Conditional Independence and Applications in Statistical Causality

Panayiota Constantinou
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
Girton College

This dissertation is submitted for the degree of
Doctor of Philosophy
May 2013

Acknowledgements

I would like to take this opportunity to thank all the people that have contributed in many different ways to the development of this thesis.

Most importantly, I would like to thank my supervisor Philip Dawid who has made this journey possible. Under your guidance and support I have learned to research, explore and discover. Your kindness and encouragement have been major driving forces for the completion of this thesis. I have learned a lot from you and I will always try to follow your example. Sincerely, it has been a privilege to have you as my supervisor.

Also, I would like to thank Susan Pitts and Richard Samworth for the friendly discussions and advice. Eva Myers for her help with computer matters and Julia Blackwell, John Shimmon and Sally Lowe for their help with administrative matters and for contributing to creating a very pleasant work environment.

A very special thank you to loan Manolescu. His love, support and help have accompanied me through the difficult periods. From the beginning until the end of this PhD, he has been my closest friend and the person who was there to share the ups and downs. Your friendship is invaluable to me. Thank you.

Many thanks to Sara Merino who has given this experience a lot of colour and positive energy. Without her Spanish sparkle lots would have been missed. Also, many thanks to my cousin Evie Efthymiou who has always been my most loyal friend and my most faithful supporter. My brother Constantinos Constantinou, my housemate Kostas Papafitsoros and my friends Revi Nicolaou, Kyriacos Leptos, Mihalis Dafermos, Filio Constantinou, Ioana Cosma, Julio Brau, Bati Sengul, Marc Briant, Kolyan Ray, Alan Sola, Robin Evans and everybody in Cambridge that made this journey such a unique and fun experience. I apologize to those unnamed.

Finally, I want to express my deepest gratitude to my parents to whom I owe everything I am today. Ευχαριστώ για όλα!

I acknowledge essential financial support from the EPSRC, the CET and the Department of Pure Mathematics and Mathematical Statistics.

Statement of Originality

I hereby declare that this dissertation entitled "Conditional Independence and Applications in Statistical Causality" is the result of my own work, conducted under the guidance of my supervisor Prof. A. Philip Dawid. It contains nothing which is the outcome of work done in collaboration with others, except where specifically indicated in the text. I further declare that the dissertation submitted is not substantially the same as any that has been submitted for a degree or diploma or other qualification at the University of Cambridge or any other university or similar institution.

Chapter 1 gives an overview of the content of the thesis and an introduction to statistical causality.

Chapter 2 is a review of all the necessary theory of measure theoretic probability adapted to our setting. Most of the material is known results or combination of known results and is cited accordingly. The review work has been done personally.

Chapter 3 is all novel material and the result of the work I have done under the guidance of my supervisor. It essentially generalises the concept of conditional independence and sets rigorous foundations for the DT framework of statistical causality.

Chapter 4 applies the obtained theory in the DT framework. The issues addressed here follow from discussions I had with my supervisor and Hui Guo.

Chapter 5 is mostly novel material unless explicitly stated. It is based on the survey paper of Dawid and Didelez (2010) and takes a formal approach in addressing the subject of sequential ignorability.

Chapter 6 consists of a summary of the thesis.

Abstract

The main goal of this thesis is to build a rigorous basis for the Decision-Theoretic (DT) framework of statistical causality. For this purpose the language and calculus of conditional independence are extended to encompass both stochastic and non-stochastic variables and the extended notion of conditional independence is applied in the DT framework to express and explore causal concepts.

Aiming to make *causal* inference for the variables of interest, the DT framework differentiates between observational and interventional regimes using a non-stochastic variable to index the regimes. Typically, we consider the regime under which data can be/is collected (observational regime) and a number of interventional regimes that we want to compare. Appreciating the fact that we mostly have access to observational data, we focus on deducing information from the observational regime for the interventional regimes. The conditions that enable us to transfer information across regimes are expressed in the language of conditional independence. In this framework, stochastic variables represent the variables of interest (e.g. phenotypes/traits/disease status of an individual etc.) and non-stochastic variables represent the different regimes under which we observe the stochastic variables (or functions that give us information on the operating regimes).

In the first part of this thesis we study *extended conditional independence* in an axiomatic framework. We introduce definitions to formalize intuitive concepts and elaborate to show that the axioms of stochastic conditional independence (classical properties of conditional independence) still hold under suitable interpretation. Measure-theoretic subtleties arise because we allow the outcomes of the same random variable to take positive probability in some regimes and zero probability in other regimes. We study discrete random variables separately and introduce *positivity* conditions to overcome technical problems.

In the second part of this thesis the language and calculus of extended conditional independence are used to address causal questions. We in-

troduce sufficient covariates and strongly sufficient covariates and show that their conditional independence properties allow identification of the Effect of Treatment on the Treated and the Average Causal Effect respectively, from the observational regime. Moreover, we address the case of dynamic treatment strategies where we are concerned with the control of a variable of interest through a sequence of consecutive actions. We consider observable, unobservable and action variables and study them under a variety of contemplated regimes. We explore conditions expressed in the language of conditional independence under which we can achieve identification of the consequence of applying a "control strategy".

Contents

Chapter 1

Introduction

1.1 Overview

The need to establish a *causal* relationship between seemingly correlated events is maybe the most challenging problem of modern statistics. It is well established knowledge that "correlation does not imply causation" and often does not even provide evidence to support causality. Especially in the presence of unobservable variables, observational data becomes inadequate to unravel any causal pathways between the variables of interest.

For example, given a simple observed correlation between a certain phenotype (observed characteristics of an individual) and a disease status, it is possible to argue in totally different directions. It could be that the phenotype causes the disease, or the disease causes the phenotype (reverse causation). It could also be that other observed factors or unobserved factors (confounding) cause both variables to change or even that the observed association is just a coincidence. Maybe time has an effect on both variables or both variables are causal to each other under different circumstances. For example, if we assume an observed association between high blood pressure and increased risk of Coronary Heart Disease (CHD), one could argue that it is high blood pressure that causes CHD to occur, or that CHD at early stages raises blood pressure or even that other factors like smoking, bad dietary habits, no exercise *etc.*, cause the effect we observe on both variables.

As the issue of inferring causality gains increasing interest, especially when the goal is to assess the effect of an intervention, the area of statistical causality offers several frameworks. In particular, the Decision Theoretic (DT) framework (Dawid, 2000, 2002) and the Potential Responses (PR) framework (Rubin, 1974, 1977, 1978),

provide two fundamentally different approaches for discussing causality. The main goal of this thesis is to set the mathematical foundations of the DT framework and to use this framework to address causal questions.

In Chapter 1, we present the two frameworks and advocate in favour of the DT framework. We see that the language of conditional independence is the most fundamental tool of the DT framework and indicate that this language needs to be extended rigorously to encompass both stochastic and non-stochastic variables. In Chapter 2, we introduce the axioms of a *separoid* and show that stochastic and variation (conditional) independence satisfy these axioms. In this chapter we provide all the measure-theoretic tools that will be used in the subsequent chapter to define and explore *extended conditional independence* (ECI). In Chapter 3, we formally study ECI and identify the conditions that allow us to deduce the axioms of a separoid for this concept. We take two different approaches, one using Bayesian arguments and one using the measure-theoretic particulars of ECI. For simplicity, we first consider discrete random variables and then discuss further generalisations.

The remaining part of the thesis focuses on applying the above theory in the area of causality to express and explore causal concepts. In Chapter 4, we introduce *sufficient covariates* and establish the conditions that allow identification of the *Average Causal Effect* and the *Effect of Treatment on the Treated*. In Chapter 5, we focus on a more complex problem, that of identifying dynamic treatment strategies. In this chapter we are concerned with controlling some variable of interest through a sequence of consecutive actions, at each stage taking into account observations made thus far. We follow closely the approach taken by Dawid and Didelez (2010) and study rigorously conditions termed *sequential randomisation* and *sequential irrelevance* that allow identification of a *control strategy*.

1.2 Formal Frameworks for Causality

Establishing a *cause* and *effect* relationship is the goal of many statistical studies, but the concept of causality itself is very subtle and beyond the sphere of classical statistics. Philosophers, statisticians, economists, epidemiologists and numerous other communities have struggled over many years to find a precise definition for what *causal* exactly means. Even today, causality admits different interpretations under different contexts. For the purposes of this thesis, an observed relationship will be considered *causal* if we believe that it holds under the application of an intervention to the variable that we suspect as causal. That is, we consider that a

variable A is *a cause of* a variable B if on intervening on variable A we observe an effect on variable B. It is important here to emphasize on the action of intervention as an external force on a variable that results in changing its status, and to distinguish the action of intervention from the case where we just observe a variable having a specific status.

