Bounds and Prediction Intervals for Individual Treatment Effects

Zhehao (Kenny) Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Thomas S. Richardson, Chair

Yanqin Fan

Ting Ye

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bounds and Prediction Intervals
for Individual Treatment Effects

Zhehao (Kenny) Zhang

Chair of the Supervisory Committee:
Thomas S. Richardson
Statistics

This dissertation investigates several problems related to bounds and prediction intervals for the individual treatment effect (ITE). While traditional causal inference has primarily focused on population-level parameters such as the average treatment effect (ATE) and the conditional average treatment effect (CATE), the ITE—often considered the ideal target for personalized decision-making – has recently garnered increasing attention. However, the ITE is generally not identifiable from the observed data, even in the context of randomized experiments. As a result, we consider the problem of bounding the ITE using prediction intervals. In particular, when the marginal distributions of potential outcomes are identifiable from a large, well-conducted randomized experiment, we aim to answer the general question: what constraints exist on the joint distribution of potential outcomes, given these known marginals?

Chapters 2 and 3 lay the theoretical foundation for addressing this question. In Chapter 2, we revisit a classical problem posed by Kolmogorov concerning the sharp upper and lower bounds for the cumulative distribution function (cdf) of the sum of two random variables with fixed marginals. Motivated in part by the challenges of bounding individual treatment effects, we focus on the *achievability* of these bounds. Specifically, we distinguish between bounds that are *achievable* and those that although they provide an infimum or supremum – and hence cannot be improved – are *not attained* by *any* distribution. We contribute new

results for the case of discrete random variables, and we also work to clarify, correct, and make more accessible several theorems in the existing literature.

In Chapter 3, we apply the insights from Chapter 2 to the difference of two random variables, with an application on individual treatment effects. We identify and address logical gaps in some prior work and illustrate our results through an example. Then we connect the problem of characterizing joint distributions with fixed marginals to the theory of couplings of probability measures. We generalize a finite version of Strassen's theorem using a max-flow/min-cut construction, which can be applied on prediction intervals (sets) for the ITE. Finally, we explore a natural extension: bounding the probability mass function (pmf) of the difference of two random variables.

In Chapter 4, we build upon the results of the previous chapters and focus on prediction intervals for individual treatment effects (ITE). For a binary treatment, we consider all three types of outcomes: binary, ordinal, and continuous. We begin by examining how to construct valid prediction intervals given known marginal distributions. We then address the converse problem: what necessary conditions must hold for a joint distribution of potential outcomes to exist such that a given prediction interval is valid? We discuss scenarios in which certain points must necessarily be included in the interval. Finally, we compare and contrast the ITE with the average treatment effect (ATE), highlighting their differing implications for causal inference.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ITE: Individual Treatment Effect.

ATE: Average Treatment Effect.

CATE: Conditional Average Treatment Effect.

CDF: Cumulative Distribution Function.

PMF: Probability Mass Function.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Thomas S. Richardson, for his unwavering guidance, support, and encouragement throughout my PhD journey. His patience and thoughtful feedback have been instrumental in my development as a researcher. His intellectual curiosity and insight have profoundly shaped both my research taste and the direction of this thesis.

I am grateful to the members of my reading committee, Yanqin Fan and Ting Ye, for their valuable time, constructive suggestions, and thoughtful critiques. I would like to thank Carlos Cinelli, who, although currently on leave and unable to be here, has guided me on various research projects and provided both valuable feedback and funding support during my PhD. I also thank Gary Chan for serving on the committee.

I would like to thank Wei Sun from Fred Hutch Cancer Research Center and Janet Baseman from UW School of Public Health for collaboration on research projects and funding support. I also want to thank my collaborators and colleagues, including but not limited to: Yilin Song, Danielle Tsao, Yikun Zhang, Jennifer Brennan, Eunice Jun, Saksham Jain, Simon Nguyen, Apara Venkat for stimulating discussions. The writing of this thesis has also benefited from valuable input and discussions with Giovanni Puccetti, Ludger Rüschendorf, James M. Robins, as well as the audience of the UW Causal Reading Group and participants of the 2024 ACIC poster session.

I am also grateful to my wonderful friends and cohort who have shared this journey with me: Alex Jiang, Kayla Irish, Zhaoqi Li, Yidan Xu, Trinity Fan, Rui Wang, Xingyu Wang, Yimin Zhao, Ronak Mehta, Shreya Prakash, James Buenfil, Jess Kunke, Jillian Fisher. Your presence made this journey lighter and more memorable. I would also like to thank all the staff in the Statistics Department for their support, as well as the professors whose courses have shaped my understanding and appreciation for the world of statistics.

Finally, I am deeply thankful to my family in China and my partner Yuan Fang for their unconditional love.

DEDICATION

to Yuan, and my parents Minyan and Mingjie

Chapter 1
INTRODUCTION

1.1 Motivation

Recent years have witnessed growing interest in individualized treatment effects (ITEs) for personalized decision-making. Unlike average or population-level treatment effects, ITEs acknowledge that different individuals may respond to the same intervention in heterogeneous ways. Although the ideal goal of causal inference might be to pin down each individual's potential outcomes, the ITE is typically not identified from observed data. This has led researchers to investigate bounds and prediction intervals for the ITE.

In this dissertation, we investigate bounds and prediction intervals on the ITE, leveraging ideas from both classical and modern approaches to bounding the joint distributions of potential outcomes. This problem is closely connected to the fundamental question of constraints on the joint distribution of two random variables when their marginal distributions are known. Notably, this is related to the line of work introduced by Kolmogorov's problem on bounding the distribution of sums and extended by Makarov [1982], Rüschendorf [1982], Frank et al. [1987], and others. Here, we extend these results to bound the pmf/cdf for the ITE under general assumptions, connect these results to the theory of probability measure coupling, and relate these bounds to answer questions about prediction intervals for the ITE.

1.2 Notation and Preliminaries

Throughout this thesis, we use the potential outcomes framework [Neyman, 1990, Rubin, 1974] and standard notation for potential outcomes. For a binary treatment variable $D \in \{0, 1\}$ and a response variable Y, we denote $Y(0)$ as the potential outcome if an individual were to take treatment $D = 0$ and $Y(1)$ as the potential outcome if an individual were to

take treatment $D = 1$. The *individual treatment effect* (ITE) is thus

$$Y(1) - Y(0).$$

For notational convenience, we occasionally use Y_1 and Y_0 as shorthand for $Y(1)$ and $Y(0)$, respectively. Although $Y(0)$ and $Y(1)$ cannot be observed at the same time for each individual, under the stable unit treatment value assumption (SUTVA), their marginal distributions can sometimes be identified or estimated from randomized experiments or observational data when the assumption of unconfoundedness holds.

Let $F_0(\cdot)$ be the cumulative distribution function (cdf) of $Y(0)$ and $F_1(\cdot)$ be the cdf of $Y(1)$. We assume these marginals are either known or consistently estimable from data (for instance, from a large well conducted randomized trial). Then our goal is to understand the different constraints on the distribution of $Y(1) - Y(0)$ induced purely by knowledge of F_0 and F_1. Such questions connect naturally to the literatures of Frchét-Hoeffding-type bounds, copulas and the coupling arguments, where one considers all joint distributions consistent with fixed marginals.

We provide novel insights into when different bounds are "tight", how to extend them to discrete or continuous cases, and how to interpret them in the language of causal inference.

1.3 Organization of the Thesis

This thesis is structured as follows:

- **Chapter 2** provides a theoretical background on bounding the sum of two random variables with given marginal distributions. We review classical Makarov bounds, the previous literature, and refine their connection to sharpness and attainability in both discrete and continuous settings.

- **Chapter 3** focuses on the difference of two random variables. We provide tight cdf/pmf bounds on the difference of two random variables with fixed marginals. We connect the bounds to the ITE prediction intervals and extend existing theorems in the coupling of probability measures to solve causal inference problems.

- **Chapter 4** applies the bounding theory specifically to the prediction intervals of ITE. We present a comprehensive characterization of when certain proposed intervals must necessarily include certain treatment effects and when they may exclude them. This leads to the construction of valid prediction intervals for $Y(1) - Y(0)$ and conditions under which a prediction interval has valid coverage.

- **Chapter 5** offers a conclusion, summarizing our findings and highlighting potential future work. We discuss open questions in bounding individual treatment effects.

The overall contributions in this thesis provide a detailed analysis of characterizing the ITE prediction intervals. We discuss constructing the best possible bounds on ITE and limitations of the prediction intervals on ITE due to the intrinsic unidentifiability nature.

Chapter 2

BOUNDS ON THE DISTRIBUTION OF A SUM OF TWO RANDOM VARIABLES: REVISITING A PROBLEM OF KOLMOGOROV

In this charpter, we revisit the following problem, proposed by Kolmogorov: given prescribed marginal distributions F and G for random variables X, Y respectively, characterize the set of compatible distribution functions for the sum $Z = X + Y$. Bounds on the distribution function for Z were first given by Makarov [1982] and Rüschendorf [1982] independently. Frank et al. [1987] provided a solution to the same problem using copula theory. However, though these authors obtain the same bounds, they make different assertions concerning their sharpness. In addition, their solutions leave some open problems in the case when the given marginal distribution functions are discontinuous. These issues have led to some confusion and erroneous statements in subsequent literature, which we correct.

2.1 Introduction

The question of the best possible bounds for the distribution function of the sum of two random variables whose individual distribution functions are fixed was originally raised by A.N. Kolmogorov and was first solved by Makarov [1982] and (independently) Rüschendorf [1982]. Using copula theory [Sklar, 1959], Frank et al. [1987] reframed this question and provided an elegant proof that the bounds were achievable in certain settings.

Makarov-type bounds are widely studied in the field of optimal transport, quantitative risk management, and banking (see Puccetti and Wang 2015, Example 4.2 and Puccetti 2024, Section 3 for a detailed discussion). Generalizations of the bounds to the sum of more than two random variables or vector-valued random variables can be found in Embrechts et al. [2013] and Li et al. [1996]. Computational aspects of the bounds have been discussed in Puccetti [2024], Hofert et al. [2017], Puccetti and Rüschendorf [2012] and Kreinovich and Ferson [2006]. Notwithstanding these extensions, here we will focus on the original bivariate problem concerning the relationship of two scalar-valued random variables with

fixed marginals.

In this chapter we revisit the connection between the construction of Frank et al. [1987] and Kolmogorov's original question. We distinguish between bounds that are *achievable* and those that although they provide an infimum or supremum – and hence cannot be improved – are *not attained* by *any* distribution. We characterize the circumstances under which the bounds are achievable.

In addition to the new results provided in the chapter, we also hope to make the theorems in the existing literature more accessible. The prior results are stated in different papers that are hard to relate because there are at least four 'dichotomies' that specify the nature of the bound and the probability (or feature of the distribution function) that is being bounded:

(i) Upper bounds versus lower bounds;

(ii) Whether the 'distribution function' of the sum is defined in terms of $P(Z < z)$ or $P(Z \leq z)$;

(iii) Whether stated bounds are on the left or right limits of the distribution function;

(iv) Whether a lower (upper) bound is considered to be 'sharp' if it represents the infimum (supremum) of the attainable probabilities versus, in addition, requiring that the bound be *achievable* in that it is attained by a joint distribution with the specified margins.

These choices interact in that, for example, a bound on the left limit of 'a distribution function evaluated at z' may correspond to a bound on $P(Z < z)$ or $P(Z \leq z)$, depending on which definition of a distribution function is used. Similarly, as we show in Section 2.3 below, though all the bounds that we give are sharp in the sense of being supremums or infimums, only certain combinations are always achievable; in particular, there are cases where the bound on a left (right) limit is attainable, but the bound on the right (left) limit is not; see Figure 2.4. We also present new results giving general conditions on the margins under which all the bounds are achievable.

The above distinctions can also be important in cases where, at first sight, it might not be expected. Obviously, if $Z = X + Y$ is a continuous random variable, then the left and right limits of its distribution function will agree (and hence distinctions (ii) and (iii) are moot). However, it is *not* the case, in general, that if X and Y are continuous random variables, then the distributions for Z at which the bounds on the distribution function for Z are attained will be continuous. As a consequence, even though X and Y may be continuous random variables, the bounds on the left and right limits of the distribution function for their sum, Z, may differ in terms of attainability.

As a further surprise, as we show in Theorem 2.3.9, if at least one of X or Y is a discrete random variable, then although the distributions for Z that attain the bounds will be discontinuous, in fact the upper and lower bounds on the left and right limits are all attainable.

In summary, we believe that the complexity within the existing literature, arising from these four dichotomies, may partly explain the origin of the suboptimal bounds stated by Williamson and Downs [1990] for example. Table A.1 in Appendix A.1 relates this chapter to the notation, terminology, and results in prior works.

The rest of the chapter is organized as follows. In Section 2.2, we review the proof of the best possible bounds proposed in Frank et al. [1987]. In Section 2.3, we revisit the connection of Frank et al. [1987]'s result to Komogorov's question. We provide characterizations of when the bounds proposed in Frank et al. [1987] for the sum of two random variables are achievable. In Theorems 2.4.1 and 2.4.2 we summarize our new results and relate them to those of Frank et al. [1987]. Figure 2.4 displays the bounds arising from the illustrative Examples 2.3.5 and 2.3.6 considered in this section.

2.2 Bounds on the sum of two random variables

Given marginal distribution functions F, G for random variables X, Y respectively, Kolmogorov's question (restated here in terms of right-continuous distribution functions) is to

find functions $\underline{J}$ and $\overline{J}$ such that for all $z \in \mathbb{R}$,

$$\underline{J}(z) = \inf P(X + Y \leq z),$$
$$\overline{J}(z) = \sup P(X + Y \leq z),$$

where the infimum and supremum are taken over all possible joint distribution functions $H(x, y)$ having the marginal cdfs $F(x)$ and $G(y)$.

First, we review some existing results on probability distributions and copulas.

Definition 2.2.1. *Let X be a random variable. The distribution function (or cumulative distribution function, cdf) F of X is defined to be $F(x) = P(X \leq x)$ for $x \in \mathbb{R}$.*

Note that under Definition 2.2.1, for any random variable X, $F(\cdot)$ is a right-continuous function. Frank et al. [1987] and Williamson and Downs [1990] used a left-continuous version of the definition of distribution functions where they replace $P(X \leq x)$ with $P(X < x)$.[1]

Definition 2.2.2 (Embrechts and Hofert 2013)**.** *Let X be a random variable with distribution function F. The generalized inverse (also known as the quantile function) $F^{-1} : [0, 1] \to \overline{\mathbb{R}} = [-\infty, \infty]$ of F is defined as:*

$$F^{-1}(u) = \inf\{x \in \mathbb{R}, F(x) \geq u\}, u \in [0, 1],$$

with $\inf \emptyset = \infty$.

Definition 2.2.3. *A two dimensional copula is a mapping C from $[0, 1]^2$ to $[0, 1]$ satisfying the conditions:*

1. *$C(a, 0) = C(0, a) = 0$ and $C(a, 1) = C(1, a) = a$, for all a in $[0, 1]$;*

[1]Some of the results we cited in this chapter was originally defined in terms of the left-continuous functions $\widetilde{F}(x) = P(X < x)$ and $\widetilde{G}(x) = P(X < x)$ (which somewhat confusingly they also call "distribution functions"). To be consistent in notation, F, G in this chapter always refer to the right-continuous distribution functions given by $F(x) = P(X \leq x)$ and likewise for G and Y. We reserve $\widetilde{F}, \widetilde{G}$ when we need to talk about the functions given by $P(X < x)$ and $P(Y < y)$, equivalently the left hand limits of $F(x)$ and $G(y)$.

2. $C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) + C(a_1, b_1) \geq 0$ *for all* a_1, a_2, b_1, b_2 *in* $[0, 1]$ *such that* $a_1 \leq a_2, b_1 \leq b_2$.

Proposition 2.2.4. *According to the definition, any copula* C *is nondecreasing in each argument, and,*

$$W(a, b) \leq C(a, b) \leq M(a, b)$$

where

$$W(a, b) = \max(a + b - 1, 0), \quad M(a, b) = \min(a, b).$$

The bounds W *and* M *are known as Fréchet-Hoeffding copula bounds.*

Theorem 2.2.5 (Sklar 1959)**.** *Consider a 2-dimensional cdf* H *with marginals* F, G*. There exists a copula* C*, such that*

$$H(x, y) = C(F(x), G(y))$$

for all x, y *in* $[-\infty, \infty]$*. If* F, G *are both continuous, then* C *is unique*[2]*; otherwise* C *is uniquely determined only on* $\operatorname{Ran} F \times \operatorname{Ran} G$*, where* $\operatorname{Ran} F, \operatorname{Ran} G$ *denote respectively the range of the cdfs* F *and* G*.*

See Sklar [1959], Embrechts and Hofert [2013], Schmidt [2007] for more discussion of general n-dimensional copulas and Sklar's Theorem.

Now we want to bound the cdf of the sum of two random variables using copulas. We reprise the argument given by Frank et al. [1987] which gives lower bounds on $P(X+Y < z)$ and upper bounds on $P(X + Y \leq z)$, and further establishes by construction that these are *achievable*.

Let H be a two-dimensional cumulative distribution function for random variables X, Y with marginals F, G respectively so that $H(x, y) = P(X \leq x, Y \leq y)$. By Sklar's theorem [Sklar, 1959], there exists a copula C such that $H(x, y) = C(F(x), G(y))$. Note that the cdf

[2]Note that the condition here relates to the cdfs F, G viewed as functions. We are not assuming that the random variables X, Y are (absolutely) continuous (with respect to Lebesgue measure). The latter is a sufficient but not necessary condition for F, G to be continuous.

for the sum of two random variables X, Y is fully characterized by their joint cdf H. Let $Z := X + Y$ and J be the cdf for Z. Then for any $z \in [-\infty, \infty]$,

$$J(z) = \iint_{x+y \leq z} dH(x, y).$$

For the copula C and marginal distribution functions F, G, let $\sigma_C(F, G)$ be the function defined by $\sigma_C(F, G)(-\infty) = 0, \sigma_C(F, G)(\infty) = 1$ and

$$\sigma_C(F, G)(z) = \iint_{x+y \leq z} dC(F(x), G(y)), \quad \text{for } -\infty < z < \infty.$$

Since $H(x, y) = C(F(x), G(y))$, $\sigma_C(F, G)(z) = J(z)$ for all $z \in [-\infty, \infty]$. We let

$$\tau_C(F, G)(z) = \sup_{x+y=z} C(F(x), G(y)),$$
$$\rho_C(F, G)(z) = \inf_{x+y=z} \breve{C}(F(x), G(y)),$$

where

$$\breve{C}(a, b) = a + b - C(a, b).$$

Theorem 2.2.6 (Frank et al. Theorem 2.14)**.** *We have the following bounds for any copula C and arbitrary given distribution functions F, G and $z \in [-\infty, \infty]$:*

$$\tau_W(F, G)(z) \leq \tau_C(F, G)(z) \leq \sigma_C(F, G)(z) \leq \rho_C(F, G)(z) \leq \rho_W(F, G)(z).^3$$

Note that τ_W, ρ_W are known functions that depend solely on the marginal cdfs F and G. We obtain the following bounds:

$$\tau_W(F, G)(z) \leq \sigma_C(F, G)(z) \leq \rho_W(F, G)(z).$$

---

[3]Proof of these bounds and visualizations can be found in Frank et al. [1987], Nelsen [2006]. Here only σ_C can be interpreted as the probability that $Z \leq z$. Therefore, there are no explicit constructions establishing whether or not the bounds on τ_W, ρ_W can be achieved.

Here we will write the bounds τ_W, ρ_W explicitly.

$$\tau_W(F,G)(z) = \sup_{x+y=z} \max(F(x)+G(y)-1,0); \tag{2.1}$$

$$\rho_W(F,G)(z) = \inf_{x+y=z} (F(x)+G(y)-\max(F(x)+G(y)-1,0)) \tag{2.2}$$

$$= \inf_{x+y=z} (F(x)+G(y)+\min(-F(x)-G(y)+1,0)) \tag{2.3}$$

$$= \inf_{x+y=z} \min(1, F(x)+G(y)) \tag{2.4}$$

$$= 1 + \inf_{x+y=z} \min(0, F(x)+G(y)-1). \tag{2.5}$$

Theorem 2.2.6 establishes the validity of the bounds in Equation (2.1) and (2.5). Thereafter, we call the bounds $\tau_W(F,G)(z)$ and $\rho_W(F,G)(z)$ the Makarov bounds on $P(X+Y \leq z)$. Frank et al. [1987] shows that these bounds are equivalent to those given in Makarov [1982], which is the first paper to prove the bounds. Independently, Rüschendorf [1982] also proves similar bounds and relates the bounds to the literature on infimal convolution [Rockafellar, 1997].[4]

2.3 Sharpness of the bounds

To investigate the tightness of the bounds, we first distinguish three notions of sharpness.

For two random variables X, Y with fixed marginals F, G, respectively, let $J(\cdot)$ be the distribution function of $X+Y$ and let $J_\ell(\cdot)$ and $J^u(\cdot)$ be bounding functions such that $J_\ell(z) \leq J(z) \leq J^u(z)$ for all $z \in \mathbb{R}$.

Definition 2.3.1 (Achievability at a point)**.** *We say the lower bound $J_\ell(\cdot)$ is* achievable at $z = z_0$ *if there exists a joint distribution H of X, Y satisfying the marginals such that under H, $J(z_0) = J_\ell(z_0)$. The upper bound $J^u(\cdot)$ is achievable at $z = z_0$ if there exists a joint distribution H of X, Y satisfying the marginals such that under H, $J(z_0) = J^u(z_0)$.*

Definition 2.3.2 (Pointwise Best-Possible)**.** *We say the lower bound $J_\ell(\cdot)$ is* pointwise best-possible *if for all $z_0 \in \mathbb{R}$ and $\epsilon > 0$, $J_\ell(z_0) + \epsilon$ will not be a valid lower bound for*

[4] In Appendix A.4, we discuss the relationship between the bounds in Rüschendorf [1982] and the Makarov bounds presented in this section.

$J(z_0)$. In other words, for all $z_0 \in \mathbb{R}$ and $\epsilon > 0$, there exists a joint distribution H of X, Y satisfying the marginals such that under H, $J(z_0) < J_\ell(z_0) + \epsilon$. The upper bound $J^u(\cdot)$ is pointwise sharp if for all $z_0 \in \mathbb{R}$ and $\epsilon > 0$, $J^u(z_0) - \epsilon$ will not be a valid lower bound for $F(z_0)$.[5]

Definition 2.3.3 (Uniformly Sharp)**.** *We say the lower bound $J_\ell(\cdot)$ is* uniformly sharp of H *if there exists a single joint distribution H of X, Y satisfying the marginals such that under H, $J(z) = J_\ell(z)$ for all $z \in \mathbb{R}$. The upper bound $J^u(\cdot)$ is uniformly sharp if there exists a single joint distribution H of X, Y satisfying the marginals such that under H, $J(z) = J^u(z)$ for all $z \in \mathbb{R}$.*

Following these definitions, if a bound is uniformly sharp, then it is achievable for all $z \in \mathbb{R}$ and also pointwise sharp. If a bound is achievable for all $z \in \mathbb{R}$, then it is pointwise sharp. However, a pointwise sharp bound may not be achievable for all $z \in \mathbb{R}$.

2.3.1 Prior results on achievability of bounds

We first state a theorem given in Nelsen [2006].[6]

Theorem 2.3.4 (Frank et al. Theorem 3.2, Nelsen Theorem 6.1.2)**.** *Let F and G be two fixed distribution functions. For any $z \in (-\infty, \infty)$:*
(i) There exists a copula C_t, dependent only on the value t of $\tau_W(F, G)$ at $z-$, such that

$$\sigma_{C_t}(F, G)(z-) = \tau_W(F, G)(z-) = t, \tag{2.6}$$

where $z-$ is the left hand limit of the functions σ_{C_t} and τ_W as they approach z.[7]

---

[5] Unlike Firpo and Ridder [2019], we differentiate between achievability and pointwise sharpness because a bound can be pointwise sharp but not necessarily achievable.

[6] Frank et al. [1987] uses an example of a degenerate distribution to show that τ_W and ρ_W can be achieved for certain F, G (where the bounds are uniformly sharp in the example). However, it is not clear that for arbitrary F, G, Kolmogorov's question is answered.

[7] Frank et al. [1987] state their result in terms of $\widetilde{F}(z) = P(Z < z)$, which is left-continuous. When translating the result, we need to be careful about the implications of this notational difference. In general, $\tau_W(\widetilde{F}, \widetilde{G})(z-) \leq \tau_W(\widetilde{F}, \widetilde{G})(z) = \tau_W(F, G)(z-) \leq \tau_W(F, G)(z)$ and $\rho_W(\widetilde{F}, \widetilde{G})(z-) \leq \rho_W(\widetilde{F}, \widetilde{G})(z) = \rho_W(F, G)(z-) \leq \rho_W(F, G)(z)$. See Appendix A.2 where we prove these results.

(ii) There exists a copula C_r, dependent only on the value r of $\rho_W(F,G)(z)$, such that

$$\sigma_{C_r}(F,G)(z) = \rho_W(F,G)(z) = r.$$

Theorem 2.3.4 rephrases the result of Frank et al. [1987] in terms of the more common definition of right continuous distribution functions. The proof of Theorem 2.3.4 is sketched in Nelsen [2006]. Embrechts et al. [2002] Theorem 5 also sketches the proof for the lower bound (2.6).

Kolmogorov's question concerns the pointwise sharp bounds on the distribution function for the sum of two random variables. Statement (ii) of Theorem 2.3.4 along with Sklar's Theorem shows that the Makarov upper bound (2.5) is achievable for all $z \in \mathbb{R}$ and thus is pointwise sharp. The achievability of the Makarov upper bounds can also be proved using a result in Rüschendorf [1983]. We provide a complete proof based on Rüschendorf [1983]'s argument in Appendix A.3. We will next consider two examples which show that statement (i) of Theorem 2.3.4 does not completely address Kolmogorov's question.

Example 2.3.5. *Let X, Y be Bernoulli random variables with $p_1 = 0.5$ and $p_2 = 0.4$. Let F, G be the distribution functions of X, Y respectively. Suppose that we are interested in obtaining a lower bound for $P(X+Y \leq 1)$. Clearly, $\tau_W(F,G)(1) = 0.6$ while $\tau_W(F,G)(1-) = 0.1$. Statement (i) of Theorem 2.3.4 does not provide lower bound $\tau_W(F,G)(1)$ on $P(X+Y \leq 1) = J(1)$.*

Example 2.3.6 (Nelsen, 2006)**.** *Let X, Y be random variables with uniform distributions on $[0,1]$. Let F, G be distribution functions of X, Y respectively. Suppose that we wish to obtain a lower bound for $P(X+Y \leq 1)$. In this example, $\tau_W(F,G)(1) = \tau_W(F,G)(1-) = 0$. We obtain the lower bound for $J(z) = P(X+Y \leq 1) = 0$. Statement (i) of Theorem 2.3.4 tells us that there exists a joint distribution such that $P(X+Y < 1) = 0$. In fact, the construction in the proof of Theorem 2.3.4, given by Nelsen will result in $P(X+Y \leq 1) = 1$ and $P(X+Y < 1) = 0$ when $X = 1-Y$. However, this leaves open the question of whether $\tau_W(F,G)(1) = 0$ is a tight bound on $P(X+Y \leq 1)$, either in the sense of being achievable, or more weakly, best possible.*

To summarize, statement (i) of Theorem 2.3.4 leaves two questions unanswered: First, as shown in Example 2.3.5, when F, G are not continuous, $\tau_W(F,G)(z-)$ can be different from $\tau_W(F,G)(z)$. Thus when F and G are not continuous Theorem 2.3.4 (i) does not provide any information regarding whether the bound $\tau_W(F,G)(z)$ is either achievable or best possible vis a vis $J(z) = P(X+Y \leq z)$. Second, even when F, G are continuous, so that $\tau_W(F,G)(z-) = \tau_W(F,G)(z)$, it is still possible that $\sigma_{C_t}(F,G)(z-) < \sigma_{C_t}(F,G)(z)$. Consequently, even in the continuous case, the existence of the copula C_t given in statement (i) of Theorem 2.3.4 merely establishes that $\tau_W(F,G)(z-)$ is achievable as a lower bound on $P(X+Y<z)$; it says nothing about the achievability (or otherwise) of $\tau_W(F,G)(z)$ as a lower bound on $J(z)$.

2.3.2 When are the bounds achievable?

The next Theorem establishes that that for all F, G (possibly discontinuous) $\tau_W(F,G)(z)$ is best possible – thus addressing the unanswered question of Kolmogorov. We will return to the issue of achievability of $\tau_W(F,G)(z)$ as a lower bound on $P(X+Y \leq z)$, in Theorems 2.3.9 (discrete case) to 2.3.15 (general case) below.

Theorem 2.3.7 (No loose end to Kolmogorov's question)**.** *Let F and G be two fixed distribution functions. For any $z \in (-\infty, \infty)$, let $s = \tau_W(F,G)(z)$. For any $\epsilon > 0$, there exists a copula $C_{s,\epsilon}$ such that*

$$\sigma_{C_{s,\epsilon}}(F,G)(z) < s + \epsilon.$$

In other words, the lower bound $\tau_W(F,G)(z)$ on $J(z)$ cannot be improved for any F, G, and is thus pointwise best possible in terms of z.

Before proving Theorem 2.3.7, we first state a useful lemma.

Lemma 2.3.8 (Firpo and Ridder Theorem 2)**.** *For fixed distribution functions F, G, the function*

$$\tau_W(F,G)(z) = \sup_{x+y=z} \max\{F(x) + G(y) - 1, 0\}$$

is right continuous and non-decreasing for all z.

Lemma 2.3.8 follows from the fact that F, G are right continuous and nondecreasing.

