permutation distributions. *Sankhya A*, 85(2):1156–1177, 2023.
- Megan L Ranney, Valerie Griffeth, and Ashish K Jha. Critical supply shortages—the need for ventilators and personal protective equipment during the covid-19 pandemic. *New England Journal of Medicine*, 382(18):e41, 2020.
- Marc Ratkovic and Dustin Tingley. Estimation and inference on nonlinear and heterogeneous effects. Technical report, 2021. URL <https://scholar.harvard.edu/files/dtingley/files/mdei.pdf>.
- Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010. doi:10.1214/09-AOS691. URL <https://doi.org/10.1214/09-AOS691>.
- Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. With malice towards none: Assessing uncertainty via equalized coverage. *arXiv preprint arXiv:1908.05428*, 2019a.
- Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019b.
- Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. Classification with valid and adaptive coverage. *arXiv preprint arXiv:2006.02544*, 2020.

- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- Haskell P Rosenthal. On the subspaces of $l^p(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.
- Walter Rudin. *Real and Complex Analysis Third edition*. McGraw-Hill Book Co., New York, 1987.
- Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.
- René Möttus Sacha Epskamp, Lourens J. Waldorp and Denny Borsboom. The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, 53(4):453–480, 2018. doi:10.1080/00273171.2018.1454823. PMID: 29658809.
- Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.
- Pedro HC Sant’Anna. Program evaluation with right-censored data. *arXiv preprint arXiv:1604.02642*, 2016.
- Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 722–726, 1999.
- Daniel O Scharfstein and James M Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634, 2002.
- Michael Schemper and Terry L Smith. A note on quantifying follow-up in studies of failure time. *Controlled clinical trials*, 17:343–346, 1996.
- Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.
- Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

- Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.
- Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. ISSN 0090-5364,2168-8966. doi:10.1214/19-AOS1857. URL <https://doi.org/10.1214/19-AOS1857>.
- Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, New York, 2007. ISBN 978-0-387-32915-4; 0-387-32915-3. doi:10.1007/978-0-387-34675-5. URL <https://doi.org/10.1007/978-0-387-34675-5>.
- Irina G Shevtsova. An improvement of convergence rate estimates in the lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010.
- Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.
- Ritesh Singh and Keshab Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in clinical research*, 2(4):145, 2011.
- Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.
- Robert A Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.
- Terry M Therneau. *A Package for Survival Analysis in R*, 2020. URL <https://CRAN.R-project.org/package=survival>. R package version 3.2-7.
- Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.
- Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.
- Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.
- John W. Tukey. A problem of Berkson, and minimum variance orderly estimators. *Ann.*

Math. Statist., 29:588–592, 1958. ISSN 0003-4851. doi:10.1214/aoms/1177706637. URL <https://doi.org/10.1214/aoms/1177706637>.

Wessel N Van Wieringen, David Kun, Regina Hampel, and Anne-Laure Boulesteix. Survival prediction using gene expression data: a review and comparison. *Computational statistics & data analysis*, 53(5):1590–1603, 2009.

M Vergano, G Bertolini, A Giannini, GR Gristina, S Livigni, G Mistraletti, L Riccioni, and F Petrini. Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: the italian perspective during the covid-19 epidemic. *Critical Care*, 24, 2020.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Pierre JM Verweij and Hans C Van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314, 1993.

Bengt von Bahr, Carl-Gustav Esseen, et al. Inequalities for the r -th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1): 299–303, 1965.

Vladimir Vovk. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 187–196. IEEE, 2002.

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

Vladimir Vovk. Transductive conformal predictors. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 348–360. Springer, 2013.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4): 791–808, 2020.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Vladimir Vovk, Ilia Nouretdinov, Alex Gammerman, et al. On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590, 2009.

- Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.
- Abraham Wald. An extension of wilks’ method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1):45–55, 1943.
- Jiantian Wang and Yingfu Li. Estimators for survival function when censoring times are known. *Communications in Statistics-theory and Methods*, 34(2):449–459, 2005.
- Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.
- Samuel S Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.
- Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1): 9–22, 2021.
- Daniela M Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51, 2010.
- Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188, 1988.
- Zihui Yan, Mengjie Cai, Xu Han, Qingguang Chen, and Hao Lu. The interaction between age and risk factors for diabetes and prediabetes: a community-based cross-sectional study. *Diabetes, Metabolic Syndrome and Obesity*, pages 85–93, 2023.
- Yachong Yang and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.
- Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox’s proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

APPENDIX A

APPENDICES FOR CHAPTER 2

A.1 Proofs

A.1.1 Proof of Theorem 3

For notational convenience, we define the error term to be

$$\Delta = \sup_{a \in [0,1]} \left(\mathbb{E} \left[\left| \frac{\hat{w}_a(X)}{w_a(X)\pi_a} - 1 \right| \mid \mathcal{I}_1 \right] + \sqrt{\frac{1 + \frac{\hat{\gamma}_a^2}{\pi_a^2} + \max(1, \frac{\hat{\gamma}_a}{\pi_a} - 1)^2}{|\mathcal{I}_2|} \cdot \log \left(\frac{1}{\delta} \right)} \right).$$

Recall that we have defined the oracle quantity

$$a(\alpha + \Delta) = \sup \left\{ a \in [0, 1] : \mathbb{P} \left\{ T < \hat{f}_a(X) \mid \mathcal{I}_1 \right\} \leq \alpha + \Delta \right\}.$$

Suppose that we can show $1 - \delta \leq \mathbb{P} \{ \hat{a} \leq a(\alpha + \Delta) \mid \mathcal{I}_1 \}$. Then we have with probability at least $1 - \delta$ that the event $\{ \hat{a} \leq a(\alpha + \Delta) \}$ holds and

$$\begin{aligned} & \mathbb{P} \left\{ T \geq \hat{f}_{\hat{a}}(X) \mid \mathcal{D} \right\} \\ & \geq \mathbb{P} \left\{ T \geq \hat{f}_{a(\alpha + \Delta)}(X) \mid \mathcal{D} \right\} \\ & \geq 1 - \alpha - \Delta \\ & = 1 - \alpha - \sup_{a \in [0,1]} \left\{ \mathbb{E} \left[\left| \frac{\hat{w}_a(X)}{w_a(X)\pi_a} - 1 \right| \mid \mathcal{I}_1 \right] + \sqrt{\frac{1 + \frac{\hat{\gamma}_a^2}{\pi_a^2} + \max(1, \frac{\hat{\gamma}_a}{\pi_a} - 1)^2}{|\mathcal{I}_2|} \cdot \log \left(\frac{1}{\delta} \right)} \right\}, \end{aligned}$$

where the first inequality is by the monotonicity of $\hat{f}_a(\cdot)$ and the second inequality uses the left-continuity of $\mathbb{P} \left\{ T < \hat{f}_a(X) \mid \mathcal{D} \right\}$ in a .

The rest of the proof is devoted to establishing $1 - \delta \leq \mathbb{P}(\hat{a} \leq a(\alpha + \Delta) \mid \mathcal{I}_1)$. Fix an arbitrary $\varepsilon > 0$. By the definition of $\hat{\alpha}(a(\alpha + \Delta) + \varepsilon)$, we have

$$\begin{aligned}
& \mathbb{P}(\hat{\alpha}(a(\alpha + \Delta) + \varepsilon) \leq \alpha \mid \mathcal{I}_1) \\
&= \mathbb{P}\left(\sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha + \Delta) + \varepsilon}(X_i) \cdot \mathbb{1}\{T_i < \hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i) \leq C_i\}\right. \\
&\quad \left. \leq \alpha \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha + \Delta) + \varepsilon}(X_i) \cdot \mathbb{1}\{\hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i) \leq C_i\} \mid \mathcal{I}_1\right) \\
&= \mathbb{P}\left(\sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha + \Delta) + \varepsilon}(X_i) \cdot (\mathbb{1}\{T_i < \hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i)\} - \alpha) \mathbb{1}\{\hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i) \leq C_i\} \leq 0 \mid \mathcal{I}_1\right).
\end{aligned} \tag{A.1}$$

For any $t > 0$, we apply Markov's inequality and get

$$\begin{aligned}
& (A.1) \leq \mathbb{E}\left[\exp\left(t \cdot \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha + \Delta) + \varepsilon}(X_i) (\alpha - \mathbb{1}\{T_i < \hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i)\})\right.\right. \\
&\quad \left.\left. \times \mathbb{1}\{\hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i) \leq C_i\}\right) \mid \mathcal{I}_1\right] \\
&= \mathbb{E}\left[\exp\left(t \cdot \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha + \Delta) + \varepsilon}(X_i) \cdot (\alpha - p_{a(\alpha + \Delta) + \varepsilon}(X_i) + p_{a(\alpha + \Delta) + \varepsilon}(X_i)\right.\right. \\
&\quad \left.\left. - \mathbb{1}\{T_i < \hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i)\}) \cdot \mathbb{1}\{\hat{f}_{a(\alpha + \Delta) + \varepsilon}(X_i) \leq C_i\}\right) \mid \mathcal{I}_1\right],
\end{aligned} \tag{A.3}$$

where we define $p_a(x) := \mathbb{P}(T < \hat{f}_a(X) \mid X = x, \mathcal{I}_1)$ for any $a \in [0, 1]$. Further conditioning

on $(X_i, C_i)_{i \in \mathcal{I}_2}$, and using the fact that $C \perp\!\!\!\perp T \mid X$ (Assumption 1), we have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left(t \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \mathbb{1} \{ \hat{f}_{a(\alpha+\Delta)}(X_i) \leq C_i \} \right. \right. \\
& \quad \left. \left. \times \left(p_{a(\alpha+\Delta)+\varepsilon}(X_i) - \mathbb{1} \{ T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) \} \right) \right) \middle| (X_i, C_i)_{i \in \mathcal{I}_2}, \mathcal{I}_1 \right] \\
& \stackrel{(a)}{\leq} \exp \left(\frac{t^2}{8} \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}^2(X_i) \cdot \mathbb{1} \{ \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) \leq C_i \} \right) \\
& \stackrel{(b)}{\leq} \exp \left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 t^2}{8} \right). \tag{A.4}
\end{aligned}$$

Above, step (a) uses the $\frac{1}{4}$ -sub-gaussianity of $p_{a(\alpha+\Delta)+\varepsilon} - \mathbb{1} \{ T_i \leq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) \}$; step (b) follows from the boundedness assumption on the estimated weights. Combining (A.1) and (A.4) leads to

$$\begin{aligned}
(A.1) & \leq \exp \left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 t^2}{8} \right) \cdot \mathbb{E} \left[\exp \left\{ t \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \mathbb{1} \{ \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) \leq C_i \} \right. \right. \\
& \quad \left. \left. \times (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i)) \right\} \middle| \mathcal{I}_1 \right]. \tag{A.5}
\end{aligned}$$

We then condition on $(X_i)_{i \in \mathcal{I}_2}$ and use the sub-gaussianity of $\mathbb{1} \{ \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) \leq C_i \} - w_{a(\alpha+\Delta)+\varepsilon}(X_i)^{-1}$ to obtain the following bound:

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ t \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) (\mathbb{1} \{ \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) \leq C_i \} - w_{a(\alpha+\Delta)+\varepsilon}(X_i)^{-1}) \right. \right. \\
& \quad \left. \left. \times (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i)) \right\} \middle| (X_i)_{i \in \mathcal{I}_2}, \mathcal{I}_1 \right] \\
& \leq \exp \left(\frac{t^2}{8} \sum_{i \in \mathcal{I}_2} (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i))^2 \cdot \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)^2 \right) \leq \exp \left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 t^2}{8} \right),
\end{aligned}$$

where we again use the boundedness of $\hat{w}_a(\cdot)$ in the last step. With the above, we bound (A.5)

as

$$(A.5) \leq \exp\left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 t^2}{4}\right) \mathbb{E}\left[\exp\left\{t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i))\right\} \middle| \mathcal{I}_1\right]. \quad (A.6)$$

Recall that we have defined for any $a \in [0, 1]$ that

$$\pi_a = \mathbb{E}_{X \sim P_X} \left[\frac{\hat{w}_a(X)}{w_a(X)} \middle| \mathcal{I}_1 \right].$$

We subsequently bound (A.6) as

$$\begin{aligned} (A.6) &\leq \exp\left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 t^2}{4}\right) \cdot \mathbb{E}\left[\exp\left\{t \sum_{i \in \mathcal{I}_2} \pi_{a(\alpha+\Delta)+\varepsilon} (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i))\right.\right. \\ &\quad \left.\left.+ t \sum_{i \in \mathcal{I}_2} \left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \right\} \middle| \mathcal{I}_1\right] \\ &\leq \exp\left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 t^2}{4}\right) \cdot \mathbb{E}\left[\exp\left\{2t \cdot \pi_{a(\alpha+\Delta)+\varepsilon} \sum_{i \in \mathcal{I}_2} (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i))\right\} \middle| \mathcal{I}_1\right]^{1/2} \\ &\quad \times \mathbb{E}\left[\exp\left\{2t \sum_{i \in \mathcal{I}_2} \left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \right\} \middle| \mathcal{I}_1\right]^{1/2}, \quad (A.7) \end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. By the definition of $a(\alpha+\Delta)$, it holds that

$$\mathbb{P}(T < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X) \mid \mathcal{I}_1) \geq \alpha + \Delta.$$

Using the above inequality, we have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ 2t \cdot \pi_{a(\alpha+\Delta)+\varepsilon} \sum_{i \in \mathcal{I}_2} (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i)) \right\} \middle| \mathcal{I}_1 \right] \\
& \leq \mathbb{E} \left[\exp \left\{ 2t \cdot \pi_{a(\alpha+\Delta)+\varepsilon} \sum_{i \in \mathcal{I}_2} (\mathbb{P}(T < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X) | \mathcal{I}_1) - \Delta - p_{a(\alpha+\Delta)+\varepsilon}(X_i)) \right\} \middle| \mathcal{I}_1 \right] \\
& = \exp(-2t\pi_{a(\alpha+\Delta)+\varepsilon}\Delta|\mathcal{I}_2|) \\
& \quad \times \mathbb{E} \left[\exp \left\{ 2t \cdot \pi_{a(\alpha+\Delta)+\varepsilon} \sum_{i \in \mathcal{I}_2} (\mathbb{P}(T < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X) | \mathcal{I}_1) - p_{a(\alpha+\Delta)+\varepsilon}(X_i)) \right\} \middle| \mathcal{I}_1 \right] \\
& \leq \exp \left(\frac{|\mathcal{I}_2| \cdot t^2 \pi_{a(\alpha+\Delta)+\varepsilon}^2}{2} - 2t \cdot \pi_{a(\alpha+\Delta)+\varepsilon} \Delta \cdot |\mathcal{I}_2| \right). \tag{A.8}
\end{aligned}$$

Above, the last inequality uses that $\mathbb{P}(T < \hat{f}_{a(\alpha+\Delta)+\varepsilon} | \mathcal{I}_1) - p_{a(\alpha+\Delta)+\varepsilon}(X_i)$ is $\frac{1}{4}$ -sub-gaussian. Note that for any $i \in \mathcal{I}_2$,

$$\begin{aligned}
& \left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} - \pi_{a(\alpha+\Delta)+\varepsilon}(\mathcal{I}_1) \right| \\
& \leq \tilde{\gamma}(a(\alpha + \Delta) + \varepsilon) := \max \{ \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon} - \pi_{a(\alpha+\Delta)+\varepsilon}, \pi_{a(\alpha+\Delta)+\varepsilon} \}.
\end{aligned}$$

We then have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ 2t \left(\sum_{i \in \mathcal{I}_2} \left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \right. \right. \right. \\
& \quad \left. \left. \left. - \mathbb{E} \left[\left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \middle| \mathcal{I}_1 \right] \right) \right\} \middle| \mathcal{I}_1 \right] \\
& \leq \exp \left(\frac{|\mathcal{I}_2| \cdot \tilde{\gamma}^2(a(\alpha + \Delta) + \varepsilon) t^2}{2} \right). \tag{A.9}
\end{aligned}$$

Combining (A.8) and (A.9), we have the following upper bound on (A.7):

$$(A.7) \leq \exp \left\{ \frac{|\mathcal{I}_2|t^2}{4} \cdot \left(\pi_{a(\alpha+\Delta)+\varepsilon}^2 + \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 + \tilde{\gamma}^2(a(\alpha+\Delta)+\varepsilon) \right) \right. \\ \left. + |\mathcal{I}_2|t \cdot \left(\mathbb{E} \left[\left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X)}{w_{a(\alpha+\Delta)+\varepsilon}(X)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \middle| \mathcal{I}_1 \right] - \pi_{a(\alpha+\Delta)+\varepsilon} \Delta \right) \right\}. \quad (A.10)$$

We now take

$$t = \frac{2(\Delta \cdot \pi_{a(\alpha+\Delta)+\varepsilon} - \mathbb{E}[\left| \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X)}{w_{a(\alpha+\Delta)+\varepsilon}(X)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \middle| \mathcal{I}_1])}{\pi_{a(\alpha+\Delta)+\varepsilon}^2 + \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 + \tilde{\gamma}^2(a(\alpha+\Delta)+\varepsilon)}.$$

This gives us

$$(A.10) \leq \exp \left(- \frac{|\mathcal{I}_2| \cdot \left(\Delta \cdot \pi_{a(\alpha+\Delta)+\varepsilon}(\mathcal{I}_1) - \mathbb{E} \left[\left| \frac{w_{a(\alpha+\Delta)+\varepsilon}(X)}{w_{a(\alpha+\Delta)+\varepsilon}(X)} - \pi_{a(\alpha+\Delta)+\varepsilon} \right| \middle| \mathcal{I}_1 \right] \right)^2}{\pi_{a(\alpha+\Delta)+\varepsilon}^2(\mathcal{I}_1) + \hat{\gamma}_{a(\alpha+\Delta)+\varepsilon}^2 + \tilde{\gamma}^2(a(\alpha+\Delta)+\varepsilon)} \right) \leq \delta.$$

The last inequality is due to the choice of Δ . As a result,

$$1 - \delta \leq \mathbb{P}(\hat{\alpha}(a(\alpha+\Delta)+\varepsilon) > \alpha \mid \mathcal{I}_1) \leq \mathbb{P}(\hat{a} < a(\alpha+\Delta)+\varepsilon \mid \mathcal{I}_1).$$

Since the above holds for any $\varepsilon > 0$, we can take $\varepsilon \rightarrow 0$ and by the continuity of the probability measure, we have $1 - \delta \leq \mathbb{P}(\hat{a} \leq a(\alpha+\Delta) \mid \mathcal{I}_1)$ and thus complete the proof.

A.1.2 Proof of Theorem 4

Instead of proving Theorem 4 directly, we prove a more general theorem that implies Theorem 4.

Theorem A.1. *Fix any $\delta, \alpha \in (0, 1)$. Under the same condition of Theorem 3, assume further that there exists a constant $r > 0$ such that*

1. $\sup_{\xi \in [a(\alpha), a(\alpha+r)+\psi]} w_\xi(x) \leq \gamma$ and $\sup_{\xi \in [a(\alpha), a(\alpha+r)+\psi]} \hat{w}_\xi(x) \leq \hat{\gamma}$, for some constants $\gamma, \hat{\gamma}, \psi > 0$;

2. for any $\eta \in [0, r]$, for any $\beta \in [\alpha, \alpha + r]$, and P_X -almost all x ,

$$\mathbb{P}(T < q_\beta(X) + \eta \mid X = x) \leq \beta + B\eta,$$

$$\mathbb{P}(T < q_\beta(X) - \eta \mid X = x) \geq \beta - B\eta,$$

for some family of oracle functions $\{q_a(\cdot)\}_{a \in [0,1]}$ and some constant $B > 0$;

3. $\sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in \mathcal{X}} \left\{ \max(B, 1) \cdot |\hat{f}_{a(\beta)}(x) - q_\beta(x)| \right\} + \hat{\gamma}\gamma \sqrt{\frac{\log(1/\delta)}{|\mathcal{I}_2|}} \leq r$.

Then with probability at least $1 - \delta$ over the draw of $\mathcal{D}$, the LPB produced by Algorithm 1 satisfies that for P_X -almost all x ,

$$\begin{aligned} & \mathbb{P}_{(X,T) \sim P}(T \geq \hat{L}(x) \mid \mathcal{D}, X = x) \\ & \geq 1 - \alpha - \sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in \mathcal{X}} \left\{ 2B \cdot |\hat{f}_{a(\beta)}(x) - q_\beta(x)| \right\} - \hat{\gamma}\gamma \sqrt{\frac{1}{|\mathcal{I}_2|} \cdot \log\left(\frac{1}{\delta}\right)}. \end{aligned}$$

To see why Theorem A.1 implies Theorem 4, note that when we take $q_\beta(x)$ to be the β -quantile of T conditional on $X = x$, and when the conditional distribution of $T \mid X$ is continuous with conditional density bounded by B , assumption 2 required by Theorem A.1 is satisfied. We now proceed to prove Theorem A.1.

Proof. Here, for notational convenience we define

$$\mathcal{E} = \sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in X_{cal}} |\hat{f}_{a(\beta)}(x) - q_\beta(x)| \text{ and } \Delta = B\mathcal{E} + \hat{\gamma}\gamma \cdot \sqrt{\frac{\log(1/\delta)}{|\mathcal{I}_2|}}.$$

By Assumption 3, $\mathcal{E} \leq r$ and $\Delta \leq r$. If we can show w.p. at least $1 - \delta$ that $a(\alpha + \Delta) \geq \hat{a}$,

then by the monotonicity of $\hat{f}_a(\cdot)$ in a ,

$$\mathbb{P}(T < \hat{f}_{\hat{a}}(X) | \mathcal{D}) \leq \mathbb{P}(T < \hat{f}_{a(\alpha+\Delta)}(X) | \mathcal{D}) \leq \alpha + \Delta,$$

where the last inequality uses the left-continuity of $\mathbb{P}(T < \hat{f}_a(X) | \mathcal{D})$. Furthermore, for a given $x \in \mathcal{X}$,

$$\begin{aligned} \mathbb{P}(T < \hat{f}_{\hat{a}}(x) | X = x, \mathcal{D}) &\leq \mathbb{P}(T < \hat{f}_{a(\alpha+\Delta)}(x) | X = x, \mathcal{D}) \\ &\leq \mathbb{P}(T < q_{\alpha+\Delta}(x) + \mathcal{E} | X = x, \mathcal{I}_2) \leq \alpha + \Delta + B\mathcal{E}, \end{aligned}$$

where the second inequality follows from the definition of $\mathcal{E}$ (and that $\Delta \leq r$), and the last inequality is due to Assumption 2 (and that both Δ and $\mathcal{E}$ are bounded by r). We have therefore arrived at our desired conclusion.

It remains to show that $\mathbb{P}(a(\alpha + \Delta) \geq \hat{a}) \geq 1 - \delta$. As before, we fix an arbitrary $\varepsilon \in (0, \psi]$.

Then for any $t > 0$,

$$\begin{aligned} &\mathbb{P}\left(\hat{a}(a(\alpha + \Delta) + \varepsilon) \leq \alpha \mid \mathcal{I}_1\right) \\ &= \mathbb{P}\left(\alpha \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \cdot \mathbb{1}\{C_i \geq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\} \right. \\ &\quad \left. - \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \cdot \mathbb{1}\{C_i \geq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) > T_i\} \geq 0 \mid \mathcal{I}_1\right) \\ &\leq \mathbb{E}\left[\exp\left\{t \cdot \left(\alpha \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \cdot \mathbb{1}\{C_i \geq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\} \right. \right. \right. \\ &\quad \left. \left. - \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \cdot \mathbb{1}\{C_i \geq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) > T_i\}\right)\right\} \mid \mathcal{I}_1\right], \quad (\text{A.11}) \end{aligned}$$

where the inequality follows from Markov's inequality. Next, we condition on $(X_i, T_i)_{i \in \mathcal{I}_2}$

and have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ t \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i) \mathbb{1}\{C_i \geq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\} \cdot (\alpha - \mathbb{1}\{T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\}) \right. \right. \\
& \quad \left. \left. - t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} \cdot (\alpha - \mathbb{1}\{T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\}) \right\} \middle| (X_i, T_i)_{i \in \mathcal{I}_2}, \mathcal{I}_1 \right] \\
& \stackrel{(a)}{\leq} \exp \left\{ \frac{t^2}{8} \sum_{i \in \mathcal{I}_2} \hat{w}_{a(\alpha+\Delta)+\varepsilon}^2(X_i) (\alpha - \mathbb{1}\{T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\})^2 \right\} \\
& \stackrel{(b)}{\leq} \exp \left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}^2 t^2}{8} \right), \tag{A.12}
\end{aligned}$$

where step (a) is due to the $\frac{1}{4}$ -sub-Gaussianity of $\mathbb{1}\{C_i \geq \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\}$, and step (b) the boundedness of $\hat{w}_{a(\alpha+\Delta)+\varepsilon}(\cdot)$. Combining (A.11) and (A.12), we have

$$\begin{aligned}
(A.11) \leq & \mathbb{E} \left[\exp \left\{ t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} (\alpha - \mathbb{1}\{T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\}) \right. \right. \\
& \left. \left. + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{8} \right\} \middle| \mathcal{I}_1 \right]. \tag{A.13}
\end{aligned}$$

Next, recall that $p_a(x) = \mathbb{P}(T < \hat{f}_a(x) \mid X = x)$. We now condition on $(X_i)_{i \in \mathcal{I}_2}$:

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} \cdot (p_{a(\alpha+\Delta)+\varepsilon}(X_i) - \mathbb{1}\{T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i)\}) \right\} \middle| (X_i)_{i \in \mathcal{I}_2}, \mathcal{I}_1 \right] \\
& \stackrel{(a)}{\leq} \exp \left\{ \frac{t^2}{8} \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}^2(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}^2(X_i)} \right\} \\
& \stackrel{(b)}{\leq} \exp \left(\frac{|\mathcal{I}_2| \cdot \hat{\gamma}^2 t^2}{8} \right),
\end{aligned}$$

where step (a) uses the sub-Gaussianity of $p_a(X_i) - \mathbb{1}\{T_i < \hat{f}_a(X_i)\}$ and step (b) is due to

the boundedness of $\hat{w}_a(\cdot)$. Combining the above, we now have

$$\begin{aligned}
(A.13) &\leq \mathbb{E} \left[\exp \left(t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} \cdot (\alpha - p_{a(\alpha+\Delta)+\varepsilon}(X_i)) + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{4} \right) \middle| \mathcal{I}_1 \right] \\
&= \mathbb{E} \left[\exp \left(t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} \cdot (\alpha - \mathbb{P}(T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) | X_i, \mathcal{I}_1)) + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{4} \right) \middle| \mathcal{I}_1 \right].
\end{aligned} \tag{A.14}$$

By Assumption 3, $\Delta \leq r$, and by the definition of $\mathcal{E}$, $|q_{\alpha+\Delta}(x) - \hat{f}_{a(\alpha+\Delta)}(x)| \leq \mathcal{E}$; Consequently,

$$\mathbb{P}(T_i < \hat{f}_{a(\alpha+\Delta)+\varepsilon}(X_i) | X_i, \mathcal{I}_1) \geq \mathbb{P}(T_i < \hat{f}_{a(\alpha+\Delta)}(X_i) | X_i, \mathcal{I}_1) \geq \mathbb{P}(T_i < q_{\alpha+\Delta}(X_i) - \mathcal{E} | X_i, \mathcal{I}_1).$$

Then,

$$\begin{aligned}
(A.14) &\leq \mathbb{E} \left[\exp \left(t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} \cdot (\alpha - \mathbb{P}(T_i < q_{\alpha+\Delta}(X_i) - \mathcal{E} | X_i, \mathcal{I}_1)) + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{4} \right) \middle| \mathcal{I}_1 \right] \\
&\stackrel{(a)}{\leq} \mathbb{E} \left[\exp \left(t \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} \cdot (\alpha - (\alpha + \Delta - B\mathcal{E})) + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{4} \right) \middle| \mathcal{I}_1 \right] \\
&= \mathbb{E} \left[\exp \left(-t(\Delta - B\mathcal{E}) \cdot \sum_{i \in \mathcal{I}_2} \frac{\hat{w}_{a(\alpha+\Delta)+\varepsilon}(X_i)}{w_{a(\alpha+\Delta)+\varepsilon}(X_i)} + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{4} \right) \middle| \mathcal{I}_1 \right]
\end{aligned} \tag{A.15}$$

$$\stackrel{(b)}{\leq} \exp \left(-\frac{|\mathcal{I}_2| t (\Delta - B\mathcal{E})}{\gamma} + \frac{|\mathcal{I}_2| \hat{\gamma}^2 t^2}{4} \right), \tag{A.16}$$

where step (a) is due to that $\mathcal{E} \leq r$ and Assumption 2; step (b) follows from the boundedness of $w_a(\cdot)$ and that $\Delta - B\mathcal{E} \geq 0$. Taking $t = \frac{2}{\gamma \hat{\gamma}^2} (\Delta - B\mathcal{E})$, we have

$$(A.14) = \exp \left(-\frac{1}{\gamma^2 \hat{\gamma}^2} \cdot (\Delta - B\mathcal{E})^2 |\mathcal{I}_2| \right) \leq \delta,$$

where the last inequality is due to the choice of Δ . As a result, we have with probability at

least $1 - \delta$ that $\hat{a}(a(\alpha + \Delta) + \varepsilon) > \alpha$, which implies that $a(\alpha + \Delta) + \varepsilon > \hat{a}$. That is,

$$\mathbb{P}(a(\alpha + \Delta) + \varepsilon > \hat{a}) \geq 1 - \delta.$$

Again, taking $\varepsilon \rightarrow 0$ and using the continuity of probability measures, we have w.p. at least $1 - \delta$ that $a(\alpha + \Delta) \geq \hat{a}$, and thus complete the proof. $\square$

APPENDIX B

APPENDICES FOR CHAPTER 3

B.1 Proofs and few additional results

B.1.1 Completing the proof of Theorem 3.6

With the calculations shown in the proof sketch (in Section 3.4.3) in place, we now need to verify two claims: we need to prove the bound (3.18) on the expected total variation distance, and we need to prove the claim in the theorem itself.

First we verify that (3.18) holds. First, by definition, $P_X \circ g \circ H(\cdot, \tilde{X})$ is absolutely continuous with respect to $P_X \circ H(\cdot, \tilde{X})$, and we can compute the Radon–Nikodym derivative relating these two distributions as

$$\frac{d(P_X \circ g \circ H(\cdot, \tilde{X}))(x)}{d(P_X \circ H(\cdot, \tilde{X}))(x)} = g(x) \cdot \frac{\int_{\mathcal{X}} H(x', \tilde{X}) dP_X(x')}{\int_{\mathcal{X}} g(x') H(x', \tilde{X}) dP_X(x')} = \frac{g(x)}{\mathbb{E}_{X' \sim P_X \circ H(\cdot, \tilde{X})} [g(X')]}.$$

Therefore, treating $\tilde{X}$ as fixed in all the following calculations, we have

$$\begin{aligned} d_{\text{TV}} \left(P_X \circ H(\cdot, \tilde{X}), P_X \circ g \circ H(\cdot, \tilde{X}) \right) &= \frac{1}{2} \mathbb{E}_{P_X \circ H(\cdot, \tilde{X})} \left[\left| \frac{g(X)}{\mathbb{E}_{X' \sim P_X \circ H(\cdot, \tilde{X})} [g(X')]} - 1 \right| \right] \\ &= \frac{\mathbb{E}_{X \sim P_X \circ H(\cdot, \tilde{X})} \left[|g(X) - \mathbb{E}_{X' \sim P_X \circ H(\cdot, \tilde{X})} [g(X')]| \right]}{2 \mathbb{E}_{P_X \circ H(\cdot, \tilde{X})} [g(X)]} \\ &\leq \frac{\mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_X \circ H(\cdot, \tilde{X})} [|g(X) - g(X')|]}{2 \mathbb{E}_{P_X \circ H(\cdot, \tilde{X})} [g(X)]} \\ &= \mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_X \circ H(\cdot, \tilde{X})} [|g(X) - g(X')|] \cdot \frac{\mathbb{E}_{P_X} [H(X, \tilde{X})]}{2 \mathbb{E}_{P_X} [g(X) H(X, \tilde{X})]}, \end{aligned}$$

where the inequality holds by Jensen's inequality. Next we need to marginalize over $\tilde{X}$, i.e., $\tilde{P}_{\tilde{X}}$. We have defined this distribution as the marginal distribution of $\tilde{X}$ when we first draw $X \sim P_X \circ g$, then draw $\tilde{X} \sim H(X, \cdot)$ (where we recall that $H(X, \cdot)$ is a density with respect to some base measure ν), meaning that $\tilde{P}_{\tilde{X}}$ has density at x given by $\mathbb{E}_{P_X \circ g} [H(X, x)] = \frac{\mathbb{E}_{P_X} [g(X)H(X, x)]}{\mathbb{E}_{P_X} [g(X)]}$. Consequently,

$$\begin{aligned} & \mathbb{E}_{\tilde{P}_{\tilde{X}}} \left[\text{dTV} \left(P_X \circ H(\cdot, \tilde{X}), P_X \circ g \circ H(\cdot, \tilde{X}) \right) \right] \\ & \leq \int_{\mathcal{X}} \frac{\mathbb{E}_{P_X} [g(X)H(X, x)]}{\mathbb{E}_{P_X} [g(X)]} \cdot \mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_{X \circ H(\cdot, x)}} [|g(X) - g(X')|] \cdot \frac{\mathbb{E}_{P_X} [H(X, x)]}{2\mathbb{E}_{P_X} [g(X)H(X, x)]} \text{d}\nu(x) \\ & \quad = \int_{\mathcal{X}} \mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_{X \circ H(\cdot, x)}} [|g(X) - g(X')|] \cdot \frac{\mathbb{E}_{P_X} [H(X, x)]}{2\mathbb{E}_{P_X} [g(X)]} \text{d}\nu(x) \\ & \quad = \frac{\mathbb{E}_{P_{\tilde{X}}} \left[\mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_{X|\tilde{X}}} [|g(X) - g(X')|] \right]}{2\mathbb{E}_{P_X} [g(X)]}, \end{aligned}$$

where for the last step, we use the marginal distribution $P_{\tilde{X}}$ of $\tilde{X}$ under the joint distribution $P_{(X, \tilde{X})}$, where we first draw $X \sim P_X$ and then $\tilde{X} \sim H(X, \cdot)$; this distribution has density $\mathbb{E}_{P_X} [H(X, x)]$ with respect to ν . Finally, standard calculations show that

$$\mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_{X|\tilde{X}}} [|g(X) - g(X')|] \leq \mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_{X|\tilde{X}}} \left[(g(X) - g(X'))^2 \right]^{1/2} = \sqrt{2 \text{Var}_{P_{X|\tilde{X}}} (g(X))},$$

which completes the proof of (3.18).

Now we need to verify the proof of the upper bound in the theorem. Returning to our calculations above, it suffices to prove that

$$\begin{aligned} & \frac{1}{2} \mathbb{E}_{P_{\tilde{X}}} \left[\mathbb{E} \left[X, X' \stackrel{\text{iid}}{\sim} P_{X|\tilde{X}} \mid |g(X) - g(X')| \right] \right] \\ & \leq \epsilon \cdot L_{g, 2\epsilon, A} + \|g\|_{\infty} \cdot P_{(X, \tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} + \|g\|_{\infty} \cdot P_X(A^c), \end{aligned}$$

for any $\varepsilon > 0$ and any $A \subseteq \mathcal{X}$. By definition of $L_{g,2\varepsilon,A}$, we can write

$$\frac{1}{2}|g(X) - g(X')| \leq \begin{cases} \frac{2\varepsilon \cdot L_{g,2\varepsilon,A}}{2}, & \text{if } X, X' \in A \text{ and } \|X - X'\| \leq 2\varepsilon, \\ \|g\|_\infty/2, & \text{otherwise,} \end{cases}$$

where for the second case we recall that g is nonnegative and so $|g(x) - g(x')| \leq \|g\|_\infty$ for any x, x' . Therefore,

$$\begin{aligned} & \frac{1}{2}\mathbb{E}_{P_{\tilde{X}}} \left[\mathbb{E} \left[X, X' \stackrel{\text{iid}}{\sim} P_{X|\tilde{X}} \mid |g(X) - g(X')| \right] \right] \\ & \leq \varepsilon \cdot L_{g,2\varepsilon,A} + \frac{\|g\|_\infty}{2} \cdot \mathbb{E}_{P_{\tilde{X}}} \left[\mathbb{P}_{X, X' \stackrel{\text{iid}}{\sim} P_{X|\tilde{X}}} \left\{ X \notin A \text{ or } X' \notin A \text{ or } \|X - \tilde{X}\| > \varepsilon \text{ or } \|X' - \tilde{X}\| > \varepsilon \right\} \right] \\ & \leq \varepsilon \cdot L_{g,2\varepsilon,A} + \frac{\|g\|_\infty}{2} \cdot 2\mathbb{E}_{P_{\tilde{X}}} \left[\mathbb{P}_{X \sim P_{X|\tilde{X}}} \{X \notin A \text{ or } \|X - \tilde{X}\| > \varepsilon\} \right] \\ & \leq \varepsilon \cdot L_{g,2\varepsilon,A} + \|g\|_\infty \cdot \mathbb{P}_{P_{(X,\tilde{X})}} \left\{ \|X - \tilde{X}\| > \varepsilon \right\} + \|g\|_\infty \cdot P_X(A^c), \end{aligned}$$

which completes the proof.