When we want to test the effect of a treatment on a disease, we need to control for other possible factors that might affect indirectly the choice for the particular treatment and the disease status. However, we can never be sure that we have identified and adjusted for all possible influencing factors. Especially when these factors are social, behavioural or physiological there is no ideal way either to find a measurement scale or to control for this measurement scale. These possibly unobserved factors that influence the variables of interest are called confounding factors, and are sometimes the reason why studies on the same set of variables produce totally different results. With the act of intervention, we establish that the alteration on the treatment variable is due to us and not confounding factors.

1.2.1 Decision Theoretic Framework

One of the frameworks statistical causality offers to express causal concepts is the *Decision Theoretic* (DT) framework introduced by Dawid (2000, 2002). The DT framework introduces different regimes that represent the setting under which data can be/is collected and the settings that we want to compare. We call the regime that gives us data the *observational regime*, and the regimes that we want to compare *interventional regimes*. We understand interventional regimes as being identical settings under which we can intervene and alter only what we suspect to be the causal variable, and then measure the effect on what we suspect to be the effect variable. Because in practice we will usually not be able to create these interventional settings (due to ethical, financial, pragmatic reasons) and so compare them, the question we ask is: Under what conditions can we draw information from the observational regime for the interventional regimes? Of course any such conditions should be justified in the context of the problem for which they are imposed.

For illustration purposes, let us assume that we are interested in assessing the effect of a binary treatment variable T on a disease variable Y. Thus

$$T = \begin{cases} 0, & \text{under control treatment} \\ 1, & \text{under active treatment} \end{cases}$$

and Y can be discrete or continuous, representing a measurement on the disease status. In this context we further consider three regimes indexed with a non-stochastic variable σ:

$$\sigma = \begin{cases} \emptyset, & \text{denotes the observational regime} \\ 0, & \text{denotes the interventional regime under control treatment} \\ 1, & \text{denotes the interventional regime under active treatment} \end{cases}$$

While T and Y are regarded as random variables that have different distributions in the different regimes, σ is a variable with no uncertainty around it and only serves as a parameter to index the regimes. To emphasize the different nature of this variable we call the latter a *decision variable*.

When the observational regime is operating, $\sigma = \emptyset$, the distribution of T will be determined by Nature. In contrast, when the interventional regimes are operating, the value of T is imposed: for $\sigma = 0$, $\mathbb{P}(T = 0) = 1$ and for $\sigma = 1$, $\mathbb{P}(T = 1) = 1$. So in the interventional regimes T is a degenerate random variable with all the probability on one value, the same as the value of σ. To express which regime is in operation we use a subscript letter under the function we are considering. For example, for $t = 0, 1$, to denote the probability of the event $\{T = t\}$ in the observational regime, we will write $\mathbb{P}_\emptyset(T = t)$. Similarly, for $t = 0, 1$, to denote the expectation of the disease status Y in the interventional regime $\sigma = t$, we will write $\mathbb{E}_t(Y)$. Analogous conventions will be adopted throughout the thesis. Precise definitions will follow.

This framework is developed to address questions of the form: Intervening on variable T, will I observe an effect on variable Y? Understanding T as a treatment variable, this question could be simplified as follows: If I take/don't take treatment will my illness get cured? Questions of this form are classified as *hypothetical questions* and aim to assess the effect of some cause. In our example assessing the effect of taking or not taking the treatment will give an indication as to whether the treatment is beneficial for the illness.

The main goal being to compare the distribution of Y in the two interventional regimes $\sigma = 0, 1$, the simplest quantity we usually seek to estimate is the *Average Causal Effect* (ACE) where

$$ACE := \mathbb{E}_1(Y) - \mathbb{E}_0(Y).$$

So the ACE is just the difference in expectations of Y under the two interventional

regimes and therefore a non-zero value admits a causal interpretation. Seeking to identify the ACE from observational data (when interventional data is not available), we look for conditions that allow us to re-express the ACE in terms of observational quantities. While formal mathematical analysis of these conditions will be given in Chapter 3 and the ACE will be revisited in Chapter 4, here we try to give an intuitive understanding.

In the simplest (though usually non-realistic) case, when we believe that no confounding is present and furthermore that the action of intervention affects the response variable the same way as if the particular choice of treatment had been naturally assigned, we could impose the following condition: for $t = 0, 1$, the conditional distribution of the disease status Y given $T = t$ is the same in the observational regime and in the interventional regime $\sigma = t$. We can use the language of conditional independence and denote this condition by $Y \perp\!\!\!\perp \sigma \mid T$. Under this condition the ACE becomes

$$\begin{aligned} ACE &= \mathbb{E}_1(Y \mid T = 1) - \mathbb{E}_0(Y \mid T = 0) \\ &= \mathbb{E}_\emptyset(Y \mid T = 1) - \mathbb{E}_\emptyset(Y \mid T = 0), \end{aligned}$$

and is fully identifiable from observational data. As mentioned above, this condition will generally not hold in practice and we will need to search for more plausible conditions that represent our problem.

1.2.2 Potential Responses Framework

An alternative framework that statistical causality offers is the *Potential Responses* (PR) framework introduced by Rubin (1974, 1977, 1978). In the setting where we are concerned with the effect of a binary treatment variable on a disease status, this framework formulates the problem as follows. Instead of a single response variable Y, we consider two potential responses (Y_0, Y_1) representing the two different outcomes under the two different treatments. So Y_0 represents the outcome of the disease status under control treatment ($T = 0$) and Y_1 represents the outcome of the disease status under active treatment ($T = 1$). We denote by $\underline{\mathbf{Y}} = (Y_0, Y_1)$ the pair of the potential responses and we understand that the components of $\mathbf{Y}$ exist simultaneously and ahead of any choice on the treatment T. However, the choice of T will determine which potential response we observe. Thus the observed response variable Y is determined by $\underline{\mathbf{Y}}$ and T as $Y = Y_T$. In this framework, both T and $\underline{\mathbf{Y}}$

are random variables and $\underline{\mathbf{Y}}$ is assumed to follow some bivariate distribution.

Working in the PR framework we can also address hypothetical questions. But in contrast with the DT framework we don't compare the distribution of the response variable in the two interventional regimes but the marginal distribution of the two potential responses. The ACE in this framework is defined by

$$ACE := \mathbb{E}(Y_1) - \mathbb{E}(Y_0).$$

If we believe that there is no confounding and that the allocation of treatment is a random procedure, we can express this using conditional independence notation and write $(Y_0, Y_1) \perp\!\!\!\perp T$. This implies that the distribution of Y_0 is the same as the conditional distribution of Y_0 given $T = 0$ and thus can be estimated from data. Similarly the distribution of Y_1 can be estimated from that of Y_1 given $T = 1$. Thus the ACE becomes identifiable from observable data.

Additionally to hypothetical questions, the PR framework can further address what are termed *counterfactual questions*. These questions are of the form: I have taken the treatment and my illness is cured. Is my illness cured because of the treatment? Addressing counterfactual questions we want to assess if the treatment is the cause of the effect we observe. They are called counterfactual questions because in order to answer them, one is asked to compare the actual scenario (in this case, the event that we have taken the treatment) with the counterfactual scenario (the one where we have not taken the treatment). There is no mathematical impediment to addressing counterfactual questions in this framework. Assuming that the observed variable Y takes value y, we can refer to the conditional distribution of Y_0 given $(T = 1, Y = y)$. However, as the following example will illustrate (Dawid, 2007b), a fundamental problem appears when trying to answer such questions.

Example 1.2.1. Assume that $\underline{\mathbf{Y}} = (Y_0, Y_1)$ has a bivariate normal distribution.

$$\underline{\mathbf{Y}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and $\underline{\mathbf{Y}} \perp\!\!\!\perp T$ (no-confounding). Then $Y_0 \mid (T = 1, Y = y)$ has the same distribution as $Y_0 \mid (T = 1, Y_1 = y)$ which in turn has (because of the assumption $\underline{\mathbf{Y}} \perp\!\!\!\perp T$) the same distribution as $Y_0 \mid Y_1 = y$. From bivariate normal theory we know that

$$Y_0 \mid Y_1 = y \sim \mathcal{N}(\mu_0 + \rho(y - \mu_1), (1 - \rho^2))$$

But Y_0 and Y_1 can't be possibly observed together and the estimation of ρ becomes problematic. Thus any question that invokes the distribution of $Y_0 \mid Y_1 = y$ cannot be answered without imposing additional untestable assumptions on ρ.