Proof. The proof of Theorem 2.3.7 uses a similar construction as in Makarov [1982]. Now suppose that for a given z, $\tau_W(F,G)(z) = s$. Following Lemma 2.3.8, for all $\epsilon > 0$, there exists $\delta > 0$ such that $\tau_W(F,G)(z+m) - s < \epsilon$ for all $0 < m < \delta$. Pick any $0 < m < \delta$ that meets the above condition that $\tau_W(F,G)(z+m) < s + \epsilon$. Then there exists a copula $C_{s,\epsilon}$ such that

$$\begin{aligned}
\sigma_{C_{s,\epsilon}}(F,G)(z) &\leq \sigma_{C_{s,\epsilon}}(F,G)((z+m)-) && (2.7)\\
&= \tau_W(F,G)((z+m)-) && (2.8)\\
&\leq \tau_W(F,G)(z+m) && (2.9)\\
&< s+\epsilon, && (2.10)
\end{aligned}$$

where (2.7) follows from the fact that for any random variable D, $P(D \leq d) \leq P(D < d+h)$ for all $h > 0$; (2.8) follows from the optimality result in part (i) of Theorem 2.3.4; (2.9) follows from the non-decreasing property of τ_W;(2.10) follows from the construction of $\tau_W(F,G)(z+m)$. This implies that for all $\epsilon > 0$, we can construct a copula $C_{s,\epsilon}$ such that $\sigma_{C_{s,\epsilon}}(F,G)(z) < \tau_W(F,G)(z) + \epsilon$, meaning that $\tau_W(F,G)(z) + \epsilon$ will not be a valid lower bound on $P(X+Y \leq z) = J(z)$. □

Theorem 2.3.7 along with Theorem 2.3.4 proves that Makarov's bounds are pointwise best-possible and thus directly address Komogorov's question. This closes the best-possible question left open by Theorem 2.3.4 (i).

2.3.3 Special case: at least one discrete variable

The following theorem states that the lower bound $\tau_W(F,G)(z)$ is achievable when we are dealing with one or more discrete random variables.

Theorem 2.3.9. *If at least one of X, Y is a discrete random variable (a random variable which may take on only a countable number of distinct values), then there exists a copula*

C_t, dependent only on the value of $t = \tau_W(F,G)(z)$, such that

$$\sigma_{C_t}(F,G)(z) = \tau_W(F,G)(z) = t.$$

In other words, the lower bound on $J(z)$ is always achievable when one of X, Y is a discrete random variable.[8]

Proof. Before proving the Theorem 2.3.9, we will first prove a useful lemma.

Lemma 2.3.10. *For any $k < z$, if $t = \tau_W(F,G)(z) > \tau_W(F,G)(z-)$, there cannot be x', y' with $x' + y' = k < z$ such that $F(x') + G(y') - 1 = t$.*

Proof. Suppose that there exist x', y' with $x' + y' = k < z$ and $F(x') + G(y') - 1 = t$, then by the definition of sup,

$$\begin{aligned} t &\leq \sup_{x+y=k} \max(F(x) + G(y) - 1, 0) && (2.11)\\ &\leq \lim_{\epsilon>0,\epsilon\to 0} \sup_{x+y=z-\epsilon} \max(F(x) + G(y) - 1, 0) && (2.12)\\ &= \tau_W(F,G)(z-) && (2.13)\\ &\leq \sup_{x+y=z} \max(F(x) + G(y) - 1, 0) = t, && (2.14) \end{aligned}$$

where the inequalities (2.12) and (2.14) follow from the fact that the function $\tau_W(F,G)(\cdot)$ is non-decreasing and (2.11),(2.13) and (2.14) are by definition of $\tau_W(F,G)(k), \tau_W(F,G)(z-), \tau_W(F,G)(z)$. Thus, we have $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$, which contradicts with the hypothesis that $\tau_W(F,G)(z) > \tau_W(F,G)(z-)$. □

Now we will prove Theorem 2.3.9. Without loss of generality, we assume X is a discrete

[8]Notice that this is Theorem 2.3.4 (i) with $z-$ replaced by z, which can be surprising because Frank et al. [1987] note that theorem 2.3.4 (i) cannot be strengthened "even" when F, G are continuous, which implicitly implies that it cannot be strengthened when F, G are not continuous. Some misconception about this point can be found in related literatures. For example, Kim [2014] stated that "If the marginal distributions of X and Y are both absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, then the Makarov upper bound and lower bound can be achieved", which is contradicted by Example 2.3.6; Similarly, Williamson and Downs [1990] stated that in order for Theorem 2.3.4 to hold, it is necessary that F, G are not both discontinuous at a point x, y such that $x + y = z$; whereas Frank et al. [1987]'s proof does not require this to hold.

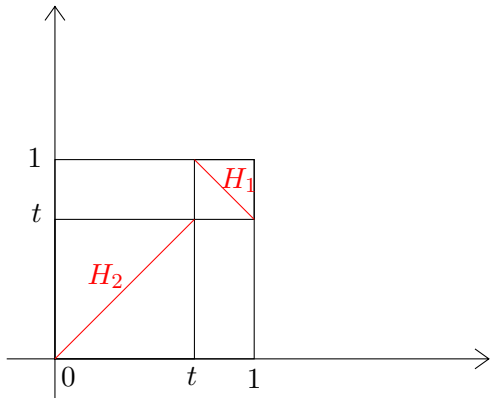

Figure 2.1: Support of C_t and mass assigned by C_t.

random variable. For $t = \tau_W(F, G)(z)$, we construct the copula

$$C_t(u, v) = \begin{cases} \text{Max}(u + v - 1, t), & (u, v) \text{ in } [t, 1] \times [t, 1], \\ \text{Min}(u, v), & \text{otherwise.} \end{cases}$$

Figure 2.1 illustrates the support of C_t and the mass assigned by C_t.[9]

Let $H_1 := \{(u, v) \in [0, 1]^2 \,|\, u + v - 1 = t\}$, $H_2 := \{(u, v) \in [0, 1]^2 \,|\, u = v < t\}$, $S_z := \{(u, v) \in [0, 1]^2 | F^{-1}(u) + G^{-1}(v) = z\}$, $S_{\bar{z}} := \{(u, v) \in [0, 1]^2 | F^{-1}(u) + G^{-1}(v) \leq z\}$, where F^{-1}, G^{-1} are the generalized inverses defined in Definition 2.2.2. Since t is a lower bound for $P(X + Y \leq z)$, C_t assigns mass at least t to the set $S_{\bar{z}}$. In particular, $H_2 \subseteq S_{\bar{z}}$. Thus, whether or not the lower bound on $P(X + Y \leq z)$ is achieved (by C_t) depends on whether the mass that C_t assigns to the set $H_1 \cap S_{\bar{z}}$ is 0.

We claim that $\iint_{S_{\bar{z}} \cap H_1} dC_t(u, v) = \iint_{S_z \cap H_1} dC_t(u, v)$. First suppose $\tau_W(F, G)(z) = \tau_W(F, G)(z-) = t$. Frank et al. [1987], showed that $P(X + Y < z) = \tau_W(F, G)(z-) = t$ under C_t. Thus, under C_t, the set $\{(u, v) \in [0, 1]^2 | F^{-1}(u) + G^{-1}(v) < z\}$ will contain all the mass in H_2 (equals t) but not any mass in H_1, so that the claim holds in this case. Now suppose $\tau_W(F, G)(z) > \tau_W(F, G)(z-)$, from Lemma 2.3.10 we know that the set $\{(u, v) \in [0, 1]^2 | F^{-1}(u) + G^{-1}(v) < z\}$ cannot contain mass in H_1. Therefore, $\iint_{S_{\bar{z}} \cap H_1} dC_t(u, v) =$

[9] In Appendix A.5, we discuss properties of the joint distribution induced by C_t and outline how to simulate from it.

$\iint_{S_z \cap H_1} dC_t(u,v)$, establishing the claim.

Since C_t assigns mass $(1-t)$ uniformly to H_1, if $S_z \cap H_1$ is empty or only contains countably many points, then $\iint_{S_z \cap H_1} dC_t(u,v) = 0$, which is sufficient to establish the claim.

Since X is discrete, x can take at most countably many values with non-zero probability under F. For a given z, there are at most countably many points (x,y) such that $x+y=z$ and $P(X=x)>0$.

Observe that

$$\{(u,v) \in [0,1]^2 | F^{-1}(u) + G^{-1}(v) = z\} = \cup_{(x,y):x+y=z} R_{xy}, \tag{2.15}$$

where $R_{xy} \equiv \{(F(x-),F(x)] \times (G(y-),G(y))]\}$ and we define the sets of form $(a,a]$ as $\{a\}$ for any $a \in \mathbb{R}$, which will arise if Y is not discrete.

We will show that for each (x,y) with $x+y=z$, $R_{xy} \cap H_1$ contains at most one point. Since by definition of C_t, $t = \sup_{x+y=z} \max\{F(x)+G(y)-1,0\}$, we have $F(x)+G(y)-1 \leq t$ for any $x+y=z$. For any $(u,v) \in R_{xy}$ with $u < F(x)$ or $v < G(y)$, it holds that $u+v-1 < t$ and thus (u,v) cannot be in H_1. Therefore $R_{xy} \cap H_1$ contains at most one point $(F(x),G(y))$.[10] As a consequence by (2.15), there exist at most countably many points in $S_z \cap H_1$. Thus, $\iint_{S_z \cap H_1} dC_t(u,v) = 0$ and

$$\sigma_{C_t}(F,G)(z) = \tau_W(F,G)(z) = t.$$

□

2.3.4 C_t not achieving the bound implies no other copula achieves the bound

Theorem 2.3.4 shows that the lower bound $\tau_W(F,G)(z-)$ on $J(z-)$ can be achieved. In fact, the proof of Theorem 2.3.4 (see Frank et al. 1987 and Nelsen 2006) shows that the copula we constructed as C_t with $t = \tau_W(F,G)(z-)$ in the proof of Theorem 2.3.9 will achieve

[10] Each rectangular region R_{xy} can touch the line $u+v-1=t$ for at most one point because all points in R_{xy} satisfy $F^{-1}(u)+G^{-1}(v)=z$ and if there is more than one point in the intersection then t is not $\sup_{x+y=z} \max\{F(x)+G(y)-1,0\}$.

the lower bound $\tau_W(F,G)(z-)$.[11] We further showed in Theorem 2.3.7 that the lower bound $\tau_W(F,G)(z)$ on $J(z)$ cannot be improved. In example 2.3.6, for $t = \tau_W(F,G)(z) = \tau_W(F,G)(z-)$, we see that $P(X+Y<z) = t$ under C_t but $P(X+Y \leq z) > t$ under C_t.

This raises a new question: if we care not merely about sharpness, but also about the achievability of the lower bound $\tau_W(F,G)$ on $J(z)$ – rather than $J(z-)$ – and if C_t does not achieve the bound $\tau_W(F,G)$, can there be other copulas that can achieve the bound $\tau_W(F,G)$ for $J(z)$? Indeed, Frank et al. [1987] and Nelsen [2006] both pointed out that there are other copulas beside C_t that achieve the lower bound $\tau_W(F,G)(z-)$ for $J(z-)$.

The corollary of the next theorem implies that for continuous F, G and an arbitrary z, in order to determine whether the lower bound $\tau_W(F,G)(z)$ on $J(z)$ can be achieved, we only need to determine whether it is achieved under C_t for $t = \tau_W(F,G)(z)$. Theorem 2.3.13 along with Theorem 2.3.11 further establishes this claim for arbitrary F, G. In other words, if the lower bound for $J(z)$ is *not* achieved under C_t, then there is *no* joint distribution that will achieve this lower bound.

Theorem 2.3.11. *Given arbitrary z and F, G, if $\tau_W(F,G)(z-) = \tau_W(F,G)(z) = t$ and the copula C_t does not achieve the lower bound $\tau_W(F,G)(z)$ on $J(z) \equiv P(X+Y \leq z)$, then no other copula can achieve this lower bound.*

Proof. Since $\tau_W(F,G)(z-) = t$, by Theorem 2.3.4(i), copula C_t achieves the bound $\tau_W(F,G)(z-)$ on $P(X+Y<z)$. That is, C_t assigns mass t to the set $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u) + G^{-1}(v) < z\}$. Since, by hypothesis, C_t does not achieve the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$, C_t assigns non-zero probability to the set $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u) + G^{-1}(v) = z)\}$. In particular, the image of the set $\{(x,y) : x+y = z\}$ under the (F,G) mapping must contain a line segment with length greater than 0 on the line $u+v-1 = t$ in the uv-plane inside the unit square[12] as illustrated in Figure 2.2; since otherwise under C_t, $P(X+Y<z) = P(X+Y \leq z)$ in which case C_t achieves the bound. Let a, b

[11] In fact, as noted above, Frank et al. [1987] consider bounds on $\widetilde{J}$, Nelsen [2006] translate the result to the standard definition J but do not provide a full proof.

[12] The first two sentences of the proof imply that when $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$ this is a necessary and sufficient condition for C_t to assign non-zero probability to the set $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u) + G^{-1}(v) = z)\}$.

be such that the line segment $\{(u,v) : u = a+s, v = 1-t-(a+s) \text{ for } s, 0 \le s \le b-a\}$ is contained in $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u)+G^{-1}(v) = z\} \cap \{(u,v) \subseteq [0,1]\times[0,1] : u+v-1 = t\}$.

Now suppose there is a copula C that achieves the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \le z)$. First, we claim that C must assign mass t to the rectangle $R_1 = [0,a]\times[0,1+t-b]$. Since R_1 is a subset of $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u)+G^{-1}(v) \le z\}$ and by hypothesis C achieves the lower bound, C cannot assign mass more than t to R_1.

Suppose C assigns mass $0 < r < t$ to R_1; see Figure 2.3. Note that we define the margins of the copula to be uniform ($C(p,1) = C(1,p) = p$, for all p in $[0,1]$). In particular, in order for $C(1,1+t-b) = 1+t-b$, C needs to assign mass $1+t-b-r$ to $[a,1]\times[0,1+t-b]$ and similarly C needs to assign mass $a-r$ to $[0,a]\times[1+t-b,1]$. As a consequence, C needs to assign mass $1-(1+t-b-r)-(a-r)-r = b-a+r-t$ to the rectangle $[a,1]\times[1+t-b,1]$. Now consider the rectangle $[a,b]\times[0,1+t-b]$. It needs to contain mass at least $t-r$ since $[a,b]\times[0,1]$ needs to contain mass $b-a$ and $[a,b]\times[1+t-b,1] \subseteq [a,1]\times[1+t-b,1]$. Similarly, $[0,a]\times[1+t-b,1+t-a]$ needs to contain mass at least $t-r$. Then C assigns mass greater than or equal to $r+2(t-r) = t+(t-r) > t$ to $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u)+G^{-1}(v) \le z\}$, which is a contradiction that C achieves the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \le z)$. Therefore, C must assign mass t to the rectangle $R_1 = [0,a]\times[0,1+t-b]$.

Next, we show that C assigns mass $b-a$ to the rectangle $[a,b]\times[1+t-b,1+t-a]$. By the hypothesis that C achieves the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \le z)$, C assigns mass $1-t$ to the set $\{(u,v) \subseteq [0,1]\times[0,1] : F^{-1}(u)+G^{-1}(v) > z)\}$, which is a subset of the union of the following three rectangles: $[0,1]\times[1+t-a,1]$, $[a,b]\times[1+t-b,1+t-a]$, $[b,1]\times[0,1]$. In order to maintain uniform margins, the first and third rectangles contain mass $a-t$ and $1-b$. So the rectangle $[a,b]\times[1+t-b,1+t-a]$ needs to contain mass at least $(1-t)-(a-t)-(1-b) = b-a$. Again from the uniformity of the margins, $[a,b]\times[1+t-b,1+t-a]$ can contain mass at most $b-a$. Thus, C assigns mass $b-a$ to the rectangle $[a,b]\times[1+t-b,1+t-a]$. Furthermore, since the rectangle $[a,1]\times[1+t-b,1]$ contains total mass $1-(1+t-b)-(a-t) = b-a$, there's no mass elsewhere in this rectangle except in $[a,b]\times[1+t-b,1+t-a]$.

Now we show that C needs to assign mass $b-a$ to the line segment $(a,1+t-a)$ to $(b,1+t-b)$ inside the square $[a,b]\times[1+t-b,1+t-a]$. Figure 2.3 shows a zoomed-in version of the rectangle $[a,b]\times[1+t-b,1+t-a]$. First, C cannot assign any mass strictly below

the line segment $(a, 1+t-a)$ to $(b, 1+t-b)$ inside the square $[a,b] \times [1+t-b, 1+t-a]$ because C already assigns mass t to R_1 and the total mass assigned by C_t to the set $\{(u,v) \subseteq [0,1] \times [0,1] : F^{-1}(u) + G^{-1}(v) < z)\}$ is t. For any rectangle $[m,n] \times [c,d]$ such that $m + c - 1 \geq t$, $a \leq m < n \leq b, 1+t-b \leq c < d \leq 1+t-a$, suppose that C assigns mass $\delta > 0$ to this rectangle. We know that C assigns mass t to the region R_1. Let E_1 be the triangular area defined by vertices $(m, 1+t-m), (a, 1+t-a), (m, 1+t-a)$ and E_2 be the triangular area defined by vertices $(m, 1+t-m), (b, 1+t-b), (b, 1+t-m)$, as depicted in Figure 2.3. Note that we have previously established that C assigns no mass to the rectangle $[a,b] \times [1+t-a, 1]$. The mass assigned by C to E_1 and R_1 is equal to the mass in rectangle $[0,m] \times [0,1]$ subtracting the mass in the rectangle $[0,a] \times [1+t-a, 1]$, which is $C(m,1) - (a-t) = m - (a-t) = m - a + t$. Similarly, the mass assigned by C to E_2 and R_1 is equal to $C(1, 1+t-m) - (1-b) = 1 + t - m - (1-b) = t - m + b$. Thus, C assigns the mass $m - a$ to E_1 and $b - m$ to E_2. Since E_1, E_2 are disjoint, C assigns the mass $b - a$ to $E_1 \cup E_2$. Then the rectangle $[m,n] \times [c,d]$ will contain mass 0, which is a contradiction. Since the choice of $[m,n] \times [c,d]$ is arbitrary, we know that C assigns mass $b - a$ to the line segment in $\mathbb{R}^2$ from $(a, 1-t-a)$ to $(b, 1-t-b)$ inside the square $[a,b] \times [1+t-b, 1+t-a]$. Finally, C assigns mass at least $t + (b-a)$ to the set $\{(u,v) \subseteq [0,1] \times [0,1] : F^{-1}(u) + G^{-1}(v) \leq z)\}$, which contradicts that C achieves the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$. Thus, when $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$, if the copula C_t does not achieve the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$, then no other copula can achieve this lower bound.

□

Corollary 2.3.12. *Given arbitrary z and continuous F, G, let $t = \tau_W(F,G)(z)$. If the copula C_t does not achieve the lower bound $\tau_W(F,G)(z)$ on $J(z) \equiv P(X+Y \leq z)$, then no other copula can achieve this lower bound.*

Proof. This follows directly from Theorem 2.3.11 since when F, G are both continuous, $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$. □

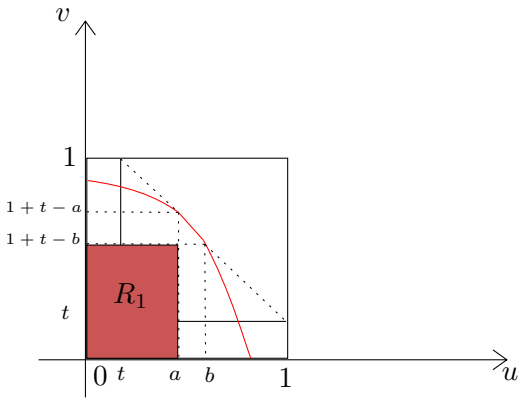

Figure 2.2: Copula C and the image of $x + y = z$ under the (F, G) mapping.

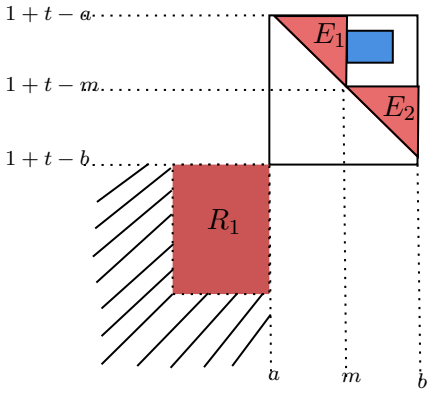

Figure 2.3: A zoom in part of Figure 2.2 with rectangle $[m, n] \times [c, d]$ colored in blue.

2.3.5 Sufficient conditions for achievability of the lower bound on $J(z)$

We will characterize another sufficient condition (different from Theorem 2.3.9) for the achievability of the lower bound on $J(z) = P(X + Y \leq z)$.

Theorem 2.3.13. *Given arbitrary z and F, G, if $\tau_W(F, G)(z) > \tau_W(F, G)(z-)$ then the copula C_t with $t = \tau_W(F, G)(z)$ will achieve the lower bound $\tau_W(F, G)(z)$ on $P(X+Y \leq z)$.*

Proof. We will prove the contrapositive: if the lower bound $t = \tau_W(F, G)(z)$ of $P(X+Y \leq z)$ is not achievable under C_t, then $\tau_W(F, G)(z) = \tau_W(F, G)(z-)$. Note that $\tau_W(F, G)(z) = \tau_W(F, G)(z-)$ holds trivially when $\tau_W(F, G)(z) = 0$. We will assume $t = \tau_W(F, G)(z) > 0$.

If the lower bound $\tau_W(F, G)(z)$ of $P(X + Y \leq z)$ is not achievable under C_t, the set $\{(u, v) \in [0, 1]^2 | F^{-1}(u) + G^{-1}(v) \leq z\}$ must contain a line segment with length greater than 0 on the line $u + v - 1 = t$ in the uv-plane. Based on Lemma 2.3.10, the set $\{(u, v) \in [0, 1]^2 | F^{-1}(u) + G^{-1}(v) = z\}$ must contain a line segment with length greater than 0 on the line $u + v - 1 = t$ in the uv-plane. The existence of this line segment implies that the image of the set $\{(x, y) : x + y = z\}$ under the (F, G) mapping must also contain a line segment with length greater than 0 on the line $u + v - 1 = t$ in the uv-plane inside the unit square. This means that there exist x^* and $\epsilon > 0$ such that $F(x) + G(z - x) - 1 = t$ for all $x \in (x^* - \epsilon, x^* + \epsilon)$ and $F(\cdot)$ is continuous and strictly increasing on $x \in (x^* - \epsilon, x^* + \epsilon)$. In

particular, for any $\delta > 0$, there exists $\epsilon^* > 0$ such that $F(x^*) - F(x^* - \epsilon^*) < \delta$. By definition of x^*, $\tau_W(F,G)(z) = F(x^*) + G(z - x^*) - 1$. Then for $\epsilon^* > 0$,

$$\tau_W(F,G)(z - \epsilon^*) = \sup_{x+y=z-\epsilon^*} \max(F(x) + G(y) - 1, 0) \geq F(x^* - \epsilon^*) + G(z - x^*) - 1.$$

And

$$\begin{aligned}
\tau_W(F,G)(z) - \tau_W(F,G)(z - \epsilon^*) &= F(x^*) + G(z - x^*) - 1 - \tau_W(F,G)(z - \epsilon^*) \\
&\leq F(x^*) + G(z - x^*) - 1 - (F(x^* - \epsilon^*) + G(z - x^*) - 1) \\
&= F(x^*) - F(x^* - \epsilon^*) < \delta.
\end{aligned}$$

Since δ is arbitrary, $\tau_W(F,G)(\cdot)$ is continuous at z and we must have $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$. □

Thus it follows from Theorem 2.3.13 that the *only* time when the lower bound on $P(X + Y \leq z)$ is *not* achievable is when the pointwise best possible bounds for $P(X + Y \leq z)$ and $P(X + Y < z)$ are the same. This result can be quite surprising: it follows that the distribution implied for $X+Y$ via the construction of C_t, namely, $\sigma_{C_t}(F,G)(z) = P(X+Y \leq z)$, is *discontinuous* at z only when $\tau_W(F,G)(z)$ is *continuous* at z, i.e. $\tau_W(F,G)(z-) = \tau_W(F,G)(z)$.

We present an example to show that we do not require the margins F, G to have uniform distributions on $[0,1]$ for the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$ to be not achievable.

Example 2.3.14. *Let*

$$F(x) = \begin{cases} 0 & x < 0, \\ x^2 & 0 \leq x < 1, \\ 1 & x \geq 1, \end{cases} \qquad G(y) = \begin{cases} 0 & y < 0, \\ 1 - (1-y)^2 & 0 \leq y < 1, \\ 1 & y \geq 1. \end{cases}$$

F is the distribution for a random variable X following a triangular distribution with $a = 0, b = c = 1$ (equivalent to Beta$(2,1)$), while G is the distribution for a random variable Y following a triangular distribution with $a = c = 0, b = 1$ (equivalent to Beta$(1,2)$).

Suppose $z = 1$. Then

$$\begin{aligned}\tau_W(F,G)(1) &= \sup_{x+y=1} \max\{F(x)+G(y)-1, 0\} \\ &= \sup_x \max\{F(x)+G(1-x)-1, 0\} \\ &= 0.\end{aligned}$$

The lower bound of 0 *corresponds to the copula C_0 constructed to achieve the lower Fréchet–Hoeffding bound (in other words, X, Y are perfectly negatively correlated). In this example, $X = 1-Y$. So under C_0, $P(X+Y=1) = P(X+Y \leq 1) = 1$ and $P(X+Y<1) = 0$.*

In contrast, it is not possible to construct a copula such that $P(X+Y \leq 1) = \tau_W(F,G)(1) =$ 0, so this lower bound is not achievable for $J(1) = P(X+Y \leq 1)$.

2.3.6 Characterization of achievability of the lower bound on $J(z)$

Theorem 2.3.15 will provide necessary and sufficient conditions for the lower bound on $J(z) = P(X+Y \leq z)$ to be achievable. Theorem 2.3.16 provides a useful summary of the results concerning both the Makarov upper and lower bounds for $J(z) = P(X+Y \leq z)$.

Theorem 2.3.15. *The Makarov lower bound $t = \tau_W(F,G)(z)$ on $P(X+Y \leq z)$ is not achievable at z if and only if there exist x^*, y^* with $x^*+y^* = z$ such that all following three conditions hold: (i) $F(x^*)+G(y^*) = \sup_{x+y=z}\{F(x)+G(y)\} \geq 1$; (ii) $F(x)+G(z-x)$ is constant for x in a neighborhood $Nr(x^*)$ of x^*; (iii) the image of the set $\{x, y : x \in Nr(x^*), y = z-x\}$ under the (F,G) mapping contains an open interval within the line segment $\{(u,v) \in [0,1]^2 \mid u+v-1 = t\}$.*[13]

Proof. If the Makarov lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$ is not achievable at z, then Theorem 2.3.13 implies that we have to have $\tau_W(F,G)(z) = \tau_W(F,G)(z-) = t$. Furthermore, when the lower bound on $P(X+Y \leq z)$ is not achievable, it is not achievable under any copula, which includes C_t. Note that by (2.6) in Theorem 2.3.4, under C_t, $P(X+Y<z) = t$. Consequently, for $S_z := \{(u,v) \in [0,1]^2 \mid F^{-1}(u)+G^{-1}(v) = z\}$ and

[13] The conditions can also be defined similarly using the neighborhood of y^*.

$H_1 := \{(u,v) \in [0,1]^2 \mid u+v-1 = t\}$, $P(X+Y=z) = \iint_{S_z \cap H_1} dC_t(u,v) > 0$, since otherwise we would also have $P(X+Y \leq z) = t$ under C_t.

If (i) does not hold, then either $\sup_{x+y=z}\{F(x)+G(y)\} < 1$ or $\sup_{x+y=z}\{F(x)+G(y)\} \geq 1$ but is not achieved on the set $x+y=z$. Note that by the definition of S_z, for any $(u,v) \in S_z$, $u+v \leq \sup_{x+y=z}\{F(x)+G(y)\}$. If $\sup_{x+y=z}\{F(x)+G(y)\} < 1$, then $t = \tau_W(F,G)(z) = 0$ by (2.1) hence H_1 is the line $u+v=1$ and thus the set S_z does not intersect H_1. On the other hand, if $\sup_{x+y=z}\{F(x)+G(y)\} \geq 1$, then again by (2.1), $1+t = \sup_{x+y=z}\{F(x)+G(y)\}$. If there do not exist x^*, y^* such that $F(x^*)+G(y^*) = \sup_{x+y=z}\{F(x)+G(y)\}$, then for any $(u,v) \in S_z$, $u+v < \sup_{x+y=z}\{F(x)+G(y)\}$, and thus $S_z \cap H_1 = \emptyset$ by definition of H_1. Hence in both sub-cases, the lower bound is achieved under C_t, which is a contradiction.

Now suppose that (i) holds. Since C_t assigns mass uniformly to the set H_1, and by hypothesis, $\iint_{S_z \cap H_1} dC_t(u,v) > 0$, there exists a line segment of positive length contained in $S_z \cap H_1$. Hence we may choose u^*, v^* in the interior of this segment, such that there exist x^*, y^* with $(F(x^*), G(y^*)) = (u^*, v^*)$ and there is a neighborhood of (u^*, v^*) in H_1 that is contained in S_z, thus establishing (iii). (ii) also needs to hold by definition of H_1.

For the converse, if (i), (ii), (iii) hold, then S_z contains a non-zero measure set in H_1. Since C_t assigns mass uniformly to the set H_1, $\iint_{S_z \cap H_1} dC_t(u,v) > 0$. This implies $P(X+Y=z) > 0$ under C_t and the lower bound $\tau_W(F,G)(z)$ will not be achievable at z under C_t. Furthermore, by contrapositive of Theorem 2.3.13, (i), (ii), (iii) imply that $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$. Finally it then follows by Theorem 2.3.11, that since C_t does not achieve the lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$, this bound will not be achievable by any copula. □

Theorem 2.3.16. *The Makarov upper $(\rho_W(F,G)(z))$ and lower $(\tau_W(F,G)(z))$ bounds on $P(X+Y \leq z)$ are pointwise best-possible,*[14]*. The Makarov upper bound is achievable for each $z \in \mathbb{R}$, but there may exist $z \in \mathbb{R}$ such that the lower bound is not achievable.*

Conversely, the Makarov upper $(\rho_W(F,G)(z-))$ and lower $(\tau_W(F,G)(z-))$ bounds on $P(X+Y < z)$ are pointwise best-possible. The Makarov lower bound is achievable for each

[14]See Definition 2.3.2.