B.1.2 Training-conditional coverage

In this subsection, we will establish training-conditional coverage guarantees for RLCP. Towards that, we start with defining miscoverage of RLCP conditioned on the training data

set $D_n = \{(X_i, Y_i)\}_{i \in [n]}$,¹

$$\alpha_{\text{tr}}(D_n) = \mathbb{P} \left\{ Y_{n+1} \notin \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid D_n \right\}.$$

1. Note that, since the score function $s = s(x, y)$ is treated as fixed, in a practical setting where s was fitted on a *pretraining* set, and so we might refer to the data points $\{(X_i, Y_i)\}_{i \in [n]}$ as the *calibration set* (as in our empirical results in Section 3.5). This means that we are interested in coverage conditional on both the pretraining and calibration datasets.

We would then like to show that $\alpha_{\text{tr}}(D_n) \lesssim \alpha$ holds with high probability, with respect to the random draw of the training data D_n . This type of result is well known to hold for split conformal prediction [Vovk, 2012]. On the other hand, for full conformal prediction, Bian and Barber [2023] prove a hardness result establishing that it is impossible to guarantee training-conditional coverage without further assumptions. Since the localized CP methods are closely related to full conformal prediction (as discussed in Section 3.6.3), it is not immediately clear whether a training-conditional guarantee can hold for a localized method.

In this section, we will establish a training-conditional coverage guarantee for RLCP. First we will show a finite-sample result, and then will discuss how to interpret it from an asymptotic point of view.

Theorem B.1. *For any $\delta > 0$, the RLCP method satisfies the training-conditional coverage guarantee*

$$\mathbb{P} \left\{ \alpha_{\text{tr}}(D_n) \leq \alpha + \frac{2}{(\delta n)^{1/3}} \mathbb{E} \left[\frac{H_2^2(\tilde{X}_{n+1})}{H_1(\tilde{X}_{n+1})^2} \right]^{1/3} \right\} \geq 1 - \delta,$$

where for $x \in \mathcal{X}$, we define

$$H_1(x) = \mathbb{E}_{P_X} [H(X, x)], \quad H_2^2(x) = \mathbb{E}_{P_X} [H(X, x)^2].$$

Now we consider an asymptotic view—when can we expect the excess error term to be vanishing as $n \rightarrow \infty$? For example, in the case $\mathcal{X} = \mathbb{R}$, under suitable regularity conditions on P_X , if at sample size n we take H to be the normalized box kernel with bandwidth h_n , then we would expect to have $H_1(\tilde{X}_{n+1}) \asymp 1$ and $H_2^2(\tilde{X}_{n+1}) \asymp h_n^{-1}$. Consequently, the upper bound on $\alpha_{\text{tr}}(D_n)$ is of the form $\alpha + O_P((nh_n)^{-1/3})$. As long as $nh_n \rightarrow \infty$, then, we expect to see asymptotic training-conditional coverage.

Proof of Theorem B.1. Fix any $\alpha^* > \alpha$. Condition on the value of $\tilde{X}_{n+1}$, and let $q^*(\tilde{X}_{n+1})$

be the $(1 - \alpha^*)$ -quantile of the distribution of $s(X, Y)$ under the distribution $(X, Y) \sim (P_X \circ H(\cdot, \tilde{X}_{n+1})) \times P_{Y|X}$. Then by our calculation of the conditional distribution of the test point (X_{n+1}, Y_{n+1}) given $\tilde{X}_{n+1}$ (as in (3.14)) we have

$$\mathbb{P} \left\{ s(X_{n+1}, Y_{n+1}) \leq q^*(\tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \right\} \geq 1 - \alpha^* \text{ almost surely.}$$

Moreover since $(X_{n+1}, Y_{n+1}, \tilde{X}_{n+1})$ is independent of the training data $D_n = \{(X_i, Y_i)\}_{i \in [n]}$ we can equivalently write this as

$$\mathbb{P} \left\{ s(X_{n+1}, Y_{n+1}) \leq q^*(\tilde{X}_{n+1}) \mid \tilde{X}_{n+1}, D_n \right\} \geq 1 - \alpha^* \text{ almost surely.}$$

We then calculate

$$\begin{aligned} \alpha_{\text{tr}}(D_n) &= \mathbb{P} \left\{ Y_{n+1} \notin \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid D_n \right\} \\ &= \mathbb{P} \left\{ s(X_{n+1}, Y_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \mid D_n \right\} \\ &= \mathbb{E} \left[\mathbb{P} \left\{ s(X_{n+1}, Y_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}, D_n \right\} \mid D_n \right] \\ &\leq \mathbb{E} \left[\mathbb{P} \left\{ s(X_{n+1}, Y_{n+1}) > q^*(\tilde{X}_{n+1}) \mid \tilde{X}_{n+1}, D_n \right\} \right. \\ &\quad \left. + \mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}, D_n \right\} \mid D_n \right] \\ &\leq \alpha^* + \mathbb{E} \left[\mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}, D_n \right\} \mid D_n \right] \\ &= \alpha^* + \mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \mid D_n \right\}. \end{aligned}$$

Moreover, by Markov's inequality, for any $\delta > 0$ we have

$$\mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \mid D_n \right\} \leq \delta^{-1} \mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \right\}$$

with probability at least $1 - \delta$. We have therefore shown that

$$\mathbb{P} \left\{ \alpha_{\text{tr}}(D_n) \leq \alpha^* + \delta^{-1} \mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \right\} \right\} \geq 1 - \delta.$$

Now we need to bound the term $\mathbb{P} \left\{ q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) \right\}$. By definition of the quantile $\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$, on the event $q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})$ we must have

$$\sum_{i=1}^n \tilde{w}_i \mathbb{1} \left\{ s(X_i, Y_i) < q^*(\tilde{X}_{n+1}) \right\} \geq 1 - \alpha,$$

for weights

$$\tilde{w}_i = \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})}.$$

Equivalently, we must have

$$\sum_{i=1}^n H(X_i, \tilde{X}_{n+1}) \mathbb{1} \left\{ s(X_i, Y_i) < q^*(\tilde{X}_{n+1}) \right\} \geq (1 - \alpha) \sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}),$$

which implies

$$\sum_{i=1}^n H(X_i, \tilde{X}_{n+1}) \left(\mathbb{1} \left\{ s(X_i, Y_i) < q^*(\tilde{X}_{n+1}) \right\} - (1 - \alpha) \right) \geq (1 - \alpha) H(X_{n+1}, \tilde{X}_{n+1}) \geq 0.$$

Next, conditional on $\tilde{X}_{n+1}$, the terms in the sum are independent, and each term i has variance

$$\begin{aligned} \text{Var} \left(H(X_i, \tilde{X}_{n+1}) \left(\mathbb{1} \left\{ s(X_i, Y_i) < q^*(\tilde{X}_{n+1}) \right\} - (1 - \alpha) \right) \mid \tilde{X}_{n+1} \right) \\ \leq \mathbb{E} \left[H(X_i, \tilde{X}_{n+1})^2 \left(\mathbb{1} \left\{ s(X_i, Y_i) < q^*(\tilde{X}_{n+1}) \right\} - (1 - \alpha) \right)^2 \mid \tilde{X}_{n+1} \right] \\ \leq \mathbb{E} \left[H(X_i, \tilde{X}_{n+1})^2 \mid \tilde{X}_{n+1} \right] = H_2^2(\tilde{X}_{n+1}). \end{aligned}$$

We can also calculate, for any $x \in \mathcal{X}$ and any $q \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[H(X_i, x) \mathbb{1}\{s(X_i, Y_i) < q\}] &= \mathbb{E}_{P_X \times P_{Y|X}}[H(X, x) \mathbb{1}\{s(X, Y) < q\}] \\ &= \mathbb{P}_{(P_X \circ H(\cdot, x)) \times P_{Y|X}}\{s(X, Y) < q\} \cdot \mathbb{E}_{P_X}[H(X, x)],\end{aligned}$$

and so, plugging in $x = \tilde{X}_{n+1}$ and $q = q^*(\tilde{X}_{n+1})$, we have

$$\begin{aligned}\mathbb{E}\left[H(X_i, \tilde{X}_{n+1}) \mathbb{1}\left\{s(X_i, Y_i) < q^*(\tilde{X}_{n+1})\right\} \mid \tilde{X}_{n+1}\right] \\ &= \mathbb{P}_{P_X \circ H(\cdot, \tilde{X}_{n+1}) \times P_{Y|X}}\left\{s(X, Y) < q^*(\tilde{X}_{n+1}) \mid \tilde{X}_{n+1}\right\} \cdot \mathbb{E}_{P_X}\left[H(X, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}\right] \\ &= \mathbb{P}\left\{s(X_{n+1}, Y_{n+1}) < q^*(\tilde{X}_{n+1}) \mid \tilde{X}_{n+1}\right\} \cdot \mathbb{E}\left[H(X_i, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}\right] \\ &\leq (1 - \alpha^*) \cdot \mathbb{E}\left[H(X_i, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}\right].\end{aligned}$$

Therefore, for each i ,

$$\begin{aligned}\mathbb{E}\left[H(X_i, \tilde{X}_{n+1}) \left(\mathbb{1}\left\{s(X_i, Y_i) < q^*(\tilde{X}_{n+1})\right\} - (1 - \alpha)\right) \mid \tilde{X}_{n+1}\right] \\ \leq -(\alpha^* - \alpha) \mathbb{E}\left[H(X_i, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1}\right] = -(\alpha^* - \alpha) \cdot H_1(\tilde{X}_{n+1}).\end{aligned}$$

Applying Chebyshev's inequality, then,

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^n H(X_i, \tilde{X}_{n+1}) \left(\mathbb{1}\left\{s(X_i, Y_i) < q^*(\tilde{X}_{n+1})\right\} - (1 - \alpha)\right) \geq 0 \mid \tilde{X}_{n+1}\right\} \\ \leq n^{-1} \cdot \frac{H_2^2(\tilde{X}_{n+1})}{(\alpha^* - \alpha)^2 H_1(\tilde{X}_{n+1})^2}.\end{aligned}$$

Combining everything, we have therefore shown that

$$\mathbb{P}\left\{q^*(\tilde{X}_{n+1}) > \hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1})\right\} \leq \frac{1}{n(\alpha^* - \alpha)^2} \cdot \mathbb{E}\left[\frac{H_2^2(\tilde{X}_{n+1})}{H_1(\tilde{X}_{n+1})^2}\right],$$

and so,

$$\mathbb{P} \left\{ \alpha_{\text{tr}}(D_n) \leq \alpha^* + \frac{1}{\delta n (\alpha^* - \alpha)^2} \cdot \mathbb{E} \left[\frac{H_2^2(\tilde{X}_{n+1})}{H_1(\tilde{X}_{n+1})^2} \right] \right\} \geq 1 - \delta.$$

Choosing

$$\alpha^* = \alpha + \frac{1}{(\delta n)^{1/3}} \mathbb{E} \left[\frac{H_2^2(\tilde{X}_{n+1})}{H_1(\tilde{X}_{n+1})^2} \right]^{1/3},$$

we have established that

$$\mathbb{P} \left\{ \alpha_{\text{tr}}(D_n) \leq \alpha + \frac{2}{(\delta n)^{1/3}} \mathbb{E} \left[\frac{H_2^2(\tilde{X}_{n+1})}{H_1(\tilde{X}_{n+1})^2} \right]^{1/3} \right\} \geq 1 - \delta,$$

as desired. □

B.1.3 Asymptotic test-conditional coverage

Since the model-free conditional coverage guarantee is impossible to attain, it is a common practice in literature to instead prove asymptotic test-conditional coverage. In this subsection, we establish such guarantees for RLCP prediction intervals.

As in Section 3.4.2, we consider an asymptotic regime where the distribution P is fixed and the marginal P_X has density f_X with respect to Lebesgue measure on $\mathbb{R}^d$ (meaning that, implicitly, dimension d is also fixed), while taking $n \rightarrow \infty$ and using a normalized box kernel H with bandwidth h_n to implement RLCP. In this section, instead of considering coverage conditional on an event $X_{n+1} \in \mathbb{B}(x_0, r_n)$ (i.e., conditioning on increasingly localized balls, as $r_n \rightarrow 0$), we will instead see that under some additional regularity assumptions we can condition on $X_{n+1} = x_0$. As before, let s be a pretrained score function. We will write $P_{S|X}$ to denote the distribution of $s(X, Y)$ conditional on X , under the distribution $(X, Y) \sim P$. We also assume that, marginally over $(X, Y) \sim P$, the distribution of $s(X, Y)$ is nonatomic.

In this regime, the following result holds for *pointwise* test-conditional coverage:

Theorem B.2. *Under the notation and assumptions above,*

$$\begin{aligned} & \left| \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} = x_0 \right\} - (1 - \alpha) \right| \\ & \leq \frac{1}{n+1} + \frac{1}{n(h_n)^d \cdot V_d \cdot \inf_{x \in \mathbb{B}(x_0, 2h_n)} f_X(x)} + 2 \sup_{x \in \mathbb{B}(x_0, 2h_n)} d_{\text{TV}}(P_{S|X}(\cdot|x_0), P_{S|X}(\cdot|x)). \end{aligned} \quad (\text{B.1})$$

In particular, if we assume $h_n \rightarrow 0$, and $h_n \cdot n^{1/d} \rightarrow \infty$, and if we also assume that $f_X(x)$ is bounded away from zero in a neighbourhood of $x = x_0$, and that that $x \mapsto P_{S|X}(\cdot|x)$ is continuous at $x = x_0$ with respect to the total variation distance, meaning that

$$\sup_{\|x-x_0\|_2 \leq \epsilon} d_{\text{TV}}(P_{S|X}(\cdot|x_0), P_{S|X}(\cdot|x)) \rightarrow 0 \text{ as } \epsilon \rightarrow 0,$$

then it holds that

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} = x_0 \right\} \rightarrow 1 - \alpha.$$

(As before, V_d is the volume of the unit ball in $\mathbb{R}^d$, while d_{TV} denotes the total variation distance.)

Proof. Let $S_i = s(X_i, Y_i)$ for $i \in [n+1]$. Since we are using a box kernel, the quantile cutoff of RLCP prediction method only uses the training samples that lies within a h_n radius ball of $\tilde{X}_{n+1}$. Define

$$I(\tilde{X}_{n+1}) = \{i \in [n] : \|X_i - \tilde{X}_{n+1}\|_2 \leq h_n\},$$

and $N_n = \sum_{i=1}^n \mathbb{1} \left\{ \|X_i - \tilde{X}_{n+1}\|_2 \leq h_n \right\} = |I(\tilde{X}_{n+1})|$ be its cardinality. Without loss of generality, let us enumerate the elements of $I(\tilde{X}_{n+1})$ as $\{i_1, \dots, i_{N_n}\}$. Hence, the quantile

cutoff can be simplified to

$$\hat{q}_{1-\alpha}(X_{n+1}, \tilde{X}_{n+1}) = \text{Quantile}_{1-\alpha} \left(S_{i_1}, \dots, S_{i_{N_n}}, S_{n+1} \right).$$

By a simple application of the tower law, we note that

$$\begin{aligned} & \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} = x_0 \right\} = \\ & \mathbb{E} \left[\mathbb{P} \left\{ S_{n+1} \leq \text{Quantile}_{1-\alpha} \left(S_{i_1}, \dots, S_{i_{N_n}}, S_{n+1} \right) \mid I(\tilde{X}_{n+1}), \tilde{X}_{n+1}, X_{n+1} = x_0 \right\} \mid X_{n+1} = x_0 \right]. \end{aligned}$$

Now, we will concentrate on the inner probability term. Conditional on $I(\tilde{X}_{n+1})$, $\tilde{X}_{n+1}$ and on $X_{n+1} = x_0$, the scores $S_{i_1}, \dots, S_{i_{N_n}}$ are sampled i.i.d. from the distribution $\tilde{P}_{S|\tilde{X}}(\cdot \mid \tilde{X}_{n+1})$, where $\tilde{P}_{S|\tilde{X}}(\cdot \mid x)$ is defined by

$$\begin{cases} X \sim P_X \circ H(\cdot, x), \\ S \mid X \sim P_{S|X}(\cdot \mid X), \end{cases}$$

while S_{n+1} is drawn from $P_{S|X}(\cdot \mid x_0)$ (independently of the other scores). Since $\tilde{P}_{S|\tilde{X}}(\cdot \mid x)$ is absolutely continuous with respect to the marginal distribution of $s(X, Y)$, it must be nonatomic. By Lemma B.3 below, we have

$$\begin{aligned} & \left| \mathbb{P} \left\{ S_{n+1} \leq \text{Quantile}_{1-\alpha} \left(S_{i_1}, \dots, S_{i_{N_n}}, S_{n+1} \right) \mid I(\tilde{X}_{n+1}), \tilde{X}_{n+1}, X_{n+1} = x_0 \right\} - (1 - \alpha) \right| \\ & \leq \frac{1}{N_n + 1} + \text{d}_{\text{TV}} \left(\tilde{P}_{S|\tilde{X}}(\cdot \mid \tilde{X}_{n+1}), P_{S|X}(\cdot \mid x_0) \right). \end{aligned}$$

Therefore,

$$\begin{aligned} & \left| \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} = x_0 \right\} - (1 - \alpha) \right| \\ & \leq \mathbb{E} \left[\frac{1}{N_n + 1} \mid X_{n+1} = x_0 \right] + \mathbb{E} \left[\text{d}_{\text{TV}} \left(\tilde{P}_{S|\tilde{X}}(\cdot \mid \tilde{X}_{n+1}), P_{S|X}(\cdot \mid x_0) \right) \mid X_{n+1} = x_0 \right]. \end{aligned}$$

To bound the second term, we note that conditional on $X_{n+1} = x_0$, we have $\tilde{X}_{n+1} \in \mathbb{B}(x_0, h_n)$ almost surely. Therefore,

$$\text{d}_{\text{TV}} \left(\tilde{P}_{S|\tilde{X}}(\cdot \mid \tilde{X}_{n+1}), P_{S|X}(\cdot \mid x_0) \right) \leq \sup_{x \in \mathbb{B}(x_0, h_n)} \text{d}_{\text{TV}} \left(\tilde{P}_{S|\tilde{X}}(\cdot \mid x), P_{S|X}(\cdot \mid x_0) \right).$$

Moreover, $\tilde{P}_{S|\tilde{X}}(\cdot \mid x)$ is a mixture of distributions $P_{S|X}(\cdot \mid X)$ where X is drawn from $P_X \circ H(\cdot, x)$, i.e., $X \in \mathbb{B}(x, h_n)$ almost surely—and so, for $x \in \mathbb{B}(x_0, h_n)$, we have $X \in \mathbb{B}(x_0, 2h_n)$ almost surely. Therefore,

$$\text{d}_{\text{TV}} \left(\tilde{P}_{S|\tilde{X}}(\cdot \mid \tilde{X}_{n+1}), P_{S|X}(\cdot \mid x_0) \right) \leq \sup_{x \in \mathbb{B}(x_0, 2h_n)} \text{d}_{\text{TV}} \left(P_{S|X}(\cdot \mid x), P_{S|X}(\cdot \mid x_0) \right).$$

Our last step is to bound $\mathbb{E} \left[\frac{1}{N_n + 1} \mid X_{n+1} = x_0 \right]$. Since the data points $X_1, \dots, X_n$ are i.i.d. samples from P_X , and are independent from $\tilde{X}_{n+1}$, we see that conditional on $\tilde{X}_{n+1}$,

$$N_n \sim \text{Binomial} \left(n, P_X(\mathbb{B}(\tilde{X}_{n+1}, h_n)) \right).$$

And, by definition, we have

$$P_X(\mathbb{B}(\tilde{X}_{n+1}, h_n)) \geq V_d(h_n)^d \cdot \inf_{x \in \mathbb{B}(x_0, 2h_n)} f_X(x).$$

Therefore,

$$\mathbb{E} \left[\frac{1}{N_n + 1} \mid X_{n+1} = x_0 \right] \geq \mathbb{E} \left[\frac{1}{\text{Binomial} \left(n, (h_n)^d \cdot V_d \cdot \inf_{x \in \mathbb{B}(x_0, 2h_n)} f_X(x) \right) + 1} \right].$$

Finally, for any $p \in [0, 1]$, $\mathbb{E} \left[\frac{1}{\text{Binomial}(n, p) + 1} \right] \leq \frac{1}{np}$, which completes the proof of the first claim of the theorem (B.1). The second claim follows by simply observing that the assumptions are sufficient to ensure that every term in the first bound (B.1) is vanishing. $\square$

Lemma B.3. *Let P_0, P_1 be distributions on $\mathbb{R}$, where P_0 is nonatomic. Let $N \geq 0$ be fixed, and suppose $A_1, \dots, A_N \sim P_0$ and $B \sim P_1$, where $A_1, \dots, A_N, B$ are independent. Then*

$$1 - \alpha - d_{\text{TV}}(P_0, P_1) \leq \mathbb{P} \{ B \leq \text{Quantile}_{1-\alpha}(A_1, \dots, A_N, B) \} \leq 1 - \alpha + \frac{1}{N+1} + d_{\text{TV}}(P_0, P_1).$$

Proof. Let $A_{N+1} \sim P_0$ be drawn independently from $A_1, \dots, A_N$. Then, by definition of the total variation distance, we have

$$\begin{aligned} & \left| \mathbb{P} \{ B \leq \text{Quantile}_{1-\alpha}(A_1, \dots, A_N, B) \} - \mathbb{P} \{ A_{N+1} \leq \text{Quantile}_{1-\alpha}(A_1, \dots, A_N, A_{N+1}) \} \right| \\ & \leq d_{\text{TV}}((A_1, \dots, A_N, A_{N+1}), (A_1, \dots, A_N, B)) = d_{\text{TV}}(A_{N+1}, B) = d_{\text{TV}}(P_0, P_1). \end{aligned}$$

Finally,

$$1 - \alpha \leq \mathbb{P} \{ A_{N+1} \leq \text{Quantile}_{1-\alpha}(A_1, \dots, A_N, A_{N+1}) \} \leq 1 - \alpha + \frac{1}{N+1},$$

as in Vovk et al. [2005], since $A_1, \dots, A_{N+1}$ are exchangeable (because they are i.i.d.) and are distinct almost surely (since P_0 is nonatomic). $\square$

B.1.4 Proving the connection between callCP and full conformal prediction

Proof of Proposition 3.7. For any dataset $(x_1, y_1), \dots, (x_{n+1}, y_{n+1})$, the new score in Section 3.6.3 for the i th data point is defined as

$$\left(T((x_1, y_1), \dots, (x_{n+1}, y_{n+1})) \right)_i = \sum_{j=1}^{n+1} w_{i,j} \mathbb{1} \{ s(x_j, y_j) < s(x_i, y_i) \},$$

where $w_{i,j} = \frac{H(x_j, x_i)}{\sum_{k=1}^{n+1} H(x_k, x_i)}$.

This implies the following equivalence between the two scores, s and T :

$$s(x_i, y_i) \leq \text{Quantile}_{1-a} \left(\sum_{j=1}^{n+1} w_{i,j} \delta_{s(x_j, y_j)} \right) \iff \left(T((x_1, y_1), \dots, (x_{n+1}, y_{n+1})) \right)_i < 1-a. \quad (\text{B.2})$$

Now, for a potential test response y , let us consider $Y_1^y, Y_2^y, \dots, Y_{n+1}^y$ as in Section 3.2.2.

Then, the corresponding transformed scores $T_1^y, T_2^y, \dots, T_{n+1}^y$, are defined as

$$T_i^y = T((X_1, Y_1^y), \dots, (X_{n+1}, Y_{n+1}^y))_i = T((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y))_i \text{ for all } i \in [n+1].$$

By the equivalence (B.2) established above, the recalibrated $\tilde{\alpha}(X_{n+1}, y)$, i.e.

$$\max \left\{ a \in \Gamma(w) : \sum_{i=1}^{n+1} \mathbb{1} \left\{ s(X_i, Y_i^y) \leq \text{Quantile}_{1-a} \left(\sum_{j=1}^{n+1} w_{i,j} \delta_{s(X_j, Y_j^y)} \right) \right\} \geq (1-\alpha)(n+1) \right\},$$

with $w_{i,j} = \frac{H(X_j, X_i)}{\sum_{k=1}^{n+1} H(X_k, X_i)}$, $i, j \in [n+1]$, and with $\Gamma(w)$ defined as in (3.8), can now be expressed as

$$\tilde{\alpha}(X_{n+1}, y) = \max \left\{ a \in \Gamma(w) : \sum_{i=1}^{n+1} \mathbb{1} \{ T_i^y < 1-a \} \geq (1-\alpha)(n+1) \right\}.$$

This means that $\tilde{\alpha}(X_{n+1}, y)$ is the largest value in $\Gamma(w)$ satisfying

$$1 - \tilde{\alpha}(X_{n+1}, y) > \text{Quantile}_{1-\alpha}(T_i^y, T_2^y, \dots, T_{n+1}^y).$$

Since $1 - T_i^y$ takes values in the grid $\Gamma(w)$, then, $T_i^y < 1 - \tilde{\alpha}(X_{n+1}, y)$ if and only if $T_i^y \leq \text{Quantile}_{1-\alpha}(T_i^y, T_2^y, \dots, T_{n+1}^y)$. Thus once again using the same equivalence (B.2), one can write $\hat{C}_n^{\text{calLCP}}(X_{n+1})$ as

$$\begin{aligned} \hat{C}_n^{\text{calLCP}}(X_{n+1}) &= \left\{ y \in \mathcal{Y} : T_{n+1}^y < 1 - \tilde{\alpha}(X_{n+1}, y) \right\} \\ &= \left\{ y \in \mathcal{Y} : T_{n+1}^y \leq \text{Quantile}_{1-\alpha}(T_i^y, T_2^y, \dots, T_{n+1}^y) \right\}. \end{aligned}$$

The set on the right-hand side is exactly the output of the full conformal prediction interval, using the score map T . This establishes the result that $\hat{C}_n^{\text{calLCP}}(X_{n+1})$ and $\hat{C}_n^{\text{fullCP}}(X_{n+1})$ are exactly same. $\square$

B.1.5 A connection between m -RLCP and baseLCP

To understand the behavior of m -RLCP from another perspective, we will look at an equivalent formulation. Define an average weight,

$$W(X_i, X_{n+1}) = \mathbb{E}_{\tilde{X}_{n+1} | X_{n+1} \sim H(X_{n+1}, \cdot)} \left[\frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \right],$$

which is the expected value of the weight $\tilde{w}_i$ placed on training point i when we run RLCP, averaged over the random draw of $\tilde{X}_{n+1}$. We then define

$$\tilde{p}(y; X_{n+1}) = \sum_{i=1}^n W(X_i, X_{n+1}) \mathbb{1} \{s(X_i, Y_i) \geq s(X_{n+1}, y)\} + W(X_{n+1}, X_{n+1}).$$

Comparing to the p-value $p(y; X_{n+1}, \tilde{X}_{n+1})$ that defines the RLCP method (see (3.20)), we can see that this construction is identical except that we use the expected weights in place of the random $\tilde{w}_i$'s. We then correspondingly define a prediction interval

$$\tilde{C}_n(X_{n+1}) = \{y \in \mathcal{Y} : \tilde{p}(y; X_{n+1}) > \alpha\}.$$

We now make a key observation: $\tilde{C}_n(X_{n+1})$ is actually equivalent to the baseLCP prediction interval $C_n^{\text{baseLCP}}(X_{n+1})$, if in place of the localizer kernel H , we use a new kernel defined by W (note, however, that we are slightly abusing the notation here, since $W(X_i, X_{n+1})$ depends not only on the two feature values X_i, X_{n+1} but also on the remaining data points in the training set).

Now we again consider m -RLCP. Observe that the derandomized p-value $\bar{p}(y; X_{n+1}, \tilde{X}_{n+1}^{[m]})$, defined in (3.21), can be viewed as a stochastic approximation to $\tilde{p}(y; X_{n+1})$ (i.e., the expected weight, $W(X_i, X_{n+1})$, is approximated by averaging m independent samples). Thus, $\hat{C}_n^{m\text{-RLCP}}(X_{n+1})$ can be viewed as an approximate version of the interval $\tilde{C}_n(X_{n+1})$, which is the interval produced by baseLCP if run with a different kernel.

Based on this connection, we see that the empirically conservative behavior of m -RLCP is not surprising; our experiments, and the experiments of Guan [2023], have demonstrated that baseLCP can often be quite conservative, and this property is then inherited by m -RLCP due to the connection between these two methods.

B.2 Details for smoothing to avoid overcoverage

In order to have a meaningful comparison between calLCP and RLCP in terms of local coverage, we should firstly have exact marginal coverage guarantee for both methods. For all methods in the conformal prediction framework (e.g., split CP), the standard theory

ensures that marginal coverage will be *at least* $1 - \alpha$, but the method can overcover (i.e., coverage strictly higher than $1 - \alpha$). In the exchangeable setting, any overcoverage is due only to two things: first, the issue of rounding (e.g., if $(1 - \alpha)(n + 1)$ is a noninteger, then calculating the $(1 - \alpha)$ -quantile for split conformal prediction as in (1.2) is necessarily slightly conservative), and second, the possibility of ties between scores.

For split and full conformal prediction (i.e., in the unweighted and unlocalized setting), Vovk et al. [2005, Proposition 2.4] propose a randomization strategy, often also referred as “smoothed conformal predictors”, to ensure exact coverage and avoid overcoverage. Taking inspiration from that randomization strategy, we will now develop the smoothed version of our method.

We will begin this section with reviewing the weighted conformal prediction (WCP) framework, and presenting its smoothed version. We will then extend these ideas to the setting of localized methods.

B.2.1 Smoothing for weighted conformal prediction (WCP)

The idea behind RLCP (and indeed, behind baseLCP and calLCP as well) is closely connected to the weighted conformal prediction (WCP) framework [Tibshirani et al., 2019]. The weighted conformal prediction (WCP) method (Tibshirani et al. [2019]) builds a prediction interval of the form

$$\hat{C}_n^{\text{WCP}}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq \text{Quantile}_{1-\alpha} \left(\sum_{i=1}^n w_i \delta_{s(X_i, Y_i)} + w_{n+1} \delta_{+\infty} \right) \right\},$$

for any pre-trained score function s . In the covariate shift setting, where we have training data $\{(X_i, Y_i)\}_{i \in [n]} \stackrel{\text{iid}}{\sim} P_X \times P_{Y|X}$ and a test point $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$, the weight

for i th data point is given by

$$w_i = \frac{\frac{d\tilde{P}_X}{dP_X}(X_i)}{\sum_{j=1}^{n+1} \frac{d\tilde{P}_X}{dP_X}(X_j)}.$$

The marginal validity of this prediction interval ensures $\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{WCP}}(X_{n+1}) \right\} \geq 1 - \alpha$.

In general, any weighted conformal prediction interval can be extremely conservative, much more so than (unweighted) conformal prediction. This is because, while computing the $(1 - \alpha)$ -quantile, $\text{Quantile}_{1-\alpha}(\cdot)$ for an unweighted empirical distribution over $n + 1$ many values, rounding means that we take the k -th smallest value for $k = \lceil (1 - \alpha)(n + 1) \rceil$; in other words, we are actually computing the quantile at level $\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}$, but this value is bounded as $\leq 1 - \alpha + \frac{1}{n+1}$. However, for WCP, the conservativeness caused by rounding when computing the quantile of a *weighted* discrete distribution can potentially be as large as the largest weight—which might be much larger than $\frac{1}{n+1}$, if the weights are concentrated on just a few samples (i.e., low effective sample size). This means that providing a smoothed version to avoid overcoverage is particularly important in the weighted setting.

To do so, we begin by returning to the p-value based interpretation. An alternative way to express the construction of WCP prediction interval would be as follows. Define a p-value for any potential response y by comparing it with the weighted empirical distribution, where the weight for i th data point is defined as above. Then, $\hat{C}_n^{\text{WCP}}(X_{n+1})$ can be also expressed as $\{y \in \mathcal{Y} : p^{\text{WCP}}(y) > \alpha\}$ where

$$p^{\text{WCP}}(y) = \sum_{i=1}^n w_i \mathbb{1} \{s(X_i, Y_i) \geq s(X_{n+1}, y)\} + w_{n+1}.$$

In the covariate shift setting described above, this p-value can be shown to follow a super-uniform distribution, i.e., $\mathbb{P} \left\{ p^{\text{WCP}}(Y_{n+1}) \leq \alpha \right\} \leq \alpha$; this claim is equivalent to the claim that $Y_{n+1} \in \hat{C}_n^{\text{WCP}}(X_{n+1})$ with probability at least $1 - \alpha$.

Now, to apply smoothing, we need to adjust this p-value so that its distribution is *exactly* uniform, which will ensure that $Y_{n+1} \in \hat{C}_n^{\text{WCP}}(X_{n+1})$ holds with probability *exactly* $1 - \alpha$. We modify the definition of the p-value to now be given by

$$p^{\text{WCP}}(y) = \sum_{i=1}^n w_i \mathbb{1} \{s(X_i, Y_i) > s(X_{n+1}, y)\} + U \left(\sum_{i=1}^n w_i \mathbb{1} \{s(X_i, Y_i) = s(X_{n+1}, y)\} + w_{n+1} \right),$$

where $U \sim \text{Uniform}[0, 1]$ is drawn independently of the data (note that typically, a single draw of U is used across all possible values $y \in \mathcal{Y}$). Under the covariate shift setting, the new p-value is now exactly uniformly distributed. This construction is analogous to the smoothed version of (unweighted) split conformal prediction, as proposed by Vovk et al. [2005]. Then the smoothed prediction interval, defined as $\hat{C}_n^{\text{WCP}}(X_{n+1}) = \{y \in \mathcal{Y} : p^{\text{WCP}}(y) > \alpha\}$, satisfies marginal coverage at level exactly $1 - \alpha$:

Proposition B.4. *Under the notation and definitions above, the smoothed WCP prediction interval has marginal coverage exactly equal to $1 - \alpha$ under the covariate shift model.*

Proof. Let $\hat{P}_{\text{score}} = \sum_{i=1}^{n+1} w_i \delta_{s(X_i, Y_i)}$ be the weighted empirical distribution of the (training and test) score values, where $w_i \propto \frac{d\tilde{P}_X}{dP_X}(X_i)$ as defined above. Following Tibshirani et al. [2019]’s formulation of the covariate shift setting through the language of weighted exchangeability (see, e.g., the proof of Tibshirani et al. [2019, Lemma 1]), we can see that the conditional distribution of the test score is given by

$$s(X_{n+1}, Y_{n+1}) \mid \hat{P}_{\text{score}} \sim \hat{P}_{\text{score}}.$$

In words, if we condition on the unordered set (or multiset, in the case of repeated values) of scores $\{s(X_i, Y_i)\}_{i \in [n+1]}$, the chance that the test score $s(X_{n+1}, Y_{n+1})$ takes any one of

these values is proportional to the weight assigned to that value. Then

$$\begin{aligned}\mathbb{P}\left\{Y_{n+1} \in \hat{C}_n^{\text{WCP}}(X_{n+1})\right\} &= 1 - \mathbb{P}\left\{p^{\text{WCP}}(Y_{n+1}) \leq \alpha\right\} \\ &= 1 - \mathbb{E}\left[\mathbb{P}\left\{p^{\text{WCP}}(Y_{n+1}) \leq \alpha \mid \hat{P}_{\text{score}}\right\}\right],\end{aligned}$$

so it suffices to show that

$$\mathbb{P}\left\{p^{\text{WCP}}(Y_{n+1}) \leq \alpha \mid \hat{P}_{\text{score}}\right\} = \alpha$$

holds almost surely. By definition of $p^{\text{WCP}}(y)$ observe that we can calculate

$$p^{\text{WCP}}(Y_{n+1}) = \sum_{i=1}^{n+1} w_i \mathbb{1}\{s(X_i, Y_i) > s(X_{n+1}, Y_{n+1})\} + U \left(\sum_{i=1}^{n+1} \mathbb{1}\{s(X_i, Y_i) = s(X_{n+1}, Y_{n+1})\} \right),$$

or equivalently, writing

$$p(s) = \mathbb{P}_{S \sim \hat{P}_{\text{score}}} \{S > s\} + U \cdot \mathbb{P}_{S \sim \hat{P}_{\text{score}}} \{S = s\},$$

we have

$$p^{\text{WCP}}(Y_{n+1}) = p(s(X_{n+1}, Y_{n+1})).$$

Now, since we know that, conditionally on $\hat{P}_{\text{score}}$ we have $s(X_{n+1}, Y_{n+1}) \sim \hat{P}_{\text{score}}$, this means that

$$\begin{aligned}\mathbb{P}\left\{p^{\text{WCP}}(Y_{n+1}) \leq \alpha \mid \hat{P}_{\text{score}}\right\} &= \mathbb{P}\left\{p(s(X_{n+1}, Y_{n+1})) \leq \alpha \mid \hat{P}_{\text{score}}\right\} \\ &= \mathbb{P}_{S \sim \hat{P}_{\text{score}}} \left\{p(S) \leq \alpha \mid \hat{P}_{\text{score}}\right\} = \alpha,\end{aligned}$$

where the last step holds by definition of $p(s)$. This completes the proof. $\square$

B.2.2 From smoothed WCP to smoothed RLCP

First, we recall the p-value based formulation of RLCP, as in Section 3.6.2: writing

$$p^{\text{RLCP}}(y) = \sum_{i=1}^n \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \mathbb{1} \{s(X_i, Y_i) \geq s(X_{n+1}, y)\} + \frac{H(X_{n+1}, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})},$$

the prediction interval is given by $\hat{C}_n^{\text{RLCP}}(X_{n+1}) = \{y \in \mathcal{Y} : p^{\text{RLCP}}(y) > \alpha\}$.