1.3 Choosing a language for causality

Due to the difficulty of counterfactual questions, a lot of research is confined to the examination of hypothetical questions. Choosing which framework is more suitable to address this type of questions has motivated a lot of discussion in the area (Dawid, 2000).

In the PR framework the problem is formulated by introducing two potential responses (Y_0, Y_1) that exist together, no matter which treatment we decide to choose. This means that if we decide to take the treatment and thus reveal Y_1, the potential response Y_0 will still exist (but not be revealed to us). Further, it will have the same value as if we had decided not to take the treatment. This setting presupposes belief in the predetermined existence of the two potential responses ahead of treatment and provokes some philosophical qualms.

Accepting the ontology of both frameworks, the element that makes the DT framework more appealing to the author is the fact that the assumptions we make regarding the relationship between the variables involved could in principle be tested. The concept of the existence of two potential responses that cannot be observed together makes any assumption concerning simultaneously Y_0 and Y_1 untestable in practice. Consider for example the case where we suspect that data is confounded by a variable U. We can think of Y as representing some measurement on an individual's weight and the binary treatment variable T as representing exercise $(T = 1)$ or no-exercise $(T = 0)$. In this scenario, one could argue that a confounding variable is the status of smoking. In practice there will be more than one confounding variable but for illustration purposes let us assume that we suspect only the smoking status to be a confounder. Denote by $U = 1$ the event of being a smoker and by $U = 0$ the event of being a non-smoker. In the DT framework this assumption is represented by writing (in conditional independence language) $Y \perp\!\!\!\perp \sigma \mid (T, U)$, thus expressing our belief that, given knowledge on the treatment T and smoking status U, the distribution of the disease Y is independent of the regime σ. In principle we could create the interventional settings and thus, test our assumption. In the PR framework on the other hand, this belief is represented by writing $T \perp\!\!\!\perp (Y_0, Y_1) \mid U$, thus expressing that treatment T is independent of the pair (Y_0, Y_1) of potential responses,

after conditioning on the smoking status U. Invoking two potential responses that can't be observed together makes this assumption untestable even given knowledge of the smoking status U.

There are many levels of comparison between the two frameworks and the reader is referred to Dawid (2007a) for a more detailed examination. In general, we consider the DT framework more straightforward in the sense that it does not evoke unnecessary elements, and has the power to express in a simple and immediate way the assumptions made. This thesis will provide all the required mathematical elements to support rigorously the foundations of the DT framework.

Chapter 2

Conditional Independence: Stochastic and non-stochastic variables separately

2.1 Conditional independence as a language for causality

Working in the DT framework, the most fundamental tool that enables us to express causal concepts is the language of conditional independence. Using this language we can state conditions to express our belief that, upon knowledge on adequate information, the extra information on the regime becomes irrelevant for making inference about the variable of interest. As examples of such conditions, for random variables Y, T, U and regime indicator σ, we have already referenced $Y \perp\!\!\!\perp \sigma \mid T$ and $Y \perp\!\!\!\perp \sigma \mid (T, U)$.

Conditional independence in the case of solely stochastic variables has been widely studied in probability theory. An analogous concept has been studied in the case of purely non-stochastic variables. Both these concepts satisfy the axioms of a *separoid* (Dawid, 2001a,b). The particularity of the DT framework is that it requires a combination of these two concepts, as it expresses conditional independence statements that contain simultaneously stochastic and non-stochastic variables. In order to study this extended setting and deduce similar calculus which will prove a powerful machinery for exploring causal concepts, it is essential to study first the two concepts separately. Once we deduce the axioms of a separoid for each case, we can use this knowledge to formally explore the extended case.

2.2 Separoids

We will present the algebraic structure of a *separoid* as it has been introduced by Dawid (2001a). First, recall some definitions from order theory.

Let V be a set with elements denoted by $x, y, \ldots$. A binary relation $\cdot \leq \cdot$ on V is called a *quasiorder* if it is reflexive and transitive. For $x, y \in V$, if $x \leq y$ and $y \leq x$, we say that x and y are *equivalent* and write $x \approx y$. For a subset $A \subseteq V$, we call z a *join* of A if it is greater than any element of A and it is a minimal element of V with that property. We call z a *meet* of A if it is smaller than any element of A and it is a maximal element of V with that property. If x and y are both joins (respectively meets) of A, then $x \approx y$. We call V a *join semilattice* if there exists a join for any nonempty finite subset. Similarly, we call V a *meet semilattice* if there exists a meet for any nonempty finite subset. When V is both a meet and join semilattice (with respect to the same quasiorder), we call V a *lattice*. For elements x and y, we denote (when they exist) their join by $x \vee y$ and their meet by $x \wedge y$.

Definition 2.2.1. Given a ternary relation $\cdot \perp\!\!\!\perp \cdot \mid \cdot$ on V, we call $\perp\!\!\!\perp$ a *separoid* (on $(V, \leq)$), or the triple $(V, \leq, \perp\!\!\!\perp)$ a *separoid*, if:

$S1$: $(V, \leq)$ is a join semilattice

and

$P1$: $x \perp\!\!\!\perp y \mid z \Rightarrow y \perp\!\!\!\perp x \mid z$,

$P2$: $x \perp\!\!\!\perp y \mid y$,

$P3$: $x \perp\!\!\!\perp y \mid z$ and $w \leq y \Rightarrow x \perp\!\!\!\perp w \mid z$,

$P4$: $x \perp\!\!\!\perp y \mid z$ and $w \leq y \Rightarrow x \perp\!\!\!\perp y \mid (z \vee w)$,

$P5$: $x \perp\!\!\!\perp y \mid z$ and $x \perp\!\!\!\perp w \mid (y \vee z) \Rightarrow x \perp\!\!\!\perp (y \vee w) \mid z$.

The following corollary shows that in $P4$ and $P5$, the choice of join does not change the property.

Corollary 2.2.2. Let $(V, \leq, \perp\!\!\!\perp)$ be a separoid and x_i, y_i, z_i be elements of V for $i = 1, 2$. Suppose that $x_1 \perp\!\!\!\perp y_1 \mid z_1$ and that $x_1 \approx x_2$, $y_1 \approx y_2$ and $z_1 \approx z_2$. Then $x_2 \perp\!\!\!\perp y_2 \mid z_2$.

Proof. See Corollary 1.2. in Dawid (2001a). □

Definition 2.2.3. We say that the triple $(V, \leq, \perp\!\!\!\perp)$ is a *strong separoid* if we strengthen $S1$ in Definition 2.2.1 to

$S1'$: $(V, \leq)$ is a lattice

and in addition to $P1$-$P5$, we require

$P6$: if $z \leq y$ and $w \leq y$, then $x \perp\!\!\!\perp y \mid z$ and $x \perp\!\!\!\perp y \mid w \Rightarrow x \perp\!\!\!\perp y \mid (z \wedge w)$.

2.3 Stochastic conditional independence

The concept of stochastic conditional independence is an example of a separoid. In order to prove this, we will recall some results from probability theory.

2.3.1 Notation and Preliminaries

Let E be a set and $\mathcal{E}$ be a set of subsets of E.

Definition 2.3.1. We say that $\mathcal{E}$ is a *σ-algebra* on E if it contains the empty set $\emptyset$ and for all $A \in \mathcal{E}$ and all sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{E}$, $A^c \in \mathcal{E}$ and $\bigcup_n A_n \in \mathcal{E}$. A measure μ on $(E, \mathcal{E})$ is a function $\mu : \mathcal{E} \to [0, +\infty]$ with $\mu(\emptyset) = 0$, such that, for any sequence $(A_n : n \in \mathbb{N})$ of disjoint elements of $\mathcal{E}$,

$$\mu(\bigcup_n A_n) = \sum_{n \in \mathbb{N}} \mu(A_n).$$

The pair $(E, \mathcal{E})$ is called a *measurable space* and the triple $(E, \mathcal{E}, \mu)$ is called a *measure space*. If $\mu(E) = 1$, we call μ a *probability measure* ($\mathbb{P}$-measure) and $(E, \mathcal{E}, \mu)$ a *probability space*. We usually denote a probability space by $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 2.3.2. We say that $\mathcal{E}$ is a *π-system* if $\emptyset \in \mathcal{E}$ and $\forall\ A, B \in \mathcal{E}$, $A \cap B \in \mathcal{E}$. We say that $\mathcal{E}$ is a *d-system* if $E \in \mathcal{E}$ and $\forall\ A, B \in \mathcal{E}$ with $A \subseteq B$ and all increasing sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{E}$, $B \setminus A \in \mathcal{E}$ and $\bigcup_n A_n \in \mathcal{E}$.