$z \in \mathbb{R}$, but there may exist $z \in \mathbb{R}$ such that the upper bound is not achievable.

The Makarov bounds are, in general, not uniformly sharp.[15]

The following tables summarize when the Makarov bounds are always achievable (for all $z \in \mathbb{R}$ and any F, G) under different definitions of distribution functions.

Bound	$P(X+Y \leq z)$	
	Always Achievable	Pointwise Best-Possible
Makarov Upper Bound $\rho_W(F,G)(z)$	✓	✓
Makarov Lower Bound $\tau_W(F,G)(z)$	✗	✓
Bound	$P(X+Y < z)$	
	Always Achievable	Pointwise Best-Possible
Makarov Upper Bound $\rho_W(F,G)(z-)$	✗	✓
Makarov Lower Bound $\tau_W(F,G)(z-)$	✓	✓

Table 2.1: The Makarov upper bound $\rho_W(F,G)(z)$ on $P(X+Y \leq z)$ and the Makarov lower bound $\tau_W(F,G)(z-)$ on $P(X+Y < z)$ are always achievable for any given marginals F, G and for all $z \in \mathbb{R}$. The achievabilities of the Makarov upper bound $\rho_W(F,G)(z-)$ on $P(X+Y < z)$ and the Makarov lower bound $\tau_W(F,G)(z)$ on $P(X+Y \leq z)$ are margin specific and depend on z. See Example 2.3.6 and Theorem 2.3.15.

2.4 Summary Theorems on Achievability and Sharpness

We now add our new results to the main result of Frank et al. [1987] given in Theorem 2.3.4. Theorems 2.4.1 and 2.4.2 extend Theorem 3.2 in Frank et al. [1987] and provide a holistic picture of the achievability and sharpness of Makarov bounds on the sum of two random variables.

Parts (i) of Theorems 2.4.1 and 2.4.2 follow directly from Frank et al. [1987] and Nelsen [2006]; parts (ii) are stated in Makarov [1982] and Rüschendorf [1982] but not explicitly stated in Frank et al. [1987]; parts (iii) and (iv) follow from our new results in this section.

[15]For given F, G, there does not exist a single joint distribution that achieves the bounds for all z. For example, the construction of the copula C_t depends on the value z in general.

Theorem 2.4.1 (Lower Makarov bounds)**.** *Let F and G be two fixed distribution functions. For any $z \in (-\infty, \infty)$:*
(i) There exists a copula C_{t^*}[16], dependent only on the value t^* of $\tau_W(F,G)$ at $z-$, such that

$$\sigma_{C_{t^*}}(F,G)(z-) = \tau_W(F,G)(z-) = t^*.$$

In other words, the lower bound on $P(X+Y<z)$ is always achievable.
(ii) We have

$$\tau_W(F,G)(z) = \inf_C \sigma_C(F,G)(z),$$

where the infimum is taken over all copulas C. In other words, $t = \tau_W(F,G)(z)$ is the infimum of $P(X+Y \le z)$ for all possible joint distributions of X, Y.
(iii) If $t^* = \tau_W(F,G)(z-) < t = \tau_W(F,G)(z)$, then there exists a copula C_t dependent only on t such that

$$\sigma_{C_t}(F,G)(z) = \tau_W(F,G)(z).$$

That is, the lower bound on $P(X+Y \le z)$ can be achieved by C_t when $\tau_W(F,G)(z-) < \tau_W(F,G)(z)$. We also have $\tau_W(F,G)(z) = \min_C \sigma_C(F,G)(z)$ where the minimun is taken over all copulas.
(iv) If there is no copula C^* such that

$$\sigma_{C^*}(F,G)(z) = \tau_W(F,G)(z),$$

then $\tau_W(F,G)(z) = \tau_W(F,G)(z-)$. (iv) is a contrapositive of (iii) which states that if the lower bound on $P(X+Y \le z)$ is not achievable by any joint distribution, then the lower bounds for $P(X+Y \le z)$ and $P(X+Y<z)$ must be the same.

Similarly, we can state the theorems for the upper bounds with a slight asymmetry.

---

[16] Previously defined in the proof of Theorem 2.3.9.

Theorem 2.4.2 (Upper Makarov bounds)**.** *Let F and G be two fixed distribution functions. For any $z \in (-\infty, \infty)$:*

(i) There exists a copula C_r, dependent only on the value r of $\rho_W(F,G)(z)$, such that

$$\sigma_{C_r}(F,G)(z) = \rho_W(F,G)(z) = r.$$

In other words, the upper bound on $P(X+Y \leq z)$ is always achievable.

(ii) We have

$$\rho_W(F,G)(z-) = \sup_C \sigma_C(F,G)(z-),$$

where the supremum is taken over all copulas C. In other words, $r^* = \rho_W(F,G)(z-)$ is the supremum of $P(X+Y<z)$ for all possible joint distributions of X, Y.

(iii) If $r^* = \rho_W(F,G)(z-) < r = \rho_W(F,G)(z)$, then there exists a copula C_{r*} dependent only on r^* such that

$$\sigma_{C_{r^*}}(F,G)(z-) = \rho_W(F,G)(z-).$$

That is, the upper bound on $P(X+Y<z)$ can be achieved by C_{r^*} when $\rho_W(F,G)(z-) < \rho_W(F,G)(z)$. We also have $\rho_W(F,G)(z) = \max_C \sigma_C(F,G)(z)$ where the maximum is taken over all copulas.

(iv) If there is no copula C^* such that

$$\sigma_{C^*}(F,G)(z-) = \rho_W(F,G)(z-),$$

then $\rho_W(F,G)(z) = \rho_W(F,G)(z-)$. (iv) is a contrapositive of (iii) which states that if the upper bound on $P(X+Y<z)$ is not achievable by any joint distribution, then the upper bounds for $P(X+Y \leq z)$ and $P(X+Y<z)$ must be the same.

2.4.1 Examples Revisited

We now illustrate these results by revisiting Examples 2.3.5 and 2.3.6. Figure 2.4 summarizes the achievability of the Makarov upper and lower bounds in the two examples.

Makarov [1982] introduced the supremum of $P(X + Y \leq z)$, and Frank et al. [1987] showed that the copula C_r, where $r = \rho_W(F, G)(z)$, attains this supremum. This naturally leads to the association of C_r with the supremum of $P(X+Y < z)$, especially when F and G are continuous, making the suprema of $P(X+Y \leq z)$ and $P(X+Y < z)$ identical. However, it is important to note that the supremum for $P(X+Y < z)$ is not always achievable, and the value obtained using the copula C_r—where $r = \rho_W(F, G)(z-) = \rho_W(F, G)(z)$—can differ from the true (unachievable) supremum by as much as 1. In other words, using the copula C_r to estimate the upper bound for $P(X + Y < z)$ is suboptimal and may significantly deviate from the optimal bound. To see this consider Example 2.3.6 displayed in the top panel of Figure 2.4. Here we see that when $z = 1$, $r = \rho_W(F, G)(z-) = \rho_W(F, G)(z) = 1$. However, in this case C_r corresponds to the degenerate joint distribution under which $X + Y = 1$, implying that $P(X+Y < z) = 0$. In other words, in this example the Makarov upper bound on $P(X + Y < 1)$ is 1, but the copula C_r which achieves the upper bound on $P(X + Y \leq 1)$ has $P(X + Y < 1) = 0$ (!) which is as different as it is possible to be from the supremum value. More generally, in this example, for $z \in (0, 1]$, the supremum for $P(X + Y < z)$ is z, yet under $C_r = C_z$ we have $P(X + Y < z) = 0$.

Note that although the copula C_r (entirely) fails to achieve the upper bound on $P(X + Y < z)$, we may use another copula C_{r^*} where $r^* < r$ to obtain a joint distribution under which $P(X + Y < z)$ is arbitrarily close to the unattainable supremum. Specifically, in this example, the copula $C_{1-\epsilon}$, for some small $\epsilon > 0$, gives a joint distribution under which $P(X + Y < 1)$ is arbitrarily close to the (unattainable) upper bound of 1. This is because under $C_{1-\epsilon}$, $1 - \epsilon = P(X + Y = 1 - \epsilon) = P(X + Y \leq 1 - \epsilon) < P(X + Y < 1)$.

Conversely, the difference between the (unachievable) infimum of the (achievable) lower bounds on $P(X+Y \leq z)$ and the value achieved under the copula C_t where $t = \tau_W(F, G)(z)$ can also be 1.

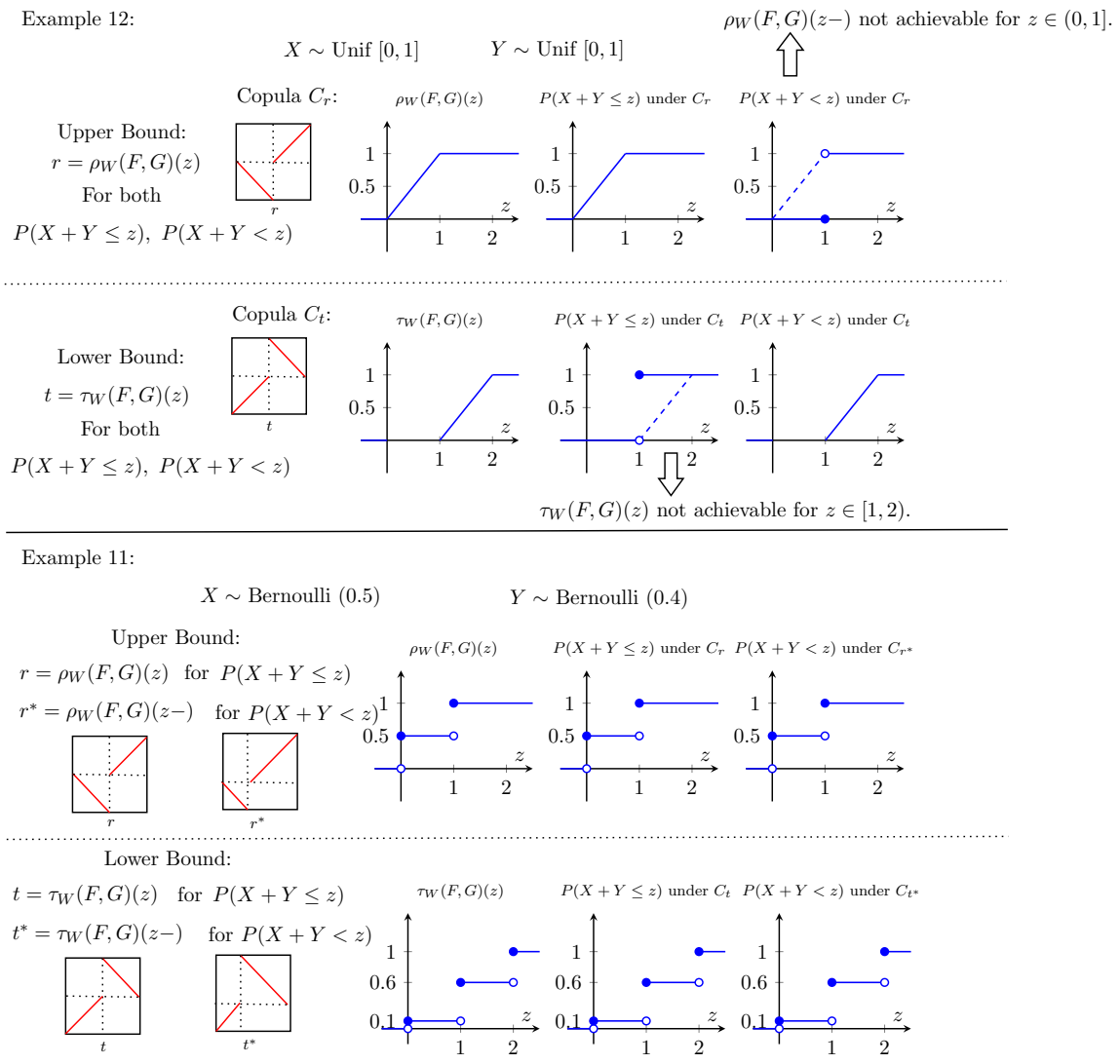

Figure 2.4: In the right two panels of Example 2.3.6, the dashed lines show the bounds $\rho_W(F,G)(z)$ and $\tau_W(F,G)(z)$ and the solid lines show what is achieved under the copulas; the difference indicates the bounds are not achievable. Hence only the upper bound on $P(X+Y \le z)$ and the lower bound on $P(X+Y \le z)$ can be achieved for all z. The upper and lower bounds for $P(X+Y < z)$ and $P(X+Y < z)$ are the same, which follows from part (iv) of Theorem 2.4.1 and 2.4.2. In Example 2.3.5, the upper and lower bounds for $P(X+Y \le z)$ and $P(X+Y < z)$ are different, all bounds on $P(X+Y \le z)$ and $P(X+Y < z)$ are achieved but under different copula constructions.

Chapter 3

BOUNDS ON THE DIFFERENCE OF TWO RANDOM VARIABLES: CDF, PMF, AND AN EXTENSION OF THE FINITE SET STRASSEN'S THEOREM

Kolmogorov's problem is closely related to inferring possible distributions for individual treatment effects (ITE) $Y_1 - Y_0$ given the marginal distributions of Y_1 and Y_0; the latter being identified from a randomized experiment. In this chapter, we first use our new insights from the previous chapter to sharpen and correct the results due to Williamson and Downs [1990], and to fill some other logical gaps. Second, we present an extension of Strassen's theorem for finite sets, providing necessary conditions under which the observed marginals admit a joint distribution that satisfies a given prediction interval (or set) constraint. The results in this chapter are particularly useful for assessing the internal consistency of observed data with assumed counterfactual structures. Finally, in the case of a binary treatment, we derive tight bounds on the probability mass function (pmf) of the ITE, extending existing results that focused primarily on cumulative distribution functions.

3.1 Introduction

The best possible bounds for the distribution function of the difference of two random variables was first studied in Williamson and Downs [1990], where the authors generalized the bounds in Frank et al. [1987] to different arithmetic operations of two random variables including subtraction, multiplication and division and claimed that these bounds are sharp. More recently, Fan and Park [2010] introduced the bounds in Williamson and Downs [1990] into the context of causal inference and concluded sharp bounds on the distribution function of the additive treatment effect contrast (which corresponds to a difference between two random variables). The bounds proposed in Fan and Park [2010] have gained widespread traction in the literature on causal inference and econometrics, in works such as Chiba 2017, Huang et al. 2017, Lu et al. 2018 and Mullahy 2018.

Building on Makarov [1982], Frank et al. [1987], and the results in our previous chapter, we first formulate sharp bounds on the distribution function of the difference of two random variables with fixed marginals; we show how these differ from those that have appeared previously. We further identify and address logical gaps in Williamson and Downs [1990] that have propagated to some of the later literature, leading to incorrect or imprecise statements. In particular, for the distribution function of the difference of two random variables, the lower bound proposed by Williamson and Downs [1990] is not sharp for measures that are not absolutely continuous with respect to the Lebesgue measure. We also identify an unnecessary exclusion in the argument given by Williamson and Downs [1990]. Then we apply the new bounds in the context of treatment effects and calculate the bounds in an illustrative example. The conditions on the joint distribution with fixed marginals are closely related to the literature on couplings of probability measures and contingency tables. We use a construction from the max-flow min-cut theorem [Ford Jr and Fulkerson, 1958] to generalize a finite version of Strassen's theorem in Koperberg [2022]. We illustrate with an example how to apply this theorem to determine a prediction interval (or set) that could be potentially valid for the individual treatment effect. Koperberg [2022] also provides important technical tools for us to bound the probability mass function of the individual treatment effect and establish the general extension of Strassen's Theorem for finite sets. Finally, although the cdf bounds apply to both continuous and discrete outcomes, a natural extension is to derive bounds on the probability mass function of the individual treatment effect when the outcome is ordinal. Individual treatment effects on ordinal outcomes have previously been studied in Lu et al. [2018]. In this chapter, we derive sharp bounds on the probability mass function of the ITE, which is a new result on best possible pmf bounds for the sum or difference of two random variables with known marginal distributions.

In Section 3.2, we give the resulting bounds for the difference of two random variables whose individual distribution functions are fixed. We then discuss the implications for some of the results given previously by Williamson and Downs [1990]. In Section 3.3, we set up the problem of causal inference and revisit the bounds in Fan and Park [2010]. We demonstrate with an example of discrete random variables that the previously stated bounds are not sharp. In section 3.4, we derive a new extension of Strassen's Theorem on finite sets using a

max-flox min-cut argument and use it to characterize conditions on marginal distributions for an ITE prediction interval to be valid. In Section 3.5, we provide general sharp bounds on the pmf of the ITE when the outcome is ordinal. This result is also generally applicable to determining the best possible pmf bounds for the sum or difference of two random variables with known marginal distributions.

3.2 Sharp bounds on the difference

Now we consider the closely related problem concerning bounds on the distribution of the difference of two random variables with fixed marginals.

As we will describe in the next section 3.3 below, bounds on the distribution function of the difference of two random variables with fixed marginals have been widely used for partial identification of individual causal effects (see Fan and Park 2010, Imbens and Menzel 2018, Firpo and Ridder 2019).

In this section, we will first focus on the theoretical results. Specifically, we will apply our previous results concerning best possible bounds on sums, but replacing one variable by its negation. Although these results are corollaries of the results in the previous chapter, we include them here to connect various results in the literature and for the convenience of readers.

Let X and Y be random variables with respective distribution functions F and G fixed. Let $\Delta = X - Y$ be the difference of random variables X, Y. Let $J_\Delta(\cdot)$ be the distribution function of Δ.

Theorem 3.2.1. *For any given value δ, best-possible bounds $\underline{J_\Delta}(\delta) \leq J_\Delta(\delta) \leq \overline{J_\Delta}(\delta)$ on*

$J_\Delta(\delta)$ are given by

$$\begin{aligned}\underline{J_\Delta}(\delta) &= \sup_{x-y=\delta} \max\{F(x) - P(Y < y), 0\} \\ &= \sup_{x-y=\delta} \max\{F(x) - G(y) + P(Y = y), 0\};\end{aligned}$$

$$\begin{aligned}\overline{J_\Delta}(\delta) &= 1 + \inf_{x-y=\delta} \min\{F(x) - P(Y < y), 0\} \\ &= 1 + \inf_{x-y=\delta} \min\{F(x) - G(y) + P(Y = y), 0\}.\end{aligned}$$

The bounds in Theorem 3.2.1 differ from those stated in Fan and Park [2010] and Williamson and Downs [1990] by the inclusion of the point mass $P(Y = y)$ term in both the upper and lower bounds. As a consequence, the lower bound reported by these authors is not best possible; see Theorem 3.2.3 and the comment below Theorem 3.2.3 for more details.

3.2.1 Proof of Theorem 3.2.1

The proof is a direct application of Theorem 2.2.6. Consider a new variable $Y' = -Y$ with cdf G'. Then from equation (2.1) and (2.5), for any δ, the bound on $P(\Delta \leq \delta) = P(X - Y \leq \delta) = P(X + Y' \leq \delta)$ is:

$$\begin{aligned}\underline{J_\Delta}(\delta) &= \sup_{x+y'=\delta} \max(F(x) + G'(y') - 1, 0), \\ \overline{J_\Delta}(\delta) &= 1 + \inf_{x+y'=\delta} \min(F(x) + G'(y') - 1, 0).\end{aligned}$$

Note that

$$\begin{aligned}G'(y') &= P(-Y \leq y') \\ &= P(Y \geq -y') \\ &= 1 - P(Y < -y') \\ &= 1 - G(-y') + P(Y = -y').\end{aligned}$$

Replace y' with $-y$ and $G'(y')$ with $1 - G(y) + P(Y = y)$, we get the best-possible bounds in Theorem 3.2.1.

Remark 3.2.2. *In the case where $P(Y = y) = 0$ for all y (for example, when the distribution function of Y is absolutely continuous with respect to the Lebesgue measure), we recover the best-possible bounds in Theorem 2 of Williamson and Downs [1990] and in Lemma 2.1 of Fan and Park [2010]. However, when G is not absolutely continuous, the bounds in Fan and Park [2010] and Williamson and Downs [1990] can be different from the bounds in Theorem 3.2.1 as the point mass $P(Y = y)$ can be nonzero for some $y \in \mathbb{R}$. In the Appendix B.1, we identify the issue in the proof of Theorem 2 in Williamson and Downs [1990]. Although we do not address it explicitly, the same issue applies to the Williamson and Downs [1990] bounds on the ratio of two random variables.*

3.2.2 Implications of the new bounds

At this point, Theorem 3.2.1 seems to imply that the lower bound in Williamson and Downs [1990] is valid but not necessarily best-possible and the upper bound might not be valid. The next theorem will establish that in fact of the upper bound on the cdf for the difference $\Delta = X - Y$ in Williamson and Downs [1990] is valid even though the proof used in Williamson and Downs [1990] is not correct.

Theorem 3.2.3. *For any random variables X and Y with respective cdfs $F(\cdot)$ and $G(\cdot)$,*

$$\inf_{x-y=\delta} \min\{F(x) - P(Y < y), 0\} = \inf_{x-y=\delta} \min\{F(x) - G(y), 0\}.$$

The proof of Theorem 3.2.3 is left to Appendix B.2. Theorem 3.2.3 implies that the upper bounds in Williamson and Downs [1990] and Fan and Park [2010] coincide with the upper bounds we proposed in Theorem 3.2.1. Since the lower bounds that we propose in Theorem 3.2.1 are greater than or equal to the lower bounds in Williamson and Downs [1990], Fan and Park [2010], Theorem 3.2.3 establishes that somewhat surprisingly all these bounds are valid, though the lower bounds are not sharp. In Section 3.3.1, we demonstrate

through an example that the lower bound proposed by Williamson and Downs [1990] and Fan and Park [2010] is not sharp.

Remark 3.2.4. *Theorem* 3 *in Williamson and Downs [1990] states an optimality result (analogous to Theorem 2.3.4 of the previous chapter) contains an unnecessary exclusion: specifically, it states that the bounds are only achievable if F and G are not both discontinuous at some x, y such that $x + y = z$. It appears that this additional unnecessary condition was added because Williamson and Downs fail to note that when F and G are discontinuous the bounds only take a strict subset of values in $[0, 1]$. Consequently, only a subset of values need to be considered.*

3.3 A causal perspective

Throughout this section, we consider a binary treatment $D = 0, 1$. Let Y_1 be the potential outcome were an individual to take the treatment and Y_0 be the potential outcome were an individual to take control. We assume the stable unit treatment value assumption (SUTVA, Rubin 1978) that there is a single version of each treatment/control and no interference among the subjects. We define our parameter of interest as the individual treatment effect: $\Delta = Y_1 - Y_0$. Fan and Park [2010] Lemma 2.1 stated the bounds on the distribution function of the individual treatment effect and claimed that they are sharp. We modify the bounds in Fan and Park [2010] based on Theorem 3.2.1. Let F_1, F_0 be the cumulative distribution function on Y_1, Y_0 respectively. Let $F_\Delta(\cdot)$ be the cdf for Δ.

Theorem 3.3.1. *For any given value δ, best-possible bounds on $F_\Delta(\delta)$ are given by*

$$\begin{aligned} F^L(\delta) &= \sup_y \max\{F_1(y) - P(Y_0 < y - \delta), 0\} \\ &= \sup_y \max\{F_1(y) - F_0(y - \delta) + P(Y_0 = y - \delta), 0\}; \end{aligned} \tag{3.1}$$

$$\begin{aligned} F^U(\delta) &= 1 + \inf_y \min\{F_1(y) - P(Y_0 < y - \delta), 0\} \\ &= 1 + \inf_y \min\{F_1(y) - F_0(y - \delta) + P(Y_0 = y - \delta), 0\}. \end{aligned} \tag{3.2}$$

Let Y denote the observed outcome variable. Under consistency, $Y = Y_0$ when $D = 0$ and $Y = Y_1$ when $D = 1$. In practice, if we are willing to assume ignorability (for example, in randomized clinical trials (RCTs)) or conditional ignorability, the marginal distributions $F_1(y)$ and $F_0(y)$ can be identified. Theorem 3.3.1 allows us to conclude best-possible bounds on the distribution function of the individual treatment effect. In the special case where Y is ordinal, Proposition 1 in Lu et al. [2018] can be recovered using Theorem 3.2.1 and Theorem 3.2.3. Lu et al. [2018] consider a special case where Y is non-negative and prove the bounds using a construction argument instead of the copula theory.

Corollary 3.3.2. *The Fan-Park upper bound is best-possible.*

Proof: This follows directly from Theorem 3.2.3 where X is replace by Y_1 and Y is replace by Y_0. □

3.3.1 Application of Theorem 3.3.1 *to cdf bounds for ITE*

Here we will present a simple example that applies our bounds in Theorem 3.3.1 and compare them with the bounds in Fan and Park [2010].

Consider the case where we have a binary treatment variable ($D = 0, 1$) and a ternary response ($Y = 0, 1, 2$). Under randomization, the relationship between the counterfactual distribution $P(Y_0, Y_1)$ and the observed distributions $\{P(Y \mid D = 0), P(Y \mid D = 1)\}$ is given by: $P(Y = i \mid D = j) = P(Y_j = i)$. Suppose we observe the marginals given in Table 3.1. We can parameterize the joint distribution with 4 parameters p, q, t, r. Note that by the Fréchet inequalities, $\max\{P(Y_0 = i) + P(Y_1 = j) - 1, 0\} \leq P(Y_0 = i, Y_1 = j) \leq \min\{P(Y_0 = i), P(Y_1 = j)\}$. We get ranges for p, q, t, r by applying the Fréchet inequalities to each quantity.

Based on the bounds proposed in (3.1) and (3.2) , we note the following alternative

	$P(Y=0 \mid D=0)=0.3$	$P(Y=1 \mid D=0)=0.2$	$P(Y=2 \mid D=0)=0.5$
$P(Y=0 \mid D=1)=0.7$	$P(Y=0 \mid D=1)-p-r$	$p \in [0, 0.2]$	$r \in [0.2, 0.5]$
$P(Y=1 \mid D=1)=0.1$	$P(Y=1 \mid D=1)-t-q$	$t \in [0, 0.1]$	$q \in [0, 0.1]$
$P(Y=2 \mid D=1)=0.2$	$1-(\ldots)$	$P(Y=1 \mid D=0)-t-p$	$P(Y=2 \mid D=0)-r-q$

Table 3.1: Application with binary treatment and ternary outcome.

expressions for $F^L(\delta)$ and $F^U(\delta)$:

$$F^L(\delta) = \max\left(\sup_y \{F_1(y) - P(Y_0 < y - \delta)\}, 0\right);$$
$$F^U(\delta) = 1 + \min\left(\inf_y \{F_1(y) - P(Y_0 < y - \delta)\}, 0\right).$$

Note that $\delta = -2$ is only possible when $Y_1 = 0, Y_0 = 2$. So this corresponds to the entry in the top right corner of Table 3.1. By the Fréchet inequalities, the bounds on $P(Y_1 = 0, Y_0 = 2)$ is given by $r \in [0.2, 0.5]$. Now consider $F_1(y) - P(Y_0 < y - \delta)$ in our example,

$$F_1(y) - P(Y_0 < y + 2) = \begin{cases} 0 & y \leq -2 \\ -0.3 & -2 < y \leq -1, \\ -0.5 & -1 < y < 0, \\ 0.2 & y = 0, \\ -0.3 & 0 < y < 1, \\ -0.2 & 1 \leq y < 2, \\ 0 & y \geq 2. \end{cases}$$

This gives $F_\Delta(-2) \in [0.2, 0.5]$, which matches the range given by Fréchet inequalities. In this case if we consider the bounds proposed in Lemma 2.1 in Fan and Park [2010], which

follow from Theorem 2 of Williamson and Downs [1990],

$$F_1(y) - F_0(y+2) = \begin{cases} 0 & y < -2, \\ -0.3 & -2 \leq y < -1, \\ -0.5 & -1 \leq y < 0, \\ -0.3 & 0 \leq y < 1, \\ -0.2 & 1 \leq y < 2, \\ 0 & y \geq 2. \end{cases}$$

The lower bounds for $F_\Delta(-2)$ is 0, which is not sharp. This example corresponds to the case where F_1 and F_0 are both discontinuous at $Y_1 = 0$ and $Y_0 = 2$.

As a second illustration of Theorem 3.3.1 in a context where the Fréchet inequalities are not directly applicable, consider the bounds for $F_\Delta(-1)$ given by Theorem 3.3.1. The individual treatment effect is less than or equal to -1 when $Y_1 = 0, Y_0 = 1$ or $Y_1 = 1, Y_0 = 2$ or $Y_1 = 0, Y_0 = 2$. Therefore, based on Table 3.1, $F_\Delta(-1) = p + q + r$. To calculate the bounds, consider $F_1(y) - P(Y_0 < y + 1)$ in our example,

$$F_1(y) - P(Y_0 < y+1) = \begin{cases} 0 & y \leq -1, \\ -0.3 & -1 < y < 0, \\ 0.4 & y = 0, \\ 0.2 & 0 < y < 1, \\ 0.3 & y = 1, \\ -0.2 & 1 < y < 2, \\ 0 & y \geq 2. \end{cases}$$

This gives bounds for $F_\Delta(-1) = p+q+r \in [0.4, 0.7]$. Again in this case, if we compute the lower bound based on Fan and Park [2010] and Williamson and Downs [1990], it is not sharp. Finally, we complete the example by obtaining the bounds $F_\Delta(0) = 1.2 + t - r \in [0.7, 1]$, $F_\Delta(1) = 1.5 - p - q - t - r \in [0.8, 1]$. $F_\Delta(2) = 1$ follows trivially by construction.

Inference on Makarov type bounds used in causal inference can be found in Fan and Park [2010] and is discussed in Fan and Park [2012] and Imbens and Menzel [2018].

Before we dive into the bounds on the probability mass function of individual treatment effect, we will first state a theorem that is going to be very useful in constructing the proof for the probability mass function bounds. We use a simplified version of Strassen's Theorem in Koperberg [2022]. We also extend this theorem to solve various problems including prediction intervals of ITE.