As we observed in the proof of Proposition 3.2, which established coverage for RLCP conditional on the “synthetic prototype” $\tilde{X}_{n+1}$, we can think of RLCP as constructing a weighted conformal prediction interval for the covariate shift problem that is (randomly) defined via the shifted feature distribution $\tilde{P}_X = P_X \circ H(\cdot, \tilde{X}_{n+1})$. In other words, in a sense we can view RLCP as an instance of WCP. This means that RLCP can be smoothed in the same way. Thus, we redefine the p-value as

$$p^{\text{RLCP}}(y) = \sum_{i=1}^n \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \mathbb{1} \{s(X_i, Y_i) > s(X_{n+1}, y)\} \\ + U \left(\sum_{i=1}^n \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \mathbb{1} \{s(X_i, Y_i) = s(X_{n+1}, y)\} + \frac{H(X_{n+1}, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \right),$$

where $U \sim \text{Uniform}[0, 1]$ is drawn independently of the data, and again define $\hat{C}_n^{\text{RLCP}}(X_{n+1}) = \{y \in \mathcal{Y} : p^{\text{RLCP}}(y) > \alpha\}$. This leads to the following marginal coverage guarantee:

Theorem B.5 (Marginal coverage for smoothed RLCP). *The smoothed RLCP method defined above satisfies the exact marginal coverage property, i.e., for any distribution P , if $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$, then*

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} = 1 - \alpha.$$

Proof. Following the same arguments as in the proof of Theorem 3.3, we can see that it

suffices to prove a result analogous to Proposition 3.2: that is, we need to establish that for smoothed RLCP, it holds that

$$\mathbb{E} \left[\alpha(X_{n+1}, \tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \right] = \alpha \text{ almost surely.}$$

As in the proof of Proposition 3.2, this result follows from reinterpreting RLCP as an instance of WCP, in the covariate shift setting induced by conditioning on $\tilde{X}_{n+1}$; since Proposition B.4 verifies exact marginal coverage under covariate shift for smoothed WCP, this completes the proof. $\square$

Consequences for coverage with respect to unknown covariate shift

In Theorem 3.6, we established a lower bound on coverage for the RLCP method, for covariate shifts of the form $P_X \circ g$ where g is unknown. In particular, in Section 3.4.3, we discussed that the result of the theorem arises from the bound

$$\mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} \geq 1 - \alpha - \mathbb{E}_{\tilde{P}_{\tilde{X}}} \left[d_{\text{TV}} \left(P_X \circ H(\cdot, \tilde{X}), P_X \circ g \circ H(\cdot, \tilde{X}) \right) \right],$$

since this expected total variation term is bounded by the excess error term in the theorem.

For the smoothed setting, this expected total variation distance bounds the deviation from $1 - \alpha$ coverage in either direction—that is, it bounds overcoverage as well as undercoverage. In particular, we have the following result:

Theorem B.6 (Robustness to covariate shift for smoothed RLCP). *The smoothed RLCP method defined above satisfies the following property: for any distribution P , for every g :*

$\mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ with $P_X \circ g$ well-defined (i.e., with $0 < \mathbb{E}_{P_X}[g(X)] < \infty$) it holds that

$$\begin{aligned} \left| \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \right\} - (1 - \alpha) \right| &\leq \mathbb{E}_{\tilde{P}_{\tilde{X}}} \left[d_{\text{TV}} \left(P_X \circ H(\cdot, \tilde{X}), P_X \circ g \circ H(\cdot, \tilde{X}) \right) \right] \\ &\leq \frac{\inf_{A \subseteq \mathcal{X}, \epsilon > 0} \left\{ \epsilon \cdot L_{g, 2\epsilon, A} + \|g\|_{\infty} \cdot P_{(X, \tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} + \|g\|_{\infty} \cdot P_X(A^c) \right\}}{\mathbb{E}_{P_X}[g(X)]}, \end{aligned}$$

where the probability on the left-hand side is taken with respect to if $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$ and $(X_{n+1}, Y_{n+1}) \sim (P_X \circ g) \times P_{Y|X}$.

We omit the proof, since it follows from analogous arguments as the results above.

This result for covariate shift also leads to a result on relaxed test-conditional coverage (just like, for the unsmoothed version of RLCP, Theorem 3.4 is a special case of Theorem 3.6): for $\{(X_i, Y_i)\}_{i \in [n+1]} \stackrel{\text{iid}}{\sim} P$, and for any subset $B \subseteq \mathcal{X}$, we have

$$\begin{aligned} \left| \mathbb{P} \left\{ Y_{n+1} \in \hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1}) \mid X_{n+1} \in B \right\} - (1 - \alpha) \right| \\ \leq \frac{\inf_{\epsilon > 0} \left\{ P_{(X, \tilde{X})} \{ \|X - \tilde{X}\| > \epsilon \} + P_X(\text{bd}_{2\epsilon}(B)) \right\}}{P_X(B)}. \end{aligned}$$

B.2.3 Smoothed versions of baseLCP and calLCP

For calLCP, Guan [2023, Theorem 2.6] proposes a smoothed version of the calLCP method, and verifies that it achieves exactly $1 - \alpha$ marginal coverage. An interesting observation is that, since we have seen in Proposition 3.7 that calLCP is actually an instance of full CP implemented with a particular choice of the score function, Guan [2023, Theorem 2.6]’s smoothing approach can also be derived by formulating calLCP as an instance of full CP, and then applying Vovk et al. [2005, Proposition 2.4]’s smoothed version of full CP. Specifically,

defining weights $w_{i,j} = \frac{H(X_j, X_i)}{\sum_{k=1}^{n+1} H(X_k, X_i)}$ as before, we compute a smoothed p-value as

$$p^{\text{calLCP}}(y) = \sum_{i=1}^n \frac{1}{n+1} \mathbb{1} \{T_i^y > T_{n+1}^y\} + U \left(\sum_{i=1}^n \frac{1}{n+1} \mathbb{1} \{T_i^y = T_{n+1}^y\} + \frac{1}{n+1} \right),$$

where $U \sim \text{Uniform}[0, 1]$ is drawn independently of the data, and where we define $T^y \in \mathbb{R}^{n+1}$ with entries

$$T_i^y = \begin{cases} \sum_{j=1}^n w_{i,j} \mathbb{1} \{s(X_j, Y_j) < s(X_i, Y_i)\} + w_{i,n+1} \mathbb{1} \{s(X_{n+1}, y) < s(X_i, Y_i)\}, & i \in [n], \\ \sum_{j=1}^n w_{n+1,j} \mathbb{1} \{s(X_j, Y_j) < s(X_{n+1}, y)\}, & i = n+1. \end{cases}$$

For baseLCP, as we have discussed, the method does not in fact offer a guarantee of marginal coverage $\geq 1 - \alpha$. Nonetheless, to treat the methods in a comparable way, we implement a smoothed version of baseLCP. Recalling our discussion in Section 3.2.3, since baseLCP is also constructed based on the WCP framework (as is RLCP), we can define a smoothed version of baseLCP by leveraging the smoothed version of WCP. To do so, first we reformulate (unsmoothed) baseLCP via a p-value: for weights $w_{i,j}$ as above, defining

$$p^{\text{baseLCP}}(y) = \sum_{i=1}^n \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \mathbb{1} \{s(X_i, Y_i) \geq s(X_{n+1}, y)\} + \frac{H(X_{n+1}, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})},$$

the baseLCP prediction interval can be equivalently written as $\hat{C}_n^{\text{baseLCP}}(X_{n+1}) = \{y \in \mathcal{Y} : p^{\text{baseLCP}}(y) > \alpha\}$. We then replace the p-value with a smoothed version,

$$p^{\text{baseLCP}}(y) = \sum_{i=1}^n \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \mathbb{1} \{s(X_i, Y_i) > s(X_{n+1}, y)\} + U \left(\sum_{i=1}^n \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \mathbb{1} \{s(X_i, Y_i) = s(X_{n+1}, y)\} + \frac{H(X_{n+1}, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \right),$$

where again $U \sim \text{Uniform}[0, 1]$ is drawn independently of the data, and then redefine $\hat{C}_n^{\text{baseLCP}}(X_{n+1}) = \{y \in \mathcal{Y} : p^{\text{baseLCP}}(y) > \alpha\}$ to obtain a smoothed interval.

B.2.4 A note on implementation

For all three methods, the p-value based formulation immediately suggests an efficient implementation to compute the prediction interval $\hat{C}_n(X_{n+1})$. Specifically, let $s_{(1)} < \dots < s_{(m)}$ be the *unique* values of the scores $\{s(X_i, Y_i)\}_{i \in [n]}$, and then choose any values $s_{(0)}^*, \dots, s_{(m)}^*$ satisfying

$$s_{(0)}^* < s_{(1)} < s_{(1)}^* < s_{(2)} < \dots < s_{(m-1)} < s_{(m-1)}^* < s_{(m)} < s_{(m)}^*.$$

Then it suffices to evaluate the (deterministic or smoothed) p-value at each of these $2m + 1$ many scores. Specifically, abusing notation and writing the p-value as a function of the score $s = s(X_{n+1}, y)$ rather than of the response value y ,

- For RLCP, for any $s \in \mathbb{R}$, define

$$p^{\text{RLCP}}(s) = \sum_{i=1}^n \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \mathbb{1}\{s(X_i, Y_i) > s\} + U \left(\sum_{i=1}^n \frac{H(X_i, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \mathbb{1}\{s(X_i, Y_i) = s\} + \frac{H(X_{n+1}, \tilde{X}_{n+1})}{\sum_{j=1}^{n+1} H(X_j, \tilde{X}_{n+1})} \right),$$

where $U \sim \text{Uniform}[0, 1]$ (for the smoothed version) or $U \equiv 1$ (for the deterministic version).

- For calLCP, for any $s \in \mathbb{R}$, define

$$p^{\text{calLCP}}(s) = \sum_{i=1}^n \frac{1}{n+1} \mathbb{1}\{T_i^s > T_{n+1}^s\} + U \left(\sum_{i=1}^n \frac{1}{n+1} \mathbb{1}\{T_i^s = T_{n+1}^s\} + \frac{1}{n+1} \right),$$

where $U \sim \text{Uniform}[0, 1]$ (for the smoothed version) or $U \equiv 1$ (for the deterministic version), and where

$$T_i^s = \begin{cases} \sum_{j=1}^n w_{i,j} \mathbb{1} \{s(X_j, Y_j) < s(X_i, Y_i)\} + w_{i,n+1} \mathbb{1} \{s < s(X_i, Y_i)\}, & i \in [n], \\ \sum_{j=1}^n w_{n+1,j} \mathbb{1} \{s(X_j, Y_j) < s\}, & i = n+1. \end{cases}$$

- For baseLCP, for any $s \in \mathbb{R}$, define

$$p^{\text{baseLCP}}(s) = \sum_{i=1}^n \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \mathbb{1} \{s(X_i, Y_i) > s\} \\ + U \left(\sum_{i=1}^n \frac{H(X_i, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \mathbb{1} \{s(X_i, Y_i) = s\} + \frac{H(X_{n+1}, X_{n+1})}{\sum_{j=1}^{n+1} H(X_j, X_{n+1})} \right),$$

where $U \sim \text{Uniform}[0, 1]$ (for the smoothed version) or $U \equiv 1$ (for the deterministic version).

(For each method, in the smoothed case, note that a single random draw of U is shared across all score values.)

Finally, for each method, writing $p(s)$ to denote $p^{\text{RLCP}}(s)$ or $p^{\text{calLCP}}(s)$ or $p^{\text{baseCP}}(s)$ as appropriate, we can observe that $p(s)$ is a nonincreasing function of s (i.e., a higher test score corresponds to a lower p-value). This means that the prediction interval can be defined as

follows:

$$\hat{C}_n(X_{n+1}) = \begin{cases} \mathcal{Y}, & p(s_{(m)}^*) > \alpha, \\ \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq s_{(m)}\}, & p(s_{(m)}^*) \leq \alpha < p(s_{(m)}), \\ \{y \in \mathcal{Y} : s(X_{n+1}, y) < s_{(m)}\}, & p(s_{(m)}) \leq \alpha < p(s_{(m-1)}^*), \\ \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq s_{(m-1)}\}, & p(s_{(m-1)}^*) \leq \alpha < p(s_{(m-1)}), \\ \dots & \\ \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq s_{(1)}\}, & p(s_{(1)}^*) \leq \alpha < p(s_{(1)}), \\ \{y \in \mathcal{Y} : s(X_{n+1}, y) < s_{(1)}\}, & p(s_{(1)}) \leq \alpha < p(s_{(0)}^*), \\ \emptyset, & p(s_{(0)}^*) < \alpha. \end{cases}$$

(See Guan [2023, Lemma 3.2] for an equivalent derivation of this implementation in the case of calLCP specifically.)

B.3 Additional experiments and implementation details

B.3.1 Quantitative analysis of the effect of randomization on RLCP prediction intervals

In Figure 3.8, we showed a visual representation of the effect of randomization on the variability of the RLCP prediction interval. In the following experiments, we aim to quantify this variability in a more precise way, to better understand the extent to which the RLCP prediction interval can vary across different random draws of $\tilde{X}_{n+1}$. In particular, we will consider the following measure of variability:

$$D(h) = \mathbb{E}_{D_n, X_{n+1}} \left[\frac{\text{MAD} \left(\lambda(\hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1})) \mid D_n, X_{n+1} \right)}{\text{Median} \left(\lambda(\hat{C}_n^{\text{RLCP}}(X_{n+1}, \tilde{X}_{n+1})) \mid D_n, X_{n+1} \right)} \right] \quad (\text{B.3})$$

where λ denotes the Lebesgue measure, and where MAD, the median absolute deviation, is defined for a random variable X as $\text{Median}(|X - \text{Median}(X)|)$. In other words, $D(h)$ measures the extent to which the width of the RLCP prediction interval can vary, relative to the typical width of this interval, given a *fixed* data set (i.e., the only distribution being considered here is the random draw of $\tilde{X}_{n+1}$). We plot (an estimate of) this measure $D(h)$ for RLCP prediction intervals against a wide range of bandwidths h , for Setting 1 and Setting 2 of the univariate simulations that were carried out in Section 3.5.1 (and analyzed in terms of the randomness of RLCP in Section 3.6.2). The results are shown in the left panel of Figure B.1. The results suggest that, at least for smaller bandwidths, a random instance of RLCP prediction interval can be approximately 10% wider or narrower than the average interval width, so the deviation is small but by no means negligible. For larger bandwidths, since the localization effect is now reduced, the variability is consequently lower.

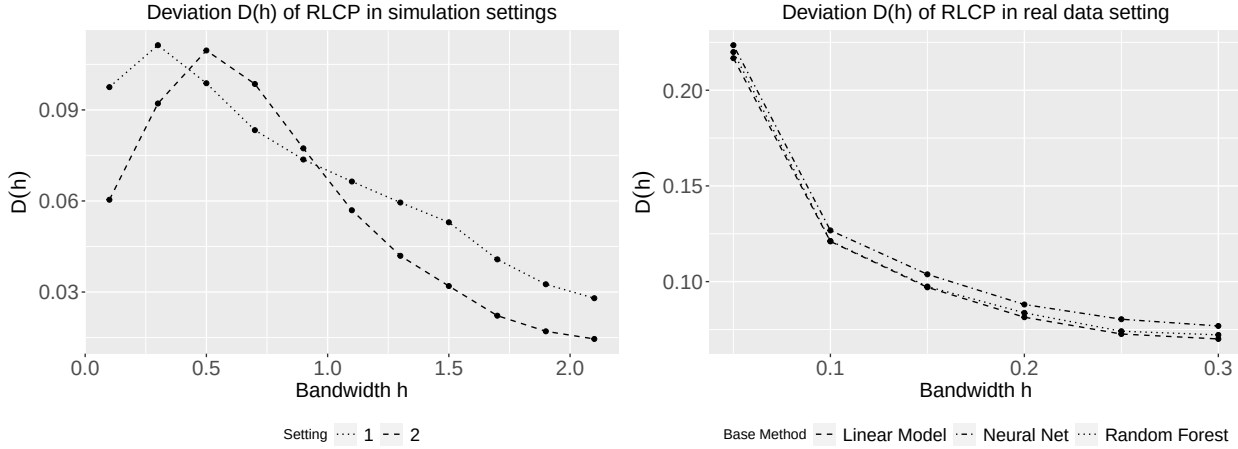


Figure B.1: Quantification of the effect of randomization on RLCP via the estimated deviation $D(h)$, for the univariate simulations (Section 3.5.1) and the real data example (Section 3.5.2). See Section B.3.1 for details.

We perform a similar analysis for our real data experiment (from Section 3.5.2), with results shown in the right panel of Figure B.1. These results show that, for our real data experiment, the RLCP prediction interval is more variable; a given draw of the RLCP prediction interval may be approximately 10–20% wider or narrower in smaller bandwidth regimes and 5–10%

wider for relatively larger bandwidths, compared to the typical width.

B.3.2 Experiments on m -RLCP

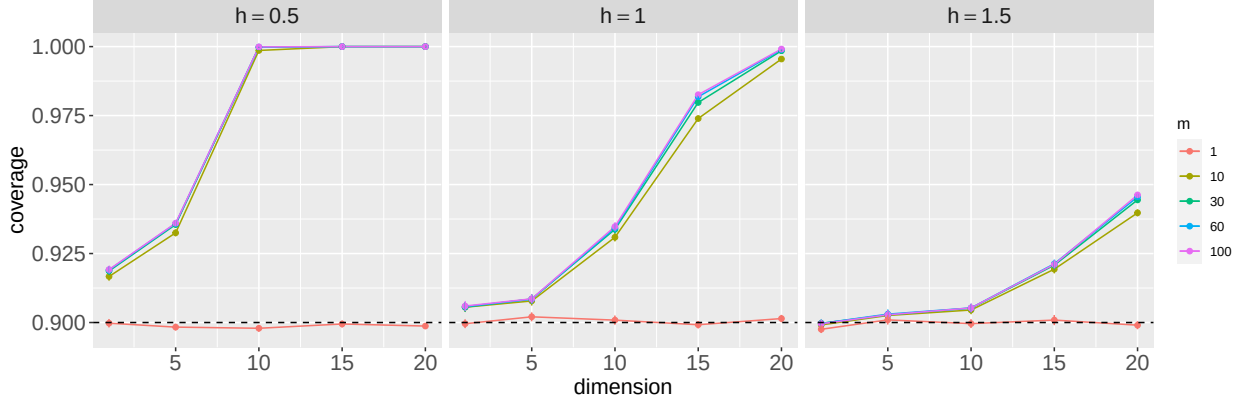


Figure B.2: Marginal coverage of m -RLCP against increasing dimension for different choices of m .

The theoretical guarantee for m -RLCP only considers the issue of potential undercoverage (i.e., it provides a lower bound on coverage). In practice, m -RLCP can be extremely conservative. In Figure B.2, we plot the marginal coverage of m -RLCP against increasing dimension d , for the simulation setting previously studied in Section 3.5.1. The plots in Figure B.2 show marginal coverage over 30 independent trials, with pretraining, calibration and test sample size each being 2000. Note that m -RLCP for $m = 1$ is equivalent to the original RLCP method, while choosing large values of m leads to a more derandomized method. While RLCP (i.e., m -RLCP with $m = 1$) shows coverage at the target level $1 - \alpha$, all larger values of m show substantial overcoverage and are thus extremely conservative, meaning that this form of aggregation does not offer a practical benefit.

B.3.3 Additional experiments on local coverage in univariate settings

Local coverage experiments with additional bandwidths

In Section 3.5.1, we showed results for our experiment comparing baseLCP, calLCP, and RLCP in the univariate setting, in terms of their local coverage properties across the range of values of the feature X . Here we show additional plots for this same experiment, with an expanded range of bandwidths, $h \in \{0.1, 0.2, 0.4, 0.8, 1.6\}$. Settings 1 and 2 are shown in Figures B.3 and B.4. All details for implementation are exactly as in Section 3.5.1.

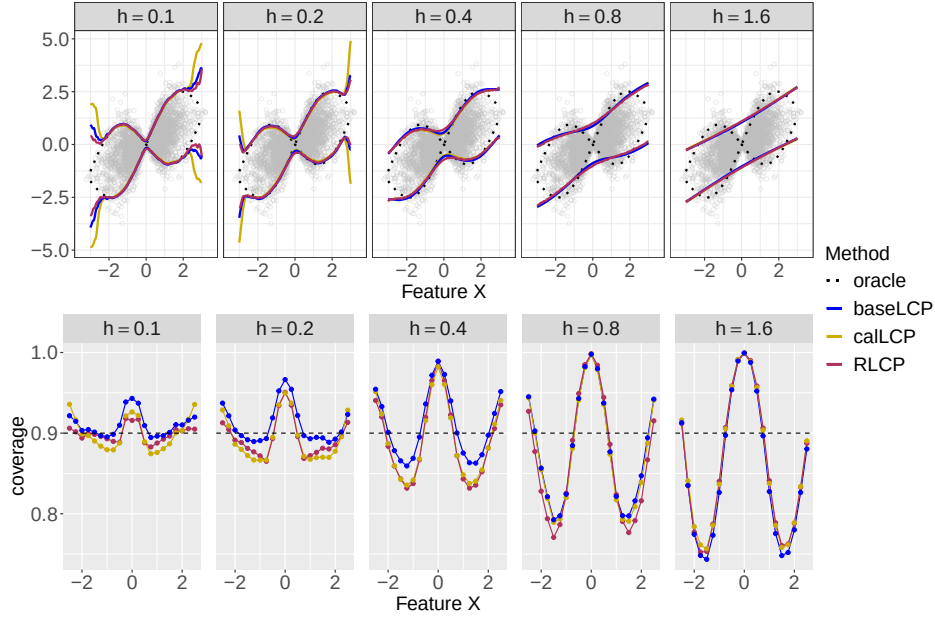


Figure B.3: Additional results for Setting 1. Top panels: the average prediction intervals produced by each method (at each feature value x , the endpoints of the prediction interval are averaged over 50 independent trials), compared to the oracle prediction interval. Bottom panels: The corresponding local coverage of each method, averaged over 50 independent trials. See Section 3.5.1 and Appendix B.3.3 for details.

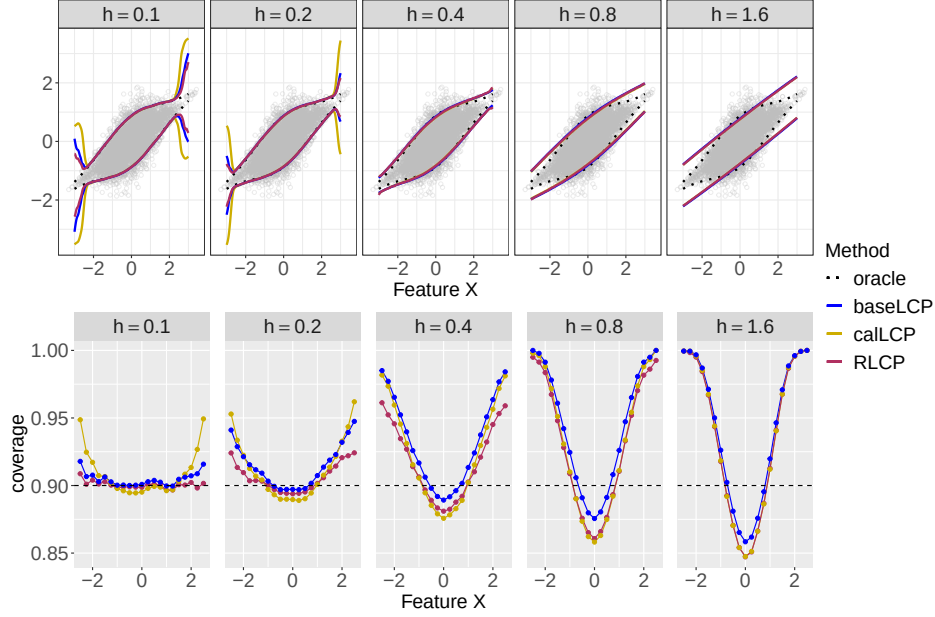


Figure B.4: Additional results for Setting 2. Top panels: the average prediction intervals produced by each method (at each feature value x , the endpoints of the prediction interval are averaged over 50 independent trials), compared to the oracle prediction interval. Bottom panels: The corresponding local coverage of each method, averaged over 50 independent trials. See Section 3.5.1 and Appendix B.3.3 for details.

Comparison with Gibbs et al. [2023]’s method

In Section 3.6.1, we have seen how localization based approaches, such as RLCP, and calibration based approaches, such as the method of Gibbs et al. [2023], fit into a more general framework of conformal algorithms. As mentioned before, the localization step of RLCP with a particular bandwidth, (Step (b) of the three-step workflow described in Section 3.6.1), and the calibration step of Gibbs et al. [2023] with respect to a particular functional class $\mathcal{F}$ (Step (c) of the workflow) can be viewed as approximately interchangeable—both approaches serve to approximate test-conditional coverage via a local relaxation. In particular, the two approaches lead to a similar trade-off between localization and sample size. Choosing a richer class of functions $\mathcal{F}$ (for the method of Gibbs et al. [2023]) or a smaller bandwidth for the kernel H (for RLCP) would have similar effects on the outcome, by placing a stronger requirement on the coverage properties of the algorithm and requiring more adaptivity to local

properties of the data; on the other hand, for both types of methods, this would come at a cost of reduced effective sample size, and therefore, the potential to return wider or less informative prediction intervals.

In order to compare these methods, and better understand how they share the same trade-off, we repeat the univariate experiment of Section 3.5.1 (Setting 1).² In Figure B.5 we plot the local coverage (as defined in Section 3.5.1) for each method, averaged over 50 independent trials. Both Gibbs et al. [2023]’s method and RLCP are implemented in two versions, “high adaptivity” and “low adaptivity”. For Gibbs et al. [2023]’s method, the function class is chosen to be more fine or more coarse,

$$\mathcal{F}_i = \{\mathbb{1}\{x \in G\} : G \in \mathcal{G}_i\} \text{ for } i = 1, 2$$

where, $\mathcal{G}_1, \mathcal{G}_2$ are partitions of $[-2.5, 2.5]$ into 8 bins of equal size or 4 bins of equal size, respectively. For RLCP, we use a Gaussian kernel with bandwidth

$$h_1 = 0.025 \text{ or } h_2 = 0.05.$$

The richer function class $\mathcal{F}_1$, and similarly the smaller bandwidth h_1 —both provide more localization and “high adaptivity”, and lead to better local coverage, as compared to $\mathcal{F}_2$ or h_2 —the choices with “low adaptivity”. For both methods, the “high adaptivity” version shows local coverage lying very close to target coverage 0.9 uniformly over the feature space, but we see larger fluctuations for the versions with “low adaptivity”.

2. The method of Gibbs et al. [2023] is run using the code provided in that paper, which is available at <https://github.com/jjcherian/conditional-conformal>.

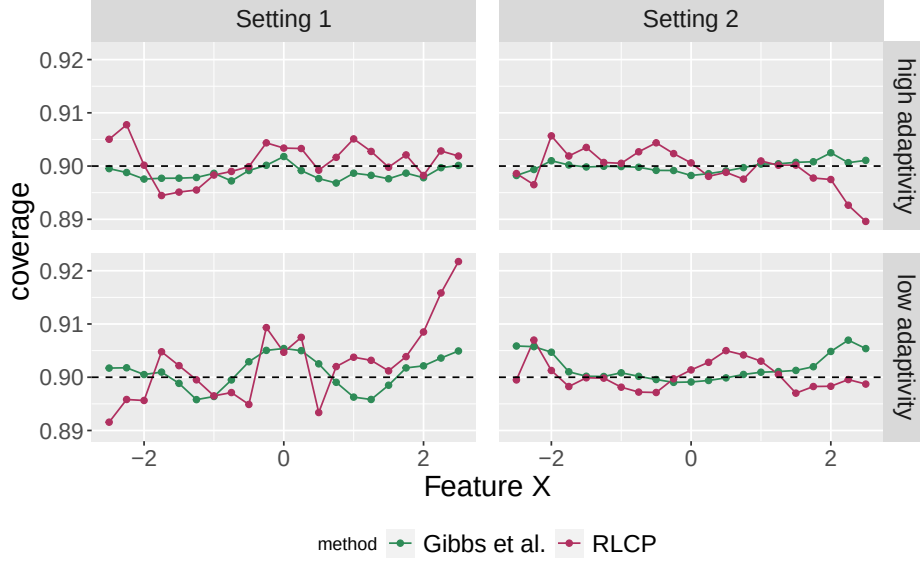


Figure B.5: Local coverage of RLCP and the method from Gibbs et al. [2023] in setting 1 for $d = 1$

B.3.4 Additional experiments on local coverage in multivariate setting

In this section, we expand our experiments on local coverage in the multivariate setting. Recall in Section 3.5.1 that we have studied coverage conditional on two sets, B_{in} and B_{out} , which are defined to characterize the centre and the tails of the multivariate normal data distribution. Here, to take a closer look at local coverage properties of RLCP as compared to calLCP, we will consider larger collections of sets by taking a finer partition of the space. Ideally, we will want to see empirical coverage on these sets to cluster around the target level 0.9—in particular, if each set B is not too small, for RLCP our coverage guarantee in Theorem 3.4 ensures approximate coverage on each B .

Specifically, we consider the data distribution given by

$$X \sim \text{Uniform}([-3, 3]^d), \quad Y|X \sim \mathcal{N}\left(\sum_{i=1}^d \frac{X_i}{2}, \sum_{i=1}^d |\sin(X_i)|\right),$$

where $d \in \{5, 10, \dots, 30\}$. We define intervals $I_i = [i, i + 1)$, and we note that $\{I_i : i = -3, -2, -2, 0, 1, 2\}$ define a partition of $[-3, 3)$. Let us write the feature x as $(x_1, x_2, \dots, x_d)$, and consider the collection of sets B of the form $B = I_i \times I_j \times I_k \times \mathbb{R}^{d-3}$,

$$\mathcal{B} = \left\{ I_i \times I_j \times I_k \times \mathbb{R}^{d-3} : i, j, k \in \{-3, -2, -2, 0, 1, 2\} \right\}.$$

This is a partition of the support of the distribution of X —we use the first 3 coordinates of x to determine which set $B \in \mathcal{B}$, it belongs to.

We compute the empirical coverage of calLCP and RLCP conditional on each set $B \in \mathcal{B}$, averaged over 30 repetitions. Each of the methods are run with 3000 training and calibration samples, and 5000 test samples with their own dimension-adaptive bandwidth choices, as defined in Section 3.5.1, which maintains a constant effective sample size of 50 across all values of dimension d . Finally, in Figure B.6 we plot the kernel density estimates of the conditional coverages on these partitions across different dimensions. While both methods show conditional coverage that is distributed near the target value 0.9, we do observe a difference between the two methods—in higher dimensions, the conditional coverage values are more widely spread around the target value 0.9 for calLCP, as compared to RLCP.

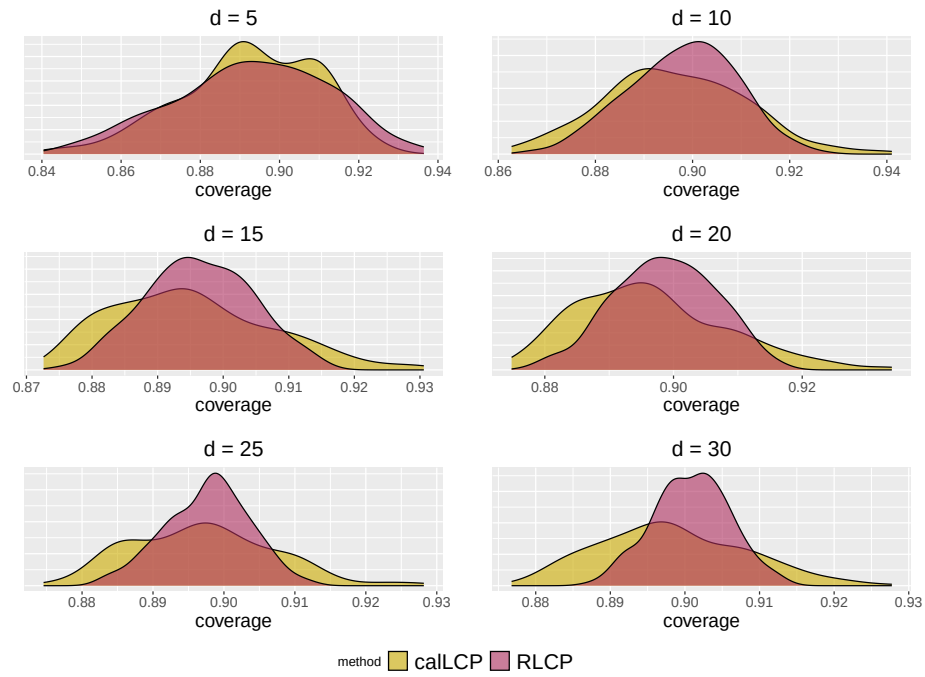


Figure B.6: Local coverage of calLCP and RLCP prediction intervals across different dimensions.

APPENDIX C

APPENDICES FOR CHAPTER 4

C.1 Proof of validity results of PairSwap-ICI test

C.1.1 Proof of Theorem 4.2

Our proof is split into three steps. First we derive some deterministic properties of the p -value. Next, we compare to the *sharp null*, where the pair X_{i_ℓ}, X_{j_ℓ} are identically distributed (rather than stochastically ordered). Finally, we examine the validity of the test under the sharp null.

Step 1: some deterministic properties of the p -value. First, fix $\alpha \in [0, 1]$ and define a function $p : \mathbb{R}^n \rightarrow [0, 1]$ as

$$p(\mathbf{x}) = \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1} \{T(\mathbf{x}^{\mathbf{s}}) \geq T(\mathbf{x})\}, \quad (\text{C.1})$$

so that the p -value in (4.5) can be written as $p = p(\mathbf{X})$. For each $\mathbf{s} \in \{\pm 1\}^L$, we can observe that the value $p(\mathbf{x}^{\mathbf{s}})$ is simply computing the quantile of $T(\mathbf{x}^{\mathbf{s}})$ among all possible swapped statistics, $(T(\mathbf{x}^{\mathbf{s}'})) : \mathbf{s}' \in \{\pm 1\}^L$. Consequently, it holds deterministically that

$$\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1} \{p(\mathbf{x}^{\mathbf{s}}) \leq \alpha\} \leq \alpha. \quad (\text{C.2})$$

(Lemma C.11 in the Appendix verifies this bound, for completeness.)