Lemma 2.3.3. $\mathcal{E}$ is a σ-algebra iff $\mathcal{E}$ is both a π-system and a d-system.

Proof. See Billingsley (1995) (p.41, Lemma 6.). □

Definition 2.3.4. Let $\mathcal{A}$ be a set of subsets of E. Define

$$\sigma(A) = \{A \subseteq E : A \in \mathcal{E} \text{ for all } \sigma\text{-algebras } \mathcal{E} \text{ containing } \mathcal{A}\}.$$

Then $\sigma(A)$ is a σ-algebra and it's called the *σ-algebra generated by $\mathcal{A}$*.

Lemma 2.3.5. (Dynkin's $\pi - system$ lemma) Let $\mathcal{A}$ be a π-system. Then any d-system containing $\mathcal{A}$ contains also the σ-algebra generated by $\mathcal{A}$.

Proof. See Billingsley (1995) (p.42, Theorem 3.2.). □

Definition 2.3.6. We say that $\mathcal{E}$ is a *ring* on E if it contains the empty set $\emptyset$ and for all $A, B \in \mathcal{E}$, $B \setminus A \in \mathcal{E}$ and $A \cup B \in \mathcal{E}$.

We can readily see that $\mathcal{E}$ is a ring implies that $\mathcal{E}$ is a π-system.

Definition 2.3.7. Let $\mathcal{A}$ be a set of subsets of E containing the empty set $\emptyset$. A *set function* is a function $\mu : \mathcal{A} \to [0, \infty]$ with $\mu(\emptyset) = 0$. Let μ be a set function. We say that μ is *countably additive* if for all sequences of disjoint sets $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$ with $\bigcup_n A_n \in \mathcal{A}$,

$$\mu(\bigcup_n A_n) = \sum_{n \in \mathbb{N}} \mu(A_n).$$

Theorem 2.3.8. (Carathéodory's extension theorem) Let $\mathcal{A}$ be a ring of subsets of E and let $\mu : \mathcal{A} \to [0, \infty]$ be a countably additive set function. Then μ extends to a measure on the σ-algebra generated by $\mathcal{A}$.

Proof. See Billingsley (1995) (p.166, Theorem 11.3.). □

Theorem 2.3.9. (Uniqueness of extension) Let μ_1, μ_2 be measures on $(E, \mathcal{E})$ with $\mu_1(E) = \mu_2(E) < \infty$. Suppose that $\mu_1 = \mu_2$ on $\mathcal{A}$, for some π-system $\mathcal{A}$ generating $\mathcal{E}$. Then $\mu_1 = \mu_2$ on $\mathcal{E}$.

Proof. See Billingsley (1995) (p.42, Theorem 3.3.). □

Definition 2.3.10. Given two measurable spaces $(\Omega, \mathcal{A})$ and $(F, \mathcal{F})$, a function $X : \Omega \to F$ is called *measurable* if $X^{-1}(B) \in \mathcal{A}$ whenever $B \in \mathcal{F}$. Here $f^{-1}(B)$ denotes the inverse image of B by X, *i.e.*, $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. When $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, measurable functions $X : \Omega \to F$ are also called random variables.

To denote the corresponding σ-algebras we may write $X : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$. The smallest σ-algebra containing all sets of the form $\{X^{-1}(B) : B \in \mathcal{F}\}$, is called the σ-algebra generated by X and is denoted by $\sigma(X)$.

Remark 2.3.11. Let $(\Omega, \mathcal{A})$ and $(F, \mathcal{F})$ be measurable spaces. For any function $f : \Omega \to F$, the inverse image preserves set operations:

$$f^{-1}(\bigcup_i B_i) = \bigcup_i f^{-1}(B_i) \quad \text{and} \quad f^{-1}(F \setminus B) = \Omega \setminus f^{-1}(B).$$

Therefore the set $\{f^{-1}(B) : B \in \mathcal{F}\}$ is a σ-algebra on Ω and the set $\{B \subseteq F : f^{-1}(B) \in \mathcal{A}\}$ is a σ-algebra on F. Thus $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{F}\}$. Similarly, for random variables $Y : (\Omega, \mathcal{A}) \to (F_Y, \mathcal{F}_Y)$ and $Z : (\Omega, \mathcal{A}) \to (F_Z, \mathcal{F}_Z)$, $\sigma(Y, Z) = \{(Y, Z)^{-1}(C) : C \in \mathcal{F}_Y \otimes \mathcal{F}_Z\}$, where $\mathcal{F}_Y \otimes \mathcal{F}_Z$ is the σ-algebra generated by the set of Cartesian products $\{A \times B : A \in \mathcal{F}_Y, B \in \mathcal{F}_Z\}$. For a collection of sets $\mathcal{F}_1$ and a collection of sets $\mathcal{F}_2$, we will write $\mathcal{F}_1 \times \mathcal{F}_2$ to denote the set of Cartesian products $\{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$. Then $\mathcal{F}_Y \otimes \mathcal{F}_Z := \sigma(\mathcal{F}_Y \times \mathcal{F}_Z)$.

Proposition 2.3.12. Let $(F_Y, \mathcal{F}_Y)$ and $(F_Z, \mathcal{F}_Z)$ be measurable spaces and let $Y : (\Omega, \sigma(Y)) \to (F_Y, \mathcal{F}_Y)$ and $Z : (\Omega, \sigma(Z)) \to (F_Z, \mathcal{F}_Z)$ be measurable functions. Also let $\mathcal{A} = \{A_Y \cap A_Z : A_Y \in \sigma(Y), A_Z \in \sigma(Z)\}$. Then $\mathcal{A}$ is a π-system and $\sigma(\mathcal{A}) = \sigma(Y, Z)$.

Proof. $\mathcal{A}$ is clearly a $\pi - system$. Now

$$\begin{aligned}\sigma(\mathcal{A}) &= \sigma\{A_Y \cap A_Z : A_Y \in \sigma(Y), A_Z \in \sigma(Z)\} \\ &= \sigma\{Y^{-1}(B_Y) \cap Z^{-1}(B_Z) : B_Y \in \mathcal{F}_Y, B_Z \in \mathcal{F}_Z\} \\ &= \sigma\{(Y, Z)^{-1}(C) : C \in \mathcal{F}_Y \times \mathcal{F}_Z\}\end{aligned}$$

and

$$\sigma(Y, Z) = \{(Y, Z)^{-1}(C) : C \in \mathcal{F}_Y \otimes \mathcal{F}_Z\}$$

Clearly $\sigma(\mathcal{A}) \subseteq \sigma(Y, Z)$. To show that $\sigma(Y, Z) \subseteq \sigma(\mathcal{A})$ consider $\mathcal{G} = \{C \subseteq F_Y \times F_Z : (Y, Z)^{-1}(C) \in \sigma(\mathcal{A})\}$. $\mathcal{G}$ is a σ-algebra that contains $\mathcal{F}_Y \times \mathcal{F}_Z$ and thus $\mathcal{F}_Y \otimes \mathcal{F}_Z$. Therefore $\sigma(Y, Z) = \{(Y, Z)^{-1}(C) : C \in \mathcal{F}_Y \otimes \mathcal{F}_Z\} \subseteq \sigma(\mathcal{A})$. □

Definition 2.3.13. Let E be a set and $A \subseteq E$. The indicator function of A is a function $\mathbb{1}_A : E \to \{0, 1\}$ defined by:

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Thus the indicator function takes the value 1 on A and 0 otherwise.

Theorem 2.3.14. (Monotone class theorem) Let $(E, \mathcal{E})$ be a measurable space and let $\mathcal{A}$ be a π-system generating $\mathcal{E}$. Let V be a vector space of bounded functions $f : E \to \mathbb{R}$ such that:

(i) $\mathbb{1}_E \in V$ and $\mathbb{1}_A \in V$ for all $A \in \mathcal{A}$

(ii) if $f_n \in V$ for all n and f is bounded with $0 \leq f_n \uparrow f$, then $f \in V$.

Then V contains every bounded measurable function.