3.4 Coupling and Strassen's theorem

The existence of probability measures with prescribed marginals has been extensively studied in the coupling literature. The classic Strassen's theorem [Strassen, 1965] gives a necessary and sufficient condition for two distributions to have a coupling for the existence of a *perfect* coupling—one whose joint distribution is supported entirely on any given subset of the product space of the support. Strassen's original proof was analytical, but later authors noted it is equivalent to a max-flow/min-cut condition and even to Hall's marriage theorem when marginal supports are finite sets [Koperberg, 2024, Hsu, 2017]. The condition articulated by Strassen's theorem is often expressed through stochastic dominance on ordered spaces, which serves as a measure-theoretic analogue of Hall's condition [Lindvall, 2002, Den Hollander, 2012]. Strassen's result is fundamental in probability and optimal transport. In computer science, it appears as a criterion for liftings of distributions [Hsu, 2017]. Strassen's theorem underpins a wide range of applications, as seen in works such as Chernozhukov et al. [2014], Marshall et al. [1979].

Many applications, however, demand *partial* couplings that match the marginals only approximately, allowing a fraction of the probability mass to fall outside the target. Such relaxations arise, for example, in approximate liftings [Hsu, 2017] and partial optimal transport [Figalli, 2010, Chapel et al., 2020]. In the next section, we will focus on the case where the supports for the marginal distributions are on finite sets A and B. While the classical proof of Strassen's theorem relies on analytical tools such as the Prokhorov metric, Koperberg [2024] demonstrates that in the finite setting, a purely combinatorial proof is possible. Building on this perspective, we further advance the theory by deriving a finite set extension

of Strassen's theorem that characterizes *exactly* when two marginals admit a coupling that places α mass on an arbitrary relation $R \subseteq A \times B$ for any $\alpha \in [0, 1]$. The proof leverages a max–flow/min-cut construction that naturally generalizes the perfect-matching argument and clarifies the geometric structure of feasible couplings.

Strassen's theorem is closely related to causal inference problems (see Fan and Park 2010, Song et al. 2024), where the marginal distributions are observed or estimated. We will later demonstrate how our generalization can be employed to construct prediction intervals for individual treatment effects in cases where the outcome is ordinal, or to bound the proportion of individuals whose treatment effect lies within a specified range. Another natural application arises in the context of the Komogorov's problem: given prescribed marginal distributions F and G for random variables X, Y respectively, characterize the set of compatible distribution functions for the sum. This problem is studied in Makarov [1982], Rüschendorf [1982], Frank et al. [1987], Zhang and Richardson [2024]. We show that our result can be used to approximate the solution for an extension of Komogorov's problem, which is to characterize the max (or min) probability mass for sum of two random variables within certain range given the marginals. Our result can also be used to bound probability mass functions with given marginals, which extends a classical result in maximal and ordered couplings.

3.4.1 Strassen's theorem, monotonicity and prediction intervals

Let A and B be sets and $R \subseteq A \times B$ a relation. Then for each $\mathrm{U} \subseteq A$ the set of neighbours of U in R is denoted by

$$\mathcal{N}_{\mathrm{R}}(\mathrm{U}) = \{\mathrm{b} \in \mathrm{B} : (\mathrm{U} \times \{\mathrm{b}\}) \cap \mathrm{R} \neq \varnothing\}.$$

Theorem 3.4.1 (Strassen's theorem for finite sets)**.** *Let* A *and* B *be finite sets,* P *and* P' *probability measures on* A *and* B *respectively and* $\mathrm{R} \subseteq A \times \mathrm{B}$ *a relation. Then there exists a coupling* $\hat{\mathrm{P}}$ *of* P *and* P' *that satisfies* $\hat{\mathrm{P}}(\mathrm{R}) = 1$ *if and only if*

$$\mathrm{P}(\mathrm{U}) \leqslant \mathrm{P}'\left(\mathcal{N}_{\mathrm{R}}(\mathrm{U})\right), \text{ for all } \mathrm{U} \subseteq A.$$

3.4.2 Characterize the marginal distributions under monotonicity

Although this thesis focuses on the case where no assumptions are made concerning the joint distribution of potential ourcomes, in some settings it may be known that treatment has non-negative effect

$$P(Y_0 \leq Y_1) = 1. \tag{3.3}$$

One use case of Theorem 3.4.1 is that it allows us to fully characterize the marginal distributions under monotonicity condition (stochastic monotonicity). Consider the case where we have a binary treatment variable $\{0, 1\}$ and a ternary response $\{0, 1, 2\}$.

	$P(Y=0 \mid D=0)$	$P(Y=1 \mid D=0)$	$P(Y=2 \mid D=0)$
$P(Y=0 \mid D=1)$		0	0
$P(Y=1 \mid D=1)$			0
$P(Y=2 \mid D=1)$			

We want to know what is the possible marginal distribution that corresponds to no one is hurt in the experiment. This means we are restricting the probabilities in the upper right corner to be 0. Consider probability measures on Y_0 and Y_1 with $A = B = \{0, 1, 2\}$ and $R = \{(0, 0), (0, 1), (0, 2), (1, 1), (1, 2), (2, 2)\}$. Applying Strassen's theorem, we get the following non-trivial constraints where $N_R(U) \neq B$:

$$P(Y=1 \mid D=0) \leq P(Y=1 \mid D=1) + P(Y=2 \mid D=1); \tag{3.4}$$

$$P(Y=2 \mid D=0) \leq P(Y=2 \mid D=1); \tag{3.5}$$

$$P(Y=1 \mid D=0) + P(Y=2 \mid D=0) \leq P(Y=1 \mid D=1) + P(Y=2 \mid D=1). \tag{3.6}$$

This further simplifies as:

$$P(Y=1 \mid D=0) \leq (P(Y=1 \mid D=1) + P(Y=2 \mid D=1); \tag{3.7}$$

$$P(Y=1 \mid D=0) + P(Y=2 \mid D=0) \leq P(Y=1 \mid D=1) + P(Y=2 \mid D=1). \tag{3.8}$$

3.4.3 Extension of Strassen's Theorem

We will first state a new theorem that extends Theorem 3.4.1 to characterize conditions on marginal distributions so that there exists a coupling with joint measures that place exactly $1-\alpha$ on any relation R. For $U \subseteq A$, we define subset $\mathcal{C}_R(U) \subseteq B$:

$$\mathcal{C}_R(U) = \{b \in B : (U \times \{b\}) \cap R \neq (U \times \{b\})\}$$

which is equivalently the subset consists of elements $b \in B$ such that there exists $a \in U$ with $(a, b) \notin R$.

Now we state an extension of Strassen's Theorem for finite sets:

Theorem 3.4.2 (Extension of Strassen's theorem)**.** *Let A and B be finite sets, P and P' probability measures on A and B respectively and $R \subseteq A \times B$ a relation. Then there exists a coupling $\hat{\mathrm{P}}$ of P and P' that satisfies $\hat{\mathrm{P}}(R) = 1-\alpha$ if and only if*

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) \geq \alpha, \text{ for all } U \subseteq A,$$

and

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq 1-\alpha, \text{ for all } U \subseteq A. \tag{3.9}$$

Proof. ($\Rightarrow$) If there exists a coupling $\hat{\mathrm{P}}$ of P and P' that satisfies $\hat{\mathrm{P}}(R) = 1-\alpha$, then mass α is assigned to the set $R^C := \{(a, b) : (a, b) \notin R\}$ by $\hat{\mathrm{P}}$. In particular, for any $U \subseteq A$, we define $T_U := B \setminus \mathcal{C}_R(U)$. By the definition of $\mathcal{C}_R(U)$, if $b \notin \mathcal{C}_R(U)$, then (a, b) belong to R for all $a \in U$. Hence in particular, $U \times T_U \subseteq R$. This gives:

$$\hat{\mathrm{P}}(U \times T_U) \leq \hat{\mathrm{P}}(R) = 1-\alpha.$$

Let V be the complement of $U \times T_U \subset A \times B$, we have:

$$\hat{\mathrm{P}}\big(V\big) = 1 - \hat{\mathrm{P}}(U \times T_U) \ \geq \ \alpha.$$

From the marginal probability, we also have

$$\hat{\mathrm{P}}(V) \leq \mathrm{P}(A \setminus U) + P'(\mathcal{C}_R(U)) = [1 - \mathrm{P}(U)] + \mathrm{P}'(\mathcal{C}_R(U)).$$

Putting it all together:

$$\alpha \leq \hat{\mathrm{P}}(V) \leq [1 - \mathrm{P}(U)] + \mathrm{P}'(\mathcal{C}_R(U)).$$

Therefore, $1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) \geq \alpha$, for all $U \subseteq A$. To show $1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq 1 - \alpha$, for all $U \subseteq A$, we will show that if equation (3.9) does not hold such that there exists $U \subseteq A$ with $P(U) > \mathrm{P}'\left(\mathcal{N}_R(U)\right) + \alpha$, then there is a contradiction. First, $\mathcal{N}_R(U) \neq B$ otherwise $\mathrm{P}'\left(\mathcal{N}_R(U)\right) = 1$ which leads to a contradiction since $P(U)$ cannot exceed 1 for any $U \subseteq A$. Then let

$$\mathcal{N}_R^C(U) := \{b \in B : (U \times \{b\}) \cap R = \emptyset\}$$

be the complement of the set $\mathcal{N}_R(U)$ in B. Then on one hand, because $U \times \mathcal{N}_R^C(U)$ is a subset of R^C, we have

$$\hat{\mathrm{P}}((U \times \mathcal{N}_R^C(U))) \leq \hat{\mathrm{P}}(R^C) = \alpha.$$

On the other hand, by Fréchet inequality, we also have

$$\begin{aligned}
\hat{\mathrm{P}}((U \times \mathcal{N}_R^C(U))) &\geq \max\{P(U) + P'(\mathcal{N}_R^C(U)) - 1, 0\} && (3.10)\\
&= \max\{P(U) + (1 - P'(\mathcal{N}_R(U))) - 1, 0\} && (3.11)\\
&= \max\{P(U) - P'(\mathcal{N}_R(U)), 0\} > \alpha, && (3.12)
\end{aligned}$$

which is a contradiction. Therefore, the existence of a coupling $\hat{\mathrm{P}}$ of P and P' that satisfies

$\hat{\mathrm{P}}(R) = 1 - \alpha$ implies that

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq 1 - \alpha, \text{ for all } U \subseteq A.$$

($\Leftarrow$) Now we assume

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) \geq \alpha, \text{ for all } U \subseteq A,$$

and

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq 1 - \alpha, \text{ for all } U \subseteq A.$$

We will prove that there exists a coupling $\hat{\mathrm{P}}$ with $\hat{\mathrm{P}}(R^c) = \alpha$ via a max-flow min-cut argument [Ford Jr and Fulkerson, 1958]. Note that this can be viewed as a generalization of Hall's Marriage theorem proved in Koperberg [2022]. We will first show that there exists a coupling $\hat{\mathrm{P}}_\alpha$ that put at least α mass in R^C and then show that there exists a coupling $\hat{\mathrm{P}}_{1-\alpha}$ put at least $(1 - \alpha)$ mass in R using a similar argument.

Build a bipartite network:

- Create a source node s, one node for each $a \in A$ (left side), one node for each $b \in B$ (right side), and a sink node t.
- Add edges $(s \to a)$ of capacity $\mathrm{P}(a)$ for each $a \in A$.
- Add edges $(b \to t)$ of capacity $\mathrm{P}'(b)$ for each $b \in B$.
- For each pair $(a, b) \notin R$ (i.e. $(a, b) \in R^c$), add an *infinite*-capacity edge $(a \to b)$.

By the max-flow min-cut theorem, there exists a flow of size greater than or equal to α from s to t if and only if every cut in the network has capacity $\geq \alpha$. A cut is described by choosing a subset $U \subseteq A$ to remain with s (and a corresponding subset of B forced to remain with t), so as not to cross any infinite-capacity edges $(a \to b)$ with $a \in U$ and $(a, b) \notin R$. But these infinite edges precisely mean we must include all $b \in B$ for which

$(a, b) \notin R$ for *some* $a \in U$, i.e. all $b \in \mathcal{C}_R(U)$. Hence the capacity of that cut becomes

$$\Big[\text{capacity from } s \text{ to } (A \setminus U)\Big] \; + \; \Big[\text{capacity from } \mathcal{C}_R(U) \text{ to } t\Big].$$

By construction, that is

$$[1 - \mathrm{P}(U)] + \mathrm{P}'(\mathcal{C}_R(U)),$$

which by hypothesis is always $\geq \alpha$. Hence no cut has capacity $< \alpha$, so the maximum flow of this graph is greater than or equal to α.

Now we will show that if a flow of capacity greater than or equal to α exists in this graph, then there exists a joint probability distribution $\hat{\mathrm{P}}_\alpha$ on $A \times B$ such that $\hat{\mathrm{P}}_\alpha(R^C) > \alpha$. A flow f of capacity k going from s to t in this network can be seen as assigning k total mass to the set of pairs $(a, b) \notin R$. Because of the capacity constraints on $(s \to a)$ and $(b \to t)$, we ensure

$$\sum_{b \in B} f(a, b) \leq \mathrm{P}(a), \quad \sum_{a \in A} f(a, b) \leq \mathrm{P}'(b).$$

First, we define the remainder marginals. For each $a \in A$, let

$$\mathrm{P}_{\mathrm{rem}}(a) = \mathrm{P}(a) - \sum_{b \in B} f(a, b),$$

and for each $b \in B$, let

$$\mathrm{P}'_{\mathrm{rem}}(b) = \mathrm{P}'(b) - \sum_{a \in A} f(a, b).$$

Following the previous argument, $\mathrm{P}_{\mathrm{rem}}(a) \geq 0$ and $\mathrm{P}'_{\mathrm{rem}}(b) \geq 0$. Observe that

$$\sum_{a \in A} \mathrm{P}_{\mathrm{rem}}(a) = \sum_{a \in A} \Big(\mathrm{P}(a) - \sum_{b \in B} f(a, b) \Big) = 1 - v,$$

for flow capacity $v > \alpha$ and similarly

$$\sum_{b \in B} \mathrm{P}'_{\mathrm{rem}}(b) = 1 - v.$$

Hence the remainder marginals match and sum to $1 - v$. Clearly there exists contin-

gency tables with given reminder marginals (for example we can take independent products $\hat{\mathrm{P}}_{\mathrm{rem}}(a,b) = \mathrm{P}_{\mathrm{rem}}(a)\mathrm{P}'_{\mathrm{rem}}(b)$ and scale each entry by $\frac{1}{1-v}$). This relates to the literature on the Gale–Ryser theorem and contingency tables; see Gale [1957], Ryser [1957], Joe [1988], Dobra and Fienberg [2000], and Ford and Fulkerson [2015] for further details. We can construct the final joint distribution by defining

$$\hat{\mathrm{P}}_{\alpha}(a,b) = f(a,b) + \hat{\mathrm{P}}_{\mathrm{rem}}(a,b).$$

It is easy to verify that f and $\hat{\mathrm{P}}_{\mathrm{rem}}$ are both nonnegative, and their row/column sums add up to $\mathrm{P}(a)$ and $\mathrm{P}'(b)$ respectively. Thus, there exists a coupling of P and P' put at least α probability mass in R^c. Similarly, we can build the graph with R using:

- Create a source node s, one node for each $a \in A$ (left side), one node for each $b \in B$ (right side), and a sink node t.
- Add edges $(s \to a)$ of capacity $\mathrm{P}(a)$ for each $a \in A$.
- Add edges $(b \to t)$ of capacity $\mathrm{P}'(b)$ for each $b \in B$.
- For each pair $(a,b) \in R$, add an *infinite*-capacity edge $(a \to b)$.

And following the same argument, the capacity of any cut becomes

$$[\text{capacity from } s \text{ to } (A \setminus U)] + [\text{capacity from } \mathcal{N}_R(U) \text{ to } t],$$

which implies that

$$[1 - \mathrm{P}(U)] + \mathrm{P}'(\mathcal{N}_R(U))$$

is always $\geq 1 - \alpha$. Thus, there exists a coupling that put at least $1 - \alpha$ probability mass in R.

Now let $\hat{\mathrm{P}}_{\alpha}$ be a joint distribution that put greater than or equal to α mass in R^C and $\hat{\mathrm{P}}_{1-\alpha}$ be a joint distribution that put greater than or equal to $1-\alpha$ mass in R. For $t \in [0,1]$,

we get a convex combination of $\hat{\mathrm{P}}_\alpha$ and $\hat{\mathrm{P}}_{1-\alpha}$ using

$$\hat{\mathrm{P}}_t = t\hat{\mathrm{P}}_\alpha + (1-t)\hat{\mathrm{P}}_{1-\alpha},$$

which is also a valid joint distribution. By choosing t appropriately, we can make $\hat{\mathrm{P}}_t$ such that $\hat{\mathrm{P}}_t = \hat{\mathrm{P}}$ with $\hat{\mathrm{P}}(R^c) = \alpha$ and $\hat{\mathrm{P}}(R) = 1-\alpha$. This completes the proof. □

Note that to construct the joint probability $\hat{P}$, it is equivalent to construct a partial flow with capacity exactly α, which can also be done using algorithms like Ford–Fulkerson algorithm with early stopping.

3.4.4 Characterize the marginal distributions for prediction intervals

Consider a similar example as in Section 3.4.2 where we have a binary treatment variable $\{0,1\}$ and a ternary response $\{0,1,2\}$. Let $A = B = \{0,1,2\}$. We further let $R = \{(0,0),(0,1),(0,2),(1,1),(1,2),(2,2)\}$.

	$P(Y=0 \mid D=0)$	$P(Y=1 \mid D=0)$	$P(Y=2 \mid D=0)$
$P(Y=0 \mid D=1)$		R^C	R^C
$P(Y=1 \mid D=1)$			R^C
$P(Y=2 \mid D=1)$			

Now instead of making the monotonicity assumption, we want to know what is the marginal distribution that corresponds to a possible non-negative 95% prediction interval without further assumptions. This means that there exists a coupling $\hat{\mathrm{P}}$ that satisfies $\hat{\mathrm{P}}(R) = 1-\alpha$.

Applying Theorem 3.4.2 with $\alpha = 0.05$, we get the following non-trivial constraints:

$$1 - P(Y=0 \mid D=0) \geq \alpha;$$
$$1 - P(Y=0 \mid D=0) - P(Y=1 \mid D=0) + P(Y=0 \mid D=1) \geq \alpha;$$
$$P(Y=0 \mid D=1) + P(Y=1 \mid D=1) \geq \alpha;$$
$$P(Y=2 \mid D=0) \leq P(Y=2 \mid D=1) + \alpha;$$
$$P(Y=1 \mid D=0) + P(Y=2 \mid D=0) \leq P(Y=1 \mid D=1) + P(Y=2 \mid D=1) + \alpha.$$

These are necessary conditions for a non-negative 95% prediction interval to be potentially valid, as we will discuss more in detail in the next chapter.

3.4.5 Bounding joint probabilities with marginal constraints

Theorem 3.4.2 implies that given a relation R, there exists a coupling $\hat{\mathrm{P}}$ of the marginals satisfying $\hat{\mathrm{P}}(R) = \alpha$ if and only if

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) \geq 1 - \alpha, \text{ for all } U \subseteq A,$$

and

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq \alpha, \text{ for all } U \subseteq A.$$

The next corollary will use this result to give sharp bounds on the joint probability with given marginals.

Corollary 3.4.3. *Let A and B be finite sets, P and P' probability measures on A and B respectively and $R \subseteq A \times B$ a relation. For any coupling $\hat{\mathrm{P}}$ of P and P', we have:*

$$\max\{0, 1 - \alpha_1\} \leq \hat{\mathrm{P}}(R) \leq \min\{1, \alpha_2\}$$

where

$$\alpha_1 = \min_{U \subseteq A} \left\{1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right)\right\};$$
$$\alpha_2 = \min_{U \subseteq A} \left\{1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right)\right\}.$$

Note that for any $U \subseteq A$, $\mathcal{C}_R(U) \cup \mathcal{N}_R(U) = B$. Therefore, $\mathrm{P}'\left(\mathcal{C}_R(U)\right) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq 1$ for all $U \subseteq A$. This implies that for all $U \subseteq A$,

$$1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) + 1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) = 2(1 - P(U)) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) + \mathrm{P}'\left(\mathcal{N}_R(U)\right) \geq 1$$

Therefore,

$$\begin{aligned}
\alpha_1 + \alpha_2 &= \min_{U \subseteq A} \left\{1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right)\right\} + \min_{U \subseteq A} \left\{1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right)\right\} \\
&\geq \min_{U \subseteq A} \{1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{C}_R(U)\right) + 1 - \mathrm{P}(U) + \mathrm{P}'\left(\mathcal{N}_R(U)\right)\} \\
&\geq 1
\end{aligned}$$

Hence $1 - \alpha_1 \leq \alpha_2$, the lower bound is always less than or equal to the upper bound. Furthermore, the bounds are achievable in the sense that there exists a coupling for which the probability mass assigned to the relation R attains the specified lower or upper bound. This result has several implications. For example, when the marginal supports are finite sets, the classical Makarov bounds [Makarov, 1982, Rüschendorf, 1982] are attainable. For further discussion, see [Zhang and Richardson, 2024]. In the subsequent sections, we demonstrate how Corollary 3.4.3 can be applied to a range of problems across probability theory and causal inference.

3.4.6 Application 1. Bounds on the pmf of the difference of two random variables with given marginals

Consider two discrete outcomes X, Y both having support on $\{0, 1, 2\}$. Given their marginal distributions, suppose that we want to find the range of possible values for $P(X - Y = 0)$

over all joint distributions of X, Y compatible with the given marginals.[1] Then we take the relation

$$R = \{(0,0),(1,1),(2,2)\} \subseteq A \times B,$$

Let P and P' denote the marginal distributions of X and Y, respectively. The marginal distributions are given in the table.

	$\Pr(X=0)=0.2$	$\Pr(X=1)=0.5$	$\Pr(X=2)=0.3$
$\Pr(Y=0)=0.1$	R		
$\Pr(Y=1)=0.3$		R	
$\Pr(Y=2)=0.6$			R

Corollary 3.4.3 gives sharp bounds on

$$\Pr(X - Y = 0) = \hat{P}(R)$$

namely

$$\max\{0, 1-\alpha_1\} \;\leq\; \hat{P}(R) \;\leq\; \min\{1, \alpha_2\},$$

where

$$\alpha_1 = \min_{U \subseteq A}\Big\{1 - P(U) + P'\big(\mathcal{C}_R(U)\big)\Big\}, \qquad \alpha_2 = \min_{U \subseteq A}\Big\{1 - P(U) + P'\big(\mathcal{N}_R(U)\big)\Big\}.$$

First, we will compute α_1 in this example. For this relation R and any $U \subseteq \{0,1,2\}$,

$$\mathcal{C}_R(U) = \begin{cases} \emptyset & \text{if } U = \emptyset \\ B \setminus U & \text{if } U \text{ is a singleton set} \\ B & \text{otherwise} \end{cases}$$

[1] We illustrate the approach by considering the probability $P(X-Y=0)$ as a representative example. By appropriately specifying the relation R, we can also derive sharp bounds on the probability mass function of the sum of two random variables with fixed marginals. The choice of 0 is without loss of generality; the method applies to any value. In causal inference, the quantity $P(Y_1 - Y_0 = 0)$ represents the probability that the individual treatment effect (ITE) is zero. In probability theory, the upper bound of $P(X=Y)$ corresponds to what is known as the maximal coupling.

When $\mathcal{C}_R(U) = B$, $P'(\mathcal{C}_R(U)) = 1$ and $1 - P(U) + P'(\mathcal{C}_R(U)) \geq 1$. When $\mathcal{C}_R(U) = \emptyset$, $1 - P(U) + P'(\mathcal{C}_R(U)) = 1$. For singleton set U, $P'(\mathcal{C}_R(U)) = P'\big(B \setminus U\big) = 1 - P'(U)$, we have

$$1 - P(U) + P'(\mathcal{C}_R(U)) = 1 - P(U) + 1 - P'(U). \tag{3.13}$$

Based on this marginal distribution, there is no element $u \subset \{0, 1, 2\}$ such that $P(u) + P'(u) \geq 1$, therefore, the quantity in 3.13 is greater than or equal to 1. Hence $\alpha_1 \geq 1$ and the lower bound is

$$\max\{0, 1 - \alpha_1\} = 0.$$

Now we will compute α_2. For the relation R, $\mathcal{N}_R(U) = \{b \in B : \exists a \in U,\ a = b\} = U$, so that $P'\big(\mathcal{N}_R(U)\big) = P'(U)$. Enumerating all subsets of A yields

U	$1 - P(U) + P'(U)$
$\varnothing$	1
$\{0\}$	$1 - 0.2 + 0.1 = 0.9$
$\{1\}$	$1 - 0.5 + 0.3 = 0.8$
$\{2\}$	$1 - 0.3 + 0.6 = 1.3$
$\{0, 1\}$	$1 - 0.7 + 0.4 = \mathbf{0.7}$
$\{0, 2\}$	$1 - 0.5 + 0.7 = 1.2$
$\{1, 2\}$	$1 - 0.8 + 0.9 = 1.1$
A	$1 - 1 + 1 = 1$

The minimum is attained at $U = \{0, 1\}$, giving $\alpha_2 = 0.70$.

Combining the two steps,

$$0 \ \leq \ \hat{P}(R) \ \leq \ 0.70.$$

Equivalently, the probability that $P(X - Y = 0)$ is at most 70% under *any* joint distribution consistent with the stated marginals—and it could be as small as zero.

3.4.7 Application 2. Characterize the set of patients with treatment effect in a certain range

We again consider the example in Section 3.4.2 with a binary treatment and an oridinal outcome $Y \in \{0, 1, 2\}$. Suppose our goal is to characterize the set of patients for whom the absolute individual treatment effect satisfies $|Y_1 - Y_0| \leq 1$. This example is similar to the previous application but we now put into context of causal inference. We define the relation R as the event that the absolute individual treatment effect (ITE) does not exceed one unit. Formally, let

$$R = \big\{(a, b) \in \{0, 1, 2\}^2 : |a - b| \leq 1\big\}.$$

The cells in the table below indicate the relation R.

	$\Pr(Y_0{=}0) = 0.2$	$\Pr(Y_0{=}1) = 0.5$	$\Pr(Y_0{=}2) = 0.3$
$\Pr(Y_1{=}0) = 0.1$	R	R	
$\Pr(Y_1{=}1) = 0.3$	R	R	R
$\Pr(Y_1{=}2) = 0.6$		R	R

Let P and P' denote the marginals of Y_0 and Y_1, respectively. As before, we have $\Pr(|Y_1 - Y_0| \leq 1) = \hat{P}(R)$. Corollary 3.4.3 yields sharp bounds

$$\max\{0, 1 - \alpha_1\} \;\leq\; \hat{P}(R) \;\leq\; \min\{1, \alpha_2\},$$

where

$$\alpha_1 = \min_{U \subseteq A}\Big\{1 - P(U) + P'\big(\mathcal{C}_R(U)\big)\Big\},$$
$$\alpha_2 = \min_{U \subseteq A}\Big\{1 - P(U) + P'\big(\mathcal{N}_R(U)\big)\Big\}.$$

First, we will compute α_1. Enumerating all subsets $U \subseteq A$, we get:

U	$\mathcal{C}_R(U)$	$1 - P(U) + P'\big(\mathcal{C}_R(U)\big)$
$\varnothing$	$\varnothing$	1
$\{0\}$	$\{2\}$	$1 - 0.2 + 0.6 = 1.4$
$\{1\}$	$\varnothing$	$1 - 0.5 + 0 = 0.5$
$\{2\}$	$\{0\}$	$1 - 0.3 + 0.1 = 0.8$
$\{0,1\}$	$\{2\}$	$1 - 0.7 + 0.6 = 0.9$
$\{0,2\}$	$\{0,2\}$	$1 - 0.5 + 0.7 = 1.2$
$\{1,2\}$	$\{0\}$	$\mathbf{1 - 0.8 + 0.1 = 0.3}$
A	$\{0,2\}$	$1 - 1 + 0.7 = 0.7$

The minimum is attained at $U = \{1, 2\}$, giving $\alpha_1 = 0.3$.

Now we will compute α_2. Similarly, enumerating subsets of A:

U	$\mathcal{N}_R(U)$	$1 - P(U) + P'\big(\mathcal{N}_R(U)\big)$
$\varnothing$	$\varnothing$	$\mathbf{1}$
$\{0\}$	$\{0,1\}$	$1 - 0.2 + 0.4 = 1.2$
$\{1\}$	$\{0,1,2\}$	$1 - 0.5 + 1 = 1.5$
$\{2\}$	$\{1,2\}$	$1 - 0.3 + 0.9 = 1.6$
$\{0,1\}$	$\{0,1,2\}$	$1 - 0.7 + 1 = 1.3$
$\{0,2\}$	$\{0,1,2\}$	$1 - 0.5 + 1 = 1.5$
$\{1,2\}$	$\{0,1,2\}$	$1 - 0.8 + 1 = 1.2$
A	$\{0,1,2\}$	$1 - 1 + 1 = 1$

The minimum value is $\alpha_2 = 1$ (attained at $U = \varnothing$ and $U = A$). Combining the α_1 and α_2 we calculated, we get

$$0.7 \leq \hat{P}\big(|Y_1 - Y_0| \leq 1\big) \leq 1$$

Thus, under any joint distribution consistent with the stated marginals, at least 70% of units experience an absolute treatment effect of 1 or less, and this proportion could be as high as 100%.

3.4.8 Application 3. Bounds on $X + Y \in [a, b]$: a geometric view and an approximation algorithm

Now we will see an example to apply corollary 3.4.3 in a non-discrete setting. We will investigate the following problem: let F and G denote the prescribed marginal cdfs of two real–valued random variables X and Y, for fixed constants $a < b$, find the maximum and minimum probability of $Z = X + Y$ in a given interval $[a, b]$ over all possible joint distributions satisfying the marginals F and G. The problem investigated herein extends the classical Kolmogorov problem concerning bounds for distribution functions with fixed marginals Makarov [1982], Rüschendorf [1982], Frank et al. [1987], Zhang and Richardson [2024]. Specifically, the original Kolmogorov problem can be viewed as the special case where the threshold a is no larger than the infimum of the support of the sum $X + Y$. To the best of our knowledge, a solution for this general case has not previously been derived.