In addition, we claim that $p(\cdot)$ is monotone in its coordinates, namely, $p(\mathbf{x})$ is nonincreasing

in each x_{i_ℓ} , and nondecreasing in each x_{j_ℓ} . To see why, for each $\mathbf{s} \in \{\pm 1\}^L$, we can calculate

$$\begin{aligned} \mathbb{1}\{T(\mathbf{x}^{\mathbf{s}}) \geq T(\mathbf{x})\} &= \mathbb{1}\left\{ \sum_{\ell: s_\ell=+1} \psi_\ell(x_{i_\ell}, x_{j_\ell}) + \sum_{\ell: s_\ell=-1} \psi_\ell(x_{j_\ell}, x_{i_\ell}) \geq \sum_{\ell=1}^L \psi_\ell(x_{i_\ell}, x_{j_\ell}) \right\} \\ &= \mathbb{1}\left\{ \sum_{\ell: s_\ell=-1} (\psi_\ell(x_{i_\ell}, x_{j_\ell}) - \psi_\ell(x_{j_\ell}, x_{i_\ell})) \leq 0 \right\}. \end{aligned}$$

By the anti-monotonicity condition (4.3) on ψ_ℓ , this function is nonincreasing in each x_{i_ℓ} , and nondecreasing in each x_{j_ℓ} , and therefore the same is true for $p(\mathbf{x})$ as well.

Step 2: compare to the sharp null. For each $i \in [n]$, let $P_i = P_{X|Z}(\cdot \mid Z_i)$ denote the null distribution of X_i (after conditioning on $\mathbf{Y}, \mathbf{Z}$). By Assumption 1, we know that $P_{i_\ell} \preceq_{\text{st}} P_{j_\ell}$ for each pair $\ell \in [L]$, where $\preceq_{\text{st}}$ denotes the stochastic ordering on distributions. Next, we also define distributions $\bar{P}_\ell$ for each $\ell \in [L]$, given by the mixture

$$\bar{P}_\ell = \frac{1}{2}P_{i_\ell} + \frac{1}{2}P_{j_\ell}.$$

In particular, then, conditional on $\mathbf{Y}, \mathbf{Z}$,

$$P_{i_\ell} \preceq_{\text{st}} \bar{P}_\ell \preceq_{\text{st}} P_{j_\ell}, \quad \ell \in [L]. \tag{C.3}$$

We will now compare the observed data values, whose distribution (conditional on $\mathbf{Y}, \mathbf{Z}$) is given by

$$\mathbf{X} = (X_1, \dots, X_n) \sim P_1 \times \dots \times P_n,$$

against a different distribution,

$$\mathbf{X}_\# = ((X_\#)_1, \dots, (X_\#)_n) \sim (P_\#)_1 \times \dots \times (P_\#)_n,$$

where the distributions $(P_{\sharp})_i$ are defined by setting

$$(P_{\sharp})_{i_{\ell}} = (P_{\sharp})_{j_{\ell}} = \bar{P}_{\ell}$$

for each $\ell \in [L]$ (and, for any index $i \in [n] \setminus \{i_1, j_1, \dots, i_L, j_L\}$ that does not belong to any of the L matched pairs, we simply take $(P_{\sharp})_i = P_i$). We can think of this alternative vector of observations as being drawn from a *sharp null*, because for each pair ℓ , the random variables $(X_{\sharp})_{i_{\ell}}, (X_{\sharp})_{j_{\ell}}$ are identically distributed (rather than stochastically ordered, as for $X_{i_{\ell}}, X_{j_{\ell}}$). In particular, this implies that for any $\mathbf{s} \in \{\pm 1\}^L$,

$$(\mathbf{X}_{\sharp})^{\mathbf{s}} \stackrel{\text{d}}{=} \mathbf{X}_{\sharp} \tag{C.4}$$

after conditioning on $\mathbf{Y}, \mathbf{Z}$.

In Step 1, we verified that the function $p(\mathbf{x})$ is nonincreasing in each $x_{i_{\ell}}$, and nondecreasing in each $x_{j_{\ell}}$. In particular, combined with the stochastic ordering (C.3), this means that $p(\mathbf{X}_{\sharp}) \preceq_{\text{st}} p(\mathbf{X})$ (conditional on $\mathbf{Y}, \mathbf{Z}$). We therefore have

$$\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} = \mathbb{P}\{p(\mathbf{X}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \mathbb{P}\{p(\mathbf{X}_{\sharp}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}.$$

From this point on, then, we only need to verify the validity of the p -value $p(\mathbf{X}_{\sharp})$ computed under the sharp null.

Step 3: validity under the sharp null. For data $\mathbf{X}_\sharp$ drawn under a sharp null, we have

$$\begin{aligned}\mathbb{P}\{p(\mathbf{X}_\sharp) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} &= \frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{P}\{p((\mathbf{X}_\sharp)^{\mathbf{s}}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \\ &= \mathbb{E} \left[\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1}_{\{p((\mathbf{X}_\sharp)^{\mathbf{s}}) \leq \alpha\}} \mid \mathbf{Y}, \mathbf{Z} \right] \leq \alpha,\end{aligned}$$

where the first step holds by (C.4), while the last step holds by the deterministic calculation (C.2) from Step 1. $\square$

C.1.2 Proof of Theorem 4.3

The proof of this result follows the same structure as the proof of Theorem 4.2.

Step 1: some deterministic properties of the p -value. Define a function $\hat{p}_M : \mathbb{R}^n \times (\{\pm 1\}^L)^M \rightarrow [0, 1]$ by

$$\hat{p}_M(\mathbf{x}; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) = \frac{1 + \sum_{m=1}^M \mathbb{1}_{\left\{T(\mathbf{x}^{\mathbf{s}^{(m)}}) \geq T(\mathbf{x})\right\}}}{1 + M}.$$

As in the proof of Theorem 4.2, this function is monotone nonincreasing in each x_{i_ℓ} , and monotone nondecreasing in each x_{j_ℓ} .

Step 2: compare to the sharp null. Define $\mathbf{X}_\sharp$ as in the proof of Theorem 4.2. Following identical arguments as in that proof, we can verify that, for any fixed $\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}$, it holds that

$$\hat{p}_M(\mathbf{X}_\sharp; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \preceq_{\text{st}} \hat{p}_M(\mathbf{X}; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)})$$

conditional on $\mathbf{Y}, \mathbf{Z}$. Since $\hat{p}_M = \hat{p}_M(\mathbf{X}; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)})$ by construction, we therefore have

$$\mathbb{P} \left\{ \hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)} \right\} \leq \mathbb{P} \left\{ \hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)} \right\}.$$

Marginalizing over the random draw of the swaps, $\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)} \stackrel{\text{iid}}{\sim} \text{Unif}(\{\pm 1\}^L)$, we therefore have

$$\mathbb{P} \{ \hat{p}_M \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \} \leq \mathbb{P} \left\{ \hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \right\}.$$

Step 3: validity under the sharp null. We now need to verify the validity of the Monte Carlo p -value, under the sharp null. Unlike the first two steps, for this step the arguments are somewhat different from the proof of Theorem 4.2.

First, let $\mathbf{s}^{(0)}$ be an additional draw from $\text{Unif}(\{\pm 1\}^L)$, sampled independently from all other random variables. Then

$$(\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \stackrel{\text{d}}{=} (\mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}),$$

where $\circ$ denotes the elementwise product, and so

$$\hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \stackrel{\text{d}}{=} \hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)})$$

conditional on $\mathbf{Y}, \mathbf{Z}$. Moreover, by construction of the sharp null data $\mathbf{X}_\#$,

$$\mathbf{X}_\# \stackrel{\text{d}}{=} (\mathbf{X}_\#)^{\mathbf{s}^{(0)}}$$

holds conditional on $\mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(0)}, \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}$, and therefore

$$\hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}) \stackrel{\text{d}}{=} \hat{p}_M((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)})$$

holds conditional on $\mathbf{Y}, \mathbf{Z}, \mathbf{s}^{(0)}, \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}$. Combining all these calculations so far, then,

$$\hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \stackrel{d}{=} \hat{p}_M\left((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\right), \quad (\text{C.5})$$

conditional on $\mathbf{Y}, \mathbf{Z}$.

Next we calculate this last p -value: by definition,

$$\begin{aligned} \hat{p}_M\left((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\right) &= \frac{1 + \sum_{m=1}^M \mathbb{1}\left\{T((\mathbf{X}_\#)^{\mathbf{s}^{(m)}}) \geq T((\mathbf{X}_\#)^{\mathbf{s}^{(0)}})\right\}}{1 + M} \\ &= \frac{\sum_{m=0}^M \mathbb{1}\left\{T((\mathbf{X}_\#)^{\mathbf{s}^{(m)}}) \geq T((\mathbf{X}_\#)^{\mathbf{s}^{(0)}})\right\}}{1 + M}, \end{aligned}$$

where the first step holds since, for each $m \in [M]$,

$$((\mathbf{X}_\#)^{\mathbf{s}^{(0)}})^{\mathbf{s}^{(0)} \circ \mathbf{s}^{(m)}} = (\mathbf{X}_\#)^{\mathbf{s}^{(0)} \circ \mathbf{s}^{(0)} \circ \mathbf{s}^{(m)}} = (\mathbf{X}_\#)^{\mathbf{s}^{(m)}}$$

by definition of the swap operation. In other words, $\hat{p}_M\left((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\right)$ simply compares the statistic $T((\mathbf{X}_\#)^{\mathbf{s}^{(0)}})$ against the $M+1$ values $T((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}), \dots, T((\mathbf{X}_\#)^{\mathbf{s}^{(M)}})$.

We therefore have

$$\mathbb{P}\left\{\hat{p}_M\left((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\right) \leq \alpha \mid \mathbf{X}_\#, \mathbf{Y}, \mathbf{Z}\right\} \leq \alpha,$$

since, conditional on $\mathbf{X}_\#, \mathbf{Y}, \mathbf{Z}$, the sign vectors $\mathbf{s}^{(0)}, \dots, \mathbf{s}^{(M)}$ are independent and identically distributed, and therefore the rank of $T((\mathbf{X}_\#)^{\mathbf{s}^{(0)}})$ among the list $T((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}), \dots, T((\mathbf{X}_\#)^{\mathbf{s}^{(M)}})$ is uniformly distributed. Marginalizing over $\mathbf{X}_\#$, therefore,

$$\mathbb{P}\left\{\hat{p}_M\left((\mathbf{X}_\#)^{\mathbf{s}^{(0)}}; \mathbf{s}^{(0)} \circ \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(0)} \circ \mathbf{s}^{(M)}\right) \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\right\} \leq \alpha.$$

Finally, combining this with our earlier calculation (C.5), we have

$$\mathbb{P} \left\{ \hat{p}_M(\mathbf{X}_\#; \mathbf{s}^{(1)}, \dots, \mathbf{s}^{(M)}) \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \right\} \leq \alpha,$$

which completes the proof. $\square$

C.2 Proof of the results from Section 4.5

In this section, we prove the results presented in Section 4.5. Throughout this appendix, we assume that the statistic T admits the form in (4.6) with $\psi(x, x') = x - x'$, and that $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$; the partial ordering $\preceq$ for Z will simply be the usual ordering $\leq$ on $\mathbb{R}$. The organization of this appendix is as follows.

- We begin in Appendix C.2.1 by proving finite sample and asymptotic upper and lower bounds on the conditional power of **PairSwap**-ICI test for any valid matching and weighting scheme, and any statistic T of the form (4.6) with a shared linear kernel $\psi(x, x') = x - x'$.
- Next, in Appendix C.2.2, we specialise these results to the oracle matching under two cases: one assuming access to oracle knowledge of μ (i.e., Theorem 4.5) and another with μ estimated from data (i.e., Theorem 4.6).
- Then, we shift our attention to the partially linear Gaussian models in (4.16). In Appendices C.2.3 and C.2.4 we prove the asymptotic behavior of conditional power for neighbor matching (Theorem 4.10) and for cross-bin matching (Theorem 4.11), respectively.
- In Appendix C.2.5 we prove the corollaries and lemmas from Section 4.5.
- In Appendix C.2.6 we introduce an oracle matching, namely isotonic median match-

ing, and discuss a key property of the same, which allows us to prove the results in Appendix C.2.2.

- Finally, in Appendix C.2.7 we prove the lemmas used in Appendices C.2.1–C.2.4.

Notation. For a matching $\{(i_1, j_1), (i_2, j_2), \dots, (i_L, j_L)\} \in \mathcal{M}_n(\mathbf{Z})$ and a vector $\mathbf{V} = (V_1, \dots, V_n) \in \mathbb{R}^n$, we define $\Delta \mathbf{V} = (\Delta_1 \mathbf{V}, \dots, \Delta_L \mathbf{V}) \in \mathbb{R}^L$ with entries $\Delta_\ell \mathbf{V} := V_{i_\ell} - V_{j_\ell}$. Given any vector $\mathbf{v}$, we write $\mathbf{v}^+$ to denote the vector with i th component $v_{i+} = \max\{v_i, 0\}$. We write $\mathbf{a} \circ \mathbf{b}$ for the Hadamard product of vectors $\mathbf{a}, \mathbf{b}$ of the same dimension, with i th component $a_i \cdot b_i$.

C.2.1 A general result on power of PairSwap-ICI test

Here, we consider any valid matching and weighting scheme, and state finite-sample upper and lower bounds on the conditional power $\mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\}$ of the PairSwap-ICI test in Theorem C.1. Further, under the asymptotic regime of Section 4.5.1, we derive asymptotic high-probability upper and lower bounds for the same quantity in Corollary C.2.

Theorem C.1. *Suppose that the triple $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ satisfies (4.12). Fix any matching $M = \{(i_1, j_1), \dots, (i_L, j_L)\} \in \mathcal{M}_n(\mathbf{Z})$, and any weight vector $\mathbf{w} \in [0, \infty)^L$, which may both depend arbitrarily on $\mathbf{Y}, \mathbf{Z}$ (but not on $\mathbf{X}$). Assume that $\|\mathbf{w}\|_2 > 0$ and fix any $\alpha \in (0, 1)$. Let*

$$\epsilon_1 := \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2}, \quad \epsilon_2 := \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_3^3}{\sigma_n^3 \|\mathbf{w}\|_2^3 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^3}.$$

Then, for any $\delta > 0$,

$$\begin{aligned}
& \Phi\left(\frac{\mathbf{w}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha_{\text{lo},\delta}) \cdot \sqrt{1 + \frac{\|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2}{2\sigma_n^2\|\mathbf{w}\|_2^2}}\right) - (C'\epsilon_1 + \delta) \\
& \leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \\
& \leq \Phi\left(\frac{\mathbf{w}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha_{\text{hi},\delta}) \cdot \sqrt{1 + \frac{\|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2}{2\sigma_n^2\|\mathbf{w}\|_2^2}}\right) + (C'\epsilon_1 + \delta),
\end{aligned} \tag{C.6}$$

where we write $\alpha_{\text{lo},\delta} = \alpha - C\delta^{-1} \cdot (\epsilon_1 + \epsilon_2)$, $\alpha_{\text{hi},\delta} = \alpha + C\delta^{-1} \cdot (\epsilon_1 + \epsilon_2)$, with $C > 0$ depending only on α and on $\mathbb{E}_{\zeta \sim P_\zeta}[\zeta^4]$, and $C' > 0$ depending only on $\mathbb{E}_{\zeta \sim P_\zeta}[\zeta^4]$.

Proof. Recall from (4.5) that the **PairSwap-ICI** test is given by

$$p = \mathbb{P}_{\mathbf{s} \sim \text{Unif}(\{\pm 1\}^L)} \{T_{\mathbf{s}} \geq T\} = \mathbb{P}_{\mathbf{s} \sim \text{Unif}(\{\pm 1\}^L)} \left\{ \mathbf{s}^\top (\mathbf{w} \circ \Delta \mathbf{X}) \geq \mathbf{1}^\top (\mathbf{w} \circ \Delta \mathbf{X}) \right\}.$$

From here, the proof is divided into four key steps.

Step 1: a CLT approximation to the p -value. Conditional on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ and with $\mathbf{s} \sim \text{Unif}(\{\pm 1\}^L)$, we see that $\mathbf{s}^\top (\mathbf{w} \circ \Delta \mathbf{X})$ is a sum of L independent random variables, where the ℓ th term $s_\ell \cdot w_\ell \cdot \Delta_\ell \mathbf{X}$ has mean zero, variance $w_\ell^2 (\Delta_\ell \mathbf{X})^2$, and third absolute moment $w_\ell^3 |\Delta_\ell \mathbf{X}|^3$ for $\ell \in [L]$. Therefore, by the Berry–Esseen theorem [Shevtsova, 2010, Theorem 1],

$$\begin{aligned}
\left| p - \bar{\Phi}\left(\frac{\mathbf{w}^\top \Delta \mathbf{X}}{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2}\right) \right| & \leq \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{\mathbf{s}^\top (\mathbf{w} \circ \Delta \mathbf{X})}{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2} \geq x \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z} \right\} - \bar{\Phi}(x) \right| \\
& \leq 0.56 \cdot \frac{\|\mathbf{w} \circ \Delta \mathbf{X}\|_3^3}{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2^3} =: \Delta_1,
\end{aligned}$$

as long as $\|\mathbf{w} \circ \Delta \mathbf{X}\|_2 > 0$. Thus

$$\mathbb{1} \left\{ \bar{\Phi} \left(\frac{\mathbf{w}^\top \Delta \mathbf{X}}{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2} \right) < \alpha - \Delta_1 \right\} \leq \mathbb{1} \{p \leq \alpha\} \leq \mathbb{1} \left\{ \bar{\Phi} \left(\frac{\mathbf{w}^\top \Delta \mathbf{X}}{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2} \right) \leq \alpha + \Delta_1 \right\},$$

or equivalently (since $\bar{\Phi}$ is strictly decreasing),

$$\begin{aligned} \mathbb{1} \left\{ \mathbf{w}^\top \Delta \mathbf{X} > \|\mathbf{w} \circ \Delta \mathbf{X}\|_2 \cdot \bar{\Phi}^{-1}(\alpha - \Delta_1) \right\} &\leq \mathbb{1} \{p \leq \alpha\} \\ &\leq \mathbb{1} \left\{ \mathbf{w}^\top \Delta \mathbf{X} \geq \|\mathbf{w} \circ \Delta \mathbf{X}\|_2 \cdot \bar{\Phi}^{-1}(\alpha + \Delta_1) \right\}. \end{aligned}$$

Moreover, when $\|\mathbf{w} \circ \Delta \mathbf{X}\|_2 = 0$, we have $p = 1$, so the above inequality holds for that case as well.

After marginalizing over the distribution of $\mathbf{X}$, then,

$$\begin{aligned} \mathbb{P} \left\{ \mathbf{w}^\top \Delta \mathbf{X} > \|\mathbf{w} \circ \Delta \mathbf{X}\|_2 \cdot \bar{\Phi}^{-1}(\alpha - \Delta_1) \mid \mathbf{Y}, \mathbf{Z} \right\} \\ \leq \mathbb{P} \{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \leq \mathbb{P} \left\{ \mathbf{w}^\top \Delta \mathbf{X} \geq \|\mathbf{w} \circ \Delta \mathbf{X}\|_2 \cdot \bar{\Phi}^{-1}(\alpha + \Delta_1) \mid \mathbf{Y}, \mathbf{Z} \right\}. \end{aligned} \tag{C.7}$$

Step 2: approximating $\|\mathbf{w} \circ \Delta \mathbf{X}\|_2$ in (C.7). Write $V_n := 2\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2$ and define

$$\Delta_2 := \left| \frac{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2}{V_n^{1/2}} - 1 \right|.$$

Then by (C.7), we have that

$$\begin{aligned} \mathbb{P} \{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} &\geq \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} > \frac{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2}{V_n^{1/2}} \cdot \bar{\Phi}^{-1}(\alpha - \Delta_1) \mid \mathbf{Y}, \mathbf{Z} \right\} \\ &\geq \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} > \bar{\Phi}^{-1}(\alpha - \Delta_1) + |\bar{\Phi}^{-1}(\alpha - \Delta_1)| \Delta_2 \mid \mathbf{Y}, \mathbf{Z} \right\}. \end{aligned}$$

We can do a similar calculation for the upper bound on the conditional power from (C.7).

Together, we have that

$$\begin{aligned} & \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} > \bar{\Phi}^{-1}(\alpha - \Delta_1) + \left| \bar{\Phi}^{-1}(\alpha - \Delta_1) \right| \Delta_2 \mid \mathbf{Y}, \mathbf{Z} \right\} \\ & \leq \mathbb{P} \{ p \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \} \leq \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} \geq \bar{\Phi}^{-1}(\alpha + \Delta_1) - \left| \bar{\Phi}^{-1}(\alpha + \Delta_1) \right| \Delta_2 \mid \mathbf{Y}, \mathbf{Z} \right\}. \end{aligned}$$

Now fix $\varepsilon_\alpha > 0$. Define the event

$$\begin{aligned} \Omega_\alpha := & \{ \bar{\Phi}^{-1}(\alpha - \Delta_1) + \left| \bar{\Phi}^{-1}(\alpha - \Delta_1) \right| \Delta_2 \leq \bar{\Phi}^{-1}(\alpha - \varepsilon_\alpha) \} \\ & \bigcap \{ \bar{\Phi}^{-1}(\alpha + \Delta_1) - \left| \bar{\Phi}^{-1}(\alpha + \Delta_1) \right| \Delta_2 \geq \bar{\Phi}^{-1}(\alpha + \varepsilon_\alpha) \}. \end{aligned} \quad (\text{C.8})$$

Then

$$\begin{aligned} & \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} > \bar{\Phi}^{-1}(\alpha - \varepsilon_\alpha) \mid \mathbf{Y}, \mathbf{Z} \right\} - \mathbb{P} \{ \Omega_\alpha^c \mid \mathbf{Y}, \mathbf{Z} \} \\ & \leq \mathbb{P} \{ p \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \} \leq \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} \geq \bar{\Phi}^{-1}(\alpha + \varepsilon_\alpha) \mid \mathbf{Y}, \mathbf{Z} \right\} + \mathbb{P} \{ \Omega_\alpha^c \mid \mathbf{Y}, \mathbf{Z} \}. \end{aligned} \quad (\text{C.9})$$

Step 3: a CLT approximation to $\mathbf{w}^\top \Delta \mathbf{X}$ in (C.7). Conditional on $\mathbf{Y}, \mathbf{Z}$, we have that $\mathbf{w}^\top \Delta \mathbf{X}$ is a sum of L independent terms, where for the ℓ th term we have

$$\mathbb{E} [w_\ell \cdot \Delta_\ell \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}] = w_\ell \cdot \Delta_\ell \mu_n(\mathbf{Y}, \mathbf{Z}),$$

and

$$\text{Var} (w_\ell \cdot \Delta_\ell \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = w_\ell^2 \cdot \text{Var} (\Delta_\ell \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = w_\ell^2 \cdot 2\sigma_n^2.$$

Moreover, we have

$$\begin{aligned}\mathbb{E} \left[|w_\ell \cdot \Delta_\ell \mathbf{X} - \mathbb{E}[w_\ell \cdot \Delta_\ell \mathbf{X} \mid \mathbf{Y}, \mathbf{Z}]|^3 \mid \mathbf{Y}, \mathbf{Z} \right] &= \mathbb{E} \left[|w_\ell \sigma_n (\zeta_{i_\ell} - \zeta_{j_\ell})|^3 \mid \mathbf{Y}, \mathbf{Z} \right] \\ &\leq 8\sigma_n^3 w_\ell^3 \mathbb{E}_{P_\zeta} \left[|\zeta|^3 \right] \leq 8\rho_4^3 \sigma_n^3 w_\ell^3,\end{aligned}$$

where we write $\rho_4 := \mathbb{E}_{P_\zeta} [\zeta^4]^{1/4}$. Therefore, another application of the Berry–Esseen theorem gives

$$\begin{aligned}\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \geq x \mid \mathbf{Y}, \mathbf{Z} \right\} - \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - x \right) \right| \\ \leq 0.56 \cdot \frac{8\sigma_n^3 \rho_4^3 \|\mathbf{w}\|_3^3}{(\sqrt{2}\sigma_n \|\mathbf{w}\|_2)^3} \leq 1.59\rho_4^3 \cdot \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2}.\end{aligned}$$

By rescaling to match the denominator in (C.9), we then have

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{\mathbf{w}^\top \Delta \mathbf{X}}{V_n^{1/2}} \geq x \mid \mathbf{Y}, \mathbf{Z} \right\} - \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \frac{x V_n^{1/2}}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \right) \right| \leq 1.59\rho_4^3 \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2}.$$

Moreover, by (C.9) we now have

$$\begin{aligned}\Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha - \varepsilon_\alpha) \cdot \frac{V_n^{1/2}}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \right) - 1.59\rho_4^3 \cdot \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} - \mathbb{P} \{ \Omega_\alpha^c \mid \mathbf{Y}, \mathbf{Z} \} \\ \leq \mathbb{P} \{ p \leq \alpha \mid \mathbf{Y}, \mathbf{Z} \} \\ \leq \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha + \varepsilon_\alpha) \cdot \frac{V_n^{1/2}}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \right) + 1.59\rho_4^3 \cdot \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + \mathbb{P} \{ \Omega_\alpha^c \mid \mathbf{Y}, \mathbf{Z} \}.\end{aligned}$$

Step 4: a concentration result, and completion of the proof. Finally, we show how to choose ε_α to control $\mathbb{P} \{ \Omega_\alpha^c \mid \mathbf{Y}, \mathbf{Z} \}$ appropriately. We will prove in Lemma C.7 that the terms $\|\mathbf{w} \circ \Delta \mathbf{X}\|_2$ and $\|\mathbf{w} \circ \Delta \mathbf{X}\|_3$ are likely to be well-behaved—in particular, $\|\mathbf{w} \circ \Delta \mathbf{X}\|_2$ concentrates near $V_n^{1/2}$, while $\|\mathbf{w} \circ \Delta \mathbf{X}\|_3$ is vanishing. Concretely, Lemma C.7 will show

that for any $\delta > 0$,

$$\mathbb{P}\left[\left\{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2^2 \in [V_n(1 - \epsilon_*) \vee 0, V_n(1 + \epsilon_*)]\right\} \cap \left\{\|\mathbf{w} \circ \Delta \mathbf{X}\|_3^2 \leq V_n \epsilon'_*\right\} \mid \mathbf{Y}, \mathbf{Z}\right] \geq 1 - \delta,$$

where

$$\epsilon_* = \sqrt{\frac{3}{\delta}} \cdot \left(2\rho_4^2 \cdot \epsilon_1 + 2(\epsilon_1^2 \epsilon_2)^{1/3}\right), \quad \epsilon'_* = \frac{(24)^{2/3} \rho_4^2}{\delta^{2/3}} \cdot \epsilon_1^{2/3} + 4\epsilon_2^{2/3}.$$

Hence, since $(1 + \epsilon_*)^{1/2} - 1 \leq \epsilon_*$ and $1 - \{(1 - \epsilon_*) \vee 0\}^{1/2} \leq \epsilon_* \wedge 1 \leq \epsilon_*$, we have

$$\mathbb{P}\left\{\Delta_1 \leq \Delta_{1,*}, \Delta_2 \leq \epsilon_* \mid \mathbf{Y}, \mathbf{Z}\right\} \geq 1 - \delta,$$

where $\Delta_{1,*} := 0.56(\epsilon'_*/\{(1 - \epsilon_*) \vee 0\})^{3/2}$ and with the convention that $a/0 = \infty$ for any $a > 0$. Now, we take

$$\begin{aligned} \varepsilon_\alpha := \max \Big\{ & \alpha - \bar{\Phi}(\bar{\Phi}^{-1}(\alpha - \Delta_{1,*}) + |\bar{\Phi}^{-1}(\alpha - \Delta_{1,*})| \epsilon_*), \\ & \bar{\Phi}(\bar{\Phi}^{-1}(\alpha + \Delta_{1,*}) - |\bar{\Phi}^{-1}(\alpha + \Delta_{1,*})| \epsilon_*) - \alpha \Big\}, \end{aligned}$$

if $\epsilon_* < 1$, and if $\epsilon_* \geq 1$, we take $\varepsilon_\alpha = \max(\alpha, 1 - \alpha)$. Also, here we take the convention that $\bar{\Phi}^{-1}(t) = \infty$ if $t \leq 0$ and $\bar{\Phi}^{-1}(t) = -\infty$ if $t \geq 1$. Then, by construction, with this choice of ε_α , the event Ω_α defined in (C.8) satisfies $\mathbb{P}\{\Omega_\alpha^c \mid \mathbf{Y}, \mathbf{Z}\} \leq \delta$. Combining both cases, we finally have that

$$\begin{aligned} & \Phi\left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha - \varepsilon_\alpha) \cdot \sqrt{1 + \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2}{2\sigma_n^2 \|\mathbf{w}\|_2^2}}\right) - 1.59\rho_4^3 \cdot \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} - \delta \\ & \leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \\ & \leq \Phi\left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha + \varepsilon_\alpha) \cdot \sqrt{1 + \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2}{2\sigma_n^2 \|\mathbf{w}\|_2^2}}\right) + 1.59\rho_4^3 \cdot \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} + \delta. \end{aligned}$$

Since by construction, $\varepsilon_\alpha \in [0, \max(\alpha, 1 - \alpha)]$, and by properties of $\bar{\Phi}$ and $\bar{\Phi}^{-1}$, there exists $C_\alpha > 0$, depending only on α , such that

$$\varepsilon_\alpha \leq C_\alpha(\epsilon_* + \epsilon'_*{}^{3/2}).$$

Therefore, choosing $C > 0$, depending only on α and ρ_4 , appropriately,

$$\varepsilon_\alpha \leq C\delta^{-1} \cdot (\epsilon_1 + \epsilon_2),$$

as required. □

Asymptotic version of the general power guarantee

Next we present an asymptotic power guarantee that follows from Theorem C.1, to help interpret the results of that theorem.

Corollary C.2. *Under the setting of Theorem C.1, suppose that the matching $M \in \mathcal{M}_n(\mathbf{Z})$ and weight vector $\mathbf{w}$ satisfy the following assumptions:*

Assumption A2. *The weight vector $\mathbf{w}$ satisfies*

$$\|\mathbf{w}\|_\infty = o_P(\|\mathbf{w}\|_2).$$

Assumption A3. *The signal satisfies*

$$\begin{aligned} \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 &= o_P\left(\left|\mathbf{w}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\right| \vee \sigma_n \|\mathbf{w}\|_2\right) \quad \text{and} \\ \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 &= o_P\left(\|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 \vee \sigma_n \|\mathbf{w}\|_2\right). \end{aligned}$$

Then for $\alpha \in (0, 1)$, we have that

$$\begin{aligned} \Phi\left(\frac{\mathbf{w}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1) &\leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \\ &\leq \Phi\left(\frac{\mathbf{w}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n\|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha)\right) + o_P(1). \end{aligned}$$

Remark 3. Before turning to the proof, we remark that the assumptions stated in the corollary statement are very natural. Since, we are under the model class (4.12), and we are using the shared linear kernel $\psi(x, x') = x - x'$, the PairSwap-ICI test revolves around the statistic

$$\mathbf{w} \circ \Delta\mathbf{X} = \mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z}) + \sigma_n(\mathbf{w} \circ \Delta\boldsymbol{\zeta}),$$

where we can interpret the first term as the signal of our test, and the second term as added noise. With this interpretation, Assumption **A1** ensures that the noise term is well-behaved by requiring that the weight vector $\mathbf{w}$ is not too spiky, while Assumption **A2** instead essentially requires that the signal term itself is not too spiky.

Proof of Corollary C.2. We start by noting that by Assumption **A1**, $\epsilon_1 = \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2} = o_P(1)$, and by Assumption **A2**,

$$\epsilon_2 = \frac{\|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3^3}{\sigma_n^3\|\mathbf{w}\|_2^3 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^3} \leq \frac{\|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3^3}{\sigma_n^3\|\mathbf{w}\|_2^3 \vee \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^3} = o_P(1).$$

Moreover, we have that

$$\begin{aligned}
& \frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha_{10,\delta}) \cdot \sqrt{1 + \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2}{2\sigma_n^2 \|\mathbf{w}\|_2^2}} \\
& \geq \frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \left\{ \bar{\Phi}^{-1}(\alpha_{10,\delta}) \left(1 + \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \right) \vee \bar{\Phi}^{-1}(\alpha_{10,\delta}) \right\} \\
& \geq \frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha_{10,\delta}) - |\bar{\Phi}^{-1}(\alpha_{10,\delta})| \cdot \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \\
& = \frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \left(1 - |\bar{\Phi}^{-1}(\alpha_{10,\delta})| \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})} \right) - \bar{\Phi}^{-1}(\alpha_{10,\delta}).
\end{aligned}$$

Now, recalling the definition of $\alpha_{10,\delta}$ from the statement of Theorem C.1, we observe that $\alpha_{10,\delta} = \alpha - o_P(1)$ for any fixed $\delta > 0$. By Assumption **A2**, if $\sqrt{2}\sigma_n \|\mathbf{w}\|_2 > \mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})$, then

$$|\bar{\Phi}^{-1}(\alpha_{10,\delta})| \cdot \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} = o_P(1),$$

and otherwise, if $2\sigma_n^2 \|\mathbf{w}\|_2 \leq \mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})$, then

$$|\bar{\Phi}^{-1}(\alpha_{10,\delta})| \cdot \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})} = o_P(1).$$

Combining both cases, since $\bar{\Phi}^{-1}$ is continuous, we see that

$$\begin{aligned}
& \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha_{10,\delta}) \cdot \sqrt{1 + \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2}{2\sigma_n^2 \|\mathbf{w}\|_2^2}} \right) - (C' \epsilon_1 + \delta) \\
& \geq \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \{1 - o_P(1)\} - \bar{\Phi}^{-1}(\alpha - o_P(1)) - o_P(1) \right) - o_P(1) - \delta \\
& = \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} \{1 - o_P(1)\} - \bar{\Phi}^{-1}(\alpha) \right) - o_P(1) - \delta \\
& = \Phi \left(\frac{\mathbf{w}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\mathbf{w}\|_2} - \bar{\Phi}^{-1}(\alpha) \right) - o_P(1) - \delta,
\end{aligned}$$

where the last equality follows by Lemma C.8. Since $\delta > 0$ was arbitrary, the desired

asymptotic lower bound follows. The asymptotic upper bound follows by a very similar argument. $\square$

C.2.2 Power of oracle and plug-in matching strategies (proof of Theorems 4.5 and 4.6)

In this section we prove Theorems 4.5 and 4.6 using the general upper and lower bounds on the conditional power of the PairSwap-ICI test from Theorem C.1.

Before turning to the proofs, we observe from Lemma 4.4 that under the model class (4.12) and with the linear kernel $\psi(x, x') = x - x'$, the oracle weights in (4.9) satisfy $w_\ell^* \propto (\mu_n(Y_{i_\ell}, Z_{i_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell}))_+$ for $\ell \in [L]$. Since the p -value of our test is independent of the scales of weights, without loss of generality we may take $\mathbf{w}^* = \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})$. When we do not have access to oracle knowledge of μ_n , we assume that we are able to learn estimates $\hat{\mu}_n$ of μ_n from an independent dataset, independent of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. With these estimates in place, the plug-in weights satisfy $\hat{w}_\ell \propto (\hat{\mu}_n(Y_{i_\ell}, Z_{i_\ell}) - \hat{\mu}_n(Y_{j_\ell}, Z_{j_\ell}))_+$, i.e., without loss of generality we may take $\hat{\mathbf{w}} = \Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})$.

Proof of Theorem 4.5

Given access to oracle knowledge of μ_n , we observe that $\hat{\mu}_n = \mu_n$ satisfies the assumptions of Theorem 4.6. Hence, Theorem 4.5 follows as an immediate corollary of Theorem 4.6. $\square$

Proof of Theorem 4.6

Our proof is split into three steps. First we state a key property of the oracle matching M^* and oracle weight $\mathbf{w}^*$. Then, we use this to deduce similar properties for plug-in matching $\hat{M}$ and the plug-in weights $\hat{\mathbf{w}}$. These in turn enable us to show that Assumptions **A1**, **A2** and **A3** of Corollary C.2 are satisfied by plug-in matching, and hence the result follows.