Proof. See Durrett (2010) (p.276, Theorem 6.1.3.). □

Proposition 2.3.15. Let $Y : (\Omega, \sigma(Y)) \to (F_Y, \mathcal{F}_Y)$ and $Z : (\Omega, \sigma(Z)) \to (F_Z, \mathcal{F}_Z)$ be surjective random variables. If the singletons $\{y\}$ are measurable in $\sigma(Y)$ (*i.e.*, $\forall\, y \in \mathcal{F}_Y$, $\{\omega \in \Omega : Y(\omega) = y\} \in \sigma(Y)$), the following are equivalent:

i) $\sigma(Y) \subseteq \sigma(Z)$.

ii) there exists $f : (F_Z, \mathcal{F}_Z) \to (F_Y, \mathcal{F}_Y)$ measurable such that $Y = f(Z)$.

Proof. i) to ii): By definition $\sigma(Z) = \{Z^{-1}(A) : A \in \mathcal{F}_Z\}$. Observe that for all $z \in F_Z$ and all $B \in \sigma(Z)$, either $\{\omega \in \Omega : Z(\omega) = z\} \cap B = \emptyset$ or $\{\omega \in \Omega : Z(\omega) = z\} \subseteq B$. Now fix $z \in F_Z$. Since Z is surjective, $\{\omega \in \Omega : Z(\omega) = z\} \neq \emptyset$. Observing that $\bigcup_{y \in F_Y} \{\omega \in \Omega : Y(\omega) = y\} = \Omega$ we can find $y \in F_Y$ such that $\{\omega \in \Omega : Z(\omega) = z\} \cap \{\omega \in \Omega : Y(\omega) = y\} \neq \emptyset$. But $\{\omega \in \Omega : Y(\omega) = y\} \in \sigma(Y) \subseteq \sigma(Z)$. Therefore $\{\omega \in \Omega : Z(\omega) = z\} \subseteq \{\omega \in \Omega : Y(\omega) = y\}$ and this $y \in F_Y$ is unique. For all such $z \in F_Z$ define $f(z) = y$. We have therefore constructed $f : F_Z \to F_Y$ where $f(Z) = Y$. To show that f is $\mathcal{F}_Z$-measurable, consider $E_Y \in \mathcal{F}_Y$. Then

$$\begin{aligned} f^{-1}(E_Y) &= \{z \in F_Z : f(z) = y \in E_Y\} \\ &= \{z = Z(\omega) \in F_Z : f(Z(\omega)) = Y(\omega) = y \in E_Y\} \\ &= Z(\{\omega \in \Omega : Y(\omega) = y \in E_Y\}). \end{aligned}$$

Since $E_Y \in \mathcal{F}_Y$, $\{\omega \in \Omega : Y(\omega) = y \in E_Y\} \in \sigma(Y) \subseteq \sigma(Z)$. Thus there exists $E_Z \in \mathcal{F}_Z$ such that $Z(\{\omega \in \Omega : Y(\omega) \in E_Y\}) = E_Z$ and we have proved that f is $\mathcal{F}_Z$-measurable.
ii) to i): Let $A \in \sigma(Y) = \{Y^{-1}(B) : B \in \mathcal{F}_Y\}$. Then there exists $B \in \mathcal{F}_Y$ such that $Y^{-1}(B) = A$, which implies that $(f \circ Z)^{-1}(B) = A$ and thus $Z^{-1}(f^{-1}(B)) = A$. Since $B \in \mathcal{F}_Y$ and f is $\mathcal{F}_Z$-measurable, $f^{-1}(B) \in \mathcal{F}_Z$. Also since $f^{-1}(B) \in \mathcal{F}_Z$ and Z is $\sigma(Z)$-measurable, $Z^{-1}(f^{-1}(B)) \in \sigma(Z)$, which concludes the proof. □

Henceforth, all random variables will denote surjective functions. Also whenever we invoke Proposition 2.3.15, we will assume that the conditions of the proposition are satisfied and when we write $f(X)$ for a random variable X, we presume that f is measurable with respect to the corresponding σ-algebras. Henceforth all functions will be assumed to take values in $[-\infty, +\infty]$. We equip $[-\infty, +\infty]$ with its Borel σ-algebra, *i.e.* the σ-algebra generated by the set of open sets.

Definition 2.3.16. Let $(E, \mathcal{E}, \mu)$ be a measure space and $f : E \to [-\infty, +\infty]$ be a measurable function. We call f integrable if $\int_E |f| d\mu < \infty$. We denote by $\mu(f) := \int_E f d\mu$ the integral of f.

To comply with standard terminology in statistical theory, for a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, we call the integral of X the expectation of X and write it $\mathbb{E}(X)$. For the remainder of this subsection, whenever no measure space is mentioned, we assume that $(E, \mathcal{E}, \mu)$ is the underlying measure space.

For a sequence of functions $(f_n : n \in \mathbb{N})$, we write $f_n \uparrow f$ if $f_n(x) \uparrow f(x)$ for any $x \in E$.

Theorem 2.3.17. (Monotone Convergence) Let f be a non-negative measurable function and let $(f_n : n \in \mathbb{N})$ be a sequence of such functions. Suppose that $f_n \uparrow f$. Then $\mu(f_n) \uparrow \mu(f)$.

Proof. See Billingsley (1995) (p.208, Theorem 16.2.). □

Note that for any function f, we can set $f^+ = f \vee 0$ and $f^- = -f \vee 0$. Then $f = f^+ - f^-$. Re-expressing functions as a difference of positive functions, we can get some variants of the monotone convergence theorem.

Example 2.3.18. Let f be a bounded, non-negative measurable function and let $(f_n : n \in \mathbb{N})$ be a sequence of such functions. Suppose that $f_n \uparrow f$. Also let g be an integrable function and set $h_n = gf_n$ and $h = gf$. Then h_n and h are integrable and $\mu(h_n) \to \mu(h)$.

Proof. Consider h_n^+ and h_n^-. Then

$$h_n^+ = (gf_n)^+ = g^+ f_n \uparrow g^+ f \quad \text{and} \quad h_n^- = (gf_n)^- = g^- f_n \uparrow g^- f.$$

By monotone convergence theorem $\mu(h_n^+) \uparrow \mu(g^+ f)$ and $\mu(h_n^-) \uparrow \mu(g^- f)$. Thus

$$\begin{aligned}
\mu(h) &= \mu(h^+ - h^-) \\
&= \mu(g^+ f) - \mu(g^- f) \quad \text{since } f \text{ bounded and } g \text{ integrable} \\
&= \lim_{n\to\infty} \mu(g^+ f_n) - \lim_{n\to\infty} \mu(g^- f_n) \\
&= \lim_{n\to\infty} \mu(g^+ f_n - g^- f_n) \quad \text{since } f_n \text{ bounded and } g \text{ integrable} \\
&= \lim_{n\to\infty} \mu(h_n)
\end{aligned}$$

□

Theorem 2.3.19. (Dominated Convergence) Let f be a measurable function and let $(f_n : n \in N)$ be a sequence of such functions. Suppose that $f_n(x) \to f(x)$ for all $x \in E$ and that $|f_n| \leq g$ for all n, for some integrable function g. Then f and f_n are integrable, for all n, and $\mu(f_n) \to \mu(f)$.

Proof. See Billingsley (1995) (p.209, Theorem 16.4.). □

2.3.2 Conditional Expectation

To define conditional independence in the most general form we will need the concept of conditional expectation. In this section we will define conditional expectation and prove all the properties that are necessary to derive the axioms of conditional independence.

All random variables in this section are considered real-valued.

Definition 2.3.20. (Conditional Expectation.) Let X be an integrable random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{A}$ be a σ-algebra. A random variable Y is called a version of the *conditional expectation of X given $\mathcal{G}$*, if it satisfies:

(i) Y is $\mathcal{G}$-measurable.

(ii) Y is integrable and $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(Y\mathbb{1}_A)$ whenever $A \in \mathcal{G}$.

It can be shown that such a random variable exists and that any two versions of the same conditional expectation are almost surely equal, in the sense that they only differ on a set of probability zero (Billingsley (1995), p.445, Section 34.). We then write $Y = \mathbb{E}(X \mid \mathcal{G})$ a.s. $[\mathbb{P}]$ or just $Y = \mathbb{E}(X \mid \mathcal{G})$ a.s. when $\mathbb{P}$ is presumed. If $\mathcal{G} = \sigma(Z)$, *i.e.* is the σ-algebra generated by a random variable Z, we will write $\mathbb{E}(X \mid Z)$ rather than $\mathbb{E}(X \mid \mathcal{G})$.