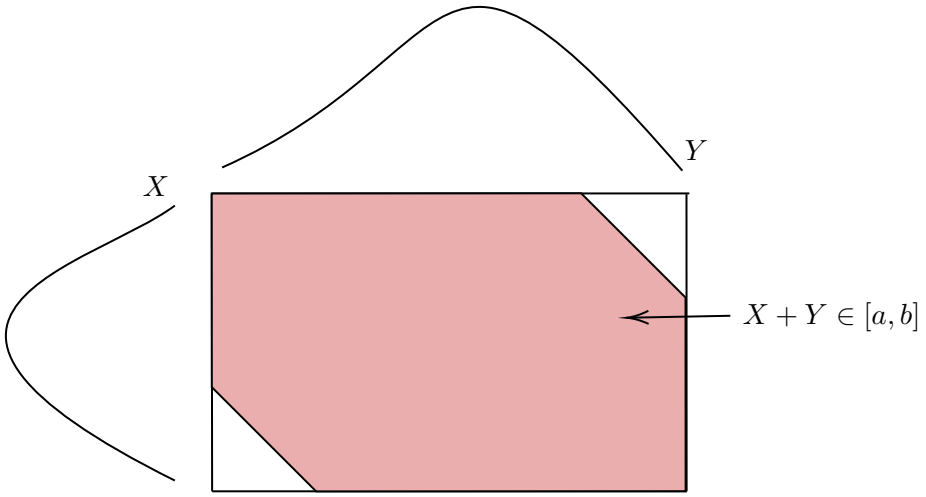

Figure 3.1: Geometry of the strip $X + Y \in [a, b]$ under arbitrary marginals F and G. The diagonal band is the region whose mass we wish to bound.

Figure 3.1 shows the strip of $X + Y \in [a, b]$ (red) in the (x, y)–plane. The upper curve sketches the density of Y (for illustration purposes we show a unimodal density, but the

result does not rely on such assumption). Intuitively, the worst–case coupling that minimizes the mass pushes probability mass away from the strip, whereas the coupling that maximizes the mass will squeeze $X + Y$ into the strip as much as possible. The idea is that we can approximate the strip using discrete states and then apply Corollary 3.4.3 to calculate the bounds. We show this intuition in Figure 3.2. With approximately chosen R, the desired probability $P(X + Y \in [a, b]) \approx P(X + Y \in R)$. Corollary 3.4.3 immediately gives

$$\max\{0, 1 - \alpha_1\} \leq P(X + Y \in R) \leq \min\{1, \alpha_2\}, \tag{3.14}$$

where $\alpha_1, \alpha_2, \mathcal{C}_R(U)$ and $\mathcal{N}_R(U)$ can be calculated the same way as before because we have discretized the continuous marginals with R.

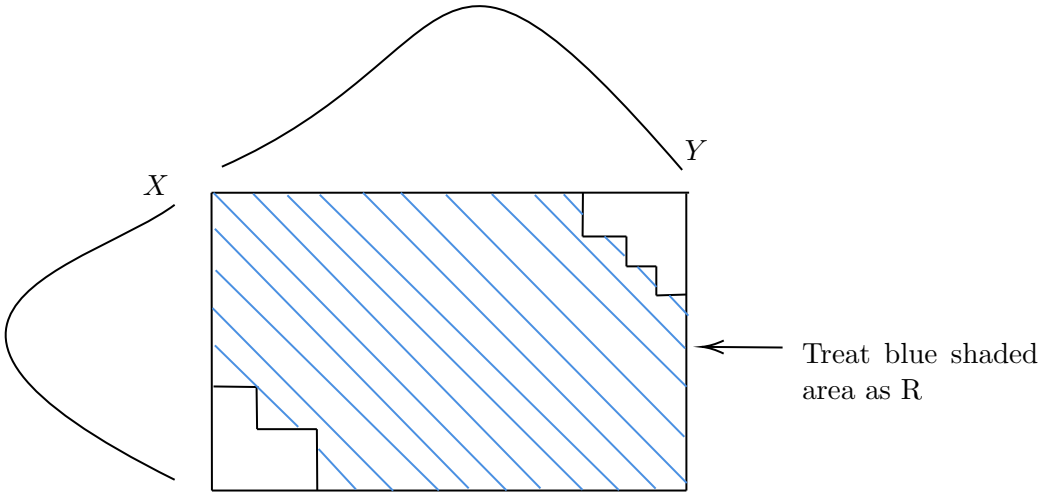

Figure 3.2: Demonstration of R that is used to approximate the strip $X + Y \in [a, b]$.

Formally, given $[a, b]$, we will construct a grid. Let $x_1 < \cdots < x_m$ and $y_1 < \cdots < y_n$ be grid points and set $\Delta F_i = F(x_i) - F(x_{i-1})$, $\Delta G_j = G(y_j) - G(y_{j-1})$ with $x_0 = -\infty$, $y_0 = -\infty$. $\Delta F_i, \Delta G_j$ can be viewed as marginal probabilities for cell (i, j). Let R be the relationship $R = \{(i, j) : [x_{i-1} + y_{i-1}, x_i + y_i] \subseteq [a, b]\}$. Applying Corollary 3.4.3 will provide

upper and lower bounds on $P(X + Y \in R)$. The grid can be refined adaptively until the bounds are sufficiently tight.

3.5 Sharp bounds on the pmf of Individual Treatment Effects

Now we get back to the question: given the marginals of Y_1, Y_0, can we derive sharp bounds on the probability mass function (pmf) of ITE, for example, what are the bounds for $P(\text{ITE} = 0)$? This is given by adding up the Fréchet inequality bounds for each $P(Y_1 = i, Y_0 = i - \delta)$.

Theorem 3.5.1. *For discrete random variables Y_1, Y_0 with given marginals and a given δ, let $[L_i, U_i] = [\max\{P(Y_1 = i) + P(Y_0 = i - \delta) - 1, 0\}, \min\{P(Y_1 = i), P(Y_0 = i - \delta)\}]$ be the Fréchet inequality bounds for the joint probability $P(Y_1 = i, Y_0 = i - \delta)$. We claim that*

$$P(Y_1 - Y_0 = \delta) \in [\sum_i L_i, \sum_i U_i]. \tag{3.15}$$

Furthermore, the bounds in (3.15) are sharp as there exists joint distributions of Y_1, Y_0 that satisfies the given marginals and achieves the upper or lower bound on $P(Y_1 - Y_0 = \delta)$.

3.5.1 Proof of Theorem 3.5.1

First, we show the bound in (3.15) is a valid bound. Notice

$$P(Y_1 - Y_0 = \delta) = \sum_i P(Y_1 = i, Y_0 = i - \delta). \tag{3.16}$$

And for all i, Fréchet inequality bounds states that

$$P(Y_1 = i, Y_0 = i - \delta) \in [L_i, U_i], \tag{3.17}$$

where

$$[L_i, U_i] = [\max\{P(Y_1 = i) + P(Y_0 = i - \delta) - 1, 0\}, \min\{P(Y_1 = i), P(Y_0 = i - \delta)\}]. \tag{3.18}$$

So the sum must be within the bound in (3.15).

Now, we will show that the bounds in (3.15) are tight in a sense that there exist joint distributions compatible with the marginal conditions that achieve the lower/upper bounds. We first start with a proposition.

Proposition 3.5.2. *If*

$$\max\{P(Y_1 = i) + P(Y_0 = j) - 1, 0\} > 0 \tag{3.19}$$

$$\textit{and} \quad \max\{P(Y_1 = k) + P(Y_0 = l) - 1, 0\} > 0 \tag{3.20}$$

then we have either $i = k$ or $j = l$. In other words, at most one Fréchet inequality lower bound can be non-zero if $i \neq k$ and $j \neq l$.

Proof. Suppose, for a contradiction that $i \neq j$, $k \neq l$ and

$$P(Y_1 = i) + P(Y_0 = j) - 1 > 0; \tag{3.21}$$

$$P(Y_1 = k) + P(Y_0 = l) - 1 > 0. \tag{3.22}$$

Summing up (3.21) and (3.22), we get

$$P(Y_1 = i) + P(Y_0 = j) + P(Y_1 = k) + P(Y_0 = l) - 2 > 0. \tag{3.23}$$

However, we also have

$$P(Y_1 = i) + P(Y_0 = j) + P(Y_1 = k) + P(Y_0 = l) \tag{3.24}$$

$$\leq \sum_i P(Y_1 = i) + \sum_j P(Y_0 = j) = 2. \tag{3.25}$$

Therefore, we got a contradiction. □

We will now show that there exist a joint distribution of Y_1, Y_0 such that $P(Y_1 = i, Y_0 = i - \delta) = \max\{P(Y_1 = i) + P(Y_0 = i - \delta) - 1, 0\} = L_i$ for all i. By Proposition 3.5.2,

$P(Y_1 = i, Y_0 = i-\delta) \neq 0$ for at most one i. There are thus two cases to consider here: when $L_i = 0$ for all i and when $L_i = 0$ for all i except one.

Case 1: when $P(Y_1 = i, Y_0 = i - \delta) = 0$ for all i. Case 1 implies that

$$\max\{P(Y_1 = i) + P(Y_0 = i - \delta) - 1, 0\} = 0 \quad \forall i \tag{3.26}$$

$$\Rightarrow \quad P(Y_1 = i) \leq 1 - P(Y_0 = i - \delta) \quad \forall i. \tag{3.27}$$

We apply Strassen's Theorem 3.4.1. Let A, B be the support of the specified margins for Y_1, Y_0 respectively. Naturally, we have $P(Y_1), P(Y_0)$ as two probability measures on A, B. Let $R = \{(i, j) : i \in A, j \in B, i - j \neq \delta\}$. Strassen's theorem states that there exists a coupling $\hat{P}$ of $P(Y_1), P(Y_0)$ with $\hat{P}(R) = 1$ if and only if

$$P(Y_1 \in U) \leqslant P\left(Y_0 \in \mathcal{N}_R(U)\right), \text{ for all } U \subseteq A, \tag{3.28}$$

where

$$\mathcal{N}_R(U) = \{b \in B : (U \times \{b\}) \cap R \neq \varnothing\}.$$

By construction of R, for any $i \in A$,

$$\mathcal{N}_R(i) = B \setminus \{i - \delta\}. \tag{3.29}$$

Therefore for any $U \subseteq A$ with more than one element, we have $\mathcal{N}_R(U) = B$ and:

$$\mathrm{P}\left(Y_0 \in \mathcal{N}_{\mathrm{R}}(\mathrm{U})\right) = \mathrm{P}\left(Y_0 \in B\right) = 1. \tag{3.30}$$

Thus the only non trivial constraints are for singleton sets U given by:

$$P(Y_1 = i) \leq \sum_{j \neq i-\delta} P(Y_0 = j) = 1 - P(Y_0 = i - \delta). \tag{3.31}$$

The constraints in (3.31) are already satisfied by the assumption in (3.27). Therefore, by Strassen's theorem, there must exist a coupling P' such that $P'(R) = 1$. Thus, the lower

bounds $L_i = 0$ for all i are achievable in case 1.

Case 2: Consider the case that $L_i = 0$ for all i except for $i = j$. We will explicitly construct a joint distribution for Y_1, Y_0 that satisfies:

$$P(Y_1 = j, Y_0 = j - \delta) = P(Y_1 = j) + P(Y_0 = j - \delta) - 1 > 0; \tag{3.32}$$

$$P(Y_1 = i, Y_0 = i - \delta) = 0 \quad \text{for all} \quad i \neq j. \tag{3.33}$$

Consider a distribution satisfies by (3.32) and (3.33). Further, let $P(Y_1 = i, Y_0 = k) = 0$ for any $i \neq j$ or $k \neq j - \delta$. Let $P(Y_1 = j, Y_0 = k) = P(Y_0 = k), k \neq j - \delta$ and $P(Y_1 = i, Y_0 = j - \delta) = P(Y_1 = i), i \neq j$. For all $i \neq j$, we have

$$\sum_k P(Y_1 = i, Y_0 = k) = P(Y_1 = i, Y_0 = j - \delta) = P(Y_1 = i). \tag{3.34}$$

For all $k \neq j - \delta$, we have

$$\sum_i P(Y_1 = i, Y_0 = k) = P(Y_1 = j, Y_0 = k) = P(Y_0 = k). \tag{3.35}$$

Also, we have

$$\begin{aligned}
\sum_k P(Y_1 = j, Y_0 = k) &= P(Y_1 = j, Y_0 = j - \delta) + \sum_{k \neq j - \delta} P(Y_1 = j, Y_0 = k) \\
&= \max\{P(Y_1 = j) + P(Y_0 = j - \delta) - 1, 0\} + \sum_{k \neq j - \delta} P(Y_0 = k) \\
&= P(Y_1 = j) + P(Y_0 = j - \delta) - 1 + (1 - P(Y_0 = j - \delta)) \\
&= P(Y_1 = j),
\end{aligned}$$

where the second equality follows definition of j. Similarly,

$$\begin{aligned}\sum_i P(Y_1 = i, Y_0 = j - \delta) &= P(Y_1 = j, Y_0 = j - \delta) + \sum_{i \neq j} P(Y_1 = i, Y_0 = j - \delta) \\ &= \max\{P(Y_1 = j) + P(Y_0 = j - \delta) - 1, 0\} + \sum_{i \neq j} P(Y_1 = j) \\ &= P(Y_1 = j) + P(Y_0 = j - \delta) - 1 + (1 - P(Y_1 = j)) \\ &= P(Y_0 = j - \delta).\end{aligned}$$

All the $P(Y_1 = i, Y_0 = j)$ are non-negative. This is a valid joint distribution that satisfies the given marginals. Thus, the lower bounds are achievable. An example of the construction of the joint probability matrix in case 2 is given in Figure 3.3.

Y_1 \ Y_0	a_0	a_1	a_2	a_3	a_4	a_5
b_0	0	0	0	b_0	0	0
b_1	a_0	a_1	a_2	m	a_4	a_5
b_2	0	0	0	b_2	0	0
b_3	0	0	0	b_3	0	0
b_4	0	0	0	b_4	0	0
b_5	0	0	0	b_5	0	0

Assume $\delta = -2$, $a_3 + b_1 - 1 = m > 0$

Figure 3.3: Example construction of the matrix in Case 2.

Next, we show that there exists a joint distribution of Y_1, Y_0 such that $P(Y_1 = i, Y_0 = i - \delta) = \min\{P(Y_1 = i), P(Y_0 = i - \delta)\} = U_i$ for all i. We will construct a joint distribution of Y_1, Y_0 using the joint probability table explicitly. We first permute the joint probability table based on whether the upper bound is achieved at $P(Y_1 = i)$ or $P(Y_0 = i - \delta)$. Let J_1 be the set of i such that $P(Y_1 = i, Y_0 = i - \delta) = P(Y_1 = i)$ for $i \in J_1$ and J_2 be the set

of j such that $P(Y_1 = j + \delta, Y_0 = j) = P(Y_0 = j)$ for $j \in J_2$. We can rearrange the joint probability table based on the partition of J_1, J_2. Let $|J_1| = n_1, |J_2| = n_2$. Let N_1 be the number of elements in the support of Y_1. Let N_2 be the number of elements in the support of Y_0. We permute the joint probability matrix of Y_1, Y_0 such that the first n_1 rows denotes the probability of elements in J_1 and the last n_2 rows denotes the probability of elements in J_2. Figure 3.4 shows an example of such permutation.

Y_1 \ Y_0	a_0	a_1	a_2	a_3	a_4	a_5
b_0			a_2		0	
b_1	0	0	0	b_1	0	0
b_2			0		a_4	
b_3	0	0	0	0	0	b_3
b_4			0		0	
b_5			0		0	

A : Matrix before permutation

Y_1 \ Y_0	a_0	a_1	a_3	a_5	a_2	a_4
b_1	0	0	b_1	0	0	0
b_3	0	0	0	b_3	0	0
b_0					a_2	0
b_2					0	a_4
b_4					0	0
b_5					0	0

A' : Matrix after permutation

Figure 3.4: Probability table permutation.

By construction, the first n_1 rows and the last n_2 columns are filled with 0's and U_i's. We have an empty $(N_1 - n_1) \times (N_2 \times n_2)$ sub-matrix to fill. We also obtain a new set of margin constraints on the sub-matrix by subtracting the marginals with what we have already filled. Notice that the row/column sum of these new constraints for the $(N_1 - n_1) \times (N_2 \times n_2)$ matrix is given by

$$s = 1 - \sum_{i \in J_1} P(Y_1 = i) - \sum_{j \in J_2} P(Y_0 = j). \tag{3.36}$$

We can fill each entry of the sub-matrix by

$$P(Y_1 = i, Y_0 = j) = (P(Y_1 = i) - I((i+\delta) \in J_2)P(Y_0 = i+\delta)) \tag{3.37}$$

$$\times (P(Y_0 = j) - I(j-\delta \in J_1)P(Y_1 = j-\delta))/s. \tag{3.38}$$

Thus, we construct a valid joint distribution of Y_1, Y_0 with $P(Y_1 = i, Y_0 = i-\delta) = \min\{P(Y_1 = i), P(Y_0 = i-\delta)\} = U_i$ for all i and satisfies the given marginals. This concludes our proof of theorem 3.5.1. We note that the same proof technique can be used to obtain the pmf bounds on the sum of two discrete random variables given the marginals.

Corollary 3.5.3. *In the case where the distribution of one potential outcome is degenerate, the individual treatment effect is fully identified. For example, if we assume $P(Y_0 = 0) = 1$, then the upper and lower bounds in Theorem 3.5.1 coincides and we can identify $P(Y_1 - Y_0 = \delta)$.*

Proof. This follows from the fact that $[L_i, U_i] = [\max\{P(Y_1 = i) + P(Y_0 = i-\delta) - 1, 0\}, \min\{P(Y_1 = i), P(Y_0 = i-\delta)\}] = [0, 0]$ when $i - \delta \neq 0$ and $[L_i, U_i] = [\max\{P(Y_1 = i) + P(Y_0 = i-\delta) - 1, 0\}, \min\{P(Y_1 = i), P(Y_0 = i-\delta)\}] = [P(Y_1 = i), P(Y_1 = i)]$ when $i - \delta = 0$. Therefore, $\sum_i L_i = \sum_i U_i = P(Y_1 = \delta)$. This is saying in the case where $P(Y_0 = 0) = 1$, the bounds in Theorem 3.5.1 tell us that $P(Y_1 - Y_0 = \delta) = P(Y_1 = \delta)$. □

3.6 Summary Remarks

In this chapter, we first sharpen and correct the cdf bounds for the difference of two random variables used in the literature. We discuss the implications of some logical gaps in Williamson and Downs [1990] that have been propagated to some of the later literature, including causal inference applications. Then we turn to a closely related problem about coupling of two measures for finite sets. We prove a very useful theorem that extends the finite version of Strassen's Theorem. The general extension on Strassen's Theorem with continuous margins could be further examined. Using the insight from finite version of Strassen's theorem, we derive the sharp pmf bounds for the sum/difference of two random variables given fixed marginals. The results presented in this chapter lay a critical founda-

tion for the study of prediction intervals for Individual Treatment Effects (ITEs), as we will demonstrate in the next chapter.

Chapter 4

A COMPLETE CHARACTERIZATION FOR PREDICTION INTERVALS ON INDIVIDUAL TREATMENT EFFECT

Individual treatment effect (ITE) is often regarded as the ideal target of inference in causal analyses and has been the focus of several recent studies. In this chapter, we describe the intrinsic limits regarding what can be learned concerning ITEs given data from large randomized experiments. We begin by examining how to construct a valid prediction interval for the ITE, focusing on when the interval is informative and when it can be bounded away from zero. We start with the binary outcome case and then extend the intuition to more general outcome types. The joint distribution over potential outcomes is only partially identified from a randomized trial. Consequently, to be valid, an ITE prediction interval must be valid for all joint distribution consistent with the observed data and hence will in general be wider than that resulting from knowledge of this joint distribution. We also give conditions on the observed data for there to exist a consistent joint distribution under which a given prediction interval (or set) would be valid. Then, we discuss Fréchet-Hoeffding bounds and bounds on the cumulative distribution function (cdf) and probability mass function (pmf) for the ITE when the outcome is binary. We further discuss the ITE prediction intervals in the setting of randomized experiments where additional covariates are observed for each individual. Finally we contrast prediction intervals for the ITE and confidence intervals for the Average Treatment Effect (ATE). This also leads to the consideration of Fisher versus Neyman null hypotheses.

4.1 Introduction

The traditional causal inference literature has been focused on population level treatment effect parameters such as the average treatment effect (ATE) and conditional average treatment effect (CATE). Although the individual treatment effect (ITE) is often regarded as the ideal parameter of interest for personalized decision making, it is not, in general identi-

fiable even if we know the outcome for an individual and have data from a large randomized experiment. Recent works by Lei and Candès [2021], Jin et al. [2023], Chernozhukov et al. [2023],Wang and Qiao [2025] discuss conformal inference methods to estimate ITE and prediction intervals for the ITE. A more recent debate from Mueller and Pearl [2022] and Dawid and Senn [2023] also explores the possibility of using ITE and bounds on ITE to help personalized decision making.

Bounds and relationships on the probability of the counterfactual treatment effect has been studied in terms of probability of causation in Robins and Greenland [1989] and Tian and Pearl [2000]. Some more recent work from Mueller et al. [2021], Sani et al. [2023] extends these ideas to learn individual responses and bounds from causal diagrams and mediation analysis. Inference and examples of probability of causation have been discussed in Dawid et al. [2016]. Fan and Park [2010] study sharp bounds on the cdf of the treatment effect using copulas and provides a cdf bound which is valid when the distribution of outcome variable is absolutely continuous with respect to the Lebesgue measure based on the previous result from Frank et al. [1987]. Mullahy [2018] applies the bounds in Fan and Park [2010] in health economics applications. Bounds and optimal policy for binary treatment and outcome have been studied in Kallus [2022a], Kallus [2022b], and Dawid and Senn [2023]. Individual treatment effects on ordinal outcomes have been studied in Lu et al. [2018]. Exact inference on individual treatment effects based on permutation has been studied in Blaker [2000], Rigdon and Hudgens [2015], Chiba [2015]. Robins [1988], Imbens and Menzel [2018] and Brennan et al. [2024] discuss the construction of confidence intervals for causal parameters.

In this chapter, we try to understand the prediction intervals and bounds for the ITE given that we have observations from well-conducted large randomized control trials (RCT). In Section 4.2, we start from the binary treatment and outcome model and provide a complete characterization of ITE prediction intervals under this simple model setting. We characterize when a degenerate interval consisting of a single value is a valid prediction interval, when is the valid prediction interval for ITE non-negative/non-positive, and when the only valid prediction interval is a trivial interval. Additionally, we give conditions on the observed data for there to exist a consistent joint distribution under which a given non-trivial prediction interval would be valid. In Section 4.3, we extend our insights to

continuous and ordinal outcomes. Our approach leverages cdf bounds for the difference of two random variables from Fan and Park [2010] and Zhang and Richardson [2024], as well as the pmf bounds and the extensions of Strassen's theorem developed in previous chapters. In Section 4.4, we provide sharp bounds on the cumulative distribution function (cdf) and probability mass function (pmf) of the ITE under binary treatment and outcome model. We explore the relationships between the cdf/pmf bounds on ITE and the Fréchet-Hoeffding bounds on two random variables. In Section 4.5, we consider the ITE prediction intervals in the setting of randomized experiments where additional covariates are observed for each individual. Lastly, we compare ITE prediction intervals and ATE confidence intervals and provide a synthetic data example to discuss the implications of them in section 4.6.

4.2 The limits to inference for individual treatment effects in binary treatment and outcome model

Throughout the chapter, we consider a binary treatment setting with treatment $D = 0, 1$. Let Y_1 be the potential outcome when receiving the treatment and Y_0 be the potential outcome when not receiving the treatment. We define the individual treatment effect as:

$$\text{ITE} = Y_1 - Y_0.$$

Let Y be the observed variable. By consistency, $Y = Y_0$ when $D = 0$ and $Y = Y_1$ when $D = 1$. We first examine what is possible to learn in the limit of a large sample size.

4.2.1 Definition of a prediction interval for the individual treatment effect (ITE)

A $(1 - \alpha)$ *prediction interval* for an individual treatment effect is an interval such that

$$P\left((Y_1 - Y_0) \in [L, R]\right) \geq 1 - \alpha.$$

Throughout the chapter, we assume α is sufficiently bounded away from 0.5.

Proposition 4.2.1. *Suppose an interval* $\mathbb{I}$ *is a valid* $(1-\alpha)\%$ *prediction interval. If* $P(\text{ITE} \in A) > \alpha$ *for some set* A*, then* $\mathbb{I} \cap A \neq \emptyset$*. If* $P(\text{ITE} \in A) \leq \alpha$ *for some set* A*, then there*

exists a $(1-\alpha)$ *valid prediction set that does not intersect with* A*. In particular, if* $\mathbb{R} \setminus A$ *is an interval, then it will be a valid* $(1-\alpha)$ *prediction interval.*

Proof. If $P(\text{ITE} \in A) > \alpha$ and $\mathbb{I} \cap A = \emptyset$, then $P(\text{ITE} \in \mathbb{I}) \leq 1 - P(\text{ITE} \in A) < 1 - \alpha$. The interval $\mathbb{I}$ does not have $1-\alpha$ coverage. If $P(\text{ITE} \in A) \leq \alpha$, then $P(\text{ITE} \in (\mathbb{R} \setminus A)) = 1 - P(\text{ITE} \in A) \geq 1 - \alpha$. $\mathbb{R} \setminus A$ is a valid $(1-\alpha)$ prediction set. □

Corollary 4.2.2. *In discrete settings, whenever* $P(ITE = i) > \alpha$*, we must have* $i \in \mathbb{I}$*.*

4.2.2 Binary Treatment and Outcome Model

To further narrow the discussion, we first consider the case in which the treatment and response are both binary. Following Copas [1973] who characterizes individual patients in the binary treatment and outcome model, we have the following four types and individual treatment effects:

Y_0	Y_1	ITE	Type
0	0	0	Never Recover (NR)
0	1	1	Helped (HE)
1	0	-1	Hurt (HU)
1	1	0	Always Recover (AR) / Immune

Notice that in this setting the individual treatment effects take three possible values: $-1, 0, 1$.

In the simple setting that we consider, since there are only 3 possible values taken by the ITE, there are 6 possible prediction intervals:

$$\{-1\},\ \ \{0\},\ \ \{1\},\ \ [-1,0],\ \ [0,1],\ \ [-1,1].$$

Note that prediction intervals $[-1,0], [0,1], [-1,1]$ here are sets correspond to $\{-1,0\}, \{0,1\}, \{-1,0,1\}$. Singleton sets can be viewed as degenerate intervals where the starting point equals to the end point. In this simple setting, there is only one prediction set that does not correspond to an interval, namely $\{-1,1\}$. We discussed it in Remark 4.2.5.

When will the valid prediction interval of minimal length for a given joint distribution not be unique?

Corollary 4.2.3. *If* $P(HE \cup HU) > \alpha$ *and* $\max\{P(HE), P(HU)\} \leq \alpha$*, then* $\{0\}$ *is not a valid* $(1-\alpha)\%$ *prediction interval but both* $[-1, 0]$ *and* $[0, 1]$ *are valid and minimal length* $(1-\alpha)\%$ *prediction intervals for the ITE.*

Under the conditions stated in Corollary 4.2.3, the set $\{0\}$ fails to provide sufficient coverage; yet each of $[-1, 0]$ and $[0, 1]$ captures a sufficiently large portion of the probability mass (at least $1-\alpha$). Both of these intervals therefore qualify as valid $(1-\alpha)\%$ prediction intervals for ITE, and each achieves the same minimal length (one unit). Consequently, there is no single "shortest" interval that strictly dominates the other. In general, even given the joint distribution of Y_0, Y_1, there could be multiple minimal-length valid prediction intervals.

Partial Identification of the Joint Distribution over Types under Randomization

Under randomization, the relationship between the counterfactual distribution $P(Y_0, Y_1)$ and the observed distributions $\{P(Y \mid D=0), P(Y \mid D=1)\}$ is given by this table:

	$P(Y=0 \mid D=0)$	$P(Y=1 \mid D=0)$
$P(Y=0 \mid D=1)$	$P(Y_0=0, Y_1=0)$	$P(Y_0=1, Y_1=0)$
$P(Y=1 \mid D=1)$	$P(Y_0=0, Y_1=1)$	$P(Y_0=1, Y_1=1)$

Here $P(Y=i \mid D=j) = P(Y_j=i)$ due to randomization.

Equivalently we may write this in terms of types:

	$P(Y=0 \mid D=0)$	$P(Y=1 \mid D=0)$
$P(Y=0 \mid D=1)$	$P(\text{NR})$	$P(\text{HU})$
$P(Y=1 \mid D=1)$	$P(\text{HE})$	$P(\text{AR})$

Proposition 4.2.4 (Fréchet inequality bounds)**.** *Also known as Boole-Fréchet inequality. For two real valued random variables* Y_0, Y_1 *and any* $(y_0, y_1) \in \mathbb{R}^2$*, suppose that we know*

$P(Y_0 = y_0) = a$ *and* $P(Y_1 = y_1) = b$, *then*

$$\max\{0, a + b - 1\} \leq P(Y_0 = y_0, Y_1 = y_1) \leq \min\{a, b\} \tag{4.1}$$

Since under randomization the joint distribution must add up to satisfy the observed distributions in the two treatment arms, we can parametrize the distribution of types using $P(AR) = t$. We then have the following solution set:

$$\left\{ \begin{array}{lcl} P(\text{AR}) & = & t, \\ P(\text{HU}) & = & P(Y{=}1 \mid D{=}0) - t, \\ P(\text{HE}) & = & P(Y{=}1 \mid D{=}1) - t, \\ P(\text{NR}) & = & 1 - P(Y{=}1 \mid D{=}0) - P(Y{=}1 \mid D{=}1) + t, \end{array} \right\},$$

where, by Proposition 4.2.4, we have

$$t \geq \max\left\{0, (P(Y{=}1 \mid D{=}0) + P(Y{=}1 \mid D{=}1)) - 1\right\},$$
$$t \leq \min\left\{P(Y{=}1 \mid D{=}0), P(Y{=}1 \mid D{=}1)\right\}.$$

4.2.3 When is the only valid prediction interval trivial?

There are circumstances in which the only valid prediction interval for the individual treatment effect is the trivial interval: $[-1, 1]$!