Our notation $\Delta\mu_n(\mathbf{Y}, \mathbf{Z})$ and $\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z})$ suppresses the dependence of the matching M for which it is computed. Since we will need both matchings $M^\star$ and $\hat{M}$ in this step, we make the dependence on the matching explicit by writing $\Delta\mu_n(\mathbf{Y}, \mathbf{Z}; M)$ and $\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}; M)$, for any matching $M \in \mathcal{M}_n(\mathbf{Z})$. Throughout this proof, we write

$$\begin{aligned}\text{Err}_2(\hat{\mu}_n, \mu_n) &= \frac{\|\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}; \hat{M}) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2}{\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2}, \\ \text{Err}_\infty(\hat{\mu}_n, \mu_n) &= \frac{\|\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}; \hat{M}) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_\infty^4}{\sigma_n^3 \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2}.\end{aligned}$$

Step 1: a key property of oracle matching and its implications. We start with a key property of oracle matching, which relates $\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M^\star)\|_2$ to the isotonic SNR: given $(\mathbf{Y}, \mathbf{Z})$, we have that

$$\widehat{\text{ISNR}}_n \leq \frac{\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M^\star)\|_2}{\sigma_n} = \frac{\sup_{M \in \mathcal{M}_n(\mathbf{Z})} \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2}{\sigma_n} \leq \sqrt{2} \widehat{\text{ISNR}}_n. \quad (\text{C.10})$$

The equality is simply due to the definition of $M^\star$ as an optimal matching (note that $\mathcal{M}_n(\mathbf{Z})$ is a finite set and so the supremum must be attained at some $M^\star \in \mathcal{M}_n(\mathbf{Z})$), while the two inequalities follow from Theorem C.4, proved later in Appendix C.2.6.

Further, we require a very similar result for the plug-in matching. To this end, we note that for any $M \in \mathcal{M}_n(\mathbf{Z})$,

$$\begin{aligned}|\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2 - \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2| &\leq \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}; M) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2 \\ &\leq \|\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}; M) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z}; M)\|_2 \\ &\leq \sqrt{2} \|\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \mu_n(\mathbf{Y}, \mathbf{Z})\|_2 = o_P(\sigma_n \widehat{\text{ISNR}}_n),\end{aligned} \quad (\text{C.11})$$

where the final step is from the assumptions of the theorem. Further, by definition of the

plug-in matching in (4.11) and by (C.11), we have that

$$\begin{aligned}\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2 &\geq \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}; M^*)\|_2 \geq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M^*)\|_2 - o_P(\widehat{\sigma_n \text{ISNR}_n}) \\ &\geq \sigma_n \widehat{\text{ISNR}_n} (1 - o_P(1)).\end{aligned}$$

By a similar argument, then, we can establish

$$\begin{aligned}\sigma_n \widehat{\text{ISNR}_n} (1 - o_P(1)) &\leq \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2 \leq \sqrt{2} \sigma_n \widehat{\text{ISNR}_n} (1 + o_P(1)), \\ \sigma_n \widehat{\text{ISNR}_n} (1 - o_P(1)) &\leq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2 \leq \sqrt{2} \sigma_n \widehat{\text{ISNR}_n},\end{aligned}\tag{C.12}$$

i.e., a property similar to (C.10) for plug-in matching with both μ_n , and $\hat{\mu}_n$. As a result, we deduce that for plug-in matching,

$$\max \left\{ \frac{\sigma_n^3 \|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \vee \|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty^4}{\sigma_n^3 \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; \hat{M})\|_2}, \text{Err}_2(\hat{\mu}_n, \mu_n), \text{Err}_\infty(\hat{\mu}_n, \mu_n) \right\} = o_P(1). \tag{C.13}$$

Step 2: establishing Assumptions A1 and A2. Henceforth we work with the matching $\hat{M}$. We have by (C.13) that

$$\begin{aligned}\frac{\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_\infty}{\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_2} &\leq \frac{\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_\infty + \|\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 - \|\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2} \\ &\leq \frac{2(\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty / \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2) + \text{Err}_2(\hat{\mu}_n, \mu_n)}{1 - \text{Err}_2(\hat{\mu}_n, \mu_n)} = o_P(1),\end{aligned}$$

i.e., Assumption **A1** holds for plug-in matching with weights $\hat{\mathbf{w}} = \Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})$. Moreover, (C.13) also yields that

$$\begin{aligned}
& \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 \\
& \leq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 + \|(\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 \\
& \leq \|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \cdot \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 + \|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \cdot \|\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 \\
& \leq 2\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \cdot (1 + \text{Err}_2(\hat{\mu}_n, \mu_n)) = o_P(\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2).
\end{aligned}$$

Next, by the Cauchy–Schwarz inequality we have that

$$\begin{aligned}
& |\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z}) - \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2| \\
& = |(\Delta\hat{\mu}_n(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n(\mathbf{Y}, \mathbf{Z}))^\top \Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})| \\
& \leq \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \cdot \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \cdot \text{Err}_2(\hat{\mu}_n, \mu_n) \\
& \leq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2 \cdot (1 + \text{Err}_2(\hat{\mu}_n, \mu_n)) \cdot \text{Err}_2(\hat{\mu}_n, \mu_n).
\end{aligned}$$

Hence, by (C.13),

$$|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z}) - \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2| = o_P(\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2), \quad (\text{C.14})$$

meaning that

$$\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2 = O_P(|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})|).$$

Combining these calculations verifies that

$$\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 = o_P(|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})|),$$

or in other words,

$$\|\hat{\mathbf{w}} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2 = o_P(|\hat{\mathbf{w}}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})|),$$

and thus the first condition of Assumption **A2** is satisfied. Next, we focus on the second condition of the same assumption. We start with writing

$$\begin{aligned}
& \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 \\
& \leq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 + \|(\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 \\
& \leq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 + \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_3 \|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty.
\end{aligned}$$

For the first term in the last expression, by (C.13) we have that

$$\begin{aligned}
\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 &= \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_6^2 \leq \left(\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2 \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_\infty^4 \right)^{1/3} \\
&\leq \left(\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^2 (2\|\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty)^4 \right)^{1/3} = o_P(\sigma_n \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2).
\end{aligned}$$

For the second term, again by (C.13),

$$\begin{aligned}
& \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_3 \|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \\
& \leq \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^{2/3} \cdot \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \\
& = \sigma_n \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \left(\frac{\|(\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}))\|_2^{2/3}}{\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^{2/3}} \times \frac{\|(\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) - \Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}))\|_\infty^{1/3}}{\sigma_n^{1/4} \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^{1/12}} \right. \\
& \quad \left. \times \frac{\|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty}{\sigma_n^{3/4} \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^{1/4}} \right) \\
& = \sigma_n \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \left(\text{Err}_2(\hat{\mu}_n, \mu_n)^{2/3} \cdot \text{Err}_\infty(\hat{\mu}_n, \mu_n)^{1/12} \cdot \frac{\|\Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty}{\sigma_n^{3/4} \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2^{1/4}} \right) \\
& = o_P(\sigma_n \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2).
\end{aligned}$$

Finally, we also note that by (C.11) and (C.12),

$$\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \geq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 - o_P(\sigma_n \widehat{\text{ISNR}}_n) \geq \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 (1 - o_P(1)).$$

Hence, combining all these calculations, $\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z}) \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 = o_P(\sigma_n \|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_2)$, or in other words, $\|\hat{\mathbf{w}} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 = o_P(\sigma_n \|\hat{\mathbf{w}}\|_2)$ as required. Therefore, the second condition of Assumption **A2** is also satisfied by plug-in matching with weights $\hat{\mathbf{w}} = \Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})$.

Step 3: applying Corollary C.2. Now that we have verified Assumptions **A1** and **A2**, we are ready to apply the corollary. By (C.12) and (C.14),

$$\left| \frac{\hat{\mathbf{w}}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\|\hat{\mathbf{w}}\|_2} - \|\hat{\mathbf{w}}\|_2 \right| = \left| \frac{\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\|\Delta\hat{\mu}_n^+(\mathbf{Y}, \mathbf{Z})\|_2} - \|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z})\|_2 \right| = o_P(\sigma_n \widehat{\text{ISNR}}_n).$$

Then, by (C.10) and (C.12),

$$\widehat{\text{ISNR}}_n(1 - o_P(1)) \leq \frac{\hat{\mathbf{w}}^\top \Delta\mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\hat{\mathbf{w}}\|_2} \leq \sqrt{2} \cdot \widehat{\text{ISNR}}_n(1 + o_P(1)).$$

By Corollary C.2, then,

$$\begin{aligned} \Phi\left(\frac{\widehat{\text{ISNR}}_n(1 - o_P(1))}{\sqrt{2}} - \bar{\Phi}^{-1}(\alpha)\right) - o_P(1) &\leq \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} \\ &\leq \Phi\left(\widehat{\text{ISNR}}_n(1 + o_P(1)) - \bar{\Phi}^{-1}(\alpha)\right) + o_P(1). \end{aligned}$$

Applying Lemma C.8 completes the proof. $\square$

C.2.3 Power of neighbor matching (proof of Theorem 4.10)

Let us define π to be any permutation of $[n]$ for which

$$Z_{\pi(1)} \leq Z_{\pi(2)} \leq \dots \leq Z_{\pi(n)},$$

and note that the set of matched pairs for neighbor matching is given by $M := \{(\pi(2\ell - 1), \pi(2\ell))\}_{\ell \in [n/2]}$. By Corollary C.2 and Lemma C.8, with $\mathbf{w} = \Delta\mathbf{Y}^+$, it suffices to prove

that

$$\left| \frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \cdot \left\{ \mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] \right\}^{1/2} \right| = o_P\left(\frac{\beta_n \sqrt{n}}{\sigma_n}\right), \quad (\text{C.15})$$

and that Assumptions **A1** and **A2** are satisfied by neighbor matching.

Step 1: establishing (C.15). Under the partially linear Gaussian model (4.16),

$$\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z}) = \Delta \mathbf{Y}^{+\top} \Delta \mu_0(\mathbf{Z}) + \beta_n \|\Delta \mathbf{Y}^+\|_2^2, \quad (\text{C.16})$$

where the first term is non-positive since $\mu_0 \in \mathcal{C}_{\text{ISO}}$. Moreover, we also have $|\Delta \mathbf{Y}^{+\top} \Delta \mu_0(\mathbf{Z})| \leq \|\Delta \mathbf{Y}^+\|_\infty \|\Delta \mu_0(\mathbf{Z})\|_1$. Since $\mu_0 \in \mathcal{C}_{\text{ISO}}$, we also have that

$$\begin{aligned} \|\Delta \mu_0(\mathbf{Z})\|_1 &= \sum_{\ell=1}^{\lfloor n/2 \rfloor} (\mu_0(Z_{\pi(2\ell)}) - \mu_0(Z_{\pi(2\ell-1)})) \leq \sum_{i=2}^n (\mu_0(Z_{\pi(i)}) - \mu_0(Z_{\pi(i-1)})) \\ &= \mu_0(Z_{\pi(n)}) - \mu_0(Z_{\pi(1)}) \leq 2\|\mu_0(\mathbf{Z})\|_\infty. \end{aligned} \quad (\text{C.17})$$

Now $\|\Delta \mathbf{Y}^+\|_\infty \leq 2$, and since $\mu_0(Z)$ is sub-Gaussian, we have $\|\mu_0(\mathbf{Z})\|_\infty = O_P(\sqrt{\log n})$.

Additionally, by Proposition C.9,

$$\left| \frac{2}{n} \|\Delta \mathbf{Y}^+\|_2^2 - \mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] \right| = o_P(1). \quad (\text{C.18})$$

Moreover, $\mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] > 0$, so it now follows that

$$\begin{aligned}
& \left| \frac{\Delta \mathbf{Y}^+{}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \left\{ \mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] \right\}^{1/2} \right| \\
& \leq \left| \frac{\Delta \mathbf{Y}^+{}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \frac{\beta_n}{\sigma_n} \|\Delta \mathbf{Y}^+\|_2 \right| + \left| \frac{\beta_n}{\sigma_n} \|\Delta \mathbf{Y}^+\|_2 - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \left\{ \mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] \right\}^{1/2} \right| \\
& \leq 4 \frac{\|\mu_0(\mathbf{Z})\|_\infty}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} + o_P \left(\frac{\beta_n \sqrt{n}}{\sigma_n} \right) = O_P \left(\frac{\sqrt{\log n}}{\sigma_n \sqrt{n}} \right) + o_P \left(\frac{\beta_n \sqrt{n}}{\sigma_n} \right) = o_P(1),
\end{aligned}$$

where in the penultimate line, since $\|\mu_0(\mathbf{Z})\|_\infty = O_P(\sqrt{\log n})$ and $\sigma_n \gg \sqrt{\frac{\log n}{n}}$, by (C.18), we have that $\|\mu_0(\mathbf{Z})\|_\infty / (\sigma_n \|\Delta \mathbf{Y}^+\|_2) = O_P(\sqrt{\log n} / (\sigma_n \sqrt{n})) = o_P(1)$; we also use the assumption that $\beta_n \gtrsim \sigma_n n^{-1/2}$, and so $\frac{\beta_n \sqrt{n}}{\sigma_n} \gtrsim 1$. This establishes (C.15).

Step 2: establishing Assumptions A1 and A2. Since Y is bounded, and $\|\Delta \mathbf{Y}^+\|_2 = O_P(\sqrt{n})$ by (C.18), Assumption **A1** holds for neighbor matching. Next, turning to Assumption **A2**, we first observe that by (C.17),

$$\begin{aligned}
\|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2 & \leq \|\Delta \mathbf{Y}^+ \circ \Delta \mu_0(\mathbf{Z})\|_2 + \beta_n \|\Delta \mathbf{Y}^+\|_4^2 \\
& \leq \|\Delta \mathbf{Y}^+\|_\infty \|\Delta \mu_0(\mathbf{Z})\|_1 + \beta_n \|\Delta \mathbf{Y}^+\|_\infty \|\Delta \mathbf{Y}^+\|_2 \\
& \leq 4 \|\mu_0(\mathbf{Z})\|_\infty + 2\beta_n \|\Delta \mathbf{Y}^+\|_2.
\end{aligned} \tag{C.19}$$

Also, by (C.15), it follows that

$$\begin{aligned}
\frac{\|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\Delta \mathbf{Y}^+{}^\top \Delta \mu_n(\mathbf{Y}, \mathbf{Z})} & \leq \frac{4 \|\mu_0(\mathbf{Z})\|_\infty + 2\beta_n \|\Delta \mathbf{Y}^+\|_2}{\|\Delta \mathbf{Y}^+\|_2 \beta_n \sqrt{n/2} \left\{ \mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] \right\}^{1/2} - o_P(\beta_n \sqrt{n} \|\Delta \mathbf{Y}^+\|_2)} \\
& = \frac{\frac{4 \|\mu_0(\mathbf{Z})\|_\infty}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} + \frac{2\beta_n}{\sigma_n}}{\frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \left\{ \mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] \right\}^{1/2} - o_P \left(\frac{\beta_n \sqrt{n}}{\sigma_n} \right)}.
\end{aligned}$$

As before, $\|\mu_0(\mathbf{Z})\|_\infty/(\sigma_n\|\Delta\mathbf{Y}^+\|_2) = o_P(1)$ and $\beta_n \gtrsim \sigma_n n^{-1/2}$, while $\mathbb{E} \left[\text{Var}_{Y \sim P_{Y|Z}}(Y) \right] > 0$ by assumption. This therefore yields

$$\frac{\|\Delta\mathbf{Y}^+ \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\Delta\mathbf{Y}^{+\top} \Delta\mu_n(\mathbf{Y}, \mathbf{Z})} = o_P(1),$$

i.e., the first condition of Assumption **A2** is satisfied by neighbor matching. Next, we note that

$$\|\Delta\mathbf{Y}^+ \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty \leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_0(\mathbf{Z})\|_\infty + \beta_n \|\Delta\mathbf{Y}^+\|_\infty^2 \leq 4(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n),$$

and also that

$$\|\Delta\mathbf{Y}^+ \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 \leq \|\Delta\mathbf{Y}^+ \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \cdot \|\Delta\mathbf{Y}^+ \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^{2/3}.$$

Thus, by (C.18) and (C.19), and recalling that $1 \gtrsim \beta_n/\sigma_n$ and $\sigma_n \gg \sqrt{\frac{\log n}{n}}$, we have that

$$\begin{aligned} \|\Delta\mathbf{Y}^+ \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 &\leq 4^{1/3}(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n)^{1/3} \cdot (4\|\mu_0(\mathbf{Z})\|_\infty + 2\beta_n\|\Delta\mathbf{Y}^+\|_2)^{2/3} \\ &= O_P\left((\sqrt{\log n} + \beta_n)^{1/3}(\sqrt{\log n} + \beta_n\sqrt{n})^{2/3}\right) = o_P(\sigma_n\sqrt{n}) = o_P(\sigma_n\|\Delta\mathbf{Y}^+\|_2). \end{aligned}$$

Therefore, the second condition of Assumption **A2** is also satisfied by neighbor matching.

This completes the proof. $\square$

C.2.4 Power of cross-bin matching (proof of Theorem 4.11)

In this section, we study the power of the PairSwap-ICI test for cross-bin matching.

Definition C.1. For any probability distribution P on $\mathbb{R}$, we define the deviance of P as

$$\text{Dev}(P) := \mathbb{E}_{q \sim \text{Unif}(0,1)} \left[\left(F^{-1}(1-q) - F^{-1}(q) \right)^2 \right], \quad (\text{C.20})$$

where F^{-1} is the quantile function of P .

We remark that deviance is a general measure of deviation: in fact, for any distribution, $\text{Dev}(P) \geq 2 \text{Var}_{X \sim P}(X)$, as we will prove formally in Lemma C.13.

With this definition in place, we first state a general concentration result for the conditional power of cross-bin matching in the following theorem.

Theorem C.3. *Under the same setting and assumptions as Theorem 4.11, the conditional power of cross-bin matching (Algorithm 3) satisfies*

$$\left| \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} - \Phi\left(\frac{\sqrt{n}\beta_n}{2\sigma_n} \cdot (\mathbb{E}[\text{Dev}(P_{Y|Z})])^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) \right| = o_P(1).$$

This result will allow us to immediately verify Theorem 4.11 (see below), but will also be useful for proving additional results later on.

Proof of Theorem 4.11

By Lemma C.13, for any distribution P on $\mathbb{R}$, it holds that $\text{Dev}(P) \geq 2\text{Var}_{X \sim P}(X)$. Therefore, $(\mathbb{E}[\text{Dev}(P_{Y|Z})])^{1/2} \geq \sqrt{2}(\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)])^{1/2}$, and so the first part of Theorem 4.11 follows as an immediate corollary of Theorem C.3.

For the second part, we note that by Lemma C.13, for any distribution P on $\mathbb{R}$ that is symmetric around its mean, it holds that $\text{Dev}(P) = 4\text{Var}_{X \sim P}(X)$. Applying this calculation to $P_{Y|Z}$, then, we have $\mathbb{E}[\text{Dev}(P_{Y|Z})] = 4\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)]$. Thus, the second part follows from Theorem C.3. $\square$

Proof of Theorem C.3

The proof for cross-bin matching follows the same key steps as in the proof for neighbor matching in Appendix C.2.3. We first show that

$$\left| \frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \cdot \left\{ \mathbb{E} \left[\text{Dev}(P_{Y|Z}) \right] \right\}^{1/2} \right| = o_P \left(\frac{\beta_n \sqrt{n}}{\sigma_n} \right) \quad (\text{C.21})$$

and then establish that Assumptions **A1** and **A2** are satisfied by cross-bin matching. Finally, we will apply Corollary C.2 with $\mathbf{w} = \Delta \mathbf{Y}^+$ to conclude the proof.

Step 1: establishing (C.21). Recall the decomposition (C.16). Since $\mu_0 \in \mathcal{C}_{\text{ISO}}$, the first term is non-positive, and we also have that $|\Delta \mathbf{Y}^{+\top} \Delta \mu_0(\mathbf{Z})| \leq \|\Delta \mathbf{Y}^+\|_\infty \|\Delta \mu_0(\mathbf{Z})\|_1$. Next, recalling the notation from Algorithm 3, and using the fact that $\mu_0 \in \mathcal{C}_{\text{ISO}}$, we have that

$$\begin{aligned} \|\Delta \mu_0(\mathbf{Z})\|_1 &\leq \sum_{k=1}^{K_n-1} \sum_{s=1}^{\lfloor m/2 \rfloor} (\mu_0(Z_{r_{k+1},s}) - \mu_0(Z_{r_{k,m+1-s}})) \\ &\leq \left\lfloor \frac{m}{2} \right\rfloor \sum_{k=1}^{K_n-1} (\mu_0(Z_{\pi((k+1)m)}) - \mu_0(Z_{\pi((k-1)m+1)})) \\ &\leq \frac{n}{K_n} (\mu_0(Z_{\pi(n)}) - \mu_0(Z_{\pi(1)})) \leq \frac{2n}{K_n} \|\mu_0(\mathbf{Z})\|_\infty. \end{aligned} \quad (\text{C.22})$$

Since $\mu_0(Z)$ is sub-Gaussian, $\|\mu_0(\mathbf{Z})\|_\infty = O_P(\sqrt{\log n})$. Also, since Y is bounded and $K_n \gg \sqrt{n \log n} / \sigma_n$, we have that

$$|\Delta \mathbf{Y}^{+\top} \Delta \mu_0(\mathbf{Z})| = o_P(\sigma_n \sqrt{n}).$$

Moreover by Proposition C.10,

$$\frac{2}{n} \|\Delta \mathbf{Y}^+\|_2^2 \xrightarrow{P} \mathbb{E} \left[\text{Dev}(P_{Y|Z}) \right]. \quad (\text{C.23})$$

Therefore, by recalling that $\beta_n \gtrsim \sigma_n n^{-1/2}$, we finally have that

$$\begin{aligned}
& \left| \frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \cdot \{\mathbb{E} [\text{Dev}(P_{Y|Z})]\}^{1/2} \right| \\
& \leq \left| \frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \frac{\beta_n \|\Delta \mathbf{Y}^+\|_2}{\sigma_n} \right| + \left| \frac{\beta_n \|\Delta \mathbf{Y}^+\|_2}{\sigma_n} - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \cdot \{\mathbb{E} [\text{Dev}(P_{Y|Z})]\}^{1/2} \right| \\
& = \left| \frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_0(\mathbf{Z})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} \right| + \left| \frac{\beta_n \|\Delta \mathbf{Y}^+\|_2}{\sigma_n} - \frac{\sqrt{n/2} \cdot \beta_n}{\sigma_n} \cdot \{\mathbb{E} [\text{Dev}(P_{Y|Z})]\}^{1/2} \right| \\
& = \frac{o_P(\sigma_n \sqrt{n})}{\sigma_n \|\Delta \mathbf{Y}^+\|_2} + o_P\left(\frac{\beta_n \sqrt{n}}{\sigma_n}\right) = o_P\left(1 + \frac{\beta_n \sqrt{n}}{\sigma_n}\right) = o_P\left(\frac{\beta_n \sqrt{n}}{\sigma_n}\right).
\end{aligned}$$

This establishes (C.21).

Step 2: establishing Assumptions A1 and A2. Since Y is bounded, we have by (C.23) that Assumption **A1** holds for the weights $\mathbf{w} = \Delta \mathbf{Y}^+$ of cross-bin matching. Next, turning to Assumption **A2**, we first observe that by (C.22),

$$\begin{aligned}
\|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2 & \leq \|\Delta \mathbf{Y}^+ \circ \Delta \mu_0(\mathbf{Z})\|_2 + \beta_n \|\Delta \mathbf{Y}^+ \circ \Delta \mathbf{Y}\|_2 \\
& \leq \|\Delta \mathbf{Y}\|_\infty (\|\Delta \mu_0(\mathbf{Z})\|_2 + \beta_n \|\Delta \mathbf{Y}^+\|_2) \\
& \leq 2(\|\Delta \mu_0(\mathbf{Z})\|_1^{1/2} \|\Delta \mu_0(\mathbf{Z})\|_\infty^{1/2} + \beta_n \|\Delta \mathbf{Y}^+\|_2) \\
& \leq 2\left(\left(\frac{4n}{K_n}\right)^{1/2} \|\mu_0(\mathbf{Z})\|_\infty + \beta_n \|\Delta \mathbf{Y}^+\|_2\right) \\
& = O_P\left(\sqrt{\frac{n \log n}{K_n}} + \beta_n \sqrt{n}\right) = o_P(\sqrt{\sigma_n} \sqrt[4]{n \log n}) + O_P(\beta_n \sqrt{n}).
\end{aligned}$$

Hence, in combination with (C.21), it follows that

$$\begin{aligned}
\frac{\|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2}{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})} &\leq \frac{o_P(\sigma_n^{-1/2}(n \log n)^{1/4}) + O_P(\frac{\beta_n \sqrt{n}}{\sigma_n})}{\frac{\beta_n \sqrt{n/2}}{\sigma_n} \|\Delta \mathbf{Y}^+\|_2 \{\mathbb{E} [\text{Dev}(P_{Y|Z})]\}^{1/2} - \|\Delta \mathbf{Y}^+\|_2 o_P(\frac{\beta_n \sqrt{n}}{\sigma_n})} \\
&= \frac{o_P(\sigma_n^{-1/2}(n \log n)^{1/4}) + O_P(\frac{\beta_n \sqrt{n}}{\sigma_n})}{O_P(\frac{\beta_n n}{\sigma_n})} \\
&= o_P(\sigma_n^{-1/2} n^{-1/4} \log^{1/4} n) + O_P(n^{-1/2}) = o_P(1),
\end{aligned}$$

since $\beta_n \gtrsim \sigma_n/\sqrt{n}$, $\sigma_n \gg \sqrt{\frac{\log n}{n}}$, and $\mathbb{E} [\text{Dev}(P_{Y|Z})] > 0$. In other words, the first condition of Assumption **A2** is satisfied by cross-bin matching. Next, since $1 \gtrsim \beta_n/\sigma_n$, we note that

$$\begin{aligned}
\|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty &\leq \|\Delta \mathbf{Y}^+ \circ \Delta \mu_0(\mathbf{Z})\|_\infty + \beta_n \|\Delta \mathbf{Y}^+\|_\infty^2 \leq 4(\|\mu_0(\mathbf{Z})\|_\infty + \beta_n) \\
&= O_P(\sqrt{\log n} + \sigma_n).
\end{aligned}$$

Since $\beta_n \lesssim \sigma_n$ and $\sigma_n \gg \sqrt{\frac{\log n}{n}}$, we therefore have that

$$\begin{aligned}
\|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_3 &\leq \|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_\infty^{1/3} \|\Delta \mathbf{Y}^+ \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^{2/3} \\
&\leq O_P(\sqrt{\log n} + \sigma_n)^{1/3} \cdot \left(o_P(\sqrt{\sigma_n} \sqrt[4]{n \log n}) + O_P(\beta_n \sqrt{n}) \right)^{2/3} \\
&= O_P(\sqrt[6]{\log n} + \sigma_n^{1/3}) \cdot \left(o_P(\sigma_n^{1/3} \sqrt[6]{n \log n}) + O_P(\beta_n^{2/3} \sqrt[3]{n}) \right) \\
&= o_P(\sigma_n \sqrt{n}) = o_P(\sigma_n \|\Delta \mathbf{Y}^+\|_2).
\end{aligned}$$

Therefore, the second condition of Assumption **A2** is also satisfied by cross-bin matching.

Step 3: completing the proof By Corollary C.2, (C.21), and Lemma C.8, we have

$$\begin{aligned}
& \left| \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} - \Phi\left(\frac{\sqrt{n}\beta_n}{2\sigma_n} \cdot (\mathbb{E}[\text{Dev}(P_{Y|Z})])^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) \right| \\
& \leq \left| \mathbb{P}\{p \leq \alpha \mid \mathbf{Y}, \mathbf{Z}\} - \Phi\left(\frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \bar{\Phi}^{-1}(\alpha)\right) \right| \\
& \quad + \left| \Phi\left(\frac{\Delta \mathbf{Y}^{+\top} \Delta \mu_n(\mathbf{Y}, \mathbf{Z})}{\sqrt{2}\sigma_n \|\Delta \mathbf{Y}^+\|_2} - \bar{\Phi}^{-1}(\alpha)\right) - \Phi\left(\frac{\sqrt{n}\beta_n}{2\sigma_n} \cdot (\mathbb{E}[\text{Dev}(P_{Y|Z})])^{1/2} - \bar{\Phi}^{-1}(\alpha)\right) \right| \\
& = o_P(1),
\end{aligned}$$

as required. $\square$

C.2.5 Proof of propositions, corollaries and lemmas from Section 4.5

Proof of Proposition 4.7

Fix a distribution $Q_{X,Y,Z} \in H_0^{\text{ICI}}$. For any test function

$$\phi : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \rightarrow [0, 1] \text{ such that } \sup_{P \in H_0^{\text{ICI}}} \mathbb{E}_P[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] \leq \alpha,$$

we have that

$$\begin{aligned}
\mathbb{E}_{P_{X,Y,Z}}[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] & \leq \mathbb{E}_{Q_{X,Y,Z}}[\phi(\mathbf{X}, \mathbf{Y}, \mathbf{Z})] + \text{d}_{\text{TV}}\left(P_{X,Y,Z}^n, Q_{X,Y,Z}^n\right) \\
& \leq \alpha + \text{d}_{\text{TV}}\left(P_{X,Y,Z}^n, Q_{X,Y,Z}^n\right).
\end{aligned}$$

The result now follows since the last inequality holds for any $Q_{X,Y,Z} \in H_0^{\text{ICI}}$.

Proof of Corollary 4.8

By Proposition 4.7, it is enough to prove that

$$\inf_{Q_{X,Y,Z} \in H_0^{\text{ICI}}} \text{d}_{\text{TV}} \left(P_{X,Y,Z}^n, Q_{X,Y,Z}^n \right) \leq L_\zeta \cdot \text{ISNR}_n.$$

Now consider any $\tilde{\mu}_n \in \mathcal{C}_{\text{ISO}}$. Similar to (4.12), define $\tilde{Q}_{X,Y,Z}$ to be the joint distribution of (X, Y, Z) when

$$X = \tilde{\mu}_n(Z) + \sigma_n \zeta, \quad (Y, Z) \sim P_{Y,Z}, \quad \zeta \sim P_\zeta.$$

Now abusing notation, we will write $P_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ to denote the conditional distribution of $\mathbf{X}$ given $(\mathbf{Y}, \mathbf{Z})$, under the joint distribution $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \sim P_{X,Y,Z}^n$. Similarly, we will write $Q_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$. Then by properties of Hellinger distance, since the marginal distribution of (Y, Z) is identical under $P_{X,Y,Z}$ as under $Q_{X,Y,Z}$ by construction, we have

$$\text{d}_{\text{TV}}(P_{X,Y,Z}^n, Q_{X,Y,Z}^n) = \mathbb{E} \left[\text{d}_{\text{TV}}(P_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}, Q_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}) \right] \leq \mathbb{E} \left[\text{H}(P_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}, Q_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}) \right],$$

where the last step holds since total variation distance is bounded by Hellinger distance [Le Cam and Yang, 2000]. Moreover, by construction, the distribution $P_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ has density

$$\prod_{i=1}^n \sigma_n^{-1} f_\zeta \left(\frac{x_i - \mu_n(Y_i, Z_i)}{\sigma_n} \right),$$

while $Q_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ has density

$$\prod_{i=1}^n \sigma_n^{-1} f_\zeta \left(\frac{x_i - \tilde{\mu}_n(Z_i)}{\sigma_n} \right).$$

Since the squared Hellinger distance is subadditive over product distributions [Canonne, 2020, Appendix C.2],

$$\begin{aligned}
H^2(P_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}, Q_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}) &\leq \sum_{i=1}^n \int_{\mathbb{R}} \left[\sqrt{\sigma_n^{-1} f_{\zeta} \left(\frac{x_i - \mu_n(Y_i, Z_i)}{\sigma_n} \right)} - \sqrt{\sigma_n^{-1} f_{\zeta} \left(\frac{x_i - \tilde{\mu}_n(Z_i)}{\sigma_n} \right)} \right]^2 dx \\
&= \sum_{i=1}^n \int_{\mathbb{R}} \left[\sqrt{f_{\zeta} \left(x_i - \frac{\mu_n(Y_i, Z_i)}{\sigma_n} \right)} - \sqrt{f_{\zeta} \left(x_i - \frac{\tilde{\mu}_n(Z_i)}{\sigma_n} \right)} \right]^2 dx \\
&\leq \sum_{i=1}^n \left(L_{\zeta} \cdot \left| \frac{\mu_n(Y_i, Z_i)}{\sigma_n} - \frac{\tilde{\mu}_n(Z_i)}{\sigma_n} \right| \right)^2 \\
&= \left(L_{\zeta} \sigma_n^{-1} \|\mu_n(\mathbf{Y}, \mathbf{Z}) - \tilde{\mu}_n(\mathbf{Z})\|_2 \right)^2,
\end{aligned}$$

where the second step holds by a change of variables in integration, and the third step holds by the condition (4.15). Returning to our calculations above, then,

$$d_{\text{TV}}(P_{X,Y,Z}^n, Q_{X,Y,Z}^n) \leq \mathbb{E} \left[L_{\zeta} \sigma_n^{-1} \|\mu_n(\mathbf{Y}, \mathbf{Z}) - \tilde{\mu}_n(\mathbf{Z})\|_2 \right].$$

Since we can choose any $\tilde{\mu}_n \in \mathcal{C}_{\text{ISO}}$, then,

$$\inf_{Q_{X,Y,Z} \in H_0^{\text{ICI}}} d_{\text{TV}}(P_{X,Y,Z}^n, Q_{X,Y,Z}^n) \leq \inf_{\tilde{\mu}_n \in \mathcal{C}_{\text{ISO}}} \mathbb{E} \left[L_{\zeta} \sigma_n^{-1} \|\mu_n(\mathbf{Y}, \mathbf{Z}) - \tilde{\mu}_n(\mathbf{Z})\|_2 \right] = L_{\zeta} \cdot \text{ISNR}_n,$$

which completes the proof. $\square$

Proof of Lemma 4.9

Under the model class (4.16), since $\mathbb{E}[Y] = 0$, we have that

$$\begin{aligned}
\sigma_n \text{ISNR}_n &\stackrel{(1)}{\leq} \mathbb{E}_{P_{Y,Z}} [\|\mu_n(\mathbf{Y}, \mathbf{Z}) - \mu_0(\mathbf{Z})\|_2] = \mathbb{E}_{P_{Y,Z}} [\|\beta_n \mathbf{Y}\|_2] \\
&\stackrel{(2)}{\leq} \beta_n \left(\mathbb{E}_{P_{Y,Z}} \left[\sum_{i=1}^n Y_i^2 \right] \right)^{1/2} = \sqrt{n} \beta_n (\text{Var}(Y))^{1/2},
\end{aligned}$$

where the inequality (1) holds since $\mu_0 \in \mathcal{C}_{\text{ISO}}$, and (2) holds by Jensen's inequality. Next, again under the model class (4.16), writing $\mathcal{G}$ for the set of Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$,

$$\begin{aligned}\sigma_n \text{ISNR}_n &= \inf_{g \in \mathcal{C}_{\text{ISO}}} \mathbb{E}_{P_{Y,Z}} [\|\mu_n(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2] \geq \inf_{g \in \mathcal{G}} \mathbb{E}_{P_{Y,Z}} [\|\mu_n(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2] \\ &= \mathbb{E}_{P_{Y,Z}} [\|\mu_n(\mathbf{Y}, \mathbf{Z}) - \mathbb{E}[\mu_n(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Z}]\|_2] = \mathbb{E}_{P_{Y,Z}} [\|\beta_n(\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{Z}])\|_2].\end{aligned}$$

Finally, since Y is a bounded random variable, we have by the weak law of large numbers that

$$\frac{1}{n} \|\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]\|_2^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - E[Y \mid Z_i])^2 = \mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)] (1 + o_P(1)).$$

Thus, $\sigma_n \text{ISNR}_n \geq \sqrt{n} \beta_n (\mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)])^{1/2} (1 + o_P(1))$. $\square$

C.2.6 An oracle matching: isotonic median matching

A key step in Appendix C.2.2 is establishing the inequalities in (C.10) for oracle matching, which clarify its relationship to the isotonic SNR, $\widehat{\text{ISNR}}_n$. In this section, we will prove these inequalities, restated here in the following theorem:

Theorem C.4. *For any $(\mathbf{Y}, \mathbf{Z}) \in \mathbb{R}^n \times \mathbb{R}^n$,*

$$\widehat{\text{ISNR}}_n \leq \frac{\sup_{M \in \mathcal{M}_n(\mathbf{Z})} \|\Delta \mu_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2}{\sigma_n} \leq \sqrt{2} \widehat{\text{ISNR}}_n.$$

In order to prove the lower bound of this result, we will give an explicit construction for an optimal matching, via a technique that we refer to as isotonic median matching (IMM). In the remainder of this section, we will first develop the definition and properties of the IMM,

and will then return to the proof of the theorem.