Remark 2.3.21. Notice that when X is $\mathcal{G}$-measurable, $\mathbb{E}[X \mid \mathcal{G}] = X$ a.s.. Thus for any integrable function $f(X)$, $\mathbb{E}[f(X) \mid X] = f(X)$ a.s.. Also by (ii) for $A = \Omega$, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X \mid \mathcal{G})]$. We will use this in the form $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X \mid Y)]$.

For the remainder of this section we assume that the underlying measure space is a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The corresponding versions of the monotone and dominated convergence theorems follow. The proofs are similar to the proofs of the archetypal forms.

Theorem 2.3.22. (Conditional Monotone Convergence) Let X be a non-negative random variable and let $(X_n : n \in \mathbb{N})$ be a sequence of such functions. Suppose that $X_n \uparrow X$ a.s.. Then $\mathbb{E}(X_n \mid \mathcal{G}) \uparrow \mathbb{E}(X \mid \mathcal{G})$ a.s..

Remark 2.3.23. Suppose that $X_n \uparrow X$ a.s. and let $(Y_n : n \in \mathbb{N})$ be a sequence of $\mathcal{G}$-measurable random variables such that $\mathbb{E}(X_n \mid \mathcal{G}) = Y_n$ a.s. for all $n \in \mathbb{N}$. Then $(Y_n : n \in \mathbb{N})$ is an a.s. increasing sequence.

Proof. For all $n \in \mathbb{N}$, define $Z_n := Y_{n+1} - Y_n$ and $A_n = \{\omega \in \Omega : Z_n(\omega) < 0\}$. Then Z_n is $\mathcal{G}$-measurable and $A_n \in \mathcal{G}$. Thus

$$\begin{aligned} 0 &\geq \mathbb{E}(Z_n \mathbb{1}_{A_n}) \\ &= \mathbb{E}(Y_{n+1}\mathbb{1}_{A_n}) - \mathbb{E}(Y_n \mathbb{1}_{A_n}) \\ &= \mathbb{E}(X_{n+1}\mathbb{1}_{A_n}) - \mathbb{E}(X_n \mathbb{1}_{A_n}) \\ &= \mathbb{E}[(X_{n+1} - X_n)\mathbb{1}_{A_n}] \\ &\geq 0 \end{aligned}$$

which implies that $\mathbb{E}(Z_n \mathbb{1}_{A_n}) = 0$. Since Z_n is negative on A_n, it follows that $\mathbb{P}(A_n) = 0$. Now

$$\mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n) = 0$$

which proves that $(Z_n : n \in \mathbb{N})$ is an increasing sequence a.s.. □

Theorem 2.3.24. (Conditional Dominated Convergence) Let X be a random variable and let $(X_n : n \in \mathbb{N})$ be a sequence of random variables. Suppose that $X_n \to X$ a.s. and that there exists an integrable random variable Y such that $|X_n| \leq Y$ a.s. for all $n \in \mathbb{N}$. Then $\mathbb{E}(X_n \mid \mathcal{G}) \to \mathbb{E}(X \mid \mathcal{G})$ a.s..

Theorem 2.3.25. Let X be an integrable random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let $\mathcal{G}_1, \mathcal{G}_2$ be σ-algebras such that $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{A}$. Then

(i) $\mathbb{E}[\mathbb{E}(X \mid \mathcal{G}_1) \mid \mathcal{G}_2] = \mathbb{E}(X \mid \mathcal{G}_1)$ a.s..

(ii) $\mathbb{E}[\mathbb{E}(X \mid \mathcal{G}_2) \mid \mathcal{G}_1] = \mathbb{E}(X \mid \mathcal{G}_1)$ a.s..

Proof. (i) Any $Y = \mathbb{E}(X \mid \mathcal{G}_1)$ a.s. is $\mathcal{G}_1$-measurable and, since $\mathcal{G}_1 \subseteq \mathcal{G}_2$, also $\mathcal{G}_2$-measurable. Thus $\mathbb{E}[\mathbb{E}(X \mid \mathcal{G}_1) \mid \mathcal{G}_2] = \mathbb{E}(X \mid \mathcal{G}_1)$ a.s. by Remark 2.3.21.
(ii) Let $Y = \mathbb{E}(X \mid \mathcal{G}_2)$ a.s. and $Z = \mathbb{E}(Y \mid \mathcal{G}_1)$ a.s.. Then Y is $\mathcal{G}_2$-measurable and Z is $\mathcal{G}_1$- and $\mathcal{G}_2$-measurable. Also for any event $A \in \mathcal{G}_1 \subseteq \mathcal{G}_2$, $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(Y\mathbb{1}_A)$ and $\mathbb{E}(Y\mathbb{1}_A) = \mathbb{E}(Z\mathbb{1}_A)$. Thus $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(Z\mathbb{1}_A)$ for all $A \in \mathcal{G}_1$ which implies that $Z = \mathbb{E}(X \mid \mathcal{G}_1)$ a.s.. □

The above theorem will mostly be used for cases where $\mathcal{G}_1$ and $\mathcal{G}_2$ are σ-algebras generated by some random variables. Thus for random variables X, Y, Z we have that

$$\mathbb{E}\{\mathbb{E}[f(X) \mid Y, Z] \mid Z\} = \mathbb{E}\{\mathbb{E}[f(X) \mid Z] \mid Y, Z\} = \mathbb{E}[f(X) \mid Z] \quad \text{a.s..}$$

Theorem 2.3.26. Let Y be a random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and suppose that $\mathcal{G} \subseteq \mathcal{A}$ is σ-algebra such that Y is $\mathcal{G}$-measurable. If X and XY are integrable functions, then

$$\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}] \quad \text{a.s..} \tag{2.3.1}$$

Proof. See Billingsley (1995) (p.447, Theorem 34.3.). □

2.3.3 Conditional Independence for Stochastic Variables

We can now define conditional independence and study its properties.

Definition 2.3.27. (Conditional Independence.) Let X, Y, Z be random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. We say that X *is (conditionally) independent of* Y *given* Z and write $X \perp\!\!\!\perp_s Y \mid Z$ if for all $A_X \in \sigma(X)$ and all $A_Y \in \sigma(Y)$,

$$\mathbb{E}(\mathbb{1}_{A_X \cap A_Y} \mid Z) = \mathbb{E}(\mathbb{1}_{A_X} \mid Z)\mathbb{E}(\mathbb{1}_{A_Y} \mid Z) \quad \text{a.s..}$$

We refer to the above definition as *stochastic conditional independence* (SCI) and we use the subscript s to emphasize that we refer to the stochastic version. Using standard tools from measure theory, we can deduce equivalent forms for the above definition, which will prove useful later.

Proposition 2.3.28. Let X, Y, Z be random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. Then the following are equivalent.

(i) $X \perp\!\!\!\perp_s Y \mid Z$.

(ii) For all $A_X \in \sigma(X)$, $\mathbb{E}[\mathbb{1}_{A_X} \mid Y, Z] = \mathbb{E}[\mathbb{1}_{A_X} \mid Z]$ a.s..

(iii) For all real, bounded and measurable functions $f(X)$, $\mathbb{E}[f(X) \mid Y, Z] = \mathbb{E}[f(X) \mid Z]$ a.s..

(iv) For all real, bounded and measurable functions $f(X), g(Y)$, $\mathbb{E}[f(X)g(Y) \mid Z] = \mathbb{E}[f(X) \mid Z]\mathbb{E}[g(Y) \mid Z]$ a.s..

Proof. (i) $\Rightarrow$ (ii): Let $A_X \in \sigma(X)$ and $w(Z)$ be a version of $\mathbb{E}[\mathbb{1}_{A_X} \mid Z]$. Then $w(Z)$ is $\sigma(Z)$-measurable and since $\sigma(Z) \subseteq \sigma(Y, Z)$, also $\sigma(Y, Z)$-measurable. To show that $\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_A] = \mathbb{E}[w(Z)\mathbb{1}_A]$ whenever $A \in \sigma(Y, Z)$, let

$$\mathcal{D}_{A_X} = \{A \in \sigma(Y, Z) : \mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_A] = \mathbb{E}[w(Z)\mathbb{1}_A]\}$$

and

$$\Pi = \{A \in \sigma(Y, Z) : A = A_X \cap A_Y \text{ for some } A_Y \in \sigma(Y), A_Z \in \sigma(Z)\}.$$

By Proposition 2.3.12, Π is a π-system and $\sigma(\Pi) = \sigma(Y, Z)$. We will show that $\mathcal{D}_{A_X}$ is a d-system that contains Π and apply Dynkin's lemma to conclude that $\mathcal{D}_{A_X}$ contains $\sigma(\Pi) = \sigma(Y, Z)$.