The interval will be trivial when the set of people of type Hurt can be larger than α, *and* the set of people of type Helped can also be larger than α.

Notice that the proportion Helped and Hurt, *both* achieve their maximum value when the proportion Always Recover t achieves its minimum value. Hence the only valid prediction interval will be trivial whenever we have both:

$$P(Y{=}1 \mid D{=}0) - \max\left\{0, (P(Y{=}1 \mid D{=}0) + P(Y{=}1 \mid D{=}1)) - 1\right\} > \alpha; \tag{4.2}$$
$$P(Y{=}1 \mid D{=}1) - \max\left\{0, (P(Y{=}1 \mid D{=}0) + P(Y{=}1 \mid D{=}1)) - 1\right\} > \alpha. \tag{4.3}$$

this may be equivalently expressed as:

$$\min\{P(Y=1 \mid D=0), 1-P(Y=1 \mid D=1)\} > \alpha; \tag{4.4}$$

$$\min\{P(Y=1 \mid D=1), 1-P(Y=1 \mid D=0)\} > \alpha. \tag{4.5}$$

Consequently, provided we have:

$$\alpha < \min\{P(Y=1 \mid D=0), P(Y=1 \mid D=1)\} \tag{4.6}$$

and

$$\max\{P(Y=1 \mid D=0), P(Y=1 \mid D=1)\} < (1-\alpha) \tag{4.7}$$

then the only valid prediction interval will be trivial. In concrete terms, if $\alpha = 0.05$ and the proportions recovering under treatment ($D = 1$) and control ($D = 0$) both lie between 5% and 95% then, given that we know the conditional distributions $P(Y|D)$, the only valid 95% prediction interval for the individual treatment effect will be $[-1, 1]$.

These results show that for randomized experiments in which the proportion of recovery in both arms lies between α and $(1-\alpha)$, the observed data is entirely uninformative regarding the ITE.

4.2.4 When is the valid prediction interval for the individual treatment effect a singleton?

Notwithstanding the results in the previous section, perhaps surprisingly, there are situations in which a valid $(1-\alpha)\%$ prediction interval is a singleton. We now characterize when this occurs. Detailed derivations are provided in the Appendix C.1. There are three cases to consider:

- $\{0\}$ is valid if the sum of AR and NR types is at least $1-\alpha$, i.e., both arms have nearly $\alpha\%$ or nearly $1-\alpha\%$ response such that the sum of the proportions of individuals

across the two arms with the less common outcome is less than α. Concretely, either

$$P(Y=1 \mid D=0) + P(Y=1 \mid D=1) \leq \alpha \tag{4.8}$$

or

$$P(Y=0 \mid D=0) + P(Y=0 \mid D=1) \leq \alpha. \tag{4.9}$$

Note however, that in such a case the average treatment effect, though less than α, may be non-zero if $P(Y=0 \mid D=0) \neq P(Y=0 \mid D=1)$. Thus, given sufficiently large sample sizes, confidence intervals for the Average Treatment Effect will not include zero. We note that the singleton ITE prediction interval $\{0\}$ is equivalent to establishing the Fisherian sharp null hypothesis [Fisher, 1936] holds for at least $(1-\alpha)$ of the population. We will further discuss this in section 4.6.

- $\{1\}$ is valid if the HE type alone can exceed $1-\alpha$. This requires

$$P(Y=1 \mid D=1) - P(Y=1 \mid D=0) \geq (1-\alpha). \tag{4.10}$$

meaning the average treatment effect is at least $1-\alpha$.

- $\{-1\}$ is valid if the HU type alone can exceed $1-\alpha$. This requires

$$P(Y=1 \mid D=0) - P(Y=1 \mid D=1) \geq (1-\alpha). \tag{4.11}$$

i.e., the average treatment effect is less than $-(1-\alpha)$.

4.2.5 When is the valid prediction interval for the individual treatment effect non-negative/non-positive?

Following the previous result, one can rule out negative (or positive) treatment effects if the proportion of HU (or HE) can never exceed α. This occurs when either $P(Y=1 \mid D=0)$

or $1 - P(Y = 1 \mid D = 1)$ is below α (and similarly for ruling out positive effects). See Appendix C.1.

Remark 4.2.5. *It is not possible to conclude* $\{-1, 1\}$ *as a best prediction set given the observed marginals* $P(Y = 1 \mid D = 1), P(Y = 1 \mid D = 0)$. *For* $\{-1, 1\}$ *to be a valid prediction set, the proportion of people of type Helped plus the proportion of people of type Hurt should always be greater than or equal to* $1 - \alpha$. *From the lower bounds on proportion of people of type Helped and the proportion of people of type Hurt, we need*

$$\begin{aligned} &P(Y = 1 \mid D = 1) + P(Y = 1 \mid D = 0) \\ &- 2\min\{P(Y = 1 \mid D = 1), P(Y = 1 \mid D = 0)\} \geq 1 - \alpha, \end{aligned} \tag{4.12}$$

equivalently, either

$$P(Y = 1 \mid D = 1) - P(Y = 1 \mid D = 0) \geq 1 - \alpha \tag{4.13}$$

or

$$P(Y = 1 \mid D = 0) - P(Y = 1 \mid D = 1) \geq 1 - \alpha. \tag{4.14}$$

However, based on the result in Section 4.2.4, in either setting, we would just conclude the singleton $\{-1\}$ *or* $\{1\}$ *respectively as the prediction set instead of* $\{-1, 1\}$. *Therefore, it is not possible to conclude* $\{-1, 1\}$ *as a best prediction set from the observed marginals* $P(Y = 1 \mid D = 1), P(Y = 1 \mid D = 0)$. *Note that* $\{-1, 1\}$ *is a theoretically possible prediction set which can be optimal for certain disributions over potential outcomes. It is just we will not be able to conclude it based on observed distribution from randomization.*

4.2.6 Visual summary of results on valid ITE prediction intervals

Based on the result in 4.2.3, 4.2.4, and 4.2.5, under randomization and in the limit of a large sample size where we observe the true marginals $P(Y = 1 \mid D = 0), P(Y = 1 \mid D = 1)$ and their counterparts, we can characterize the corresponding ITE prediction intervals. These intervals are valid no matter the joint distribution over Y_0 and Y_1 (provided it is compatible

with $P(Y \mid D)$. These intervals are also "the best we can do" without additional assumptions or information, e.g. from a cross-over study.

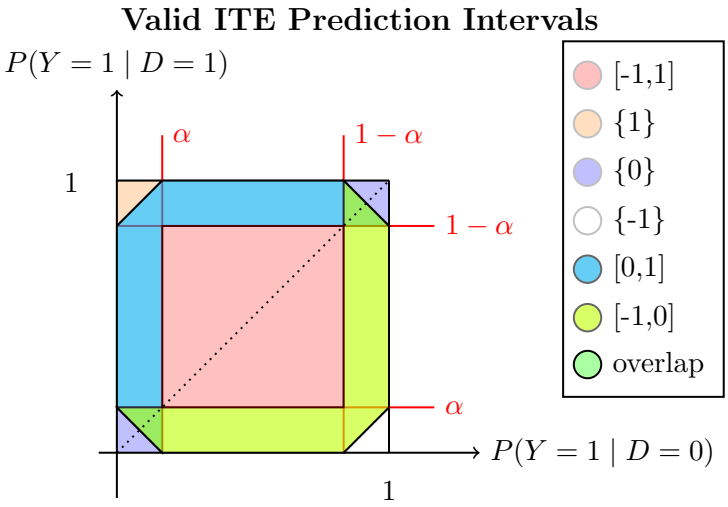

Figure 4.1: Shortest ITE intervals given different marginal distributions.

Notice that we have two overlapping triangles each with area $\frac{1}{2}\alpha^2$ in Figure 4.1 where both $[0, 1]$ and $[-1, 0]$ can serve as valid $\alpha-$level prediction interval for individual treatment effects. Assume that for the same length of intervals, intervals with higher coverage are "better". Then we can further decompose these triangular areas to obtain the "best" prediction intervals; see Figure 4.2.

4.2.7 Necessary conditions for a given prediction interval to be valid and of minimal length

Suppose we are given a prediction interval and an observed distribution from an RCT, we can consider when does there exist a joint distribution over potential outcomes that is compatible with the observed distributions and for the prediction interval to be valid and of minimal length. In Appendix C.2, we derive the precise constraints on the marginal distributions $P(Y = 1 \mid D = 0)$ and $P(Y = 1 \mid D = 1)$ that guarantee each of the intervals can serve as a valid (or "best") prediction interval for the individual treatment effect for some joint distribution over potential outcome. The necessary conditions take the form of simple inequalities relating the two response probabilities; they characterize scenarios in

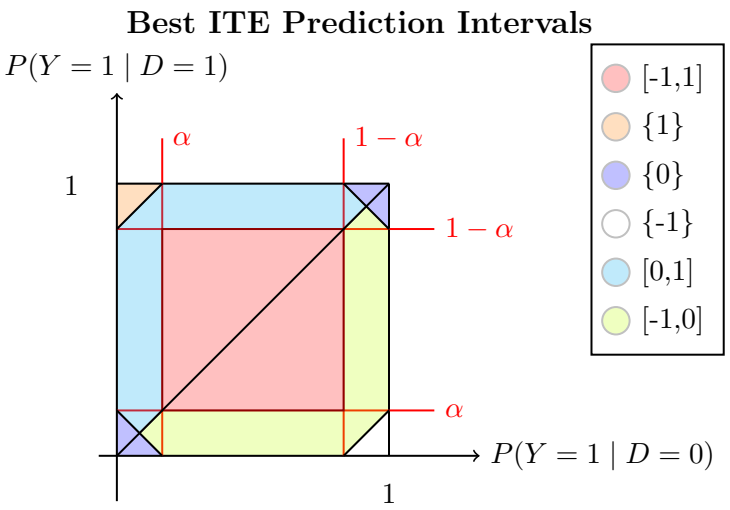

Figure 4.2: The prediction interval with the highest coverage among those that are valid and have minimal length.

which each the true treatment effect with probability at least $1 - \alpha$ lies within the given set for som population compatible with $P(Y \mid D)$. Full details, derivations, and proofs appear in Appendix C.2.

4.2.8 Visualization of Necessary Conditions given an ITE prediction interval to be the best

Figure 4.3 and 4.4 provide visualizations of the necessary conditions on $P(Y = 1 \mid D = 0)$ and $P(Y = 1 \mid D = 1)$ for a given prediction interval to be valid/best. Note that if for a given point in the unit square, we consider the intervals (from Figure 4.3 and 4.4) containing $(P(Y = 1 \mid D = 1), P(Y = 1 \mid D = 0))$ and then select the longest, we recover Figure 4.2.

4.3 Beyond binary outcomes

We now turn to the case where the outcome is continuous or ordinal, rather than binary. Our goal is to address questions analogous to those explored in the binary setting. In particular, we begin by investigating how to construct valid prediction intervals based solely on the marginal distributions, without imposing assumptions on the joint distribution of the potential outcomes. We then examine the conditions under which these intervals can be

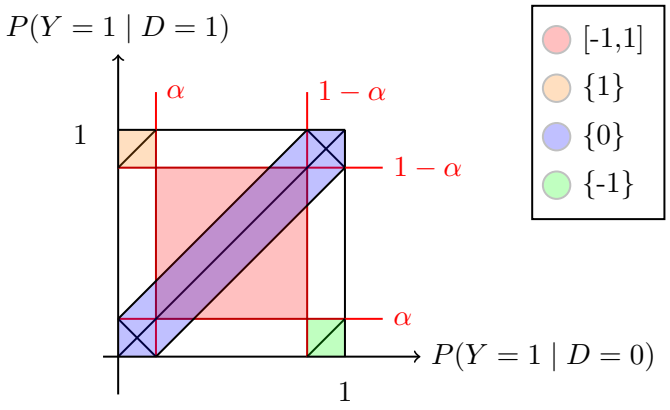

Figure 4.3: Necessary condition on the marginal distributions for a given prediction interval, respectively, $[-1, 1], \{1\}, \{0\}, \{-1\}$, to be valid and best for some joint distribution $P(Y_0, Y_1)$ compatible with $P(Y \mid D)$; see also 4.4.

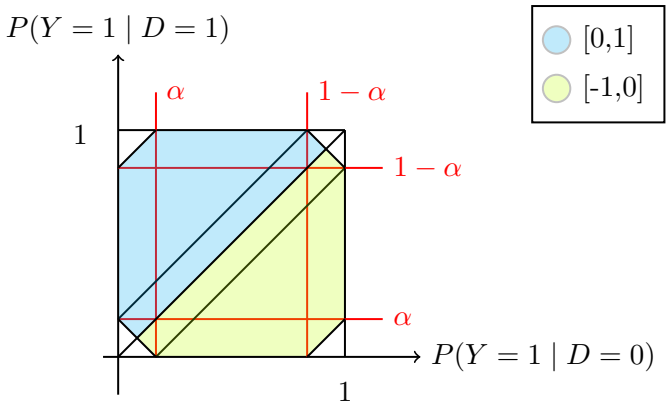

Figure 4.4: Necessary condition on the marginal distributions for a given prediction interval $[0, 1], [-1, 0]$ to be valid and best for some joint distribution $P(Y_0, Y_1)$ compatible with $P(Y \mid D)$.

bounded away from zero. Further, we solve the problem: given a prediction interval (or set), what conditions on the observed data ensure the existence of a consistent joint distribution under which the interval is valid? We first focus on the continuous outcome case.

4.3.1 A conservative interval for continuous outcomes

The marginal distributions of Y_1 and Y_0 are identified from randomized experiments. Let $[L_0, R_0]$ be an interval such that $P\left(Y_0 \in [L_0, R_0]\right) \geq 1 - \alpha/2$. Let $[L_1, R_1]$ be an interval such that $P\left(Y_1 \in [L_1, R_1]\right) \geq 1 - \alpha/2$. Then

$$P\left((Y_1 - Y_0) \in [L_1 - R_0, R_1 - L_0]\right) \geq 1 - \alpha. \tag{4.15}$$

Proof. Note that if $Y_1 - Y_0$ is not in $[L_1 - R_0, R_1 - L_0]$, then either Y_1 is not in $[L_1, R_1]$ or Y_0 is not in $[L_0, R_0]$. Put into set notation,

$$\{Y_1 - Y_0 \notin [L_1 - R_0,\ R_1 - L_0]\} \subseteq \{Y_1 \notin [L_1, R_1]\} \cup \{Y_0 \notin [L_0, R_0]\}.$$

Applying the union bound to that set containment gives

$$P\Big(Y_1 - Y_0 \notin [L_1 - R_0, R_1 - L_0]\Big) \leq P\big(Y_1 \notin [L_1, R_1]\big) + P\big(Y_0 \notin [L_0, R_0]\big).$$

By hypothesis, $P\left(Y_0 \notin [L_0, R_0]\right) < \alpha/2$ and $P\left(Y_1 \notin [L_1, R_1]\right) < \alpha/2$, therefore,

$$P\big(Y_1 - Y_0 \notin [L_1, R_1]\big) < \alpha.$$

Thus,

$$P\left((Y_1 - Y_0) \in [L_1 - R_0, R_1 - L_0]\right) \geq 1 - \alpha.$$

□

This result is similar to the "naive" conformal inference prediction interval for the ITE given in Lei and Candès [2021] Section 4.1. Though we further assume an infinite sample size and do not use covariates.

4.3.2 Points that must be included in every valid $(1 - \alpha)$ prediction interval

One might argue that the bounds in (4.15) are too conservative. We will take a different perspective to see what are the points that must be included in the interval. Note that to

be valid, an ITE prediction interval must be valid for all joint distributions consistent with the observed data, and hence will in general be wider than that resulting from knowledge of this joint distribution. Let $[L_0', R_0']$ be an interval such that

$$L_0' := \min\Big\{\ell \in \mathbb{R} : P\big(Y_0 < \ell\big) > \alpha\Big\}$$

and

$$R_0' := \max\Big\{\ell \in \mathbb{R} : P\big(Y_0 > \ell\big) > \alpha\Big\}.$$

In other words, L_0' and R_0' are the α-quantile and $(1-\alpha)$-quantile of $Y_i(0)$. Similarly we can define

$$L_1' := \min\Big\{\ell \in \mathbb{R} : P\big(Y_1 < \ell\big) > \alpha\Big\}$$

and

$$R_1' := \max\Big\{\ell \in \mathbb{R} : P\big(Y_1 > \ell\big) > \alpha\Big\}.$$

Then a valid $(1-\alpha)$ prediction interval for the ITE must include these points:

- $R_1' - L_0'$,
- $L_1' - R_0'$.

This follows because otherwise there exists a joint distribution of Y_1, Y_0 such that there is more than α mass outside of the prediction interval. This is because the only constraint imposed on the joint distribution by the marginals is given by the Fréchet inequalities,

$$P(Y_1 > R_1', Y_0 < L_0') \leq \min\{P(Y_1 > R_1'), P(Y_0 < L_0')\}$$

By construction the minimum is greater than α, meaning there exists a joint distribution that $P(Y_1 > R_1', Y_0 < L_0') > \alpha$. Hence under this joint distribution, $P(Y_1 - Y_0 > R_1' - L_0') > \alpha$, therefore, any interval that does not include $R_1' - L_0'$ will not be a valid α level prediction interval.

Even small tails in marginal distributions of Y_1, Y_0 can force the prediction interval to expand considerably at both extremes. We illustrate this in Figure 4.5.

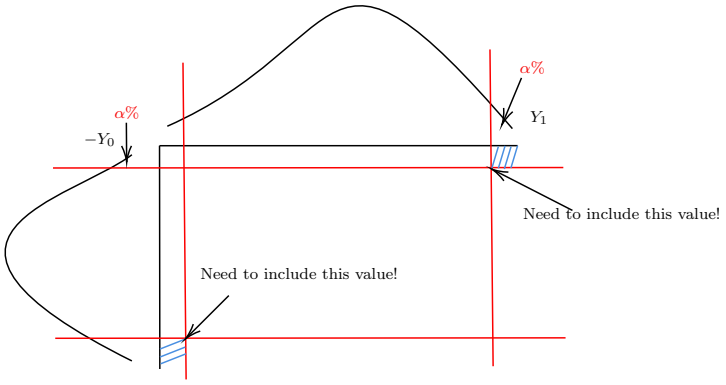

Figure 4.5: Illustration of the continuous Y case. In order to maintain a $1-\alpha$ coverage probability for the ITE, the prediction interval must include the key quantile differences on both the left and right tails of the outcome distributions.

4.3.3 Can we obtain a prediction interval bounded away from zero?

In the previous section, we gave conditions under which a point is included in every valid ITE prediction interval. Here we now give necessary and sufficient conditions for a valid interval to exclude the half line $(0, \infty)$ (or $(-\infty, 0)$). Though the arguments extend to any constant, not just zero, we now address the question: can a valid prediction interval for the individual treatment effect (ITE) be bounded away from zero? Suppose there exists a joint distribution of the potential outcomes such that $P(Y_1 - Y_0 \leq 0) > \alpha$. Then, without further assumptions on the joint distribution, the left endpoint of any valid $(1-\alpha)$ prediction interval must be less than or equal to zero. This is closely related to Kolmogorov's problem and the results in Fan and Park [2010], Zhang and Richardson [2024], which we discussed in detail in Chapter 2 and 3. In particular, consider the following upper bound on $P(Y_1 - Y_0 \leq 0)$:

$$\begin{aligned} F^U(0) &= 1 + \inf_y \min\left\{F_1(y) - P(Y_0 < y),\, 0\right\} \\ &= 1 + \inf_y \min\left\{F_1(y) - F_0(y) + P(Y_0 = y),\, 0\right\}. \end{aligned} \tag{4.16}$$

If this upper bound exceeds α, then zero must lie within the prediction interval, i.e., the left endpoint cannot be strictly greater than zero.

Similarly, if $P(Y_1 - Y_0 \geq 0) > \alpha$, then the right endpoint of any valid prediction interval must be at least zero. This probability can be written as

$$P(Y_1 - Y_0 \geq 0) = 1 - P(Y_1 - Y_0 < 0),$$

so we may use a lower bound on $P(Y_1 - Y_0 < 0)$ to assess whether the right endpoint must include zero. The lower bound is given by:

$$F^L(0^-) = \sup_y \max\{F_1(y) - F_0(y), 0\}. \tag{4.17}$$

If this lower bound is less than $1 - \alpha$, then the right endpoint must be at least zero.

Therefore, if both the left endpoint must be less than or equal to zero and the right endpoint must be greater than or equal to zero, the prediction interval necessarily includes zero and cannot be bounded away from it.

In fact, if the distribution of $Y_1 - Y_0$ is absolutely continuous with respect to Lebesgue measure so $P(Y_1 - Y_0 = 0) = 0$, then we can combine the conditions in 4.16 and 4.17 to state that for a valid $(1-\alpha)$ prediction interval for the ITE to exclude zero, either the lower bound on $P(Y_1 - Y_0 \leq 0)$ to exceed $1 - \alpha$ or the upper bound to be less than α—conditions that are only met in highly extreme cases.

4.3.4 Ordinal Outcome Cases

Ordinal outcomes with more than two categories arise in a variety of settings, such as when a continuous outcome is discretized into multiple bins. In some cases, it is reasonable to endow the ordinal outcomes with a metric. Under such assumptions, one can derive bounds on the individual treatment effect (ITE) using the probability mass function (pmf) bounds introduced in Section 3.5.

When no explicit metric is defined on the outcome space, alternative bounds on types may be constructed using the extension of the finite-version Strassen's theorem presented in

Section 3.4. We will not focus on this scenario here, as it pertains to a different topic than bounding the ITE. However, the problems can be addressed using similar techniques. Here, we introduce an algorithm to construct valid ITE prediction intervals when the outcome is ordinal and a metric is assumed.

Algorithm 1 Getting a valid ITE prediction interval for ordinal outcome

Suppose that both Y_1, Y_0 takes values in $0, 1, 2, ..., n$ and we have the marginal pmfs on Y_1 and Y_0. The range of ITE is $[-n, n]$.

for $i = 1, \ldots, n$ **do**

Mark the cells that make $Y_1 - Y_0 \geq n - i$ as R.

Use Theorem 3.4.2 to check if the marginals satisfy the condition for $\hat{P}(R) = \alpha/2$.

If yes, break the loop and set the right end point for the interval $U = n - i + 1$.

If not, continue the loop for next i.

end for

for $i = 1, \ldots, n + U$ **do**

Mark the cells that make $Y_1 - Y_0 \in [-n + i, U]$ as R.

Use Corollary 3.4.3 to check if the upper bound for $\hat{P}(R)$ is greater than $1 - \alpha$.

If not, break the loop and set the left end point for the interval $L = -n + i - 1$.

If yes, continue the loop with $i \leftarrow i + 1$

end for

The resulting interval $[L, U]$ is constructed such that no joint distribution of the potential outcomes assigns more than $\alpha/2$ probability[1] to the region where the ITE exceeds U, and no joint distribution assigns more than α probability to the region where the ITE is outside $[L, U]$. Consequently, $[L, U]$ forms a valid $(1 - \alpha)$ prediction interval for the individual treatment effect under any admissible joint distribution. Note that, since Theorem 3.4.2 allows us to verify whether there exists a coupling that places exactly $(1 - \alpha)$ probability mass on the set R, the resulting prediction interval is generally sharper than conservative intervals derived solely from marginal constraints. For continuous outcomes, one can employ

[1]This probability can be adjusted depending on whether a symmetric prediction interval is desired, or if an asymmetric interval better reflects the application context.

a discretization-based approximation to construct prediction intervals for the individual treatment effect (ITE) to improve the bounds given in Section 4.3.1.

In Appendix C.3, we further explore the case of bounding ITE prediction intervals away from zero in the ordinal outcome setting. We consider when the ITE prediction interval will be trivial and provide necessary conditions such that there exists a joint distribution of potential outcomes for a given ITE prediction interval to be valid.

4.4 Fréchet-Hoeffding bound on the pmf and cdf of ITE under binary treatment and outcome model

We first consider the binary treatment and binary outcome model where we know the marginals for Y_0 is $p, 1-p$ and the marginals for Y_1 is $q, 1-q$. We can parameterize the joint density using $P(Y_0 = 1, Y_1 = 0) = t$ and Table 4.1.

	$P(Y=0 \mid D=0) = p$	$P(Y=1 \mid D=0) = 1-p$
$P(Y=0 \mid D=1) = q$	$t \in [\max\{0, p+q-1\}, \min\{p, q\}]$	$q - t$
$P(Y=1 \mid D=1) = 1-q$	$p - t$	$1 - q - p + t$

Table 4.1: Binary treatment and outcome model.

Definition 4.4.1 (Fréchet-Hoeffding bounds)**.** *For two real valued random variables Y_1, Y_0 and any $y_1, y_0 \in \mathbb{R}$, suppose that we know $P(Y_0 \leq y_0) = a$ and $P(Y_1 \leq y_1) = b$, then*

$$\max\{0, a+b-1\} \leq P(Y_0 \leq y_0, Y_1 \leq y_1) \leq \min\{a, b\}. \tag{4.18}$$

Note: Here we make the distinction that Fréchet-Hoeffding bounds are bounding the joint cdf of Y_1, Y_0 while the Fréchet inequality bounds are bounding the joint pmf of Y_1, Y_0.

Definition 4.4.2 (Comonotonicity)**.** *The bivariate random vector $Y = (Y_0, Y_1) \in \mathbb{R}^2$ is called comonotonic if*

$$P(Y_0 \leq y_0, Y_1 \leq y_1) = \min\{P(Y_0 \leq y_0), P(Y_1 \leq y_1)\} \tag{4.19}$$

for all $y_0, y_1 \in \mathbb{R}$. *In this case, the bivariate distribution functions* $F(y_1, y_0) = P(Y_0 \leq y_0, Y_1 \leq y_1)$ *achieves the Fréchet-Hoeffding upper bound and we call the two random variables* Y_0, Y_1 *perfectly positively dependent.*

Definition 4.4.3 (Countermonotonicity)**.** *The bivariate random vector* $Y = (Y_0, Y_1) \in \mathbb{R}^2$ *is called countermonotonic if*

$$P(Y_0 \leq y_0, Y_1 \leq y_1) = \max\{0, P(Y_0 \leq y_0) + P(Y_1 \leq y_1) - 1\} \tag{4.20}$$

for all $y_0, y_1 \in \mathbb{R}$. *In this case, the bivariate distribution functions* $F(y_1, y_0) = P(Y_0 \leq y_0, Y_1 \leq y_1)$ *achieves the Fréchet-Hoeffding lower bound and we call the two random variables* Y_0, Y_1 *perfectly negatively dependent.*

In the binary treatment and outcome model, a perfect positive dependence of Y_1, Y_0 is achieved by $t = \min\{p, q\}$ and a perfect negative dependence of Y_1, Y_0 is achieved by $t = \max\{0, p + q - 1\}$.

Proposition 4.4.4. *When the treatment and outcome are both binary, the sharp pmf bounds on* $Y_1 - Y_0$ *are achieved at the Fréchet-Hoeffding bounds.*

From Table 4.1, we have $P(\text{ITE} = -1) = q - t$, $P(\text{ITE} = 0) = 1 - p - q + 2t$, $P(\text{ITE} = 1) = p - t$. The upper and lower bounds for each ITE value are reached at the Fréchet-Hoeffding bound on the joint distribution of Y_1, Y_0 (i.e. when t takes minimum or maximum value).

We further observe that for $P(\text{ITE} = 1)$, $P(\text{ITE} = -1)$, we can obtain pmf bounds by applying Fréchet inequality bounds on each cell directly. And for $P(\text{ITE} = 0)$, we can first obtain Fréchet inequality bounds on the two cells $t \in [\max\{0, p+q-1\}, \min\{p, q\}]$ and $1-q-p+t \in [\max\{0, 1-p-q\}, \min\{1-p, 1-q\}]$ and then add up the lower and upper bounds to obtain $1-q-p+2t \in [\max\{0, p+q-1\} + \max\{0, 1-p-q\}, \min\{p, q\} + \min\{1-p, 1-q\}]$. This simplifies to the bounds given by $1 - p - q + 2t, t \in [\max\{0, p + q - 1\}, \min\{p, q\}]$ as before. We will generalize this observation in section 3.5.

Proposition 4.4.5. *When the treatment and outcome are both binary, the sharp bounds on cdf of $Y_1 - Y_0$ are achieved at the Fréchet-Hoeffding bounds.*

Fréchet-Hoeffding bound on the joint distribution of Y_1, Y_0 implies that $t \in [\max\{0, p + q - 1\}, \min\{p, q\}]$. Since we can parameterize the joint probability with one parameter t, we have $P(\text{ITE} = -1) = q - t$, $P(\text{ITE} = 0) = 1 - p - q + 2t$. In the binary case, the cdf for the ITE is characterized by the values $F(-1) = P(\text{ITE} = -1) = q - t$, $F(0) = P(\text{ITE} \leq 0) = P(\text{ITE} = -1) + P(\text{ITE} = 0) = 1 - p + t$, $F(1) = 1$.

As noted in Section 2.1 of Fan and Park [2010], sharp bounds on the cdf of the individual treatment effect are not achieved at the Fréchet-Hoeffding lower and upper bounds (perfectly positive/ perfectly negative dependence) for the distribution of Y_1, Y_0. As a special case, the sharp bounds on cdf of ITE are achieved at the Fréchet-Hoeffding bounds in binary treatment and outcome model.

Proposition 4.4.6. *In general, we can obtain valid pmf bounds from the cdf bounds. For example, if the ITE takes integer values, then $P(\text{ITE} = i) = P(\text{ITE} \leq i) - P(\text{ITE} < i) = F(i) - F(i-1)$. If we know $F(i) \in [a, b]$ and $F(i-1) \in [c, d]$, we can obtain that*

$$P(\text{ITE} = i) \in [a - d, b - c]. \tag{4.21}$$

When the treatment and outcome are both binary, the pmf bounds obtained from the sharp bounds on the cdf of $Y_1 - Y_0$ using the above method are sharp.