From this point on, assume that $Z_1 \leq \dots \leq Z_n$ without loss of generality. Given $a = (a_1, \dots, a_m) \in \mathbb{R}^m$, we define the median of this vector by

$$\text{Med}(a) = \begin{cases} a_{(k+1)} & \text{if } m = 2k + 1 \text{ for some } k \in \mathbb{N}_0, \\ \frac{a_{(k)} + a_{(k+1)}}{2} & \text{if } m = 2k \text{ for some } k \in \mathbb{N}, \end{cases}$$

where $a_{(1)} \leq \dots \leq a_{(m)}$ denote the order statistics of $a_1, \dots, a_m$. Given $\mathbf{Z}$, we define the cone of vectors in $\mathbb{R}^n$ that obey isotonic relation with $\mathbf{Z}$ as

$$\widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z}) := \{v \in \mathbb{R}^n : \text{for all } 1 \leq i \leq j \leq n, (z_i - z_j)(v_i - v_j) \geq 0\}.$$

Moreover, given $(\mathbf{Y}, \mathbf{Z})$ and any vector $u \in \mathbb{R}^n$, we define $\tilde{u} \in \mathbb{R}^n$ with i th component given by

$$\tilde{u}_i := \max_{j \in [n]: Z_j \leq Z_i} \min_{k \in [n]: Z_k \geq Z_i} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k}). \quad (\text{C.24})$$

We claim that $\tilde{u} \in \widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z})$ —the intuition is that we can think of $\tilde{u}$ as a projection of u onto the cone $\widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z})$ (in fact, it is the projection with respect to the ℓ_1 norm, although we will not establish this claim formally as it is not needed here). To verify that $\tilde{u} \in \widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z})$, observe that for $1 \leq i_1 \leq i_2 \leq n$,

$$\begin{aligned} \tilde{u}_{i_1} &= \max_{j \in [n]: Z_j \leq Z_{i_1}} \min_{k \in [n]: Z_k \geq Z_{i_1}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k}) \\ &\leq \max_{j \in [n]: Z_j \leq Z_{i_1}} \min_{k \in [n]: Z_k \geq Z_{i_2}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k}) \\ &\leq \max_{j \in [n]: Z_j \leq Z_{i_2}} \min_{k \in [n]: Z_k \geq Z_{i_2}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k}) = \tilde{u}_{i_2}, \end{aligned}$$

which proves the claim. Next, since $\tilde{u} \in \widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z})$, there exist $m \in [n+1]$, integers $1 = n_1 <$

$\dots < n_m = n + 1$ and real numbers $r_1 < \dots < r_{m-1}$ such that

$$\tilde{u}_i = \sum_{t=1}^{m-1} r_t \mathbb{1} \{n_t \leq i < n_{t+1}\}, \quad (\text{C.25})$$

i.e., $\tilde{u}$ is a piecewise constant vector, with $m - 1$ components.

The following lemma establishes some additional properties of this vector.

Lemma C.5. *In the representation (C.25) of the vector $\tilde{u}$, for each $t \in [m - 1]$,*

$$r_t = \text{Med}((u_\ell)_{n_t \leq \ell \leq n_{t+1}-1}), \quad (\text{C.26})$$

and moreover, for each $i \in \{n_t, n_t + 1, \dots, n_{t+1} - 1\}$,

$$\text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_i}) \geq r_t \geq \text{Med}((u_\ell)_{\ell: Z_i \leq Z_\ell \leq Z_{n_{t+1}-1}}). \quad (\text{C.27})$$

Proof. First, since $\tilde{u} \in \widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z})$, and $\tilde{u}_{n_t-1} < \tilde{u}_{n_t}$ for each $t = 2, \dots, m - 1$, we have $Z_{n_t-1} < Z_{n_t}$ for each $t = 2, \dots, m - 1$ and consequently

$$n_t \leq \ell \leq n_{t+1} - 1 \iff Z_{n_t} \leq Z_\ell \leq Z_{n_{t+1}-1}$$

for each $t \in [m - 1]$. Therefore, establishing (C.26) is equivalent to proving that

$$r_t = \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_{n_{t+1}-1}}),$$

which is an immediate consequence of the inequalities in (C.27) (applied with $i = n_{t+1} - 1$ on the left and with $i = n_t$ on the right). Therefore we now only need to prove these inequalities.

Upper bound on r_t . First we will prove that

$$r_t \leq \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_i}) \quad (\text{C.28})$$

for all $t \in [m-1]$ and all $i \in \{n_t, \dots, n_{t+1}-1\}$. By definition, we have

$$r_t = \tilde{u}_{n_t} = \max_{j \in [n]: Z_j \leq Z_{n_t}} \min_{k \in [n]: Z_k \geq Z_{n_t}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k}).$$

As noted above, $Z_j < Z_{n_t}$ for any $j < n_t$. Therefore we can write

$$\{j \in [n] : Z_j \leq Z_{n_t}\} = \{j \in [n] : j < n_t\} \cup \{j \in [n] : Z_j = Z_{n_t}\},$$

and hence,

$$r_t = \max \left\{ \underbrace{\max_{j < n_t} \min_{k \in [n]: Z_k \geq Z_{n_t}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k})}_{=: r_{t,1}}, \underbrace{\min_{k \in [n]: Z_k \geq Z_{n_t}} \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_k})}_{=: r_{t,2}} \right\},$$

where we take the convention that the maximum over an empty set is equal to $-\infty$ (to handle the case $t = 1$, i.e., $\{j \in [n] : j < n_1\} = \emptyset$, so that $r_{t,1} = -\infty$ and $r_t = r_{t,2}$). If instead $t \geq 2$, then we claim that $r_{t,1} \leq r_{t,2}$, so that $r_t = r_{t,2}$ again. Suppose for a contradiction that $r_{t,1} > r_{t,2}$. Let $\hat{j} \in [n_t - 1]$ and $\hat{k} \in [n]$ be such that $Z_{\hat{k}} \geq Z_{n_t}$ and

$$r_{t,1} = \min_{k \in [n]: Z_k \geq Z_{n_t}} \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_k}) \text{ and } r_{t,2} = \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_{\hat{k}}}). \quad (\text{C.29})$$

Then, $\text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{\hat{k}}}) \geq r_{t,1} > \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_{\hat{k}}})$. Since $\{j : j < n_t\} = \{j : Z_j \leq Z_{n_t-1}\}$ (due to Z being a nondecreasing vector with $Z_{n_t-1} < Z_{n_t}$ as established

above), it holds that $Z_{\hat{j}} \leq Z_{n_t-1} < Z_{n_t}$ and thus, by Lemma C.14,

$$\text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{\hat{k}}}) \leq \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{n_t-1}}).$$

Now, we also have

$$r_{t,1} = r_t > \tilde{u}_{n_t-1} \geq \min_{k \in [n]: Z_k \geq Z_{n_t-1}} \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_k}) \quad (\text{C.30})$$

Moreover, by noting that $\{k : Z_k \geq Z_{n_t-1}\} = \{k : Z_k \geq Z_{n_t}\} \cup \{k : Z_k = Z_{n_t-1}\}$ and by (C.29), it follows that

$$\begin{aligned} & \min_{k \in [n]: Z_k \geq Z_{n_t-1}} \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_k}) \\ &= \min \left\{ \min_{k \in [n]: Z_k \geq Z_{n_t}} \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_k}), \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{n_t-1}}) \right\} \\ &= \min \left\{ r_{t,1}, \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{n_t-1}}) \right\} = \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{n_t-1}}), \end{aligned}$$

where the last step follows by (C.30). Thus, we have that

$$\text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{\hat{k}}}) \geq r_{t,1} > \text{Med}((u_\ell)_{\ell: Z_{\hat{j}} \leq Z_\ell \leq Z_{n_t-1}}),$$

which is a contradiction. Thus we again have $r_t = r_{t,2}$ in this case. Therefore, for all $t \in [m-1]$,

$$r_t = r_{t,2} = \min_{k \in [n]: Z_k \geq Z_{n_t}} \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_k}) \leq \min_{i \in \{n_t, \dots, n_{t+1}-1\}} \text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_i}).$$

This establishes the desired upper bound (C.28) on r_t .

Lower bound on r_t . Next we will prove that

$$r_t \geq \text{Med}((u_\ell)_{\ell: Z_i \leq Z_\ell \leq Z_{n_{t+1}-1}}) \quad (\text{C.31})$$

for all $t \in [m-1]$ and all $i \in \{n_t, \dots, n_{t+1}-1\}$. This proof is not simply a symmetric version of the proof of the upper bound, since the vector $\tilde{u}$ is defined by taking the maximum of the minimum (and not vice versa).

By definition, we have

$$r_t = \tilde{u}_{n_{t+1}-1} = \max_{j \in [n]: Z_j \leq Z_{n_{t+1}-1}} \min_{k \in [n]: Z_k \geq Z_{n_{t+1}-1}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_k}).$$

If $t = m-1$ then $n_{t+1}-1 = n$, and so $\{k \in [n] : Z_k \geq Z_{n_{t+1}-1}\} = \{k \in [n] : Z_k \geq Z_n\}$.

Therefore, for $t = m-1$ we have

$$\begin{aligned} r_t &= \max_{j \in [n]: Z_j \leq Z_n} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_n}) = \max_{j \in [n]} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_n}) \\ &\geq \max_{j \in \{n_t, \dots, n\}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_\ell \leq Z_n}). \end{aligned}$$

This proves the claim for the case $t = m-1$.

From this point on we assume $t \in [m-2]$. As noted above, $Z_k > Z_{n_{t+1}-1}$ for any $k \geq n_{t+1}$.

Therefore we can write

$$\{k \in [n] : Z_k \geq Z_{n_{t+1}-1}\} = \{k \in [n] : k \geq n_{t+1}\} \cup \{k \in [n] : Z_k = Z_{n_{t+1}-1}\},$$

and hence,

$$r_t = \max_{j \in [n]: Z_j \leq Z_{n_{t+1}-1}} \min \left\{ \underbrace{\min_{k \geq n_{t+1}} \text{Med}((u_\ell)_\ell: Z_j \leq Z_\ell \leq Z_k)}_{=: r_{t,j,1}}, \underbrace{\text{Med}((u_\ell)_\ell: Z_j \leq Z_\ell \leq Z_{n_{t+1}-1})}_{=: r_{t,j,2}} \right\}.$$

Fix any j with $Z_j \leq Z_{n_{t+1}-1}$, and suppose that $r_{t,j,1} < r_{t,j,2}$. Then

$$\text{Med}((u_\ell)_\ell: Z_j \leq Z_\ell \leq Z_k) < \text{Med}((u_\ell)_\ell: Z_j \leq Z_\ell \leq Z_{n_{t+1}-1})$$

for some $k \geq n_{t+1}$. Then by Lemma C.14, we must have

$$\text{Med}((u_\ell)_\ell: Z_j \leq Z_\ell \leq Z_k) \geq \text{Med}((u_\ell)_\ell: Z_{n_{t+1}-1} < Z_\ell \leq Z_k).$$

Now let $t' \geq t+1$ be such that $n_{t'} \leq k < n_{t'+1}$. Since $Z_{n_{t+1}-1} < Z_{n_{t+1}}$ as explained above, we have by Lemma C.14 that

$$\begin{aligned} \text{Med}((u_\ell)_\ell: Z_{n_{t+1}-1} < Z_\ell \leq Z_k) &= \text{Med}((u_\ell)_\ell: Z_{n_{t+1}} \leq Z_\ell \leq Z_k) \\ &\geq \min \left\{ \text{Med}((u_\ell)_\ell: Z_{n_{t+1}} \leq Z_\ell < Z_{n_{t+2}}), \dots, \text{Med}((u_\ell)_\ell: Z_{n_{t'-1}} \leq Z_\ell < Z_{n_{t'}}), \text{Med}((u_\ell)_\ell: Z_{n_{t'}} \leq Z_\ell \leq Z_k) \right\} \\ &\geq \min\{r_{t+1}, \dots, r_{m-1}\} = r_{t+1}, \end{aligned}$$

where the second inequality applies the upper bound (C.28) on each $r_{t+1}, \dots, r_{m-1}$. But we also have

$$r_t \geq \min\{r_{t,j,1}, r_{t,j,2}\} = r_{t,j,1} \geq r_{t+1},$$

which is a contradiction.

Therefore we have $r_{t,j,1} \geq r_{t,j,2}$ for all j such that $Z_j \leq Z_{n_{t+1}-1}$, and so

$$\begin{aligned} r_t = \max_{j \in [n]: Z_j \leq Z_{n_{t+1}-1}} r_{t,j,2} &= \max_{j \in [n]: Z_j \leq Z_{n_{t+1}-1}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_{n_{t+1}-1}}) \\ &\geq \max_{j \in \{n_t, \dots, n_{t+1}-1\}} \text{Med}((u_\ell)_{\ell: Z_j \leq Z_{n_{t+1}-1}}). \end{aligned}$$

This proves the lower bound (C.31), and thus completes the proof of the lemma. $\square$

We are now in a position to define the isotonic median matching for any vector $u \in \mathbb{R}^n$. Let m be as defined in (C.25). For $t \in [m-1]$, define

$$\begin{aligned} P_t &:= \{\ell \in \{n_t, n_t+1, \dots, n_{t+1}-1\} : u_\ell > r_t = \tilde{u}_\ell\}, \\ N_t &:= \{\ell \in \{n_t, n_t+1, \dots, n_{t+1}-1\} : u_\ell < r_t = \tilde{u}_\ell\}. \end{aligned}$$

Let us order the indices in P_t as $\ell_{t,1}^+ < \dots < \ell_{t,n_t^+}^+$ and the indices in N_t as $\ell_{t,1}^- < \dots < \ell_{t,n_t^-}^-$. Defining $\bar{n}_t := n_t^+ \wedge n_t^-$, we consider the collection of ordered pairs

$$\mathcal{C}_t := \{(\ell_{t,1}^+, \ell_{t,1}^-), \dots, (\ell_{t,\bar{n}_t}^+, \ell_{t,\bar{n}_t}^-)\}.$$

Finally, we consider $\widetilde{M}(u) := \bigcup_{t=1}^{m-1} \mathcal{C}_t$ and call it the isotonic median matching (IMM). In the lemma below, we prove that this is a valid matching.

Lemma C.6. *Suppose that the elements of $\{u_\ell : \ell \in [n]\}$ are all distinct. Then for any $t \in [m-1]$, we have $n_t^+ = n_t^-$ and $u_\ell = \tilde{u}_\ell$ for all $\ell \in \{n_t, n_t+1, \dots, n_{t+1}-1\} \setminus (P_t \cup N_t)$. Moreover, $\widetilde{M}(u) \in \mathcal{M}_n(\mathbf{Z})$.*

Proof. Fix $t \in [m-1]$. When $n_{t+1} - n_t$ is even, we have by (C.26) (in Lemma C.5) that $n_t^+ = n_t^- = (n_{t+1} - n_t)/2$ and $\{n_t, n_t+1, \dots, n_{t+1}-1\} \setminus (P_t \cup N_t) = \emptyset$. On the other hand, when $n_{t+1} - n_t$ is odd, we have $n_t^+ = n_t^- = (n_{t+1} - n_t - 1)/2$ and $\{n_t, n_t+1, \dots, n_{t+1}-$

$1\} \setminus (P_t \cup N_t) =: \{\ell_0\}$ is a singleton set. Moreover, $u(\ell_0) = \tilde{u}(\ell_0)$ which proves the first two claims.

In order to prove that $\widetilde{M}(u)$ is a valid matching, we need to show that $Z_{\ell_{t,j}^+} \leq Z_{\ell_{t,j}^-}$ for any $t \in [m-1]$ and $j \in [\bar{n}_t]$. To see this, suppose for a contradiction that $Z_{\ell_{t,j_0}^+} > Z_{\ell_{t,j_0}^-}$ for some $t \in [m-1]$ and some minimal $j_0 \in [\bar{n}_t]$. Since $\tilde{u} \in \widehat{\mathcal{C}}_{\text{ISO}}(\mathbf{Z})$, by (C.25), we have $\ell \geq n_t$ whenever $Z_\ell \geq Z_{n_t}$. Consequently, this gives us

$$\begin{aligned} |\{\ell : Z_{n_t} \leq Z_\ell \leq Z_{\ell_{t,j_0}^-} \text{ and } u_\ell < r_t\}| &\geq |\{\ell : n_t \leq \ell \leq \ell_{t,j_0}^- \text{ and } u_\ell < r_t\}| = j_0 \\ |\{\ell : Z_{n_t} \leq Z_\ell \leq Z_{\ell_{t,j_0}^-} \text{ and } u_\ell > r_t\}| &< |\{\ell : n_t \leq \ell \leq \ell_{t,j_0}^+ \text{ and } u_\ell > r_t\}| = j_0, \end{aligned}$$

where the second inequality follows by recalling that $\ell_{t,1}^+ < \dots < \ell_{t,n_t}^+$. Since the elements of $\{u_\ell : \ell \in [n]\}$ are all distinct, this further implies that

$$\text{Med}((u_\ell)_{\ell: Z_{n_t} \leq Z_\ell \leq Z_{\ell_{t,j_0}^-}}) < r_t,$$

which contradicts the result (C.27) proved in Lemma C.5. $\square$

With the definition and properties of the IMM now in place, we are now ready to prove Theorem C.4. This result establishes the claim (C.10) in Appendix C.2.2, which was used for proving the power guarantees for both oracle and plugin matching in Theorems 4.5 and 4.6.

Proof of Theorem C.4. First we prove the upper bound. Fix any matching $M \in \mathcal{M}_n(\mathbf{Z})$. By definition, $\sigma_n \widehat{\text{ISNR}}_n = \inf_{g \in \mathcal{C}_{\text{ISO}}} \|\mu_n(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2$. We claim that

$$\left(\mu_n(Y_{i_\ell}, Z_{i_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell}) \right)_+ \leq |\mu_n(Y_{i_\ell}, Z_{i_\ell}) - g(Z_{i_\ell})| + |g(Z_{j_\ell}) - \mu(Y_{j_\ell}, Z_{j_\ell})| \quad (\text{C.32})$$

for every $\ell \in [L]$. To see this, we may assume that $\mu_n(Y_{i_\ell}, Z_{i_\ell}) > \mu_n(Y_{j_\ell}, Z_{j_\ell})$ since otherwise the left-hand side is zero, and the inequality trivially holds. Then

$$\begin{aligned}
(\mu_n(Y_{i_\ell}, Z_{i_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell}))_+ &= \mu_n(Y_{i_\ell}, Z_{i_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell}) \\
&= \mu_n(Y_{i_\ell}, Z_{i_\ell}) - g(Z_{i_\ell}) + g(Z_{j_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell}) + g(Z_{i_\ell}) - g(Z_{j_\ell}) \\
&\leq |\mu_n(Y_{i_\ell}, Z_{i_\ell}) - g(Z_{i_\ell})| + |g(Z_{j_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell})| + (g(Z_{i_\ell}) - g(Z_{j_\ell})) \\
&\leq |\mu_n(Y_{i_\ell}, Z_{i_\ell}) - g(Z_{i_\ell})| + |g(Z_{j_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell})|,
\end{aligned}$$

where the last step holds since we must have $Z_{i_\ell} \leq Z_{j_\ell}$ (since $M \in \mathcal{M}_n(\mathbf{Z})$) and therefore $g(Z_{j_\ell}) - g(Z_{i_\ell}) \leq 0$ (since $g \in \mathcal{C}_{\text{ISO}}$). This proves the claim (C.32), and therefore, it follows that

$$\begin{aligned}
\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2^2 &= \sum_{\ell=1}^L \left(\mu_n(Y_{i_\ell}, Z_{i_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell}) \right)_+^2 \\
&\leq \sum_{\ell=1}^L \left(|\mu_n(Y_{i_\ell}, Z_{i_\ell}) - g(Z_{i_\ell})| + |g(Z_{j_\ell}) - \mu_n(Y_{j_\ell}, Z_{j_\ell})| \right)^2 \\
&\leq 2 \sum_{\ell=1}^L \left\{ \mu_n(Y_{i_\ell}, Z_{i_\ell}) - g(Z_{i_\ell}) \right\}^2 + 2 \sum_{\ell=1}^L \left\{ \mu_n(Y_{j_\ell}, Z_{j_\ell}) - g(Z_{j_\ell}) \right\}^2 \\
&\leq 2 \sum_{i=1}^n \left\{ \mu_n(Y_i, Z_i) - g(Z_i) \right\}^2 = 2 \|\mu_n(\mathbf{Y}, \mathbf{Z}) - g(\mathbf{Z})\|_2^2 \leq 2\sigma_n^2 \widehat{\text{ISNR}}_n^2.
\end{aligned}$$

This completes the proof of the upper bound.

Next we turn to the lower bound. It is sufficient to show that, for any $\epsilon > 0$, we can find some $M \in \mathcal{M}_n(\mathbf{Z})$ such that $\|\Delta\mu_n^+(\mathbf{Y}, \mathbf{Z}; M)\|_2^2 \geq \sigma_n^2 \widehat{\text{ISNR}}_n^2 - \epsilon$. Fixing any $\epsilon > 0$, first we note that we can find some $u \in \mathbb{R}^n$ such that the coordinates of u are all distinct and $\|\mu_n(\mathbf{Y}, \mathbf{Z}) - u\|_2 < \epsilon/(\sqrt{2} + 1)$. Consider the IMM for this vector u , $\widetilde{M}(u) := \{(i_\ell, j_\ell)\}_{\ell \in \tilde{L}}$, where $\widetilde{M}(u) \in \mathcal{M}_n(\mathbf{Z})$ by Lemma C.6. Furthermore by Lemma C.6 and (C.25), we have for

any $\ell \in [\tilde{L}]$ that $u_{i_\ell} > \tilde{u}_{i_\ell} = \tilde{u}_{j_\ell} > u_{j_\ell}$. Hence,

$$\begin{aligned} \|\Delta u^+\|_2^2 &= \sum_{\ell=1}^{\tilde{L}} (u_{i_\ell} - u_{j_\ell})^2 = \sum_{\ell=1}^{\tilde{L}} (u_{i_\ell} - \tilde{u}_{i_\ell} + \tilde{u}_{j_\ell} - u_{j_\ell})^2 \\ &\geq \sum_{\ell=1}^{\tilde{L}} \{(u_{i_\ell} - \tilde{u}_{i_\ell})^2 + (\tilde{u}_{j_\ell} - u_{j_\ell})^2\} \\ &= \sum_{i=1}^n (u_i - \tilde{u}_i)^2 = \|u - \tilde{u}\|_2^2, \end{aligned}$$

where the penultimate equality holds because $u_i = \tilde{u}_i$ for $i \in [n] \setminus \{i_1, j_1, \dots, i_{\tilde{L}}, j_{\tilde{L}}\}$, by Lemma C.6. Next, we observe that

$$\begin{aligned} \|\Delta u^+\|_2 - \|\Delta \mu_n^+(\mathbf{Y}, \mathbf{Z}; \widetilde{M}(u))\|_2 &\leq \|\Delta u^+ - \Delta \mu_n^+(\mathbf{Y}, \mathbf{Z}; \widetilde{M}(u))\|_2 \\ &\leq \|\Delta u - \Delta \mu_n(\mathbf{Y}, \mathbf{Z}; \widetilde{M}(u))\|_2 \\ &\leq \sqrt{2} \|u - \mu_n(\mathbf{Y}, \mathbf{Z})\|_2 \leq \sqrt{2} \cdot \frac{\epsilon}{\sqrt{2} + 1}, \end{aligned}$$

and that by (4.13),

$$\sigma_n \widehat{\text{ISNR}}_n \leq \|\mu_n(\mathbf{Y}, \mathbf{Z}) - \tilde{u}\|_2 \leq \|u - \tilde{u}\|_2 + \frac{\epsilon}{\sqrt{2} + 1}.$$

Therefore,

$$\begin{aligned} \|\Delta \mu_n^+(\mathbf{Y}, \mathbf{Z}; \widetilde{M}(u))\|_2 &\geq \|\Delta u^+\|_2 - \frac{\sqrt{2}\epsilon}{\sqrt{2} + 1} \geq \|u - \tilde{u}\|_2 - \frac{\sqrt{2}\epsilon}{\sqrt{2} + 1} \\ &\geq \left(\sigma_n \widehat{\text{ISNR}}_n - \frac{\epsilon}{\sqrt{2} + 1} \right) - \frac{\sqrt{2}\epsilon}{\sqrt{2} + 1} = \sigma_n \widehat{\text{ISNR}}_n - \epsilon, \end{aligned}$$

which completes the proof of the lower bound. $\square$

C.2.7 Proof of lemmas from Appendices C.2.1—C.2.4

Lemmas used to prove general power guarantees in Appendix C.2.1

Lemma C.7. *Under the setting and assumptions of Theorem C.1, for any $\delta > 0$, we have that*

$$\mathbb{P}\left\{\frac{\|\mathbf{w} \circ \Delta \mathbf{X}\|_2^2}{(2\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2)} \in [(1 - \epsilon_*) \vee 0, 1 + \epsilon_*] \quad \text{and} \right. \\ \left. \|\mathbf{w} \circ \Delta \mathbf{X}\|_3^2 \leq \left(2\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2\right) \cdot \epsilon'_* \mid \mathbf{Y}, \mathbf{Z}\right\} \geq 1 - \delta$$

where

$$\epsilon_* = \sqrt{\frac{3}{\delta}} \cdot \left(2\rho_4^2 \epsilon_1 + 2(\epsilon_1^2 \epsilon_2)^{1/3}\right), \quad \epsilon'_* = \frac{(24)^{2/3} \rho_4^2}{\delta^{2/3}} \cdot \epsilon_1^{2/3} + 4\epsilon_2^{2/3},$$

and we write $\rho_4 := \mathbb{E}_{P_\zeta} [\zeta^4]^{1/4}$.

Proof. Recall the notation from the hypothesis in statement of Theorem C.1:

$$\epsilon_1 = \frac{\|\mathbf{w}\|_\infty}{\|\mathbf{w}\|_2}, \quad \epsilon_2 = \frac{\|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_3^3}{\sigma_n^3 \|\mathbf{w}\|_2^3 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^3}.$$

Under the model class (4.12), $\Delta \mathbf{X} = \Delta \mu_n(\mathbf{Y}, \mathbf{Z}) + \sigma_n \Delta \zeta$, and thus, we can write

$$\|\mathbf{w} \circ \Delta \mathbf{X}\|_2^2 = \sigma_n^2 \|\mathbf{w} \circ \Delta \zeta\|_2^2 + 2\sigma_n (\mathbf{w} \circ \Delta \zeta)^\top (\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})) + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2. \quad (\text{C.33})$$

For the first term, we have

$$\begin{aligned}
\mathbb{E} \left[\sigma_n^2 \|\mathbf{w} \circ \Delta \boldsymbol{\zeta}\|_2^2 \mid \mathbf{Y}, \mathbf{Z} \right] &= 2\sigma_n^2 \|\mathbf{w}\|_2^2, \\
\text{Var} \left(\sigma_n^2 \|\mathbf{w} \circ \Delta \boldsymbol{\zeta}\|_2^2 \mid \mathbf{Y}, \mathbf{Z} \right) &\leq \sum_{\ell=1}^L w_\ell^4 \sigma_n^4 \mathbb{E}_{\zeta, \zeta' \sim P_\zeta} \left[|\zeta - \zeta'|^4 \right] \\
&\leq 16\rho_4^4 \sigma_n^4 \|\mathbf{w}\|_4^4 \leq 16\rho_4^4 \sigma_n^4 \|\mathbf{w}\|_2^2 \|\mathbf{w}\|_\infty^2 \leq 16\rho_4^4 \sigma_n^4 \|\mathbf{w}\|_2^4 \cdot \epsilon_1^2,
\end{aligned}$$

and therefore by Chebyshev's inequality, for every $A > 0$,

$$\mathbb{P} \left\{ \left| \sigma_n^2 \|\mathbf{w} \circ \Delta \boldsymbol{\zeta}\|_2^2 - 2\sigma_n^2 \|\mathbf{w}\|_2^2 \right| \geq A \mid \mathbf{Y}, \mathbf{Z} \right\} \leq \frac{16\rho_4^4 \sigma_n^4 \|\mathbf{w}\|_2^4 \cdot \epsilon_1^2}{A^2}.$$

For the second term in (C.33), we have that

$$\mathbb{E} \left[2\sigma_n (\mathbf{w} \circ \Delta \boldsymbol{\zeta})^\top (\mathbf{w} \circ \Delta \mu(\mathbf{Y}, \mathbf{Z})) \mid \mathbf{Y}, \mathbf{Z} \right] = 0$$

and moreover, by Hölder's inequality,

$$\begin{aligned}
&\text{Var} \left(2\sigma_n (\mathbf{w} \circ \Delta \boldsymbol{\zeta})^\top (\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})) \mid \mathbf{Y}, \mathbf{Z} \right) \\
&= 4 \sum_{\ell=1}^L w_\ell^2 \sigma_n^2 \text{Var} (\Delta_\ell \boldsymbol{\zeta} \mid \mathbf{Y}, \mathbf{Z}) \cdot (w_\ell \Delta_\ell \mu_n(\mathbf{Y}, \mathbf{Z}))^2 \\
&= 8\sigma_n^2 \sum_{\ell=1}^L w_\ell^2 (w_\ell \Delta_\ell \mu_n(\mathbf{Y}, \mathbf{Z}))^2 \\
&\leq 8\sigma_n^2 \|\mathbf{w}\|_6^2 \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_3^2 \\
&\leq 8\sigma_n^2 \left(\|\mathbf{w}\|_2^2 \|\mathbf{w}\|_\infty^4 \right)^{1/3} \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_3^2 \\
&= 8\sigma_n^2 \|\mathbf{w}\|_2^2 \left(\sigma_n^3 \|\mathbf{w}\|_2^3 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^3 \right)^{2/3} \cdot (\epsilon_1^2 \epsilon_2)^{2/3} \\
&\leq 8\sigma_n^2 \|\mathbf{w}\|_2^2 \left(\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2 \right) \cdot (\epsilon_1^2 \epsilon_2)^{2/3},
\end{aligned}$$

where the penultimate step uses the definitions of ϵ_1^2 and ϵ_2 . Another application of Chebyshev's inequality then gives us for every $B > 0$ that

$$\begin{aligned} \mathbb{P} \left\{ \left| 2\sigma_n(\mathbf{w} \circ \Delta\boldsymbol{\zeta})^\top (\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})) \right| \geq B \mid \mathbf{Y}, \mathbf{Z} \right\} \\ \leq \frac{8\sigma_n^2 \|\mathbf{w}\|_2^2 (\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2) \cdot (\epsilon_1^2 \epsilon_2)^{2/3}}{B^2}. \end{aligned}$$

Together, then,

$$\begin{aligned} \mathbb{P} \left\{ \left| \|\mathbf{w} \circ \Delta\mathbf{X}\|_2^2 - (2\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2) \right| \geq A + B \mid \mathbf{Y}, \mathbf{Z} \right\} \\ \leq \frac{16\sigma_n^4 \rho_4^4 \|\mathbf{w}\|_2^4 \cdot \epsilon_1^2}{A^2} + \frac{8\sigma_n^2 \|\mathbf{w}\|_2^2 (\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2) \cdot (\epsilon_1^2 \epsilon_2)^{2/3}}{B^2}. \end{aligned}$$

Next, we choose

$$\begin{aligned} A^2 &= \frac{3}{\delta} \cdot 16\rho_4^4 \sigma_n^4 \|\mathbf{w}\|_2^4 \cdot \epsilon_1^2, \quad \text{and} \\ B^2 &= \frac{3}{\delta} \cdot 8\sigma_n^2 \|\mathbf{w}\|_2^2 \left(\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2 \right) \cdot (\epsilon_1^2 \epsilon_2)^{2/3}. \end{aligned}$$

Then

$$\mathbb{P} \left\{ \frac{\|\mathbf{w} \circ \Delta\mathbf{X}\|_2^2}{(2\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2)} \in [(1 - \epsilon_*) \vee 0, 1 + \epsilon_*] \mid \mathbf{Y}, \mathbf{Z} \right\} \geq 1 - \frac{2\delta}{3}.$$

Next we bound $\|\mathbf{w} \circ \Delta\mathbf{X}\|_3^3$. We start by writing

$$\begin{aligned} \|\mathbf{w} \circ \Delta\mathbf{X}\|_3 &\leq \sigma_n \|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_3 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_3 \\ &= \sigma_n \|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_3 + \left(\sigma_n^3 \|\mathbf{w}\|_2^3 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2^3 \right)^{1/3} \cdot \epsilon_2^{1/3} \\ &\leq \sigma_n \|\mathbf{w} \circ \Delta\boldsymbol{\zeta}\|_3 + (\sigma_n \|\mathbf{w}\|_2 + \|\mathbf{w} \circ \Delta\mu_n(\mathbf{Y}, \mathbf{Z})\|_2) \cdot \epsilon_2^{1/3}. \end{aligned}$$

Further, we have that

$$\begin{aligned}\mathbb{E} \left[\|\mathbf{w} \circ \Delta \boldsymbol{\zeta}\|_3^3 \mid \mathbf{Y}, \mathbf{Z} \right] &= \sum_{\ell=1}^L w_\ell^3 \mathbb{E} \left[|\Delta_\ell \boldsymbol{\zeta}|^3 \mid \mathbf{Y}, \mathbf{Z} \right] \\ &\leq \|\mathbf{w}\|_3^3 \cdot \mathbb{E}_{\zeta, \zeta' \sim P_\zeta^{\text{iid}}} \left[|\zeta - \zeta'|^3 \right] \leq \|\mathbf{w}\|_3^3 \cdot 8\rho_3^3\end{aligned}$$

where $\rho_3 = \mathbb{E}_{\zeta \sim P_\zeta} [|\zeta|^3]^{1/3}$. Since $\rho_3 \leq \rho_4$, we further have that

$$\mathbb{E} \left[\|\mathbf{w} \circ \Delta \boldsymbol{\zeta}\|_3^3 \mid \mathbf{Y}, \mathbf{Z} \right] \leq 8\rho_4^3 \|\mathbf{w}\|_2^2 \cdot \|\mathbf{w}\|_\infty \leq 8\rho_4^3 \|\mathbf{w}\|_2^3 \cdot \epsilon_1.$$

By Markov's inequality, then,

$$\mathbb{P} \left\{ \sigma_n^3 \|\mathbf{w} \circ \Delta \boldsymbol{\zeta}\|_3^3 \leq \frac{24\sigma_n^3 \rho_4^3 \|\mathbf{w}\|_2^3 \cdot \epsilon_1}{\delta} \mid \mathbf{Y}, \mathbf{Z} \right\} \geq 1 - \frac{\delta}{3}.$$

Together, then, we have

$$\mathbb{P} \left\{ \|\mathbf{w} \circ \Delta \mathbf{X}\|_3 \leq \frac{\sqrt[3]{24}\sigma_n \rho_4 \|\mathbf{w}\|_2 \epsilon_1^{1/3}}{\delta^{1/3}} + (\sigma_n \|\mathbf{w}\|_2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2) \epsilon_2^{1/3} \mid \mathbf{Y}, \mathbf{Z} \right\} \geq 1 - \frac{\delta}{3},$$

which we can relax to

$$\begin{aligned}\mathbb{P} \left\{ \|\mathbf{w} \circ \Delta \mathbf{X}\|_3^2 \leq \frac{2(24)^{2/3} \sigma_n^2 \rho_4^2 \|\mathbf{w}\|_2^2 \epsilon_1^{2/3}}{\delta^{2/3}} + 4(\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2) \epsilon_2^{2/3} \mid \mathbf{Y}, \mathbf{Z} \right\} \\ \geq 1 - \frac{\delta}{3}.\end{aligned}$$

Thus

$$\mathbb{P} \left\{ \|\mathbf{w} \circ \Delta \mathbf{X}\|_3^2 \leq \left(2\sigma_n^2 \|\mathbf{w}\|_2^2 + \|\mathbf{w} \circ \Delta \mu_n(\mathbf{Y}, \mathbf{Z})\|_2^2 \right) \cdot \epsilon'_* \mid \mathbf{Y}, \mathbf{Z} \right\} \geq 1 - \frac{\delta}{3}$$

as required. $\square$

Lemma C.8. Consider any real-valued sequence $(x_n)_{n \in \mathbb{N}}$, and $y \in \mathbb{R}$. If $\epsilon_n \xrightarrow{P} 0$ and $\epsilon'_n \xrightarrow{P} 0$ as $n \rightarrow \infty$, then

$$|\Phi(x_n(1 + \epsilon_n) + y + \epsilon'_n) - \Phi(x_n + y)| \xrightarrow{P} 0.$$

Proof. First, since Φ is $\frac{1}{\sqrt{2\pi}}$ -Lipschitz, we have that

$$\begin{aligned} |\Phi(x_n(1 + \epsilon_n) + y + \epsilon'_n) - \Phi((x_n + y) \cdot (1 + \epsilon_n))| &\leq \frac{1}{\sqrt{2\pi}} |\epsilon'_n - y\epsilon_n| \\ &\leq \frac{|\epsilon'_n| + |y||\epsilon_n|}{\sqrt{2\pi}} = o_P(1). \end{aligned}$$

Hence, it suffices to show that

$$|\Phi((x_n + y) \cdot (1 + \epsilon_n)) - \Phi(x_n + y)| \xrightarrow{P} 0.$$

In fact, without loss of generality, we may take $y = 0$. By the mean value theorem, there exists a real-valued sequence of $(\bar{x}_n)_{n \in \mathbb{N}}$, where $\bar{x}_n$ lies between x_n and $x_n(1 + \epsilon_n)$, such that

$$|\Phi(x_n(1 + \epsilon_n)) - \Phi(x_n)| = \phi(\bar{x}_n) \cdot |x_n \epsilon_n| = \phi(|\bar{x}_n|) |x_n| |\epsilon_n|.$$

Define the event $A_n := \{|\bar{x}_n| \geq |x_n|/2\}$ and note that $\mathbb{P}(A_n) \geq \mathbb{P}(|\epsilon_n| \leq 1/2)$. Since $\epsilon_n \xrightarrow{P} 0$ as $n \rightarrow \infty$, we have $\mathbb{P}(A_n) \rightarrow 1$ as $n \rightarrow \infty$. Moreover, writing $K := \sup_{x \in \mathbb{R}} |x| \exp(-x^2/8) \in (0, \infty)$, we have on A_n that,

$$\phi(|\bar{x}_n|) \cdot |x_n| \leq \exp(-x_n^2/8) \cdot |x_n| \leq K.$$

Finally fix $\delta > 0$ and observe that

$$\begin{aligned}
\mathbb{P}\{|\Phi(x_n(1 + \epsilon_n)) - \Phi(x_n)| \geq \delta\} &\leq \mathbb{P}\left[\{|\Phi(x_n(1 + \epsilon_n)) - \Phi(x_n)| \geq \delta\} \cap A_n\right] + \mathbb{P}(A_n^c) \\
&\leq \mathbb{P}\left[\{|\epsilon_n| \geq \delta/K\} \cap A_n\right] + \mathbb{P}(A_n^c) \\
&\leq \mathbb{P}(|\epsilon_n| \geq \delta/K) + \mathbb{P}(A_n^c) \rightarrow 0,
\end{aligned}$$

as $n \rightarrow \infty$. Since $\delta > 0$ was arbitrary, the result follows. $\square$

Concentration results for neighbor matching

We now prove a concentration result for the neighbor matching in the asymptotic regime from Section 4.5.3. Let $(Y_1, Z_1), (Y_2, Z_2), \dots \stackrel{\text{iid}}{\sim} P_{Y,Z}$, where $P_{(Y,Z)}$ is a distribution on $[-1, 1] \times \mathbb{R}$, and let $Z_{(n,1)} \leq \dots \leq Z_{(n,n)}$ be the order statistics of $Z_1, \dots, Z_n$. Abusing notation, we will write $Y_{(n,i)}$ to denote the Y value corresponding to $Z_{(n,i)}$ —that is,

$$(Y_{(n,1)}, \dots, Y_{(n,n)}) \mid (Z_{(n,1)}, \dots, Z_{(n,n)}) \sim P_{Y|Z}(\cdot \mid Z_{(n,1)}) \times \dots \times P_{Y|Z}(\cdot \mid Z_{(n,n)}).$$

Then for neighbor matching at sample size n , we are interested in the quantity

$$\text{NM}_n := \frac{2}{n} \sum_{i=1}^{\lfloor n/2 \rfloor} (Y_{(n,2i-1)} - Y_{(n,2i)})_+^2.$$

In the following result, we show that this quantity concentrates around the mean of the conditional variance of $P_{Y|Z}$:

Proposition C.9. *Under the assumptions and definitions stated above,*

$$|\text{NM}_n - \mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)]| = o_P(1).$$

Proof. The proof is split into four key steps.