To show that $\mathcal{D}_{A_X}$ contains Π, let $A_{Y,Z} = A_Y \cap A_Z$ such that $A_Y \in \sigma(Y)$ and $A_Z \in \sigma(Z)$. Then

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_A] &= \mathbb{E}\{\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_Y \cap A_Z} \mid Z]\} \\
&= \mathbb{E}\{\mathbb{1}_{A_Z}\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_Y} \mid Z]\} \\
&= \mathbb{E}[\mathbb{1}_{A_Z}\mathbb{E}(\mathbb{1}_{A_X} \mid Z)\mathbb{E}(\mathbb{1}_{A_Y} \mid Z)] \quad \text{by (i)} \\
&= \mathbb{E}\{\mathbb{E}[\mathbb{E}(\mathbb{1}_{A_X} \mid Z)\mathbb{1}_{A_Y \cap A_Z} \mid Z]\} \\
&= \mathbb{E}[\mathbb{E}(\mathbb{1}_{A_X} \mid Z)\mathbb{1}_A].
\end{aligned}$$

To show that $\mathcal{D}_{A_X}$ is a d-system, first notice that $\Omega \in \mathcal{D}_{A_X}$. Also, for $A_1, A_2 \in \mathcal{D}_{A_X}$ such that $A_1 \subseteq A_2$,

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_2 \setminus A_1}] &= \mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_2}] - \mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_1}] \\
&= \mathbb{E}[w(Z)\mathbb{1}_{A_2}] - \mathbb{E}[w(Z)\mathbb{1}_{A_1}] \quad \text{since } A_1, A_2 \in \mathcal{D}_{A_X}
\end{aligned}$$

$$= \mathbb{E}[w(Z)\mathbb{1}_{A_2\setminus A_1}].$$

Now consider $(A_n : n \in \mathbb{N})$, an increasing sequence in $\mathcal{D}_{A_X}$. Then $A_n \uparrow \cup_k A_k$ and $\mathbb{1}_{A_X}\mathbb{1}_{A_n} \uparrow \mathbb{1}_{A_X}\mathbb{1}_{\cup_k A_k}$ pointwise. Thus by monotone convergence $\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_n}] \uparrow \mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{\cup_k A_k}]$. Also $w(Z)\mathbb{1}_{A_n} \uparrow w(Z)\mathbb{1}_{\cup_k A_k}$ pointwise. Thus by monotone convergence, $\mathbb{E}[w(Z)\mathbb{1}_{A_n}] \uparrow \mathbb{E}[w(Z)\mathbb{1}_{\cup_k A_k}]$.
We can now see that:

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{\cup_n A_n}] &= \lim_n \mathbb{E}[\mathbb{1}_{A_X}\mathbb{1}_{A_n}] \\ &= \lim_n \mathbb{E}[w(Z)\mathbb{1}_{A_n}] \\ &= \mathbb{E}[w(Z)\mathbb{1}_{\cup_n A_n}].\end{aligned}$$

(ii) $\Rightarrow$ (iii): We will prove (iii) using monotone class theorem. Consider

$$V = \{f \text{ real, bounded and measurable} : \mathbb{E}[f(X) \mid Y, Z] = \mathbb{E}[f(X) \mid Z] \quad \text{a.s.}\}.$$

By linearity of expectation V is a vector space of real and bounded functions. Now by (ii) $\mathbb{1}_{A_X} \in V$ for all $A_X \in \sigma(X)$. Let $f_n(X) \in V$ for all $n \in \mathbb{N}$ and $f(X)$ bounded such that $0 \leq f_n(X) \uparrow f(X)$. Using conditional monotone convergence,

$$\begin{aligned}\mathbb{E}[f(X) \mid Y, Z] &= \lim_{n\to\infty} \mathbb{E}[f_n(X) \mid Y, Z] \\ &= \lim_{n\to\infty} \mathbb{E}[f_n(X) \mid Z] \text{ since } f_n(X) \in V \\ &= \mathbb{E}[f(X) \mid Z].\end{aligned}$$

Thus $f(X) \in V$ and we have shown (iii)
(iii) $\Rightarrow$ (iv): Let $f(X), g(Y)$ be real, bounded and measurable functions. Then

$$\begin{aligned}\mathbb{E}[f(X)g(Y) \mid Z] &= \mathbb{E}\{\mathbb{E}[f(X)g(Y) \mid Z, Y] \mid Z\} \quad \text{a.s.} \\ &= \mathbb{E}\{g(Y)\mathbb{E}[f(X) \mid Z, Y] \mid Z\} \quad \text{a.s.} \\ &= \mathbb{E}\{g(Y)\mathbb{E}[f(X) \mid Z] \mid Z\} \quad \text{a.s. by (iii)} \\ &= \mathbb{E}[f(X) \mid Z]\mathbb{E}[g(Y) \mid Z] \quad \text{a.s.}\end{aligned}$$

(iv) $\Rightarrow$ (i): Let $A_X \in \sigma(X)$ and $A_Y \in \sigma(Y)$. Then $\mathbb{1}_{A_X}$ is a real, bounded and $\sigma(X)$-measurable function and $\mathbb{1}_{A_Y}$ is a real, bounded and $\sigma(Y)$-measurable function. Thus (i) is a special case of (iv). □

In the following we will use the notation $X \preceq Y$ to mean that $X = f(Y)$ for some measurable function f. The descriptive terms given to the axioms are those given by Pearl (1997).

Theorem 2.3.29. (Axioms of Conditional Independence.) Let X, Y, Z, W be random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. Then the following properties hold.

$P1^s$. [Symmetry] $X \perp\!\!\!\perp_s Y \mid Z \Rightarrow Y \perp\!\!\!\perp_s X \mid Z$.

$P2^s$. $X \perp\!\!\!\perp_s Y \mid Y$.

$P3^s$. [Decomposition] $X \perp\!\!\!\perp_s Y \mid Z$ and $W \preceq Y \Rightarrow X \perp\!\!\!\perp_s W \mid Z$.

$P4^s$. [Weak Union] $X \perp\!\!\!\perp_s Y \mid Z$ and $W \preceq Y \Rightarrow X \perp\!\!\!\perp_s Y \mid (W, Z)$.

$P5^s$. [Contraction] $X \perp\!\!\!\perp_s Y \mid Z$ and $X \perp\!\!\!\perp_s W \mid (Y, Z) \Rightarrow X \perp\!\!\!\perp_s (Y, W) \mid Z$.

Proof. $P1^s$). Follows directly from Definition 2.3.27.

$P2^s$). Let $f(X), g(Y)$ be real, bounded and measurable functions. Then

$$\begin{aligned}\mathbb{E}[f(X)g(Y) \mid Y] &= g(Y)\mathbb{E}[f(X) \mid Y] \quad \text{a.s.} \\ &= \mathbb{E}[f(X) \mid Y]\mathbb{E}[g(Y) \mid Y] \quad \text{a.s.}\end{aligned}$$

which proves $P2^s$.

$P3^s$). Let $f(X)$ be a real, bounded and measurable function. Since $W \preceq Y$, it follows from Proposition 2.3.15 that $\sigma(W) \subseteq \sigma(Y)$ and thus $\sigma(W, Z) \subseteq \sigma(Y, Z)$. Then

$$\begin{aligned}\mathbb{E}[f(X) \mid W, Z] &= \mathbb{E}\{\mathbb{E}[f(X) \mid Y, Z] \mid W, Z\} \quad \text{a.s.} \\ &= \mathbb{E}\{\mathbb{E}[f(X) \mid Z] \mid W, Z\} \quad \text{a.s. since } X \perp\!\!\!\perp_s Y \mid Z \\ &= \mathbb{E}[f(X) \mid Z] \quad \text{a.s.}\end{aligned}$$

which proves $P3^s$.

$P4^s$). Let $f(X)$ be a real, bounded and measurable function. Since $W \preceq Y$, it follows from Proposition 2.3.15 that $\sigma(W) \subseteq \sigma(Y)$ and thus $\sigma(Y, W, Z) = \sigma(Y, Z)$. Then

$$\begin{aligned}\mathbb{E}[f(X) \mid Y, W, Z] &= \mathbb{E}[f(X) \mid Y, Z] \\ &= \mathbb{E}[f(X) \mid Z] \quad \text{a.s. since } X \perp\!\!\!\perp_s Y \mid Z \\ &= \mathbb{E}[f(X) \mid W, Z] \quad \text{a.s. by } P3^s\end{aligned}$$

which proves $P4^s$.