Using the cdf bounds for $Y_1 - Y_0$ in Proposition 4.4.5, we have

$$P(\text{ITE} = 1) = F(1) - F(0),$$

with $F(1) = 1, F(0) = 1 - p + t \in [1 - p + \max\{0, p + q - 1\}, 1 - p + \min\{p, q\}]$. Threrefore,

$$P(\text{ITE} = 1) \in [p - \min\{p, q\}, p - \max\{0, p + q - 1\}].$$

Similarly, $P(\text{ITE} = 0) = F(0) - F(-1)$ for $F(-1) \in [q - \min\{p, q\}, q - \max\{0, p + q - 1\}]$.

Thus,

$$P(\text{ITE} = 0) \in [1 - p - q + 2\max\{0, p + q - 1\}, 1 - p - q + 2\min\{p, q\}].$$

And

$$P(\text{ITE} = -1) = F(-1) = q - t, t \in [\max\{0, p + q - 1\}, \min\{p, q\}].$$

These bounds agree with the sharp bounds derived in Proposition 4.4.4. However, in general, the pmf bounds derived from the cdf bounds using (4.21) are not sharp. See the sharp pmf bounds in Section 3.5.

4.5 Prediction Intervals for ITE with Covariates

In this section, we illustrate the construction of ITE prediction intervals in the context of randomized experiments where additional covariates are observed for each individual. As we will demonstrate through an example, the prediction interval for the ITE conditional on covariates is not necessarily shorter than its marginal counterpart. In general, ITE prediction intervals and conditional prediction intervals are not directly comparable, as they address different inferential targets and rely on distinct sources of variability.

4.5.1 Problem Setup

Consider i.i.d. random samples $\{(D_i, X_i, Y_i)\}_{i=1}^n$ of n individuals, where $D_i \in \{0, 1\}$ is a binary treatment indicator, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T \in \mathbb{R}^p$ is a vector of observed covariates for each individual i, and $Y_i \in \mathbb{R}$ is the observed outcome for individual i under the potential outcome framework.

We assume that each individual receives the treatment independently with equal probability $P(D_i = 1 \mid X_i) = \pi$ for all i, where $0 < \pi < 1$ is a known constant. To simplify notation, we suppress the subscript i in what follows. Note that this is equivalent to saying the treatment is randomized regardless of the observed covariates. Each individual has two potential outcomes Y_1 and Y_0 and we observe $Y = DY_1 + (1 - D)Y_0$. We assume the stable unit treatment value assumption (SUTVA) that there is a single version of each

treatment/control and no interference among the subjects. The individual treatment effect is defined as:

$$\text{ITE} = Y_1 - Y_0$$

4.5.2 Example where a subset is more homogeneous

Here we present a simple numeric illustration of a randomized experiment with two binary covariates, $X_1, X_2 \in \{0, 1\}$, and a binary outcome Y. Suppose there are 10000 subjects in the treatment arm and 10000 subjects in the control arm, allocated 1:1 at random. Table 4.2 displays how many subjects fall into each (X_1, X_2) cell, along with the number of observed outcome $Y = 1$ and the corresponding probability that $Y = 1$.

Covariates		Treatment ($D = 1$)			Control ($D = 0$)		
X_1	X_2	n_T	$Y = 1$	$\hat{p}_T$	n_C	$Y = 1$	$\hat{p}_C$
0	0	3000	1200	0.40	3000	1000	0.33
0	1	3000	1000	0.33	3000	800	0.27
1	0	3000	3000	1.00	3000	0	0.00
1	1	1000	800	0.80	1000	200	0.2
Total	–	10000	6000	0.60	10000	2000	0.20

Table 4.2: Synthetic data in which the treatment arm has overall 60% of $Y = 1$ and the control arm has 20% of $Y = 1$. Within certain subgroups (e.g. conditioning on $X_1 = 1$), the treatment effect is more homogeneous, but conditioning further (e.g. on $(X_1 = 1, X_2 = 1)$) breaks the homogeneity again.

4.5.3 Key observations

- Without covariate adjustment, the treatment arm has probability $Y = 1$ equal to 0.6 and the control arm has probability $Y = 1$ equal to 0.2.

- Condition on $X_1 = 1$, in treatment, among those with $X_1 = 1$ (rows with $(X_1 = 1, X_2 = 0)$ and $(X_1 = 1, X_2 = 1)$), we have $n_T = 3000 + 1000 = 4000$ individuals, with $3000 + 800 = 3800$ individuals with $Y = 1$, i.e. 95% of individuals with $X_1 = 1$ in the treatment arm have positive outcomes. In control arm, among $X_1 = 1$, $n_C =$

$3000 + 1000 = 4000$ with $0 + 200 = 200$ individuals with $Y = 1$, i.e. 5% of individuals with $X_1 = 1$ in the control arm have positive outcomes.

- Condition on $(X_1 = 1, X_2 = 1)$, we are restricting further to just the last row. In this case, 80% of individuals in the treatment arm have positive outcomes while 20% of individuals in the control arm have positive outcomes.

4.5.4 Constructing ITE prediction intervals

Now given this experiment data, suppose we want to construct a 90% prediction interval for this people with $X_1 = 1, X_2 = 1$. First, under randomization we have:

$$P(Y = i \mid D = j, X_1 = 1) = P(Y_j = i \mid X_1 = 1)$$

We can write the following two-way table:

	$\hat{P}(Y=0 \mid D=0, X_1=1) = 0.95$	$\hat{P}(Y=1 \mid D=0, X_1=1) = 0.05$
$\hat{P}(Y=0 \mid D=1, X_1=1) = 0.05$	$P(Y_1 = 0, Y_0 = 0 \mid X_1 = 1)$	$P(Y_1 = 0, Y_0 = 1 \mid X_1 = 1)$
$\hat{P}(Y=1 \mid D=1, X_1=1) = 0.95$	$P(Y_1 = 1, Y_0 = 0 \mid X_1 = 1)$	$P(Y_1 = 1, Y_0 = 1 \mid X_1 = 1)$

Table 4.3: Binary Treatment and Outcome Model with condition on $X_1 = 1$

By Frechet inequality, we can conclude that

$$P\left(Y_1 - Y_0 \in \{1\} \mid X_1 = 1\right) = P(Y_1 = 1, Y_0 = 0 \mid X_1 = 1) \geq 0.95 + 0.95 - 1 = 0.9$$

Note that if we instead condition on both $X_1 = 1$ and $X_2 = 1$, we will get:

Based on this table, we can only conclude that:

$$P\left(Y_1 - Y_0 \in \{-1, 0, 1\} \mid X_1 = 1, X_2 = 1\right) \geq 0.9$$

	$\hat{P}(Y=0 \mid D=0, X_1=1, X_2=1) = 0.8$	$\hat{P}(Y=1 \mid D=0, X_1=1, X_2=1) = 0.2$
$\hat{P}(Y=0 \mid D=1, X_1=1, X_2=1) = 0.2$	$P(Y_1 = 0, Y_0 = 0 \mid X_1 = 1, X_2=1)$	$P(Y_1 = 0, Y_0 = 1 \mid X_1 = 1, X_2=1)$
$\hat{P}(Y=1 \mid D=1, X_1=1, X_2=1) = 0.8$	$P(Y_1 = 1, Y_0 = 0 \mid X_1 = 1, X_2=1)$	$P(Y_1 = 1, Y_0 = 1 \mid X_1 = 1, X_2=1)$

Table 4.4: Binary Treatment and Outcome Model condition on $X_1 = 1$ and $X_2 = 1$

since the proportion of ITE $= 0$ and ITE $= -1$ condition on $X_1 = 1, X_2 = 1$ can both be greater than 0.2.

4.5.5 Discussion of implications

In this example, it maybe paradoxical that for an individual with $X_1 = 1, X_2 = 1$, we can conclude that this person's individual treatment effect is 1 with probability 90% condition only on $X_1 = 1$ but instead get a trivial prediction interval condition on $X_1 = 1$ and $X_2 = 1$. What extra information does X_2 give? Additionally conditioning on $X_2 = 1$ picks out a subgroup in which the treatment effect is not as tightly concentrated as it is in the overall $X_1 = 1$ group. That would indicate that X_2 is related to heterogeneity in the effect. Conditioning on $X_2 = 1$ and $X_1 = 1$ defines a subgroup in which the treatment effect is more heterogeneous. In this case, do we really want to make decisions based on the narrower prediction interval? And as we can see, prediction intervals for ITE may not shrink if we condition on more covariates.

4.5.6 Relationship between conditional ITE and ITE

We will consider the simple case where $X_1 \in \{0, 1\}$. First, we have

$$P\big(Y_1 - Y_0 = 1\big) = P\big(Y_1 - Y_0 = 1,\, X_1 = 1\big) + P\big(Y_1 - Y_0 = 1,\, X_1 = 0\big).$$

By the law of total probability, we can also write

$$\begin{aligned} P\big(Y_1 - Y_0 = 1\big) = P\Big(Y_1 - Y_0 = 1 \;\Big|\; X_1 = 1\Big)\, P(X_1 = 1) \\ + P\Big(Y_1 - Y_0 = 1 \;\Big|\; X_1 = 0\Big)\, P(X_1 = 0). \end{aligned}$$

Thus, even if

$$P\Big(Y_1 - Y_0 = 1 \;\Big|\; X_1 = 1\Big) \geq 0.9,$$

it does **not** necessarily imply

$$P\Big(Y_1 - Y_0 = 1\Big) \geq 0.9,$$

because the overall (unconditional) probability also depends on $P(X_1 = 1)$ and $P(Y_1 - Y_0 = 1 \mid X_1 = 0)$. If $P(X_1 = 1)$ is small or if $P(Y_1 - Y_0 = 1 \mid X_1 = 0)$ is small, the unconditional probability may be much less than 0.9.

4.6 Discussion of ATE versus ITE

In this section, we discuss the relationship between the average treatment effect (ATE) and the individual treatment effect (ITE). We focus on the implications for prediction intervals and hypothesis testing. The ATE is a well-defined parameter and is identifiable under standard assumptions. Consequently, as the sample size increases, confidence intervals for the ATE will eventually shrink to a point. In contrast, the individual treatment effect (ITE) is not a parameter in the classical sense: it is a random quantity defined at the unit level. Although prediction intervals for the ITE may become narrower with larger sample sizes, they do not, in general, converge to zero width. This reflects the inherent uncertainty in predicting individual level responses, even when the population-level effect is precisely estimated.

The Neyman null hypothesis posits a zero effect on average, while the Fisher null hypothesis posits no effect for *every* individual. By construction, a true Fisher null implies the Neyman null; if all individual effects are zero, their average must be zero. A true Fisher null also implies that $\{0\}$ will be a valid prediction interval for any level α. A true Neyman null implies that confidence interval for ATE will contain zero with probability $(1 - \alpha)\%$.

When performing hypothesis testing, $\{0\}$ being a valid prediction interval can be considered as evidence in support of the Fisher null while confidence interval does not contain 0 can be considered as evidence against Neyman null. However, in finite samples, a failure to reject the Fisher null does not imply the ATE is in fact zero, and rejecting the Neyman null (i.e. finding a nonzero estimated ATE) does not automatically imply that $\{0\}$ cannot

be a valid prediction interval.

It is possible to construct a valid prediction interval for ITE as $\{0\}$, even when the data yield a confidence interval for the ATE that excludes zero. Formally, if $\alpha = 0.05$, it is possible that 95% of the individuals have a zero treatment effect, while the ATE analysis, being an aggregate statement, detects a statistically meaningful difference overall from the rest of the 5% population. This implies, perhaps paradoxically, we can have evidence against Neyman null but also evidence in support of Fisher null.

Non-individualized decision policies can consequently outperform individualized decision policies in some partial identification settings; see Cui [2021]. Even with an ATE significantly different from zero, substantial uncertainty at the individual level (reflected in wide ITE intervals) can yield a scenario where assigning the same treatment to everyone is more reliable than any personalized rule that attempts to exploit covariate information. This outcome highlights how partial identification and weakly informative data can obscure the individual-level signal, despite showing clear evidence of an overall effect.

4.6.1 Synthetic data example

Consider a randomized trial for $n = 50000$ participants with binary outcomes, divided into treatment and control groups using a coin flip. Suppose that treatment cures 3% of the population, hurts 1% of the population, while having no effect on 96% of the populations. We observe the following marginals in Table 4.5.

	$\hat{P}(Y=0 \mid D=0) = 0.9896$	$\hat{P}(Y=1 \mid D=0) = 0.0104$
$\hat{P}(Y=0 \mid D=1) = 0.9697$	$P(\text{NR})$	$P(\text{HU})$
$\hat{P}(Y=1 \mid D=1) = 0.0303$	$P(\text{HE})$	$P(\text{AR})$

Table 4.5: Marginal distributions estimated from the numerical simulation.

We estimate the average treatment effect to be 0.0198 with 95% confidence interval $[0.0174, 0.0223]$. This yields a confidence interval for the average treatment effect excluding zero. However, based on the result in Section 4.2.4, we will conclude a 95% prediction

interval for ITE as a singleton $\{0\}$.

Although this situation might seem paradoxical, it simply reflects the fact that population-level inferences about the mean effect can diverge from inferences about individual-level effects under uncertainty. On the one hand, one can accumulate enough information to claim that the *average* effect is nonzero, while on the other hand, one lacks the precision required to identify which individuals truly benefit.

4.6.2 Further discussions

Even with unconfoundedness (or randomization), the difficulty of estimating ITE lies in the unknown structure of potential outcomes Y_1, Y_0. If Y_1 and Y_0 are independent, then the fundamental problem of causal inference does not exist. For example, the introduction section of Yin et al. [2022] can be confusing. Similarly, the section on numerical experiments in Lei and Candès [2021] also assumes the independence of Y_1 and Y_0.

A complementary strategy under partial identification is to condition on covariates X, seeking subpopulations where treatment effects are more homogeneous. Our results can be extended to this scenario when the treatment is conditionally independent of the potential outcomes. For example, if there exist covariates for which $P(Y = 1 \mid D = 1, X)$ or $P(Y = 1 \mid D = 0, X)$ are close to 0 or 1, the uncertainty about individual-level effects in that subgroup may be greatly reduced, enabling tighter bounds on $P\big((Y_1 - Y_0) \in [L, R] \mid X\big)$. However, we still need to be cautious with such conditional individual treatment effect prediction intervals as it could disagree with the conditional average treatment effect (CATE) estimation results.

Chapter 5

CONCLUSION

In this dissertation, we explored several problems related to prediction intervals for the individual treatment effect (ITE). In the second chapter, we addressed some missing pieces in the existing literature about the Kolmogorov's problem: finding the best possible bounds for the distribution function of the sum of two random variables with fixed marginals. In the third chapter, we addressed logical gaps and missing pieces from this line of inquiry to the context of causal inference. We further investigated the coupling of two probability measures on finite sets and derived best possible bounds for the probability mass function under fixed marginals. Finally, we synthesized these insights to address the construction and interpretation of prediction intervals for the ITE, bridging foundational probabilistic results with practical questions in causal inference.

In particular, under binary treatment, we aimed to answer the following problems for binary, ordinal, and continuous outcomes:

- How to construct a valid prediction interval for $Y(1) - Y(0)$ given the marginal distribution of potential outcomes $Y(1)$ and $Y(0)$? The prediction interval needs to be compatible for any joint distribution between $Y(1)$ and $Y(0)$ satisfying the given marginals.

- Given the marginal distribution of potential outcomes $Y(1)$ and $Y(0)$ and a point Y', does Y' need to be included in all prediction intervals if we do not make assumptions about the joint distribution of potential outcomes? In the discrete/binary case, this problem can be formulated as does there exist a joint distribution of $Y(1)$ and $Y(0)$ satisfying the marginals such that $P(Y(1) - Y(0) = Y') > \alpha$? Alternatively, in all cases, we can also formulate a cdf problem: is Y' larger/smaller than the left/right end point of any prediction interval? Does there exist a joint distribution of $Y(1)$ and

$Y(0)$ satisfying the marginals such that $P(Y(1) - Y(0) < Y') > \alpha$?

- Given the marginal distribution of potential outcomes $Y(1)$ and $Y(0)$ and and interval $[L, R]$, does there exist a joint distribution of $Y(1)$ and $Y(0)$ satisfying the given marginals such that $[L, R]$ is a valid $1 - \alpha$ level prediction interval for the ITE?

Table 5.1 provides an overview of where each question is addressed in this dissertation, organized by outcome type.

Outcome types:	Binary	Ordinal	Continuous
Valid prediction intervals	Section 4.2	Section 4.3.4	Section 4.3.1
Must included points	Section 4.2	Appendix C.3	Section 4.3.2
Necessary conditions	Appendix C.2	Appendix C.3	Future work

Table 5.1: Summary of questions addressed in this dissertation, categorized by outcome types.

One of the questions concerning probability bounds within a fixed interval given known marginals remains partially unresolved. While numerical approaches such as discretization provide approximate solutions, a complete analytical characterization is still open. Future research could also further explore estimation, inference, and covariate adjustment for prediction intervals on the individual treatment effect (ITE), building upon the foundational results established in this dissertation.

BIBLIOGRAPHY

Alexander Barvinok. Matrices with prescribed row and column sums. *Linear Algebra and its Applications*, 436(4):820–844, 2012.

Helge Blaker. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*, 28(4):783–798, 2000.

Jennifer Brennan, Sébastien Lahaie, Adel Javanmard, Nick Doudchenko, and Jean Pouget-Abadie. Causal bootstrap for general randomized designs. *arXiv preprint arXiv:2410.21464*, 2024.

Richard A Brualdi, Herbert John Ryser, et al. *Combinatorial matrix theory*, volume 39. Springer, 1991.

Undral Byambadalai, Tatsushi Oka, and Shota Yasui. Estimating distributional treatment effects in randomized experiments: machine learning for variance reduction. *arXiv preprint arXiv:2407.16037*, 2024.

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. 2014.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Toward personalized inference on individual treatment effects. *Proceedings of the National Academy of Sciences*, 120(7): e2300458120, 2023.

Yasutaka Chiba. Exact tests for the weak causal null hypothesis on a binary outcome in randomized trials. *Journal of Biometrics & Biostatistics*, 6(244), 2015.

Yasutaka Chiba. A note on exact confidence interval for causal effects on a binary outcome in randomized trials. *Statistics in Medicine*, 35(10):1739–1741, 2016.

Yasutaka Chiba. Sharp nonparametric bounds and randomization inference for treatment effects on an ordinal outcome. *Statistics in medicine*, 36(25):3966–3975, 2017.

JB Copas. Randomization models for the matched and unmatched 2× 2 tables. *Biometrika*, 60(3):467–476, 1973.

Yifan Cui. Individualized decision-making under partial identification: three perspectives, two optimality results, and one paradox. *arXiv preprint arXiv:2110.10961*, 2021.

A Philip Dawid and Stephen Senn. Personalised decision-making without counterfactuals. *arXiv preprint arXiv:2301.11976*, 2023.

A Philip Dawid, Monica Musio, and Stephen E Fienberg. From statistical evidence to evidence of causality. 2016.

Frank Den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/CouplingLectures. pdf)*, 3, 2012.

Adrian Dobra and Stephen E Fienberg. Bounding entries in multi-way contingency tables given a set of marginal totals. In *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000*, pages 3–16. Springer, 2000.

Paul Embrechts and Marius Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432, 2013.

Paul Embrechts, Alexander McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223, 2002.

Paul Embrechts, Giovanni Puccetti, and Ludger Rüschendorf. Model uncertainty and var aggregation. *Journal of Banking & Finance*, 37(8):2750–2764, 2013.

Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010.

Yanqin Fan and Sang Soo Park. Confidence intervals for the quantile of treatment effects in randomized experiments. *Journal of Econometrics*, 167(2):330–344, 2012.

Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

Sergio Firpo and Geert Ridder. Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234, 2019.

Ronald Aylmer Fisher. Design of experiments. *British Medical Journal*, 1(3923):554, 1936.

Lester Randolph Ford and Delbert Ray Fulkerson. Flows in networks. 2015.

LR Ford Jr and DR Fulkerson. Network flow and systems of representatives. *Canadian Journal of Mathematics*, 10:78–84, 1958.

Maurice J Frank, Roger B Nelsen, and Berthold Schweizer. Best-possible bounds for the distribution of a sum—a problem of kolmogorov. *Probability theory and related fields*, 74 (2):199–211, 1987.

David Gale. A theorem on flows in networks. In *Classic Papers in Combinatorics*, pages 259–268. Springer, 1957.

Marius Hofert, Amir Memartoluie, David Saunders, and Tony Wirjanto. Improved algorithms for computing worst value-at-risk. *Statistics & Risk Modeling*, 34(1-2):13–31, 2017.

Justin Hsu. *Probabilistic couplings for probabilistic reasoning*. University of Pennsylvania, 2017.

Emily J Huang, Ethan X Fang, Daniel F Hanley, and Michael Rosenblum. Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics*, 18(2):308–324, 2017.

Guido Imbens and Konrad Menzel. A causal bootstrap. Technical report, National Bureau of Economic Research, 2018.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

Harry Joe. Majorization, entropy and paired comparisons. *The Annals of Statistics*, pages 915–925, 1988.

Nathan Kallus. What's the harm? sharp bounds on the fraction negatively affected by treatment. *Advances in Neural Information Processing Systems*, 35:15996–16009, 2022a.

Nathan Kallus. Treatment effect risk: Bounds and inference. *arXiv preprint arXiv:2201.05893*, 2022b.

Ju Hyun Kim. Identifying the distribution of treatment effects under support restrictions. *arXiv preprint arXiv:1410.5885*, 2014.

Twan Koperberg. Couplings and matchings: Combinatorial notes on strassen's theorem. *arXiv preprint arXiv:2202.02092*, 2022.

Twan Koperberg. Couplings and matchings: Combinatorial notes on strassen's theorem. *Statistics & Probability Letters*, 209:110089, 2024.

Vladik Kreinovich and Scott Ferson. Computing best-possible bounds for the distribution of a sum of several variables is np-hard. *International Journal of Approximate Reasoning*, 41(3):331–342, 2006.

Sungwon Lee. Partial identification and inference for conditional distributions of treatment effects. *arXiv preprint arXiv:2108.00723*, 2021.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Haijun Li, Marco Scarsini, and Moshe Shaked. Bounds for the distribution of a multivariate sum. *Lecture Notes-Monograph Series*, pages 198–212, 1996.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Jiannan Lu, Peng Ding, and Tirthankar Dasgupta. Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43(5):540–567, 2018.

GD Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26 (4):803–806, 1982.

Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: theory of majorization and its applications. 1979.

Radko Mesiar and Carlo Sempi. Ordinal sums and idempotents of copulas. *Aequationes mathematicae*, 79:39–52, 2010.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

Scott Mueller and Judea Pearl. Personalized decision making–a conceptual introduction. *arXiv preprint arXiv:2208.09558*, 2022.

Scott Mueller, Ang Li, and Judea Pearl. Causes of effects: Learning individual responses from population data. *arXiv preprint arXiv:2104.13730*, 2021.

John Mullahy. Individual results may vary: Inequality-probability bounds for some health-outcome treatment effects. *Journal of Health Economics*, 61:151–162, 2018.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

Jerzy Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):463–480, 1990. Originally presented in 1923.

Yu V Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

Giovanni Puccetti. The beautiful art of rearranging matrices. *Available at SSRN 4818004*, 2024.

Giovanni Puccetti and Ludger Rüschendorf. Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics*, 236(7):1833–1840, 2012.

Giovanni Puccetti and Ruodu Wang. Extremal dependence concepts. *Statistical Science*, 30(4):485 – 517, 2015.

J Rigdon, WW Loh, and MG Hudgens. Ri2by2: Randomization inference for treatment effects on a binary outcome. *R package version*, 1, 2014.

Joseph Rigdon and Michael G Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in medicine*, 34(6):924–935, 2015.

Joseph Rigdon, Wen Wei Loh, and Michael G Hudgens. Response to comment on "randomization inference for treatment effects on a binary outcome". *Statistics in medicine*, 36(5):876, 2017.

James Robins and Sander Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, pages 1125–1138, 1989.

James M Robins. Confidence intervals for causal parameters. *Statistics in medicine*, 7(7): 773–785, 1988.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

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.

Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

Ludger Rüschendorf. Solution of a statistical optimization problem by rearrangement methods. *Metrika*, 30(1):55–61, 1983.

Ludger Rüschendorf. Fréchet-bounds and their applications. In *Advances in Probability Distributions with Given Marginals: beyond the copulas*, pages 151–187. Springer, 1991.

Herbert J Ryser. Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics*, 9:371–377, 1957.

Numair Sani, Atalanti A Mastakouri, and Dominik Janzing. Bounding probabilities of causation through the causal marginal problem. *arXiv preprint arXiv:2304.02023*, 2023.

Thorsten Schmidt. Coping with copulas. *Copulas-From theory to application in finance*, 3: 34, 2007.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

M Sklar. Fonctions de répartition à N dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

Yilin Song, Richard Guo, KC Chan, and Thomas S Richardson. The instrumental variable model with categorical instrument, treatment and outcome. *arXiv preprint arXiv:2405.09510*, 2024.

Jörg Stoye. Partial identification of spread parameters. *Quantitative Economics*, 1(2): 323–357, 2010.

Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

Baozhen Wang and Xingye Qiao. Conformal inference of individual treatment effects using conditional density estimates. *arXiv preprint arXiv:2501.14933*, 2025.

Robert C Williamson and Tom Downs. Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds. *International journal of approximate reasoning*, 4(2):89–158, 1990.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2022.

Zhehao Zhang and Thomas S Richardson. Bounds on the distribution of a sum of two random variables: Revisiting a problem of kolmogorov with application to individual treatment effects. *arXiv preprint arXiv:2405.08806*, 2024.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Table on prior work and relation to this chapter

	Def. of cdf	Def. of Sharpness	Lower Bound on $P(Z < z)$	Lower Bound on $P(Z \leq z)$	Upper Bound on $P(Z < z)$	Upper Bound on $P(Z \leq z)$
Makarov [1982]	$\widetilde{F}$	sup / inf	✓	✓	✓	✓
Rüschendorf [1982]	F	attainable	✓			✓
Frank et al. [1987]	$\widetilde{F}$	attainable	✓			✓
Nelsen [2006]	F	attainable	✓			✓
This chapter	F	sup/inf	✓	✓	✓	✓
		attainable	✓	Thm. 2.4.1	Thm. 2.4.2	✓

Table A.1: Concordance with prior work illustrating the different definitions and bounds provided in relation to the results in this chapter. 'sup/inf' indicates that the given bound is the best-possible, i.e. sup or inf over all joint distributions consistent with the prescribed marginals, whereas 'attainable' means there is at least one joint distribution achieving that bound. A ✓ means the property (best-possible or attainability) holds for all choices of marginals, a blank means it may fail for some marginals, and Theorems 2.4.1-2.4.2 provide conditions ensuring these bounds are indeed attained.

A.2 Relationship between bounds defined using left or right continuous cdfs

Note that we previously defined the left-continuous distribution functions $\widetilde{F}(x) := P(X < x)$ and $\widetilde{G}(y) := P(Y < y)$, so that indeed $\widetilde{F}(x) = F(x-)$ and $\widetilde{G}(x) = G(x-)$. Recall that the lower bound in our setup is given by

$$\tau_W(F, G)(z) \;=\; \sup_{x+y=z} \max\big(F(x) + G(y) - 1,\, 0\big),$$

and the upper bound is

$$\rho_W(F,G)(z) \;=\; 1 + \inf_{x+y=z} \min\big(0,\; F(x)+G(y)-1\big).$$

We want to prove that for fixed z,

$$\tau_W(\widetilde{F},\widetilde{G})(z-) \;\le\; \tau_W(\widetilde{F},\widetilde{G})(z) \;=\; \tau_W(F,G)(z-) \;\le\; \tau_W(F,G)(z),$$

and

$$\rho_W(\widetilde{F},\widetilde{G})(z-) \;\le\; \rho_W(\widetilde{F},\widetilde{G})(z) \;=\; \rho_W(F,G)(z-) \;\le\; \rho_W(F,G)(z).$$

The outer inequalities are given by monotonicity arguments. In particular, for fixed F, G, the map $z \mapsto \tau_W(F,G)(z)$ is non-decreasing, so $\tau_W(F,G)(z-) \le \tau_W(F,G)(z)$, and similarly for $\tau_W(\widetilde{F},\widetilde{G})$, $\rho_W(F,G)$, and $\rho_W(\widetilde{F},\widetilde{G})$. It remains to show

$$\tau_W(\widetilde{F},\widetilde{G})(z) = \tau_W(F,G)(z-) \quad \text{and} \quad \rho_W(\widetilde{F},\widetilde{G})(z) = \rho_W(F,G)(z-).$$

First, we focus on $\tau_W(\widetilde{F},\widetilde{G})(z) = \tau_W(F,G)(z-)$. Based on the definition of τ_W, we want to show that

$$\sup_{x+y=z} \max(\widetilde{F}(x)+\widetilde{G}(y)-1,0) = \sup_{x+y=z} \max(F(x-)+G(y-)-1,0) = \sup_{x+y=z-} \max(F(x)+G(y)-1,0)$$

Because adding a constant and taking max with 0 does not change equality, it suffices to show

$$\sup_{x+y=z} (F(x-)+G(y-)) = \sup_{x+y=z-} (F(x)+G(y)).$$

We will prove this equality by showing two inequalities. First, we have

$$\sup_{x+y=z-} (F(x)+G(y)) := \lim_{h^*\to 0, h^*>0} \sup_{x+y=z-h^*} (F(x)+G(y)) \tag{A.1}$$

$$= \lim_{h^*\to 0, h^*>0} \sup_{x+y=z} (F(x-\frac{h^*}{2})+G(y-\frac{h^*}{2})). \tag{A.2}$$

Since F, G are non-decreasing, for all x, y and $h^* > 0$,

$$F(x - \frac{h^*}{2}) + G(y - \frac{h^*}{2}) \leq \lim_{h \to 0, h > 0} (F(x - \frac{h}{2}) + G(y - \frac{h}{2})).$$

Taking the sup on both sides gives:

$$\sup_{x+y=z} F(x - \frac{h^*}{2}) + G(y - \frac{h^*}{2}) \leq \sup_{x+y=z} \lim_{h \to 0, h > 0} (F(x - \frac{h}{2}) + G(y - \frac{h}{2})).$$

Therefore,

$$\sup_{x+y=z-} (F(x) + G(y)) = \lim_{h^* \to 0, h^* > 0} \sup_{x+y=z} (F(x - \frac{h^*}{2}) + G(y - \frac{h^*}{2})) \leq \sup_{x+y=z} (F(x-) + G(y-))$$

For the second direction, since F and G are non-decreasing, for any $\epsilon > 0$, there exists $h^* > 0$ such that for all $h < h^*$:

$$F(x - \frac{h}{2}) \geq F(x-) - \frac{\epsilon}{2}, \quad G(y - \frac{h}{2}) \geq G(y-) - \frac{\epsilon}{2}.$$

This implies for all $h < h^*$:

$$F(x - \frac{h}{2}) + G(y - \frac{h}{2}) \geq F(x-) + G(y-) - \epsilon. \tag{A.3}$$

Taking the supremum over (x, y) gives:

$$\sup_{x+y=z} (F(x - \frac{h}{2}) + G(y - \frac{h}{2})) \geq \sup_{x+y=z} (F(x-) + G(y-)) - \epsilon$$

for all $h < h^*$. The $\lim_{h \downarrow 0}$ of the left-hand side (which is equal to (A.2)) is at least the right-hand side. Since $\epsilon > 0$ was arbitrary, this gives

$$\sup_{x+y=z-} (F(x) + G(y)) \geq \sup_{x+y=z} (F(x-) + G(y-)).$$

Therefore, we have

$$\sup_{x+y=z-}(F(x)+G(y)) = \sup_{x+y=z}(F(x-)+G(y-)).$$

A completely analogous argument applies to the upper bound ρ_W, just replacing the 'sup' by 'inf'. We want to prove that

$$\inf_{x+y=z}\min(1, F(x-)+G(y-)) = \inf_{x+y=z-}\min(1, F(x)+G(y)).$$

Again it suffices to show that

$$\inf_{x+y=z}(F(x-)+G(y-)) = \inf_{x+y=z-}(F(x)+G(y)).$$

Since F, G are non-decreasing, we note that

$$\inf_{x+y=z} F(x-\frac{h^*}{2})+G(y-\frac{h^*}{2}) \leq \inf_{x+y=z}\lim_{h\to 0, h>0}(F(x-\frac{h}{2})+G(y-\frac{h}{2})).$$

Therefore,

$$\inf_{x+y=z-}(F(x)+G(y)) = \lim_{h^*\to 0, h^*>0}\inf_{x+y=z}(F(x-\frac{h^*}{2})+G(y-\frac{h^*}{2})) \leq \inf_{x+y=z}(F(x-)+G(y-)).$$

Using equation (A.3), and taking the infimum over (x, y) gives:

$$\inf_{x+y=z}(F(x-\frac{h}{2})+G(y-\frac{h}{2})) \geq \inf_{x+y=z}(F(x-)+G(y-)) - \epsilon$$

for all $h < h^*$. The $\lim_{h\downarrow 0}$ of the left-hand side is at least the right-hand side. Since $\epsilon > 0$ was arbitrary, this gives

$$\inf_{x+y=z-}(F(x)+G(y)) \geq \inf_{x+y=z}(F(x-)+G(y-)),$$

and

$$\inf_{x+y=z-}(F(x)+G(y)) = \inf_{x+y=z}(F(x-)+G(y-)).$$

A.3 Achievability of the bounds

Puccetti and Rüschendorf [2012] comment that a general result in Rüschendorf [1983] implies the achievability of the infimum and supremum of functionals (lower bound on $P(X+Y < z)$ and upper bound on $P(X+Y \leq z)$). Here for the sake of completeness, we give the argument in detail. This provides a (non-constructive) proof of the result in Frank et al. [1987]. To prove the achievability of the bounds, we start with a definition.