Step 1: a concentration step. First, we note that conditional on $Z_1, \dots, Z_n$, the sum $\sum_{i=1}^{\lfloor n/2 \rfloor} (Y_{(n,2i-1)} - Y_{(n,2i)})_+^2$ is the sum of $\lfloor n/2 \rfloor$ independent terms, with each summand lying in $[0, 4]$. Hence, for any $\delta > 0$, by Hoeffding's inequality we have for $n \geq 2$ that

$$\mathbb{P} \left\{ \left| \text{NM}_n - \mathbb{E} [\text{NM}_n \mid Z_1, \dots, Z_n] \right| \geq 4 \cdot \frac{2}{n} \sqrt{\frac{\lfloor n/2 \rfloor \cdot \log(2/\delta)}{2}} \mid \mathbf{Z} \right\} \leq \delta.$$

In particular,

$$\left| \text{NM}_n - \mathbb{E} [\text{NM}_n \mid Z_1, \dots, Z_n] \right| = O_P(n^{-1/2}). \quad (\text{C.34})$$

Step 2: comparing to i.i.d. pairs. Since $(y_1, y_2) \mapsto (y_1 - y_2)_+^2$ is 4-Lipschitz with respect to the ℓ_1 norm on $[-1, 1]^2$, for any two distributions Q, Q' on $[-1, 1]^2$ we have

$$\left| \mathbb{E}_{(Y, Y') \sim Q} \left[(Y - Y')_+^2 \right] - \mathbb{E}_{(Y, Y') \sim Q'} \left[(Y - Y')_+^2 \right] \right| \leq 4 \text{d}_W(Q, Q'),$$

where $\text{d}_W(\cdot, \cdot)$ is the L_1 -Wasserstein distance between distributions. In particular, in the special case that $Q = Q_1 \times Q_2$ and $Q' = Q'_1 \times Q'_2$ are both product distributions, then we have

$$\left| \mathbb{E}_{(Y, Y') \sim Q_1 \times Q_2} \left[(Y - Y')_+^2 \right] - \mathbb{E}_{(Y, Y') \sim Q'_1 \times Q'_2} \left[(Y - Y')_+^2 \right] \right| \leq 4(\text{d}_W(Q_1, Q'_1) + \text{d}_W(Q_2, Q'_2)).$$

Now, fix any $i \in \{1, \dots, \lfloor n/2 \rfloor\}$ and observe that

$$(Y_{(n,2i-1)}, Y_{(n,2i)}) \mid (Z_1, \dots, Z_n) \sim P_{Y|Z}(\cdot \mid Z_{(n,2i-1)}) \times P_{Y|Z}(\cdot \mid Z_{(n,2i)}).$$

Hence, for any $z \in \mathbb{R}$,

$$\begin{aligned}
& \left| \mathbb{E} \left[\left(Y_{(n,2i-1)} - Y_{(n,2i)} \right)_+^2 \mid Z_1, \dots, Z_n \right] - \text{Var}_{Y \sim P_{Y|Z(\cdot|z)}}(Y) \right| \\
&= \left| \mathbb{E} \left[\left(Y_{(n,2i-1)} - Y_{(n,2i)} \right)_+^2 \mid Z_1, \dots, Z_n \right] - \mathbb{E}_{Y, Y' \stackrel{\text{iid}}{\sim} P_{Y|Z(\cdot|z)}} \left[(Y - Y')_+^2 \right] \right| \\
&\leq 4 \{ \text{dW} \left(P_{Y|Z(\cdot|Z_{(n,2i-1)})}, P_{Y|Z(\cdot|z)} \right) + \text{dW} \left(P_{Y|Z(\cdot|Z_{(n,2i)})}, P_{Y|Z(\cdot|z)} \right) \}.
\end{aligned}$$

Now let $Z'_1, Z'_2, \dots \stackrel{\text{iid}}{\sim} P_Z$ be drawn independently of the data. Then, applying the calculations above with $z = Z'_{(\lfloor n/2 \rfloor, i)}$,

$$\begin{aligned}
& \left| \mathbb{E} \left[\left(Y_{(n,2i-1)} - Y_{(n,2i)} \right)_+^2 \mid Z_1, \dots, Z_n \right] - \text{Var}_{Y \sim P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})}}(Y) \right| \\
&\leq 4 \{ \text{dW} \left(P_{Y|Z(\cdot|Z_{(n,2i-1)})}, P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})} \right) + \text{dW} \left(P_{Y|Z(\cdot|Z_{(n,2i)})}, P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})} \right) \}.
\end{aligned}$$

For $n \geq 3$, summing the inequality above over $i = 1, \dots, \lfloor n/2 \rfloor$, we obtain

$$\begin{aligned}
& \left| \mathbb{E} [\text{NM}_n \mid Z_1, \dots, Z_n] - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \text{Var}_{Y \sim P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})}}(Y) \right| \\
&\leq \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left| \mathbb{E} \left[\left(Y_{(n,2i-1)} - Y_{(n,2i)} \right)_+^2 \mid Z_1, \dots, Z_n \right] - \text{Var}_{Y \sim P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})}}(Y) \right| + \frac{4}{n} \\
&\leq \frac{8}{n-2} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{j=2i-1}^{2i} \text{dW} (P_{Y|Z(\cdot|Z_{(n,j)})}, P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})}) + \frac{4}{n}. \tag{C.35}
\end{aligned}$$

Step 3: concentration around the mean of conditional variance. Note that

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \text{Var}_{Y \sim P_{Y|Z(\cdot|Z'_{(\lfloor n/2 \rfloor, i)})}}(Y) = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \text{Var}_{Y \sim P_{Y|Z(\cdot|Z'_i)}}(Y),$$

since by definition, $Z'_{(\lfloor n/2 \rfloor, 1)}, \dots, Z'_{(\lfloor n/2 \rfloor, \lfloor n/2 \rfloor)}$ is simply a permutation of $Z'_1, \dots, Z'_{\lfloor n/2 \rfloor}$. Moreover, since $Z'_1, Z'_2, \dots \stackrel{\text{iid}}{\sim} P_Z$, the random variables $\text{Var}_{Y \sim P_{Y|Z}(\cdot | Z'_i)}(Y)$ are independent and identically distributed, with mean $\mathbb{E}[\text{Var}(P_{Y|Z})]$ and variance at most 1. Therefore, by the central limit theorem, we have

$$\left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \text{Var}_{Y \sim P_{Y|Z}(\cdot | Z'_i)}(Y) - \mathbb{E}[\text{Var}_{Y \sim P_{Y|Z}}(Y)] \right| = O_P(n^{-1/2}).$$

Combining this with (C.34) and (C.35), we have

$$\begin{aligned} & \left| \text{NM}_n - \mathbb{E}[\text{Var}(P_{Y|Z})] \right| \\ & \leq \frac{8}{n-2} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{j=2i-1}^{2i} \text{dW}(P_{Y|Z}(\cdot | Z_{(n,j)}), P_{Y|Z}(\cdot | Z'_{(\lfloor n/2 \rfloor, i)})) + O_P(n^{-1/2}). \end{aligned}$$

Step 4: the Wasserstein distance. Finally, by Lemma C.15 (applied with $K_n = \lfloor n/2 \rfloor$, and $B_{n,k} = \{2k-1, 2k\}$ for $k = 1, \dots, K_n$), we have

$$\frac{8}{n-2} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{j=2i-1}^{2i} \text{dW}(P_{Y|Z}(\cdot | Z_{(n,j)}), P_{Y|Z}(\cdot | Z'_{(\lfloor n/2 \rfloor, i)})) = o_P(1).$$

This completes the proof. □

Concentration results for cross-bin matching

The aim of this subsection is to prove a concentration result for cross-bin matching under the asymptotic framework of Section 4.5.3, analogous to Proposition C.9 for neighbor matching in Section C.2.7. To state and prove the result, we first introduce some more notation for

cross-bin matching at sample size n with K_n bins. Let $m_n = \lfloor n/K_n \rfloor$, and let

$$A_{n,1} = \{1, \dots, m_n\}, \dots, A_{n,K_n} = \{(K_n - 1)m_n + 1, \dots, K_n m_n\}$$

be an uniform partition of $\{1, \dots, K_n m_n\}$ into K_n bins. Given $y, y' \in [-1, 1]^{m_n}$, let $y_{(1)} \leq \dots \leq y_{(m_n)}$ and $y'_{(1)} \leq \dots \leq y'_{(m_n)}$ denote the sorted values of y and y' , and define $f_n : [-1, 1]^{m_n} \times [-1, 1]^{m_n} \rightarrow [0, \infty)$ by

$$f_n(y, y') = \sum_{i=1}^{\lfloor m_n/2 \rfloor} (y_{(m_n+1-i)} - y'_{(i)})_+^2. \quad (\text{C.36})$$

Building on the notation defined at the beginning of Section C.2.7, for any $k \in [K_n]$ we define $Y_{(n, A_{n,k})} := (Y_{(n, (k-1)m_n+1)}, \dots, Y_{(n, km_n)})$. Our quantity of interest can then be written as

$$\text{CBM}_n := \frac{2}{n} \|\Delta \mathbf{Y}^+\|_2^2 = \frac{2}{n} \sum_{k=1}^{K_n-1} f_n(Y_{(n, A_{n,k})}, Y_{(n, A_{n,k+1})}).$$

Similar to neighbor matching, the following proposition establishes concentration of the aforementioned quantity around the mean conditional deviance:

Proposition C.10. *Under the assumptions and definitions stated above, if $K_n \rightarrow \infty$ and $n/K_n \rightarrow \infty$, then*

$$\left| \text{CBM}_n - \mathbb{E} \left[\text{Dev}(P_{Y|Z}) \right] \right| = o_P(1).$$

Proof of Proposition C.10. Similar to the concentration result for neighbor matching, this proof is also split into four key steps.

Step 1: a concentration step. Our first step is a concentration result: we claim that for any $\delta \in (0, 1]$, we have

$$\mathbb{P}\left\{\left|\text{CBM}_n - \mathbb{E}[\text{CBM}_n \mid \mathbf{Z}]\right| \geq \sqrt{\frac{4 \log(4/\delta)}{K_n}} \mid \mathbf{Z}\right\} \leq \delta. \quad (\text{C.37})$$

To prove this, we start by writing

$$\text{CBM}_n = \frac{2}{n} \sum_{k=1}^{\lfloor K_n/2 \rfloor} f_n(Y_{(n,A_n,2k-1)}, Y_{(n,A_n,2k)}) + \frac{2}{n} \sum_{k=1}^{\lfloor (K_n-1)/2 \rfloor} f_n(Y_{(n,A_n,2k)}, Y_{(n,A_n,2k+1)}).$$

After conditioning on $\mathbf{Z}$, the first term is a sum of $\lfloor K_n/2 \rfloor$ independent terms because the sets $\{A_{n,k} : k \in [K_n]\}$ are pairwise disjoint. Furthermore, each term $2f_n(Y_{(n,A_n,2k-1)}, Y_{(n,A_n,2k)})/n$ takes values in $[0, 8 \cdot \lfloor m_n/2 \rfloor/n] \subseteq [0, 4/K_n]$. Therefore by Hoeffding's inequality,

$$\mathbb{P}\left\{\left|\frac{2}{n} \sum_{k=1}^{\lfloor K_n/2 \rfloor} f_n(Y_{(n,A_n,2k-1)}, Y_{(n,A_n,2k)}) - \frac{2}{n} \sum_{k=1}^{\lfloor K_n/2 \rfloor} \mathbb{E}\left[f_n(Y_{(n,A_n,2k-1)}, Y_{(n,A_n,2k)}) \mid \mathbf{Z}\right]\right| \geq \frac{4}{K_n} \sqrt{\frac{\lfloor K_n/2 \rfloor \cdot \log(4/\delta)}{2}} \mid \mathbf{Z}\right\} \leq \frac{\delta}{2}.$$

We can prove a similar concentration for the second term, and together with a union bound this establishes (C.37). It follows that

$$|\text{CBM}_n - \mathbb{E}[\text{CBM}_n \mid \mathbf{Z}]| = O_P(K_n^{-1/2}).$$

Step 2: comparing to i.i.d. terms. Since $(y, y') \mapsto f_n(y, y')$ is 4-Lipschitz with respect to the ℓ_1 norm on $[-1, 1]^{mn} \times [-1, 1]^{mn}$ (see Lemma C.17), we have for any distributions Q, Q' on $[-1, 1]^{mn} \times [-1, 1]^{mn}$ that

$$|\mathbb{E}_{(Y, Y') \sim Q}[f_n(Y, Y')] - \mathbb{E}_{(\tilde{Y}, \tilde{Y}') \sim Q'}[f_n(\tilde{Y}, \tilde{Y}')]| \leq 4\text{d}_W(Q, Q'),$$

where $d_W(\cdot, \cdot)$ is the L_1 -Wasserstein distance between distributions. In particular, in the special case where $Q = Q_1 \times \cdots \times Q_{2m_n}$ and $Q' = Q'_1 \times \cdots \times Q'_{2m_n}$, we have

$$|\mathbb{E}_{(Y, Y') \sim Q} [f_n(Y, Y')] - \mathbb{E}_{(\tilde{Y}, \tilde{Y}') \sim Q'} [f_n(\tilde{Y}, \tilde{Y}')] | \leq 4 \sum_{j=1}^{2m_n} d_W(Q_j, Q'_j).$$

Next, fix any $k \in [K_n - 1]$. By construction, we have

$$(Y_{(n, A_{n, k})}, Y_{(n, A_{n, k+1})}) \mid \mathbf{Z} \sim P_{Y|Z}(\cdot \mid Z_{(n, A_{n, k})}) \times P_{Y|Z}(\cdot \mid Z_{(n, A_{n, k+1})}),$$

where, abusing notation, we write

$$P_{Y|Z}(\cdot \mid Z_{(n, A_{n, k})}) = P_{Y|Z}(\cdot \mid Z_{(n, (k-1)m_n+1)}) \times \cdots \times P_{Y|Z}(\cdot \mid Z_{(n, km_n)}),$$

Now let $Z'_1, Z'_2, \dots \stackrel{\text{iid}}{\sim} P_Z$ be drawn independently of the data. Then

$$\begin{aligned} & \left| \mathbb{E}[\text{CBM}_n \mid \mathbf{Z}] - \frac{2}{n} \sum_{k=1}^{K_n-1} \mathbb{E}_{Y, Y' \stackrel{\text{iid}}{\sim} (P_{Y|Z}(\cdot \mid Z'_{(K_n, k)}))^{m_n}} [f_n(Y, Y')] \right| \\ & \leq \frac{2}{n} \sum_{k=1}^{K_n-1} \left| \mathbb{E} \left[f_n \left(Y_{(n, A_{n, k})}, Y_{(n, A_{n, k+1})} \right) \mid \mathbf{Z} \right] - \mathbb{E}_{Y, Y' \stackrel{\text{iid}}{\sim} (P_{Y|Z}(\cdot \mid Z'_{(K_n, k)}))^{m_n}} [f_n(Y, Y')] \right| \\ & \leq \frac{8}{n} \sum_{k=1}^{K_n-1} \sum_{i \in A_{n, k} \cup A_{n, k+1}} d_W \left(P_{Y|Z}(\cdot \mid Z_{(n, i)}), P_{Y|Z}(\cdot \mid Z'_{(K_n, k)}) \right). \end{aligned}$$

Step 3: concentration around the mean of conditional deviance. By Lemma C.18, there exists a universal constant $C > 0$ such that

$$\begin{aligned}
& \left| \frac{2}{n} \sum_{k=1}^{K_n-1} \mathbb{E}_{Y, Y' \stackrel{\text{iid}}{\sim} (P_{Y|Z}(\cdot | Z'_{(K_n, k)}))^{m_n}} [f_n(Y, Y')] - \frac{m_n}{n} \sum_{k=1}^{K_n} \text{Dev} \left(P_{Y|Z}(\cdot | Z'_k) \right) \right| \\
&= \left| \frac{2}{n} \sum_{k=1}^{K_n-1} \mathbb{E}_{Y, Y' \stackrel{\text{iid}}{\sim} (P_{Y|Z}(\cdot | Z'_{(K_n, k)}))^{m_n}} [f_n(Y, Y')] - \frac{m_n}{n} \sum_{k=1}^{K_n} \text{Dev} \left(P_{Y|Z}(\cdot | Z'_{(K_n, k)}) \right) \right| \\
&\leq \sum_{i=1}^{K_n-1} \left| \frac{2}{n} \mathbb{E}_{Y, Y' \stackrel{\text{iid}}{\sim} (P_{Y|Z}(\cdot | Z'_{(K_n, k)}))^{m_n}} [f_n(Y, Y')] - \frac{m_n}{n} \text{Dev} \left(P_{Y|Z}(\cdot | Z'_{(K_n, k)}) \right) \right| \\
&\quad + \frac{m_n}{n} |\text{Dev} \left(P_{Y|Z}(\cdot | Z'_{(K_n, K_n)}) \right)| \\
&\leq (K_n - 1) \cdot \frac{m_n}{n} \cdot \frac{C}{\sqrt{m_n}} + \frac{4m_n}{n} \leq \frac{C}{\sqrt{m_n}} + \frac{4m_n}{n}.
\end{aligned}$$

Moreover, since $1/(K_n + 1) \leq m_n/n \leq 1/K_n$, since $Z'_1, \dots, Z'_{K_n} \stackrel{\text{iid}}{\sim} P_Z$, we have by the central limit theorem that

$$\begin{aligned}
& \left| \frac{m_n}{n} \sum_{k=1}^{K_n} \text{Dev} \left(P_{Y|Z}(\cdot | Z'_k) \right) - \mathbb{E} [\text{Dev}(P_{Y|Z})] \right| \\
&\leq \left(\frac{1}{K_n} - \frac{m_n}{n} \right) 4K_n + \left| \frac{1}{K_n} \sum_{k=1}^{K_n} \text{Dev} \left(P_{Y|Z}(\cdot | Z'_k) \right) - \mathbb{E} [\text{Dev}(P_{Y|Z})] \right| = O_P(K_n^{-1/2}).
\end{aligned}$$

Combining these bounds with our conclusions from Steps 1 and 2, then,

$$\begin{aligned}
& \left| \text{CBM}_n - \mathbb{E} [\text{Dev}(P_{Y|Z})] \right| \leq \frac{8}{n} \sum_{k=1}^{K_n-1} \sum_{i \in A_{n, k}} \text{dW} \left(P_{Y|Z}(\cdot | Z_{(n, i)}), P_{Y|Z}(\cdot | Z'_{(K_n, k)}) \right) \\
&+ \frac{8}{n} \sum_{k=1}^{K_n-1} \sum_{i \in A_{n, k+1}} \text{dW} \left(P_{Y|Z}(\cdot | Z_{(n, i)}), P_{Y|Z}(\cdot | Z'_{(K_n, k)}) \right) + O_P(K_n^{-1/2}) + O(m_n^{-1/2}).
\end{aligned}$$

Step 4: the Wasserstein distance. Finally, we bound the remaining terms. First,

$$\frac{8}{n} \sum_{k=1}^{K_n-1} \sum_{i \in A_{n,k}} d_W \left(P_{Y|Z}(\cdot | Z_{(n,i)}), P_{Y|Z}(\cdot | Z'_{(K_n,k)}) \right) = o_P(1),$$

by Lemma C.15 below (applied with sets $B_{n,k} = A_{n,k}$ for $k \in [K_n - 1]$, and $B_{n,K_n} = \emptyset$).

Similarly,

$$\frac{8}{n} \sum_{k=1}^{K_n-1} \sum_{i \in A_{n,k+1}} d_W \left(P_{Y|Z}(\cdot | Z_{(n,i)}), P_{Y|Z}(\cdot | Z'_{(K_n,k)}) \right) = o_P(1),$$

by Lemma C.15 again (applied with sets $B_{n,k} = A_{n,k+1}$ for $k \in [K_n - 1]$, and $B_{n,K_n} = \emptyset$).

This completes the proof. $\square$

C.3 Auxiliary lemmas

Lemma C.11. *It holds that deterministically, for any $\alpha \in [0, 1]$,*

$$\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1} \{p(\mathbf{x}^{\mathbf{s}}) \leq \alpha\} \leq \alpha,$$

where $p(\cdot)$ is as defined in (C.1).

Proof. Consider a bijection $\sigma : [2^L] \rightarrow \{-1, 1\}^L$ such that $T(\mathbf{x}^{\sigma(1)}) \geq \dots \geq T(\mathbf{x}^{\sigma(2^L)})$, and let $r \in \{0, 1, \dots, 2^L\}$ be such that $\alpha \in [r/2^L, (r+1)/2^L)$. Then, since we have

$\sum_{k=1}^{2^L} \mathbb{1} \left\{ T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)}) \right\} \in [2^L]$ for each $j \in [2^L]$, we have deterministically that

$$\begin{aligned}
\frac{1}{2^L} \sum_{\mathbf{s} \in \{\pm 1\}^L} \mathbb{1} \{p(\mathbf{x}^{\mathbf{s}}) \leq \alpha\} &= \frac{1}{2^L} \sum_{j=1}^{2^L} \mathbb{1} \left\{ p(\mathbf{x}^{\sigma(j)}) \leq \alpha \right\} \\
&= \frac{1}{2^L} \sum_{j=1}^{2^L} \mathbb{1} \left\{ \frac{1}{2^L} \sum_{k=1}^{2^L} \mathbb{1} \left\{ T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)}) \right\} \leq \alpha \right\} \\
&= \frac{1}{2^L} \sum_{j=1}^{2^L} \mathbb{1} \left\{ \sum_{k=1}^{2^L} \mathbb{1} \left\{ T(\mathbf{x}^{\sigma(k)}) \geq T(\mathbf{x}^{\sigma(j)}) \right\} \leq r \right\} \\
&\leq \frac{1}{2^L} \sum_{j=1}^{2^L} \mathbb{1} \left\{ \sum_{k=1}^{2^L} \mathbb{1} \{k \leq j\} \leq r \right\} = \frac{1}{2^L} \sum_{j=1}^{2^L} \mathbb{1} \{j \leq r\} = \frac{r}{2^L} \leq \alpha,
\end{aligned}$$

as required. $\square$

Lemma C.12. *For any $p, q \in (0, 1)$, we have*

$$H^2(\text{Ber}(p), \text{Ber}(q)) \leq \frac{2(p-q)^2}{(p+q)(2-p-q)} \leq \min \left\{ \frac{(p-q)^2}{p(1-p)}, \frac{(p-q)^2}{q(1-q)} \right\}.$$

Proof. For any $p, q \in (0, 1)$,

$$\begin{aligned}
H^2(\text{Ber}(p), \text{Ber}(q)) &= (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2 \\
&= \frac{(p-q)^2}{(\sqrt{p} + \sqrt{q})^2} + \frac{(p-q)^2}{(\sqrt{1-p} + \sqrt{1-q})^2} \\
&\leq \frac{(p-q)^2}{p} + \frac{(p-q)^2}{(1-p)} = \frac{(p-q)^2}{p(1-p)}.
\end{aligned}$$

By symmetry, the bound $\frac{2(p-q)^2}{(p+q)(2-p-q)} \leq \frac{(p-q)^2}{q(1-q)}$ also holds. $\square$

Lemma C.13. *For any distribution P on $\mathbb{R}$,*

$$2 \operatorname{Var}_{X \sim P}(X) \leq \operatorname{Dev}(P) \leq 4 \operatorname{Var}_{X \sim P}(X),$$

where $\operatorname{Dev}(\cdot)$ is the deviance of a distribution, defined in (C.20). Moreover, if P is symmetric around its mean, then $\operatorname{Dev}(P) = 4 \operatorname{Var}_{X \sim P}(X)$.

Proof. Let F^{-1} denote the quantile function of P (i.e., the generalized inverse of its CDF).

For any $q \in (0, 1)$, by monotonicity of F , we have

$$\{F^{-1}(1 - q) - F^{-1}(1/2)\} \cdot \{F^{-1}(1/2) - F^{-1}(q)\} \geq 0.$$

Therefore,

$$\begin{aligned} \operatorname{Dev}(P) &= \mathbb{E}_{q \sim \operatorname{Unif}(0,1)} \left[(F^{-1}(1 - q) - F^{-1}(1/2) + F^{-1}(1/2) - F^{-1}(q))^2 \right] \\ &\geq \mathbb{E}_{q \sim \operatorname{Unif}(0,1)} \left[(F^{-1}(1 - q) - F^{-1}(1/2))^2 + (F^{-1}(1/2) - F^{-1}(q))^2 \right] \\ &= 2 \mathbb{E}_{q \sim \operatorname{Unif}(0,1)} \left[(F^{-1}(q) - F^{-1}(1/2))^2 \right] \geq 2 \operatorname{Var}_{q \sim \operatorname{Unif}(0,1)} (F^{-1}(q)) = 2 \operatorname{Var}_{X \sim P}(X), \end{aligned}$$

since when $q \sim \operatorname{Unif}(0, 1)$ we have $F^{-1}(q) \sim P$ by construction. For the upper bound, writing $\mu = \mathbb{E}_{X \sim P}[X]$,

$$\begin{aligned} \operatorname{Dev}(P) &= \mathbb{E}_{q \sim \operatorname{Unif}(0,1)} \left[(F^{-1}(1 - q) - \mu + \mu - F^{-1}(q))^2 \right] \\ &\leq \mathbb{E}_{q \sim \operatorname{Unif}(0,1)} \left[2(F^{-1}(1 - q) - \mu)^2 + 2(F^{-1}(q) - \mu)^2 \right] \\ &= 4 \mathbb{E}_{q \sim \operatorname{Unif}(0,1)} \left[(F^{-1}(q) - \mu)^2 \right] = 4 \operatorname{Var}_{q \sim \operatorname{Unif}(0,1)} (F^{-1}(q)) = 4 \operatorname{Var}_{X \sim P}(X), \end{aligned} \tag{C.38}$$

Next we turn to the special case that P is symmetric around its mean. In this case, symmetry of the distribution around its mean ensures that $F^{-1}(1 - q) - \mu = \mu - F^{-1}(q)$ holds for

almost every $q \in [0, 1]$ (more precisely, this can fail only at discontinuities of the CDF of P). Therefore,

$$(F^{-1}(1-q) - \mu + \mu - F^{-1}(q))^2 = 2(F^{-1}(1-q) - \mu)^2 + 2(F^{-1}(q) - \mu)^2$$

for almost every $q \in [0, 1]$, and so the inequality in the step (C.38) above is in fact an equality. This proves $\text{Dev}(P) = 4 \text{Var}_{X \sim P}(X)$, as desired. $\square$

Lemma C.14. *Suppose $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$. Let $c = (a, b) \in \mathbb{R}^{m+n}$ be the concatenation of the two vectors. If $\text{Med}(c) < \text{Med}(a)$, then $\text{Med}(c) \geq \text{Med}(b)$; similarly, if $\text{Med}(c) > \text{Med}(a)$, then $\text{Med}(c) \leq \text{Med}(b)$.*

Proof. We will assume that $\text{Med}(c) < \text{Med}(a)$ and $\text{Med}(c) < \text{Med}(b)$, and will show that this leads to a contradiction. (The second claim is proved analogously and so we omit the proof.)

We first calculate

$$|\{i \in [m] : a_i \leq \text{Med}(c)\}| \leq |\{i \in [m] : a_i < \text{Med}(a)\}| \leq \begin{cases} \frac{m}{2}, & \text{if } m \text{ is even,} \\ \frac{m-1}{2}, & \text{if } m \text{ is odd,} \end{cases}$$

where the first step holds since we have assumed $\text{Med}(c) < \text{Med}(a)$, and the second step uses the definition of the median. Similarly,

$$|\{j \in [n] : b_j \leq \text{Med}(c)\}| \leq |\{j \in [n] : b_j < \text{Med}(b)\}| \leq \begin{cases} \frac{n}{2}, & \text{if } n \text{ is even,} \\ \frac{n-1}{2}, & \text{if } n \text{ is odd.} \end{cases}$$

On the other hand, again by definition of the median,

$$\begin{aligned} & |\{i \in [m] : a_i \leq \text{Med}(c)\}| + |\{j \in [n] : b_j \leq \text{Med}(c)\}| \\ &= |\{k \in [m+n] : c_k \leq \text{Med}(c)\}| \geq \begin{cases} \frac{m+n}{2}, & \text{if } m+n \text{ is even,} \\ \frac{m+n+1}{2}, & \text{if } m+n \text{ is odd.} \end{cases} \end{aligned}$$

So far, then, we have shown that

$$\frac{m - \mathbb{1}_{m \text{ is odd}}}{2} + \frac{n - \mathbb{1}_{n \text{ is odd}}}{2} \geq \frac{m+n + \mathbb{1}_{m+n \text{ is odd}}}{2}.$$

If either m or n is odd (or both), this immediately yields a contradiction. The remaining case is that m and n are both even, and

$$|\{i \in [m] : a_i \leq \text{Med}(c)\}| = |\{i \in [m] : a_i < \text{Med}(a)\}| = \frac{m}{2}$$

and

$$|\{j \in [n] : b_j \leq \text{Med}(c)\}| = |\{j \in [n] : b_j < \text{Med}(b)\}| = \frac{n}{2}.$$

Letting $a_{(1)} \leq \dots \leq a_{(m)}$ denote the order statistics of a , and similarly letting $b_{(1)} \leq \dots \leq b_{(n)}$ denote the order statistics of b , we therefore have

$$a_{(m/2)} \leq \text{Med}(c) < a_{(m/2+1)} \text{ and } b_{(n/2)} \leq \text{Med}(c) < b_{(n/2+1)}.$$

In particular, writing $c_{(1)} \leq \dots \leq c_{(m+n)}$ for the order statistics of c , this means that

$$c_{((m+n)/2)} = \max\{a_{(m/2)}, b_{(n/2)}\} \text{ and } c_{((m+n)/2+1)} = \min\{a_{(m/2+1)}, b_{(n/2+1)}\}.$$

By definition of the median, then,

$$\begin{aligned} \text{Med}(c) &= \frac{c((m+n)/2) + c((m+n)/2+1)}{2} = \frac{\max\{a_{(m/2)}, b_{(n/2)}\} + \min\{a_{(m/2+1)}, b_{(n/2+1)}\}}{2} \\ &\geq \min\left\{\frac{a_{(m/2)} + a_{(m/2+1)}}{2}, \frac{b_{(n/2)} + b_{(n/2+1)}}{2}\right\} = \min\{\text{Med}(a), \text{Med}(b)\}. \end{aligned}$$

This contradicts our assumption, and completes the proof. $\square$

Lemma C.15. *Let $P_{(Y,Z)}$ be a distribution on $[-1, 1] \times \mathbb{R}$, and let $Z_1, Z'_1, Z_2, Z'_2, \dots$ be independent draws from the corresponding marginal P_Z on $\mathbb{R}$. Let (K_n) be a sequence of positive integers with $K_n \rightarrow \infty$ as $n \rightarrow \infty$, and for each n , let $B_{n,1}, \dots, B_{n,K_n} \subseteq [n]$ be a collection of disjoint sets such that¹*

$$\max_{k \in [K_n]} \max_{i \in B_{n,k}} \left| \frac{i}{n} - \frac{k}{K_n} \right| \rightarrow 0.$$

Then, as $n \rightarrow \infty$

$$\mathbb{E} \left[\frac{1}{n} \sum_{k=1}^{K_n} \sum_{i \in B_{n,k}} d_W(P_{Y|Z}(\cdot | Z_{(n,i)}), P_{Y|Z}(\cdot | Z'_{(K_n,k)})) \right] \rightarrow 0.$$

Remark 4. *The intuition is that, at each value of n , we should have $\frac{i}{n} \approx \frac{k}{K_n}$ and therefore $Z_{(n,i)} \approx Z'_{(K_n,k)}$, for all k and all $i \in B_{n,k}$.*

Proof. Let F_Z^{-1} be the quantile function corresponding to P_Z , so that if $U \sim \text{Unif}[0, 1]$ then $F_Z^{-1}(U) \sim P_Z$. Now define $\tilde{P}_{Y|U}(\cdot | u) = P_{Y|Z}(\cdot | F_Z^{-1}(u))$, i.e., the conditional distribution of Y given U where $Z = F_Z^{-1}(U)$ and we sample Y conditional on Z from $P_{Y|Z}$. Since F_Z^{-1}

1. If $B_{n,k} = \emptyset$ for some $k \in [K_n]$, then we interpret $\max_{i \in B_{n,k}} |i/n - k/K_n| = 0$.

is monotone, it suffices to prove that

$$\mathbb{E} \left[\frac{1}{n} \sum_{k=1}^{K_n} \sum_{i \in B_{n,k}} d_W(\tilde{P}_{Y|U}(\cdot | U_{(n,i)}), \tilde{P}_{Y|U}(\cdot | V_{(K_n,k)})) \right] \rightarrow 0,$$

where $U_1, V_1, U_2, V_2, \dots$ are independent and identically distributed $\text{Unif}[0, 1]$ random variables. Writing $F_{Y|U}$ for the distribution function corresponding to the conditional distribution $\tilde{P}_{Y|U}$, we have

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{n} \sum_{k=1}^{K_n} \sum_{i \in B_{n,k}} d_W(\tilde{P}_{Y|U}(\cdot | U_{(n,i)}), \tilde{P}_{Y|U}(\cdot | V_{(K_n,k)})) \right] \\ &= \mathbb{E} \left[\frac{1}{n} \sum_{k=1}^{K_n} \sum_{i \in B_{n,k}} \int_{-1}^1 |F_{Y|U}(t | U_{(n,i)}) - F_{Y|U}(t | V_{(K_n,k)})| dt \right] \\ &= \int_{-1}^1 \mathbb{E} \left[\frac{1}{n} \sum_{k=1}^{K_n} \sum_{i \in B_{n,k}} |F_{Y|U}(t | U_{(n,i)}) - F_{Y|U}(t | V_{(K_n,k)})| \right] dt, \end{aligned} \quad (\text{C.39})$$

where the last step follows by Fubini's theorem.

Now, we define some additional notation: for $n \in \mathbb{N}$ and $j \in [n]$, let

$$V_{n,j} = \begin{cases} V_{(K_n,k)} & \text{if } U_j = U_{(n,i)} \text{ for some } i \in B_{n,k}, \\ U_j & \text{if } U_j = U_{(n,i)} \text{ for some } i \in [n] \setminus (B_{n,1} \cup \dots \cup B_{n,K_n}). \end{cases} \quad (\text{C.40})$$

Then for every $t \in [0, 1]$ and $n \in \mathbb{N}$,

$$\sum_{k=1}^{K_n} \sum_{i \in B_{n,k}} |F_{Y|U}(t | U_{(n,i)}) - F_{Y|U}(t | V_{(K_n,k)})| = \sum_{j=1}^n |F_{Y|U}(t | U_j) - F_{Y|U}(t | V_{n,j})|.$$

Moreover, $|F_{Y|U}(t | U_j) - F_{Y|U}(t | V_{n,j})|$ is identically distributed for each $j = 1, \dots, n$

(when n is fixed). Hence,

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{K_n}\sum_{i\in B_{n,k}}|F_{Y|U}(t|U_{(n,i)})-F_{Y|U}(t|V_{(K_n,k)})|\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n|F_{Y|U}(t|U_j)-F_{Y|U}(t|V_{n,j})|\right] \\ &= \mathbb{E}[|F_{Y|U}(t|U_1)-F_{Y|U}(t|V_{n,1})|].\end{aligned}$$

By Lemma C.16 below,

$$|F_{Y|U}(t|U_1)-F_{Y|U}(t|V_{n,1})| = o_P(1).$$

Since $F_{Y|U}(t|\cdot)$ is a bounded function, we deduce that

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{K_n}\sum_{i\in B_{n,k}}|F_{Y|U}(t|U_{(n,i)})-F_{Y|U}(t|V_{(K_n,k)})|\right] \rightarrow 0$$

for every $t \in [-1, 1]$. The result now follows by (C.39) and the dominated convergence theorem. $\square$

Lemma C.16. *Let $U_1, V_1, U_2, V_2, \dots \stackrel{\text{iid}}{\sim} \text{Unif}[0, 1]$, and let (K_n) be a sequence of positive integers with $K_n \rightarrow \infty$ as $n \rightarrow \infty$. Furthermore, for each n , let $B_{n,1}, \dots, B_{n,K_n} \subseteq \{1, \dots, n\}$ be a collection of disjoint sets such that*

$$\max_{k \in [K_n]} \max_{i \in B_{n,k}} \left| \frac{i}{n} - \frac{k}{K_n} \right| \rightarrow 0.$$

Then, for any Borel measurable $g : [0, 1] \rightarrow \mathbb{R}$, we have $|g(U_1) - g(V_{n,1})| = o_P(1)$, where $V_{n,1}$ is defined as in (C.40).

Proof. Let $\epsilon > 0$. For any Borel measurable $h : [0, 1] \rightarrow \mathbb{R}$, we can write

$$\begin{aligned} & \mathbb{P} \{ |g(U_1) - g(V_{n,1})| > \epsilon \} \\ & \leq \mathbb{P} \{ |h(U_1) - h(V_{n,1})| > \epsilon \} + \mathbb{P} \{ g(U_1) \neq h(U_1) \} + \mathbb{P} \{ g(V_{n,1}) \neq h(V_{n,1}) \}. \end{aligned} \tag{C.41}$$

By Lusin's theorem (restated for convenience in Theorem C.19), we can find a continuous function $h : [0, 1] \rightarrow \mathbb{R}$ such that

$$\mathbb{P}_{U \sim \text{Unif}[0,1]} \{ g(U) \neq h(U) \} \leq \frac{\epsilon}{4}.$$

Next, for each $k = 1, \dots, K_n$, let $\mathcal{E}_{n,k}$ denote the event that $U_1 = U_{(n,i)}$ for some $i \in B_{n,k}$, which has probability $\frac{|B_{n,k}|}{n}$. Further, define $\mathcal{E}_{n,*}$ to be the event that $U_1 = U_{(n,i)}$ for some $i \in [n] \setminus (B_{n,1} \cup \dots \cup B_{n,K_n})$. By construction of $V_{n,1}$, on the event $\mathcal{E}_{n,k}$ we have $V_{n,1} = V_{(K_n,k)}$, and on $\mathcal{E}_{n,*}$ we have $V_{n,1} = U_1$. Therefore, letting π_n denote a permutation

of $[K_n]$ such that $V_{(K_n,k)} = V_{\pi_n(k)}$ for $k \in [K_n]$, we have

$$\begin{aligned}
\mathbb{P}\{g(V_{n,1}) \neq h(V_{n,1})\} &= \sum_{k=1}^{K_n} \mathbb{P}\{\mathcal{E}_{n,k}, g(V_{(K_n,k)}) \neq h(V_{(K_n,k)})\} + \mathbb{P}\{\mathcal{E}_{n,*}, g(U_1) \neq h(U_1)\} \\
&\stackrel{(*)}{=} \sum_{k=1}^{K_n} \mathbb{P}\{\mathcal{E}_{n,k}\} \cdot \mathbb{P}\{g(V_{(K_n,k)}) \neq h(V_{(K_n,k)})\} + \mathbb{P}\{\mathcal{E}_{n,*}, g(U_1) \neq h(U_1)\} \\
&\leq \sum_{k=1}^{K_n} \frac{|B_{n,k}|}{n} \mathbb{P}\{g(V_{(K_n,k)}) \neq h(V_{(K_n,k)})\} + \mathbb{P}\{g(U_1) \neq h(U_1)\} \\
&= \sum_{k=1}^{K_n} \frac{|B_{n,k}|}{n} \mathbb{P}\{g(V_{\pi_n(k)}) \neq h(V_{\pi_n(k)})\} + \mathbb{P}\{g(U_1) \neq h(U_1)\} \\
&= \sum_{k=1}^{K_n} \frac{|B_{n,\pi_n^{-1}(k)}|}{n} \mathbb{P}\{g(V_k) \neq h(V_k)\} + \mathbb{P}\{g(U_1) \neq h(U_1)\} \\
&\leq \sum_{k=1}^{K_n} \frac{|B_{n,k}|}{n} \cdot \frac{\epsilon}{4} + \frac{\epsilon}{4} \leq \frac{\epsilon}{2},
\end{aligned}$$

where the step marked $(*)$ holds because $\mathcal{E}_{n,k}$ depends only on $(U_1, \dots, U_n)$, while $V_{(K_n,k)}$ depends only on $(V_1, \dots, V_{K_n})$, and therefore $\mathbb{1}_{\mathcal{E}_{n,k}} \perp\!\!\!\perp V_{(K_n,k)}$, while the penultimate step holds since $U_1 \sim \text{Unif}[0, 1]$, and $V_k \sim \text{Unif}[0, 1]$ for each k . Finally, let $\hat{F}_{U,n} = \frac{1}{n} \sum_{i=1}^n \delta_{U_i}$ and $\hat{F}_{V,K_n} = \frac{1}{n} \sum_{i=1}^{K_n} \delta_{V_i}$ be the empirical distribution of $U_1, \dots, U_n$ and $V_1, \dots, V_{K_n}$ respectively. We may assume that $U_1, \dots, U_n$ are distinct, since this holds almost surely, so if $i \in [n]$ is such that $U_1 = U_{(n,i)}$, then we have

$$\hat{F}_{U,n}(U_1) = \frac{i}{n}.$$

Let F_{Unif} denote the distribution function of the $\text{Unif}[0, 1]$ distribution, i.e., $F_{\text{Unif}}(t) = t$ for

$t \in [0, 1]$. If i further satisfies $i \in B_{n,k}$ for some $k \in [K_n]$, then $V_{n,1} = V_{(K_n,k)}$, so

$$\begin{aligned} |U_1 - V_{n,1}| &\leq \left| U_1 - \frac{i}{n} \right| + \left| V_{n,1} - \frac{k}{K_n} \right| + \left| \frac{i}{n} - \frac{k}{K_n} \right| \\ &\leq \|\widehat{F}_{U,n} - F_{\text{Unif}}\|_\infty + \|\widehat{F}_{V,K_n} - F_{\text{Unif}}\|_\infty + \max_{k \in [K_n]} \max_{i \in B_{n,k}} \left| \frac{i}{n} - \frac{k}{K_n} \right|. \end{aligned}$$

If instead we have $i \in [n] \setminus (B_{n,1} \cup \dots \cup B_{n,K_n})$, then this bound holds trivially, since in that case $V_{n,1} = U_1$ by definition. Since the first two terms are each $o_P(1)$ by the Glivenko–Cantelli theorem, while the last term $\max_{k \in [K_n]} \max_{i \in B_{n,k}} \left| \frac{i}{n} - \frac{k}{K_n} \right|$ is vanishing by assumption, it follows that $V_{n,1} \xrightarrow{P} U_1$, so by the continuous mapping theorem, $\mathbb{P} \{ |h(U_1) - h(V_{n,1})| > \epsilon \} \leq \epsilon/4$ for sufficiently large n . The result therefore follows from (C.41). $\square$

Lemma C.17. *The function f_n from (C.36) satisfies the Lipschitz property*

$$|f_n(y, y') - f_n(\tilde{y}, \tilde{y}')| \leq 4\|y - \tilde{y}\|_1 + 4\|y' - \tilde{y}'\|_1$$

for $(y, y'), (\tilde{y}, \tilde{y}') \in [-1, 1]^{mn} \times [-1, 1]^{mn}$.

Proof of Lemma C.17. Since $t \mapsto t_+^2$ is 4-Lipschitz on $[-2, 2]$, we have for each $i = 1, \dots, \lfloor m/2 \rfloor$ that

$$\left| (y_{(m_n+1-i)} - y'_{(i)})_+^2 - (\tilde{y}_{(m_n+1-i)} - \tilde{y}'_{(i)})_+^2 \right| \leq 4|y_{(m_n+1-i)} - \tilde{y}_{(m_n+1-i)}| + 4|y'_{(i)} - \tilde{y}'_{(i)}|.$$

Therefore,

$$|f_n(y, y') - f_n(\tilde{y}, \tilde{y}')| \leq 4 \sum_{i=1}^{\lfloor mn/2 \rfloor} |y_{(m_n+1-i)} - \tilde{y}_{(m_n+1-i)}| + 4 \sum_{i=1}^{\lfloor mn/2 \rfloor} |y'_{(i)} - \tilde{y}'_{(i)}|.$$

Writing $y_{()} = (y_{(1)}, \dots, y_{(m_n)})$ for the sorted vector of y , and same for $y'_{()}, \tilde{y}_{()}$, and $\tilde{y}'_{()}$, we deduce that

$$|f_n(y, y') - f_n(\tilde{y}, \tilde{y}')| \leq 4\|y_{()} - \tilde{y}_{()}\|_1 + 4\|y'_{()} - \tilde{y}'_{()}\|_1.$$

Finally, we write $\hat{P}_y := \sum_{i=1}^{m_n} \delta_{y_i}$ and $\hat{P}_{\tilde{y}} := \sum_{i=1}^{m_n} \delta_{\tilde{y}_i}$. Then, from the definition of the L_1 -Wasserstein distance $d_W(\cdot, \cdot)$, it holds that

$$\|y_{()} - \tilde{y}_{()}\|_1 = d_W(\hat{P}_y, \hat{P}_{\tilde{y}}) \leq \|y - \tilde{y}\|_1,$$

and a similar inequality also holds for $y', \tilde{y}'$. This completes the proof. $\square$

Lemma C.18. *Let Q be any distribution on $[-1, 1]$, and let $A = (A_1, \dots, A_{m_n}) \sim Q^{m_n}$ and $B = (B_1, \dots, B_{m_n}) \sim Q^{m_n}$. Then, with f_n as defined in (C.36), there exists a universal constant $C > 0$ such that*

$$\left| \frac{2}{m_n} \mathbb{E}[f_n(A, B)] - \text{Dev}(Q) \right| \leq \frac{C}{\sqrt{m_n}}.$$

Proof of Lemma C.18. First let $A_{(1)} \leq \dots \leq A_{(m_n)}$ and $B_{(1)} \leq \dots \leq B_{(m_n)}$ be the order statistics of A and B , and write $A_{()} = (A_{(1)}, \dots, A_{(m_n)})$ and $B_{()} = (B_{(1)}, \dots, B_{(m_n)})$ for the corresponding sorted vectors. Then

$$f_n(A, B) = f_n(A_{()}, B_{()})$$

since $(y, y') \mapsto f_n(y, y')$ is invariant to permutations of y , and invariant to permutations of y' .

Next, let F_Q^{-1} denote the quantile function corresponding to Q . Let $U_1, \dots, U_{m_n} \stackrel{\text{iid}}{\sim} \text{Unif}[0, 1]$, so we can assume without loss of generality that $A_i = F_Q^{-1}(U_i)$ for $i \in [m_n]$. Define $A' = (A'_1, \dots, A'_{m_n})$ by $A'_i = F_Q^{-1}(1 - U_i)$ for $i \in [m_n]$, and let $A'_{()}$ denote the corresponding

sorted vector. Then by Lemma C.17,

$$\begin{aligned}
|\mathbb{E}[f_n(A, B)] - \mathbb{E}[f_n(A, A')]| &\leq \mathbb{E}[|f_n(A, B) - f_n(A, A')|] \\
&= \mathbb{E}[|f_n(A_{()} , B_{()}) - f_n(A_{()} , A'_{()})|] \leq 4\mathbb{E}[\|A'_{()} - B_{()}\|_1].
\end{aligned} \tag{C.42}$$

Next, let $U_{(1)} \leq \dots \leq U_{(m_n)}$ be the order statistics of $U_1, \dots, U_{m_n}$. Then $A_{(i)} = F_Q^{-1}(U_{(i)})$ and $A'_{(i)} = F_Q^{-1}(1 - U_{(m_n+1-i)})$ for each $i \in [m_n]$. Now, for $i \in [m_n]$, define the event

$$\mathcal{E}_i := \begin{cases} \{U_{(i)} < 1/2\} & \text{if } i \leq \lfloor m_n/2 \rfloor, \\ \emptyset & \text{if } i = \lceil m_n/2 \rceil \text{ when } m_n/2 \notin \mathbb{N}, \\ \{U_{(i)} \geq 1/2\} & \text{if } i \geq \lceil m_n/2 \rceil + 1. \end{cases}$$

Thus, it follows that

$$f_n(A', A) = \sum_{i=1}^{\lfloor m_n/2 \rfloor} (A'_{(m_n+1-i)} - A_{(i)})_+^2 = \sum_{i=1}^{\lfloor m_n/2 \rfloor} \{F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)})\}^2 \mathbb{1}_{\mathcal{E}_i}$$

and

$$\begin{aligned}
f_n(A, A') &= \sum_{i=1}^{\lfloor m_n/2 \rfloor} (A_{(m_n+1-i)} - A'_{(i)})_+^2 \\
&= \sum_{i=1}^{\lfloor m_n/2 \rfloor} \{F_Q^{-1}(U_{(m_n+1-i)}) - F_Q^{-1}(1 - U_{(m_n+1-i)})\}^2 \mathbb{1}_{\mathcal{E}_{m_n+1-i}} \\
&= \sum_{i=\lceil m_n/2 \rceil + 1}^{m_n} \{F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)})\}^2 \mathbb{1}_{\mathcal{E}_i}.
\end{aligned}$$

Moreover, since $(A, A') \stackrel{d}{=} (A', A)$, we have $f_n(A, A') \stackrel{d}{=} f_n(A', A)$. This further implies that

$$\mathbb{E} [f_n(A, A')] = \mathbb{E} \left[\frac{f_n(A', A) + f_n(A, A')}{2} \right] = \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^{m_n} \{F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)})\}^2 \cdot \mathbb{1}_{\mathcal{E}_i} \right].$$

By definition, $\mathbb{E} [\{F_Q^{-1}(U_i) - F_Q^{-1}(1 - U_i)\}^2] = \text{Dev}(Q)$ for each $i \in [m_n]$. Hence,

$$\begin{aligned} & \left| \mathbb{E} [f_n(A, A')] - \frac{m_n}{2} \text{Dev}(Q) \right| \\ &= \left| \mathbb{E} [f_n(A, A')] - \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^{m_n} \{F_Q^{-1}(1 - U_i) - F_Q^{-1}(U_i)\}^2 \right] \right| \\ &= \left| \mathbb{E} [f_n(A, A')] - \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^{m_n} \{F_Q^{-1}(1 - U_{(i)}) - F_Q^{-1}(U_{(i)})\}^2 \right] \right| \leq 2 \mathbb{E} \left[\sum_{i=1}^{m_n} \mathbb{1}_{\mathcal{E}_i^c} \right], \end{aligned} \tag{C.43}$$

where the last step follows by recalling that Q is supported on $[-1, 1]$. Next, we define

$I_1 := \sum_{i=1}^{m_n} \mathbb{1}_{\{U_i < 1/2\}}$. Then by definition of the events $\mathcal{E}_i$, we have

$$\sum_{i=1}^{\lfloor m_n/2 \rfloor} \mathbb{1}_{\mathcal{E}_i^c} = \sum_{i=1}^{\lfloor m_n/2 \rfloor} \mathbb{1}_{U_{(i)} \geq 1/2} = \lfloor m_n/2 \rfloor - \sum_{i=1}^{\lfloor m_n/2 \rfloor} \mathbb{1}_{U_{(i)} < 1/2} = \max\{\lfloor m_n/2 \rfloor - I_1, 0\},$$

and similarly

$$\sum_{i=\lfloor m_n/2 \rfloor + 1}^{m_n} \mathbb{1}_{\mathcal{E}_i^c} = \sum_{i=\lfloor m_n/2 \rfloor + 1}^{m_n} \mathbb{1}_{U_{(i)} < 1/2} = \max\{I_1 - \lceil m_n/2 \rceil, 0\}.$$

Therefore,

$$\begin{aligned} \sum_{i=1}^{m_n} \mathbb{1}_{\mathcal{E}_i^c} &\leq 1 + \sum_{i=1}^{\lfloor m_n/2 \rfloor} \mathbb{1}_{\mathcal{E}_i^c} + \sum_{i=\lfloor m_n/2 \rfloor + 1}^{m_n} \mathbb{1}_{\mathcal{E}_i^c} \\ &= 1 + \max\{\lfloor m_n/2 \rfloor - I_1, 0\} + \max\{I_1 - \lceil m_n/2 \rceil, 0\} \leq 1 + \left| I_1 - \frac{m_n}{2} \right|. \end{aligned}$$

Since $I_1 = \sum_{i=1}^{m_n} \mathbb{1}_{\{U_i < 1/2\}} \sim \text{Bin}(m_n, 1/2)$ and $m_n \in \mathbb{N}$, by the Cauchy–Schwarz inequality we further have that

$$\mathbb{E} \left[\sum_{i=1}^{m_n} \mathbb{1}_{\mathcal{E}_i^c} \right] \leq 1 + \mathbb{E} \left[\left| I_1 - \frac{m_n}{2} \right| \right] \leq 1 + \sqrt{\text{Var}(I_1)} = 1 + \frac{\sqrt{m_n}}{2} \leq \frac{3\sqrt{m_n}}{2}. \quad (\text{C.44})$$

Next we turn to bounding $\|A'_0 - B_0\|_1$. Write $\hat{P}_{A'}$ and $\hat{P}_B$ for the empirical distributions of A' and B , respectively. Then

$$\|A'_0 - B_0\|_1 = m_n d_W(\hat{P}_{A'}, \hat{P}_B) \leq m_n d_W(\hat{P}_{A'}, Q) + m_n d_W(\hat{P}_B, Q).$$

By bounds on the empirical 1-Wasserstein distance [Lei, 2020, Theorem 3.1], we have

$$\mathbb{E} \left[d_W(\hat{P}_{A'}, Q) \right] \leq C' m_n^{-1/2}$$

for a universal constant $C' > 0$, and same for $\hat{P}_B$. We conclude from this together with (C.42), (C.43) and (C.44) that

$$\left| \mathbb{E} [f_n(A, B)] - \frac{m_n}{2} \text{Dev}(Q) \right| \leq 8C' \sqrt{m_n} + 3\sqrt{m_n},$$

as required. □

Lemma C.19 (A version of Lusin’s theorem). *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a measurable function and let ν be a probability measure on $\mathbb{R}$. Then, given any $\epsilon > 0$, there is a compactly supported continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that $\nu\{x : f(x) \neq g(x)\} < \epsilon$.*

Proof. The proof follows by combining Theorem 2.18 and Theorem 2.24 of Rudin [1987]. □

APPENDIX D

APPENDICES FOR CHAPTER 5

D.1 Proofs

D.1.1 Proof of Theorem 5.1

Fix $\delta > 0$. Let A_δ denote the set $\{\tilde{x} : \mathbb{E}_{X \sim P_X} [H(X, \tilde{x}; h)] \geq \delta\}$, and $T_n(\tilde{x}) := \frac{\sum_{i=1}^n Y_i H(X_i, \tilde{x}; h)}{\sum_{i=1}^n H(X_i, \tilde{x}; h)}$.

With these definitions in place, we observe that

$$\left\{ \mathbb{E} [Y_{n+1} \mid \tilde{X}_{n+1}] \notin \hat{C}_{n,\delta}(\tilde{X}_{n+1}) \right\} \iff \left\{ \left| T_n(\tilde{X}_{n+1}) - \mathbb{E}[Y_{n+1} \mid \tilde{X}_{n+1}] \right| > \frac{4\sqrt{2}B}{h^d \sqrt{n\alpha} \delta} \right\}.$$

Therefore,

$$\begin{aligned} & \mathbb{P} \left\{ \mathbb{E} [Y_{n+1} \mid \tilde{X}_{n+1}] \notin \hat{C}_{n,\delta}(\tilde{X}_{n+1}) \right\} \\ & \leq \mathbb{P} \left\{ \mathbb{E} [Y_{n+1} \mid \tilde{X}_{n+1}] \notin \hat{C}_{n,\delta}(\tilde{X}_{n+1}) \mid \tilde{X}_{n+1} \in A_\delta \right\} \\ & \quad + \mathbb{P} \left\{ \mathbb{E} [Y_{n+1} \mid \tilde{X}_{n+1}] \notin \hat{C}_{n,\delta}(\tilde{X}_{n+1}), \tilde{X}_{n+1} \in A_\delta^c \right\} \\ & \leq \mathbb{E}_{\tilde{X}_{n+1}} \left[\mathbb{P}^{Pn} \left\{ \left| T_n(\tilde{X}_{n+1}) - \mathbb{E}[Y_{n+1} \mid \tilde{X}_{n+1}] \right| > \frac{4\sqrt{2}B}{h^d \sqrt{n\alpha} \delta} \mid \tilde{X}_{n+1} \right\} \mid \tilde{X}_{n+1} \in A_\delta \right] \\ & \quad + \mathbb{P}_{\tilde{X}_{n+1}} \left\{ \mathbb{E}_{X \sim P_X} [H(X, \tilde{X}_{n+1})] < \delta \right\}. \end{aligned}$$

Hence, it suffices to show that for any fixed $\tilde{x} \in A_\delta$,

$$\mathbb{P}^{Pn} \left\{ \left| T_n(\tilde{x}) - \mathbb{E}[Y_{n+1} \mid \tilde{X}_{n+1} = \tilde{x}] \right| \leq \frac{4\sqrt{2}B}{h^d \sqrt{n\alpha} \delta} \right\} \geq 1 - \alpha.$$

To see this, fix any $\tilde{x} \in A_\delta$. Recall that $X_{n+1} \mid \tilde{X}_{n+1} \sim P_X \circ H(\cdot, \tilde{X}_{n+1}; h)$ and thus,

$$\begin{aligned} \mathbb{E} \left[Y_{n+1} \mid \tilde{X}_{n+1} = \tilde{x} \right] &= \mathbb{E}_{X_{n+1} \mid \tilde{X}_{n+1} = \tilde{x}} [\mathbb{E} [Y_{n+1} \mid X_{n+1}]] \\ &= \frac{\mathbb{E}_{X_{n+1} \sim P_X} [\mathbb{E} [Y_{n+1} \mid X_{n+1}] H(X_{n+1}, \tilde{x}; h)]}{\mathbb{E}_{X_{n+1} \sim P_X} [H(X_{n+1}, \tilde{x}; h)]} = \frac{\mathbb{E}_{(X,Y) \sim P} [Y H(X, \tilde{x}; h)]}{\mathbb{E}_{(X,Y) \sim P} [H(X, \tilde{x}; h)]}. \end{aligned}$$

Hence, by Chebyshev's inequality, for any $\alpha \in (0, 1)$,

$$\mathbb{P}_{P^n} \left\{ \left| \frac{1}{n} \sum_{i=1}^n H(X_i, \tilde{x}; h) - \mathbb{E}_{(X,Y) \sim P} [H(X, \tilde{x}; h)] \right| \geq \left(\frac{2 \text{Var}_{(X,Y) \sim P} (H(X, \tilde{x}; h))}{n\alpha} \right)^{1/2} \right\} \leq \alpha/2,$$

and similarly, since $Y \in [0, 1]$,

$$\mathbb{P}_{P^n} \left\{ \left| \frac{1}{n} \sum_{i=1}^n Y_i H(X_i, \tilde{x}; h) - \mathbb{E}_{(X,Y) \sim P} [Y H(X, \tilde{x}; h)] \right| \geq \left(\frac{2 \text{Var}_{(X,Y) \sim P} (Y H(X, \tilde{x}; h))}{n\alpha} \right)^{1/2} \right\} \leq \alpha/2.$$

Moreover, since H follows the form given in (5.1), $\text{Var}_{(X,Y) \sim P} (H(X, \tilde{x}; h)) \leq \frac{B^2}{h^{2d}}$ for any $\tilde{x}$.

Therefore, with probability at least $1 - \alpha$,

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n H(X_i, \tilde{x}; h) - \mathbb{E}_{(X,Y) \sim P} [H(X, \tilde{x}; h)] \right| &\leq \frac{\sqrt{2}B}{h^d \sqrt{n\alpha}}, \quad \text{and} \\ \left| \frac{1}{n} \sum_{i=1}^n Y_i H(X_i, \tilde{x}; h) - \mathbb{E}_{(X,Y) \sim P} [Y H(X, \tilde{x}; h)] \right| &\leq \frac{\sqrt{2}B}{h^d \sqrt{n\alpha}}. \end{aligned}$$

Consequently, for any $n \geq N_{\delta,h}$, whenever $\tilde{x} \in A_\delta$, the following holds with probability at least $1 - \alpha$:

$$\left| T_n(\tilde{x}) - \frac{\mathbb{E}_{(X,Y) \sim P} [Y H(X, \tilde{x}; h)]}{\mathbb{E}_{(X,Y) \sim P} [H(X, \tilde{x}; h)]} \right| \leq \frac{\frac{2\sqrt{2}B}{h^d \sqrt{n\alpha}}}{\mathbb{E}_{(X,Y) \sim P} [H(X, \tilde{x}; h)] - \frac{\sqrt{2}B}{h^d \sqrt{n\alpha}}} \leq \frac{4\sqrt{2}B}{h^d \sqrt{n\alpha} \delta},$$

as required. This completes the proof. $\square$

D.1.2 Proof of Theorem 5.2

The proof follows the same reasoning as in Theorem 5.1. Fix $\delta > 0$, let A_δ denote the set $\{x : \mathbb{P}_{X \sim P_X} \{X \in H_\infty^-(x, h)\} \geq \delta h^d\}$, and for any $x \in \mathbb{R}^d$, define

$$T_n(x) := \frac{\sum_{i=1}^n Y_i \mathbb{1} \{X_i \in H_\infty^-(x, h)\}}{\sum_{i=1}^n \mathbb{1} \{X_i \in H_\infty^-(x, h)\}}.$$

Thus, we have that

$$\begin{aligned} & \left\{ \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h) \right] < \hat{L}_n(X_{n+1}) \right\} \\ & \subseteq \left\{ \left| T_n(X_{n+1}) - \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h) \right] \right| > \frac{2\sqrt{2}}{h^d \sqrt{n\alpha} \delta} \right\}. \end{aligned}$$

Moreover, it follows that

$$\begin{aligned} & \mathbb{P} \left\{ \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h) \right] < \hat{L}_n(X_{n+1}) \right\} \\ & \leq \mathbb{P} \left\{ \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h) \right] < \hat{L}_n(X_{n+1}) \mid X_{n+1} \in A_\delta \right\} \\ & \quad + \mathbb{P} \left\{ \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h) \right] < \hat{L}_n(X_{n+1}), X_{n+1} \in A_\delta^c \right\} \\ & \leq \mathbb{E}_{X_{n+1}} \left[\mathbb{P}_{P^n} \left\{ \left| T_n(X_{n+1}) - \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(X_{n+1}, h) \right] \right| > \frac{2\sqrt{2}}{h^d \sqrt{n\alpha} \delta} \mid X_{n+1} \right\} \mid X_{n+1} \in A_\delta \right] \\ & \quad + \mathbb{P}_{X_{n+1}} \left\{ \mathbb{P}_{X \sim P_X} \{X \in H_\infty^-(X_{n+1}, h)\} < \delta h^d \right\}. \end{aligned}$$

Thus, it remains to show that for any fixed $x \in A_\delta$,

$$\mathbb{P}_{P^n} \left\{ \left| T_n(x) - \mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(x, h) \right] \right| \leq \frac{2\sqrt{2}}{h^d \sqrt{n\alpha} \delta} \right\} \geq 1 - \alpha.$$

To see this, fix any $x \in A_\delta$ and observe that

$$\begin{aligned}
\mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(x, h) \right] &= \mathbb{E}_{\tilde{X}_{n+1} \mid \tilde{X}_{n+1} \in H_\infty^-(x, h)} \left[\mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \right] \right] \\
&= \frac{\mathbb{E}_{\tilde{X}_{n+1} \sim P_X} \left[\mathbb{E} \left[\tilde{Y}_{n+1} \mid \tilde{X}_{n+1} \right] \mathbb{1} \left\{ \tilde{X}_{n+1} \in H_\infty^-(x, h) \right\} \right]}{\mathbb{E}_{\tilde{X}_{n+1} \sim P_X} \left[\mathbb{1} \left\{ \tilde{X}_{n+1} \in H_\infty^-(x, h) \right\} \right]} \\
&= \frac{\mathbb{E}_{(X, Y) \sim P} \left[Y \mathbb{1} \left\{ X \in H_\infty^-(x, h) \right\} \right]}{\mathbb{P}_{(X, Y) \sim P} \left\{ X \in H_\infty^-(x, h) \right\}}.
\end{aligned}$$

Applying Chebyshev's inequality, we observe that for any $\alpha \in (0, 1)$,

$$\mathbb{P}^{P^n} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1} \left\{ X_i \in H_\infty^-(x, h) \right\} - \mathbb{P}_{X \sim P_X} \left\{ X \in H_\infty^-(x, h) \right\} \right| \geq \sqrt{\frac{2}{n\alpha}} \right\} \leq \alpha/2,$$

and similarly, since $Y \in [0, 1]$,

$$\mathbb{P}^{P^n} \left\{ \left| \frac{1}{n} \sum_{i=1}^n Y_i \mathbb{1} \left\{ X_i \in H_\infty^-(x, h) \right\} - \mathbb{E}_{(X, Y) \sim P} \left[Y \mathbb{1} \left\{ X \in H_\infty^-(x, h) \right\} \right] \right| \geq \sqrt{\frac{2}{n\alpha}} \right\} \leq \alpha/2.$$

Therefore, with probability at least $1 - \alpha$,

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n \mathbb{1} \left\{ X_i \in H_\infty^-(x, h) \right\} - \mathbb{P}_{X \sim P_X} \left\{ X \in H_\infty^-(x, h) \right\} \right| &\leq \sqrt{\frac{2}{n\alpha}}, \quad \text{and} \\
\left| \frac{1}{n} \sum_{i=1}^n Y_i \mathbb{1} \left\{ X_i \in H_\infty^-(x, h) \right\} - \mathbb{E}_{(X, Y) \sim P} \left[Y \mathbb{1} \left\{ X \in H_\infty^-(x, h) \right\} \right] \right| &\leq \sqrt{\frac{2}{n\alpha}}.
\end{aligned}$$

Consequently, for any $x \in \mathbb{R}^d$, the following holds with probability at least $1 - \alpha$:

$$\mathbb{P} \left\{ T_n(x) \geq \frac{\mathbb{E}_{(X, Y) \sim P} \left[Y \mathbb{1} \left\{ X \in H_\infty^-(x, h) \right\} \right] - \sqrt{\frac{2}{n\alpha}}}{\mathbb{P}_{X \sim P_X} \left\{ X \in H_\infty^-(x, h) \right\} + \sqrt{\frac{2}{n\alpha}}} \right\} \geq 1 - \alpha.$$

Since $x \in A_\delta$, this further implies that

$$\mathbb{P} \left\{ T_n(x) \geq \frac{\mathbb{E}_{(X,Y) \sim P} [Y \mathbb{1} \{X \in H_\infty^-(x, h)\}]}{\mathbb{P}_{(X,Y) \sim P} \{X \in H_\infty^-(x, h)\}} - \frac{2\sqrt{2}}{h^d \sqrt{n\alpha} \delta} \right\} \geq 1 - \alpha,$$

or equivalently

$$\mathbb{P} \left\{ T_n(x) \geq \mathbb{E} [Y_{n+1} \mid X_{n+1} \in H_\infty^-(x, h)] - \frac{2\sqrt{2}}{h^d \sqrt{n\alpha} \delta} \right\} \geq 1 - \alpha,$$

as required. This completes the proof. □