Definition A.3.1. *A function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is lower semicontinuous if, for all $(x, y) \in \mathbb{R}^2$,*

$$\liminf_{(x',y')\to(x,y)} \varphi(x', y') \geq \varphi(x, y).$$

Given marginal distribution functions F and G for random variables X and Y respectively, let $\mathcal{M}(F, G)$ denote the set of all joint distribution functions on (X, Y) that have the given marginals. Rüschendorf [1983] proved that the set $\mathcal{M}(F, G)$ is convex, tight, and closed. Therefore, by Prokhorov's theorem [Prokhorov, 1956], $\mathcal{M}(F, G)$ is compact with respect to the weak topology.

Proposition A.3.2. *For a measurable function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ and the compact set $\mathcal{M}(F, G)$,*

$$\inf_{H\in\mathcal{M}(F,G)} \int \varphi \, dH$$

achieves its minimum in $\mathcal{M}(F, G)$ when φ is lower semicontinuous.

Proof. Since $\mathcal{M}(F, G)$ is compact with respect to the weak topology and φ is lower semicontinuous, we can apply the Portmanteau theorem. Specifically, for any sequence $\{H_n\} \subset \mathcal{M}(F, G)$ that converges weakly to some $H \in \mathcal{M}(F, G)$, we have

$$\int \varphi \, dH \leq \liminf_{n\to\infty} \int \varphi \, dH_n.$$

This means the mapping $H \mapsto \int \varphi \, dH$ is lower semicontinuous on $\mathcal{M}(F, G)$. Since the infimum of a lower semicontinuous function on a compact set is attained, there exists $H^* \in$

$\mathcal{M}(F,G)$ such that

$$\inf_{H\in\mathcal{M}(F,G)}\int \varphi\, dH = \int \varphi\, dH^*.$$

□

Proposition A.3.3. *For any given $z \in \mathbb{R}$, the function $\varphi(X,Y) = \mathbb{1}_{\{X+Y<z\}}$ is lower semicontinuous.*

Proof. Consider the function $\varphi(x,y) = \mathbb{1}_{\{x+y<z\}}$.

Case 1: If $x + y < z$, then $\varphi(x,y) = 1$. For any sequence $(x_n, y_n) \to (x,y)$, we have $x_n + y_n < z$ for sufficiently large n. Thus, $\varphi(x_n, y_n) = 1$ eventually, and

$$\liminf_{(x',y')\to(x,y)} \varphi(x',y') = 1 = \varphi(x,y).$$

Case 2: If $x + y \geq z$, then $\varphi(x,y) = 0$. For any sequence $(x_n, y_n) \to (x,y)$, $\varphi(x_n, y_n) \geq 0 = \varphi(x,y)$. Therefore,

$$\liminf_{(x',y')\to(x,y)} \varphi(x',y') \geq \varphi(x,y).$$

In both cases, the condition for lower semicontinuity is satisfied. Hence, φ is lower semicontinuous. □

Therefore, by Proposition A.3.2 for all $z \in \mathbb{R}$, there exists a joint distribution function $H_z(x,y)$ such that $P(X+Y<z)$ under H_z equals

$$\inf_{H\in\mathcal{M}(F,G)} P_H(X+Y<z),$$

where the infimum is taken over all joint distribution functions $H(x,y)$ with marginals $F(x)$ and $G(y)$.

An analog to Proposition A.3.2 will give: $\sup_{H\in\mathcal{M}(F,G)} \int \varphi\, dH$ achieves its maximum in compact set $\mathcal{M}(F,G)$ when φ is upper-semicontinuous. Thus, for all z, the upper bound on $P(X+Y\leq z)$ is achievable since $\varphi = \mathbb{1}_{\{X+Y\leq z\}}$ is upper semicontinuous. □

Remark A.3.4. *Note this does not address achievability of upper bounds on $P(X+Y<z)$ or lower bounds on $P(X+Y\leq z)$ since their functions are not respectively upper and lower*

semi-continuous.

A.4 Relating Rüschendorf and Makarov Bounds

In Section 2, Proposition 1 of Rüschendorf [1982], the upper bound M (which we denote by $\rho_W(F,G)(z)$) on $A_2(t)$ (corresponding to $P(X+Y \leq z)$ in our notation) is given by

$$F_1 \wedge F_2(t) := \inf_x \Big(F_1(x-) + F_2(t-x) \Big),$$

which in our notation this bound would be

$$\inf_x (F(x-) + G(z-x)) = \inf_{x+y=z} (F(x-) + G(y)).$$

In Theorem 2.2.6, we state the upper bound as

$$\inf_{x+y=z} \min\big(1, F(x) + G(y)\big).$$

There are two notable differences between these formulations. First, the Makarov bound explicitly takes the minimum with 1 to ensure that the upper bound does not exceed 1, thereby avoiding a trivial bound.[1] Second, the expression in Rüschendorf [1982] uses the left-hand limit $F_1(x-)$ instead of $F(x)$ as used in our bound $\inf_x \big(F(x)+G(z-x)\big)$. Although it may not be immediately obvious, one can show that the two formulations are indeed equivalent.

Proposition A.4.1. *For any $z \in \mathbb{R}$ and distribution functions F, G,*

$$\inf_x (F(x) + G(z-x)) = \inf_x (F(x-) + G(z-x))$$

Proof. We will prove this equality by showing two inequalities.

[1] The Makarov lower bound for $P(X+Y<z)$ differs from the lower bound in Rüschendorf [1982] only by taking the maximum with 0; we do not elaborate on this minor distinction here.

Since F is a distribution function, $F(x-) \leq F(x)$ for all x, thus

$$F(x-) + G(z-x) \leq F(x) + G(z-x).$$

Taking the infimum over all x preserves the inequality:

$$\inf_x (F(x-) + G(z-x)) \leq \inf_x (F(x) + G(z-x)).$$

For the second direction, consider a point x and a sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $x_n \uparrow x$ (i.e., $x_n < x$ for all n and $\lim_{n\to\infty} x_n = x$) and $\lim_{n\to\infty} F(x_n) = F(x-)$. Consider the expression $F(x_n) + G(z-x_n)$. By definition of infimum, for all n,

$$\inf_{x^*} (F(x^*) + G(z-x^*)) \leq F(x_n) + G(z-x_n).$$

Taking the limit as $n \to \infty$, we get

$$\inf_{x^*} (F(x^*) + G(z-x^*)) \leq F(x-) + \lim_{n\to\infty} G(z-x_n).$$

As $z - x_n$ converges to $z - x$ from the right, we have $\lim_{n\to\infty} G(z-x_n) = G(z-x)$ by right continuity of G. Therefore,

$$\inf_{x^*} (F(x^*) + G(z-x^*)) \leq F(x-) + G(z-x).$$

Taking the infimum over x on the right side yields

$$\inf_{x^*} (F(x^*) + G(z-x^*)) \leq \inf_x (F(x-) + G(z-x)).$$

Therefore, we have:

$$\inf_x (F(x) + G(z-x)) = \inf_x (F(x-) + G(z-x)).$$

where the right hand side here is the upper bound in Rüschendorf [1982] in our notation. □

In fact, there is another equivalent way to represent this upper bound as we will show next.

Corollary A.4.2. *For any $z \in \mathbb{R}$ and distribution functions F, G,*

$$\inf_{x+y=z}(F(x)+G(y)) = \inf_{x+y=z}(F(x)+G(y-)) = \inf_{y}(F(z-y)+G(y-)).$$

Proof. This follows from the symmetry of F, G in the infimum expression and we can exchange x, y in the sum. □

Note that the conclusions in Proposition A.4.1 and Corollary A.4.2 do not hold when the inf operators are replaced with sup operators.

A.5 Simulation under the special copula

We note that the special copula C_t which achieves the lower bound on $P(X+Y<z)$ for $t = \sup_{x+y=z} \max(F(x)+G(y)-1, 0)$ in Frank et al. [1987], Nelsen [2006] is degenerate in the sense that it does not have a density with respect to the two-dimensional Lebesgue measure over the unit square. Recall that it is defined as:

$$C_t(u,v) = \begin{cases} \text{Max}(u+v-1, t), & (u,v) \text{ in } [t,1]\times[t,1], \\ \text{Min}(u,v), & \text{otherwise.} \end{cases}$$

Though possibly not immediately obvious, under this copula, $V = G(Y)$ is almost surely a deterministic (piecewise-defined) function of $U = F(X)$ (and vice versa).[2] Thus, we can easily simulate the joint distribution from the copula with a single draw. We introduce the following algorithm:

---

[2]This does not mean X is a deterministic function of Y unless they are both absolutely continuous with respect to the Lebesgue measure.

Algorithm 2 Simulate Joint Distribution from the Copula C_t

Compute $t = \sup_{x+y=z} \max(F(x) + G(y) - 1, 0)$

for $i = 1, \ldots, n$ **do**

Draw $u_i \sim \text{Unif}[0, 1]$.

Compute

$$v_i = u_i \mathbb{1}\{u_i \leq t\} + (1 + t - u_i)\mathbb{1}\{u_i > t\}.$$

Set $x_i = F^{-1}(u_i)$ and $y_i = G^{-1}(v_i)$, where the generalized inverses are defined in Definition 2.2.2.

Take (x_i, y_i) as a generated sample.

end for

Similarly, the joint distribution under copula C_r, which achieves the upper bound on $P(X + Y \leq z)$ with $r = \inf_{x+y=z} \min\big(1, F(x) + G(y)\big)$ is defined as:

$$C_r(u, v) = \begin{cases} \text{Max}(u + v - r, 0), & (u, v) \text{ in } [0, r] \times [0, r], \\ \text{Min}(u, v), & \text{otherwise.} \end{cases}$$

C_r can be simulated similarly using the deterministic function $v_i = (r - u_i)\mathbb{1}\{u_i \leq r\} + u_i \mathbb{1}\{u_i > r\}$. Here, copulas C_t and C_r are ordinal sums [Mesiar and Sempi, 2010, Nelsen, 2006] of the singular copulas $W(u, v) = \max(u + v - 1, 0)$ and $M(u, v) = \min(u, v)$ corresponding to the Fréchet-Hoeffding lower and upper bounds.

As noted in Frank et al. [1987], copulas that achieve the bounds in certain scenarios are not unique. For example, $\text{Min}(u, v)$ in $[0, t] \times [0, t]$ of C_t can be replaced by uv. Similarly, $\text{Min}(u, v)$ in $[r, 1] \times [r, 1]$ of C_r can also be replaced by uv. However, each of these copulas is an ordinal sum of two copulas, with at least one being degenerate. As a result, the induced joint distribution may fail to be continuous—even when the marginal distributions are continuous. Consequently, the Makarov bounds need not always be attainable under these copulas (see Theorem 2.3.16). In contrast, when at least one of the marginals is discrete, the Makarov bounds are always attained (see Theorem 2.3.9).

Appendix B

APPENDIX TO CHAPTER 3

B.1 Revisit Theorem 2 in Williamson and Downs [1990]

This dissertation was in part motivated by the observation that authors were using the lower bound given in Williamson and Downs [1990] that were claimed to be sharp when this was clearly not true. Here we identify the error present in Williamson and Downs [1990]'s proof. The proof of Theorem 2 of Williamson and Downs [1990] states "Let $Y' = -Y$. Then $F_{Y'}(y) = 1 - F(-y)$". Since Williamson and Downs [1990] use the left-continuous version definition of cdf, in our notation,

$$\begin{aligned}
\tilde{G}'(y) &= P(Y' < y) \\
&= P(-Y < y) \\
&= P(Y > -y) \\
&= 1 - P(Y \leq -y) \\
&= 1 - \tilde{G}(-y) - P(Y = -y),
\end{aligned}$$

where $\tilde{G}'(y) = P(Y' < y)$. This statement is also not correct if we use the right-continuous version definition of cdf, as we have

$$\begin{aligned}
G'(y) &= P(Y' \leq y) \\
&= P(-Y \leq y) \\
&= P(Y \geq -y) \\
&= 1 - P(Y < -y) \\
&= 1 - G(-y) + P(Y = -y).
\end{aligned}$$

In fact, it is easy to show that $G'(y) = 1 - G(-y)$ if and only if $\tilde{G}'(y) = 1 - \tilde{G}(-y)$.

In this dissertation, we only focus on the sum and difference of two random variables, however, a similar mistake also appears in the argument given for the bounds on the cumulative distribution function for the quotient of two random variables stated in Theorem 2 of Williamson and Downs [1990].

Corollary A.4.2 in Chapter 1 provides intuition to prove Theorem 3.2.3, as detailed in Appendix B.2.

B.2 Proof of Theorem 3.2.3

Here we prove, as stated in Theorem 3.2.3, that the upper bound on the cdf for the difference $\Delta = X - Y$ given by Williamson and Downs [1990] is valid even though the proof in Williamson and Downs [1990] is not correct.

It is sufficient to show that for any random variables X and Y with respective cdfs $F(\cdot)$ and $G(\cdot)$,

$$\inf_{x-y=\delta} \min\{F(x) - P(Y < y), 0\} = \inf_{x-y=\delta} \min\{F(x) - G(y), 0\}. \tag{B.1}$$

Recall that Proposition A.4.2 shows

$$\inf_{x+y=z} (F(x) + G(y)) = \inf_{x+y=z} (F(x) + G(y-)).$$

Consider a new variable $Y' = -Y$ with cdf G'. Then for any δ, applying Proposition A.4.2 we get:

$$\inf_{x+y'=\delta} (F(x) + G'(y')) = \inf_{x+y'=\delta} (F(x) + G'(y'-)).$$

For $y' = -y$, we see that

$$\begin{aligned} G'(y'-) &= P(Y' < y') \\ &= P(-Y < y') \\ &= P(Y > -y') \\ &= 1 - P(Y \leq -y') \\ &= 1 - G(-y') \\ &= 1 - G(y). \end{aligned}$$

And similarly $G'(y') = P(-Y \leq y') = P(Y \geq -y') = 1 - P(Y < -y') = 1 - P(Y < y)$, so

$$\inf_{x-y=\delta}(F(x) - P(Y < y) + 1) = \inf_{x-y=\delta}(F(x) + G(y) + 1).$$

Because adding or subtracting a constant and taking minimum do not affect the relevant infimum. Thus, the two expressions in (B.1) coincide.

Appendix C

APPENDIX TO CHAPTER 4

C.1 When is a given prediction interval valid?

C.1.1 When is $\{0\}$ *a valid prediction interval for the ITE?*

A valid interval consists of $\{0\}$ when the proportion of people of type Always Recover plus the proportion of type Never Recover is known to be greater than or equal to $(1-\alpha)$.

This implies that:

$$\begin{aligned} &1 - P(Y=1 \mid D=0) - P(Y=1 \mid D=1) \\ &+ 2\max\{0, (P(Y=1 \mid D=0) + P(Y=1 \mid D=1)) - 1\} \geq (1-\alpha), \end{aligned}$$

equivalently,

$$\begin{aligned} &\max\{1 - P(Y=1 \mid D=0) - P(Y=1 \mid D=1), \\ &P(Y=1 \mid D=0) + P(Y=1 \mid D=1) - 1\} \geq (1-\alpha), \end{aligned}$$

or in other words, either

$$P(Y=1 \mid D=0) + P(Y=1 \mid D=1) \leq \alpha,$$

or

$$P(Y=0 \mid D=0) + P(Y=0 \mid D=1) \leq \alpha.$$

When is $\{1\}$ *a valid prediction interval for the ITE?*

A valid interval consists of $\{1\}$ when the proportion of people of type Helped is always greater than or equal to $(1-\alpha)$.

This will occur when:

$$P(Y=1 \mid D=1) - P(Y=1 \mid D=0) \geq (1-\alpha).$$

In other words, the Average Treatment Effect is required to be greater than $(1-\alpha)$.

When is $\{-1\}$ *a valid prediction interval for the ITE?*

A valid interval consists of $\{-1\}$ when the proportion of people of type Hurt is always greater than or equal to $(1-\alpha)$.

This will occur when:

$$P(Y=1 \mid D=0) - P(Y=1 \mid D=1) \geq (1-\alpha).$$

In other words, the Average Treatment Effect is required to be less than $-(1-\alpha)$.

C.1.2 When is the valid prediction interval for the individual treatment effect non-negative/non-positive?

We now consider when a valid $(1-\alpha)\%$ prediction interval rules out a negative/positive result. There are two cases to consider:

When is $[0,1]$ *a valid prediction interval for the ITE?*

A valid interval consists of $[0,1]$ when the proportion of people of type Helped plus proportion of people of type Always Recover plus proportion of people of type Never Recover is always greater than or equal to $(1-\alpha)$. In other words, the proportion of people of type Hurt should always be less than α. This implies that

$$P(Y=1 \mid D=0) - t < \alpha$$

for all t. That is,

$$P(Y = 1 \mid D = 0) - \max\{0, (P(Y{=}1 \mid D{=}0) + P(Y{=}1 \mid D{=}1)) - 1\} < \alpha,$$

equivalently, either

$$P(Y = 1 \mid D = 0) < \alpha,$$

or

$$1 - P(Y = 1 \mid D = 1) < \alpha.$$

When is $[-1, 0]$ *a valid prediction interval for the ITE?*

A valid interval consists of $[-1, 0]$ when the proportion of people of type Hurt plus proportion of people of type Always Recover plus proportion of people of type Never Recover is always greater than or equal to $(1 - \alpha)$. In other words, the proportion of people of type Helped should always be less than α. This implies that

$$P(Y = 1 \mid D = 1) - t < \alpha$$

for all t. That is,

$$P(Y = 1 \mid D = 1) - \max\{0, (P(Y{=}1 \mid D{=}0) + P(Y{=}1 \mid D{=}1)) - 1\} < \alpha,$$

equivalently, either

$$P(Y = 1 \mid D = 1) < \alpha,$$

or

$$1 - P(Y = 1 \mid D = 0) < \alpha.$$

C.2 Necessary conditions for a given prediction interval to be valid and of minimal length

C.2.1 Necessary conditions for $\{1\}$ *being a valid prediction interval*

Suppose we know $\{1\}$ is a valid prediction interval, what are the necessary conditions on the marginal distributions $P(Y = 1 \mid D = 0)$ and $P(Y = 1 \mid D = 1)$? This means that there exists t such that the proportion of Helped is greater than $1 - \alpha$. That is:

$$P(Y = 1 \mid D = 1) - t \geq 1 - \alpha \tag{C.1}$$

for some

$$\max\{0, P(Y = 1 \mid D = 1) + P(Y = 1 \mid D = 0) - 1\} \leq t \leq \min\{P(Y = 1 \mid D = 1), P(Y = 1 \mid D = 0)\}.$$

We obtain the constraints

$$P(Y = 1 \mid D = 1) \geq 1 - \alpha;$$
$$P(Y = 1 \mid D = 0) \leq \alpha.$$

For any marginal distribution that satisfies these constraints, there exists a t that satisfies (C.1).

C.2.2 Necessary conditions for $\{-1\}$ *being a valid prediction interval*

Similarly, for

$$P(Y = 1 \mid D = 0) \geq 1 - \alpha,$$
$$P(Y = 1 \mid D = 1) \leq \alpha,$$

there exists t such that

$$P(Y = 1 \mid D = 0) - t \geq 1 - \alpha.$$

C.2.3 Necessary conditions for $\{0\}$ *being a valid prediction interval*

Suppose we know $\{0\}$ is a valid prediction interval, what are the necessary conditions on the marginal distributions $P(Y = 1 \mid D = 0)$ and $P(Y = 1 \mid D = 1)$? This means the proportion of Never Recover plus Always Recover is greater than or equal to $1 - \alpha$. That is:

$$1 - P(Y = 1 \mid D = 1) - P(Y = 1 \mid D = 0) + 2t \geq 1 - \alpha. \tag{C.2}$$

For any

$$|P(Y = 1 \mid D = 1) - P(Y = 1 \mid D = 0)| \leq \alpha,$$

there exists a t that satisfies $(C.2)$.

C.2.4 Necessary conditions for $[-1, 1]$ *being the best prediction interval*

Suppose we know $[-1, 1]$ is the best prediction interval we can get. This means there exists t such that the proportion of people of type Hurt and the proportion of people of type Helped are both greater than or equal to α. That is:

$$P(Y = 1 \mid D = 1) - t \geq \alpha;$$
$$P(Y = 1 \mid D = 0) - t \geq \alpha.$$

We obtain the constraints

$$\alpha \leq P(Y = 1 \mid D = 1) \leq 1 - \alpha; \tag{C.3}$$
$$\alpha \leq P(Y = 1 \mid D = 0) \leq 1 - \alpha. \tag{C.4}$$

For any marginal distribution that satisfies these constraints, there exists a t that satisfies both $(C.3), (C.4)$.

C.2.5 Necessary conditions for $[-1, 0]$ *being the best prediction interval*

Suppose we know $[-1, 0]$ is the best prediction interval we can get. First, it needs to be valid. This means there exists t such that the proportion of people of type Hurt, Always Recovered and Never Recovered needs to be greater than or equal to $1 - \alpha$. That is, the proportion of people of type Helped is less than or equal to α. Then it needs to be best (we cannot conclude $\{-1\}, \{0\}$ as prediction intervals and $[-1, 0]$ has better coverage than $[0, 1]$), that means the proportion of people of type Hurt is greater than or equal to the proportion of people of type Helped, the proportion of people of type Hurt is less than $1 - \alpha$, the proportion of people of type Always Recover plus the proportion of people of type Never Recover is less than $1 - \alpha$. That is:

$$P(Y = 1 \mid D = 1) - t \leq \alpha; \tag{C.5}$$

$$P(Y = 1 \mid D = 0) - t \geq P(Y = 1 \mid D = 1) - t; \tag{C.6}$$

$$P(Y = 1 \mid D = 0) - t < 1 - \alpha; \tag{C.7}$$

$$1 - P(Y = 1 \mid D = 1) - P(Y = 1 \mid D = 0) + 2t < 1 - \alpha. \tag{C.8}$$

We obtain the constraints

$$\begin{aligned} P(Y = 1 \mid D = 0) &\geq P(Y = 1 \mid D = 1); \\ 1 - P(Y = 1 \mid D = 0) + P(Y = 1 \mid D = 1) &> \alpha; \\ P(Y = 1 \mid D = 0) + P(Y = 1 \mid D = 1) &> \alpha; \\ 2 - P(Y = 1 \mid D = 0) - P(Y = 1 \mid D = 1) &> \alpha. \end{aligned}$$

For any marginal distribution that satisfies these constraints, there exists a t that satisfies both $(C.5) - (C.8)$.

C.2.6 Necessary conditions for $[0, 1]$ *being a best prediction interval*

Similarly, suppose we know $[0, 1]$ is the best prediction interval we can get. This means there exists t such that the proportion of people of type Hurt is less than or equal to α,

the proportion of people of type Helped is greater than or equal to the proportion of people of type Hurt, the proportion of people of type Helped is less than $1-\alpha$, the proportion of people of type Always Recover plus the proportion of people of type Never Recover is less than $1-\alpha$. That is:

$$P(Y=1 \mid D=0) - t \leq \alpha; \tag{C.9}$$

$$P(Y=1 \mid D=1) - t \geq P(Y=1 \mid D=0) - t; \tag{C.10}$$

$$P(Y=1 \mid D=1) - t < 1-\alpha; \tag{C.11}$$

$$1 - P(Y=1 \mid D=1) - P(Y=1 \mid D=0) + 2t < 1-\alpha. \tag{C.12}$$

We obtain the constraints

$$\begin{aligned} P(Y=1 \mid D=1) &\geq P(Y=1 \mid D=0); \\ 1 - P(Y=1 \mid D=1) + P(Y=1 \mid D=0) &> \alpha; \\ P(Y=1 \mid D=0) + P(Y=1 \mid D=1) &> \alpha; \\ 2 - P(Y=1 \mid D=0) - P(Y=1 \mid D=1) &> \alpha. \end{aligned}$$

For any marginal distribution that satisfies these constraints, there exists a t that satisfies both $(C.9) - (C.12)$.

C.3 Understanding prediction intervals for individual treatment effect when the outcome is ordinal

We now return to the question originally considered in section 4.2 on prediction intervals with ordinal outcome.

C.3.1 When will the prediction interval be trivial?

What is the condition on the observed distribution under which without additional knowledge the only interval that we can be sure is valid is a trivial prediction interval? Let's assume Y_1, Y_0 takes value $\{0, 1, 2, ..., n-1\}$. Then a trivial prediction interval would be $[-n+1, n-1]$. If it is possible that $P(\text{ITE} = n-1)$ and $P(\text{ITE} = -n+1)$ can both be

greater than α, then we can only provide a trivial prediction interval. A necessary condition for $P(\text{ITE} = n-1)$ and $P(\text{ITE} = -n+1)$ to be greater than α is that the Fréchet inequality upper bounds for both $P(Y_0 = n-1, Y_1 = 0)$ and $P(Y_0 = 0, Y_1 = n-1)$ are greater than α. That is,

$$\begin{aligned} & \min\{P(Y_1 = 0), P(Y_0 = n-1)\} > \alpha \\ \text{and} \quad & \min\{P(Y_1 = n-1), P(Y_0 = 0)\} > \alpha. \end{aligned}$$

Here we only consider prediction intervals. We could also consider non-overlapping prediction intervals or prediction sets.

C.3.2 When can we tell that the prediction interval need not include 0*?*

We want to answer the question: what is the marginal condition that allow us to conclude the prediction interval not include 0. A necessary condition for this is $P(\text{ITE} = 0) < \alpha$. Based on Theorem 3.5.1, the necessary condition requires that

$$\sum_i \min\{P(Y_1 = i), P(Y_0 = i)\} < \alpha.$$

Alternatively, if we interpret the question as whether the left endpoint of the prediction interval is less than or equal to 0 and the right endpoint is greater than or equal to 0, then the same result from Section 4.3.3 applies. In particular, the conditions derived using cdf bounds for the continuous outcome setting can also be applied to the ordinal outcome setting, allowing us to assess whether the prediction interval needs to contain zero.

C.3.3 Necessary conditions for a prediction interval or set to be valid

Given a prediction interval or, more generally, any prediction set in the ordinal outcome setting, we can determine whether there exists a joint distribution of the potential outcomes under which the prediction set is valid. This follows directly from the extension of the finite version of Strassen's Theorem, as presented in Theorem 3.4.2. Notably, this result enables us to characterize necessary conditions for validity without assuming any metric structure

on the ordinal outcomes. This can be particularly useful in settings where we are interested in validating a small number of prediction on certain types, but do not wish to impose comparability of ordinal outcomes in terms of differences.

VITA

Zhehao (Kenny) Zhang was born in Shanghai, China. He received his dual B.S. degrees in Mathematics and Statistical Science from the College of Creative Studies at the University of California, Santa Barbara. He earned his Ph.D. in Statistics from the University of Washington, Seattle, in 2025.