$P5^s$). Let $f(X)$ be a real, bounded and measurable function. Then

$$\begin{aligned}\mathbb{E}[f(X) \mid Y, W, Z] &= \mathbb{E}[f(X) \mid Y, Z] \quad \text{a.s. since } X \perp\!\!\!\perp_s W \mid (Y, Z) \\ &= \mathbb{E}[f(X) \mid Z] \quad \text{a.s. since } X \perp\!\!\!\perp_s Y \mid Z\end{aligned}$$

which proves $P5^s$. □

In the above theorem we have shown that stochastic conditional independence satisfies the axioms of a separoid. Denoting by V the set of all random variables defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and equipping V with the quasiorder $\preceq$, we can readily see that $(V, \preceq)$ is a join semilattice. Thus the triple $(V, \preceq, \perp\!\!\!\perp)$ is a separoid. However, $(V, \preceq, \perp\!\!\!\perp)$ is not a strong separoid (Dawid, 1979b).

Using the axioms of conditional independence we can derive many further properties of SCI in an automatic way, without calling on the specific measure-theoretic properties of the random variables involved. A nice example is the "nearest neighbour" property of a Markov Chain (Dawid, 1979a) given below.

Example 2.3.30 ("Nearest Neighbour Property of a Markov Chain")**.** Let X_1, X_2, X_3, X_4, X_5 be random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ and suppose that

(i) $X_3 \perp\!\!\!\perp_s X_1 \mid X_2$,

(ii) $X_4 \perp\!\!\!\perp_s (X_1, X_2) \mid X_3$,

(iii) $X_5 \perp\!\!\!\perp_s (X_1, X_2, X_3) \mid X_4$.

Then $X_3 \perp\!\!\!\perp_s (X_1, X_5) \mid (X_2, X_4)$.

Proof. Applying $P1^s$ to (i), we obtain

$$X_1 \perp\!\!\!\perp_s X_3 \mid X_2 \tag{2.3.2}$$

and applying $P4^s$ to (ii), we obtain

$$X_4 \perp\!\!\!\perp_s (X_1, X_2) \mid (X_2, X_3). \tag{2.3.3}$$

Applying $P3^s$ and $P1^s$ in turn to (2.3.3), we obtain

$$X_1 \perp\!\!\!\perp_s X_4 \mid (X_2, X_3) \tag{2.3.4}$$

and applying $P5^s$ to (2.3.2) and (2.3.4), we obtain

$$X_1 \perp\!\!\!\perp_s (X_3, X_4) \mid X_2. \tag{2.3.5}$$

Applying $P4^s$ to (2.3.5), we obtain

$$X_1 \perp\!\!\!\perp_s (X_3, X_4) \mid (X_2, X_4) \tag{2.3.6}$$

and applying $P3^s$ and $P1^s$ in turn to (2.3.6), we obtain

$$X_3 \perp\!\!\!\perp_s X_1 \mid (X_2, X_4). \tag{2.3.7}$$

Applying $P4^s$ to (iii), we obtain

$$X_5 \perp\!\!\!\perp_s (X_1, X_2, X_3) \mid (X_1, X_2, X_4) \tag{2.3.8}$$

and applying $P3^s$ and $P1^s$ in turn to (2.3.8), we obtain

$$X_3 \perp\!\!\!\perp_s X_5 \mid (X_1, X_2, X_4). \tag{2.3.9}$$

The result follows by applying $P5^s$ to (2.3.7) and (2.3.9). □

Another useful property we can obtain using the axioms is proved in the following proposition.

Proposition 2.3.31. Let X, Y, Z be random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. Then $X \perp\!\!\!\perp_s Y \mid Z$ implies that $(X, Z) \perp\!\!\!\perp_s Y \mid Z$.

Proof. Applying $P1^s$ to $X \perp\!\!\!\perp_s Y \mid Z$, we obtain

$$Y \perp\!\!\!\perp_s X \mid Z. \tag{2.3.10}$$

By $P2^s$, we obtain

$$Y \perp\!\!\!\perp_s (X, Z) \mid (X, Z). \tag{2.3.11}$$

Applying $P3^s$ to (2.3.11), we obtain

$$Y \perp\!\!\!\perp_s Z \mid (X, Z) \tag{2.3.12}$$

and applying $P5^s$ to (2.3.10) and (2.3.12), we obtain

$$Y \perp\!\!\!\perp_s (X, Z) \mid Z. \tag{2.3.13}$$

The result follows by applying $P1^s$ to (2.3.13). □

2.4 Variation independence

In this section we will study a different concept of conditional independence that involves non-stochastic variables. We will call this concept *variation independence* and we will see that, although it has a completely different interpretation from stochastic conditional independence, it still satisfies the axioms of a separoid.

Let $\mathcal{S}$ be a set with elements denoted by σ and V be the set of all functions defined on $\mathcal{S}$, taking values in some arbitrary space E. The elements of V will be denoted by capital letters such as $X, Y, \ldots$. Here, $X, Y, \ldots$ are non-stochastic functions so we do not now need any underlying σ-algebra or $\mathbb{P}$-measure. We denote by $\preceq$ a quasiorder on V and we write $X \preceq Y$, if $Y(\sigma_1) = Y(\sigma_2) \Rightarrow X(\sigma_1) = X(\sigma_2)$. The equivalence classes for this quasiorder correspond to partitions of $\mathcal{S}$. Then $(V, \preceq)$ forms a join semilattice, since a join $X \vee Y$ is the function $(X, Y) \in V$. We define the *(unconditional) image of* Y as $Y(\mathcal{S})$ and write it $R(Y)$. The *conditional image of* X*, given* $Y = y$ is $R(X \mid Y = y) := \{X(\sigma) : \sigma \in \mathcal{S}, Y(\sigma) = y\}$. For simplicity of notation we will sometimes write $R(X \mid y)$ instead of $R(X \mid Y = y)$, and $R(X \mid Y)$ for the function $R(X \mid Y = .)$.

Definition 2.4.1. We say that X *is variation (conditionally) independent of* Y *given* Z (on Ω) and write $X \perp\!\!\!\perp_v Y \mid Z$ $[\mathcal{S}]$ (or, if $\mathcal{S}$ is understood, just $X \perp\!\!\!\perp_v Y \mid Z$) if for any $(y, z) \in R(Y, Z)$, $R(X \mid y, z) = R(X \mid z)$.

Proposition 2.4.2. The following are equivalent.

(i) $X \perp\!\!\!\perp_v Y \mid Z$.

(ii) The function $R(X \mid Y, Z)$ of (Y, Z) is a function of Z alone.

(iii) For any $z \in R(Z)$, $R(X, Y \mid z) = R(X \mid z) \times R(Y \mid z)$.

Proof. (i) $\Rightarrow$ (ii): It readily follows from (i) since $R(X \mid y, z) = R(X|z)$ which is a function of z only.

(ii) $\Rightarrow$ (i): Let $a(Z) := R(X \mid Y, Z)$ and let $(y, z) \in R(Y, Z)$. Then

$$\begin{aligned} R(X \mid z) &= \{X(\sigma) : \sigma \in \mathcal{S}, Z(\sigma) = z\} \\ &= \bigcup_{y' \in Y(\mathcal{S})} \{X(\sigma) : \sigma \in \mathcal{S}, (Y, Z)(\sigma) = (y', z)\} \\ &= \bigcup_{y' \in Y(\mathcal{S})} R(X \mid y', z) \\ &= \bigcup_{y' \in Y(\mathcal{S})} a(z) \\ &= a(z) \\ &= R(X \mid y, z). \end{aligned}$$

(i) $\Rightarrow$ (iii): Let $z \in R(Z)$. Then

$$\begin{aligned} R(X, Y \mid z) &= \{(X, Y)(\sigma) : \sigma \in \mathcal{S}, Z(\sigma) = z\} \\ &= \bigcup_{y \in R(Y|z)} \bigcup_{x \in R(X|y,z)} \{(x, y)\} \\ &= \bigcup_{y \in R(Y|z)} \bigcup_{x \in R(X|z)} \{(x, y)\} \quad \text{by (i)} \\ &= R(X \mid z) \times R(Y \mid z). \end{aligned}$$

(iii) $\Rightarrow$ (i): Let $(y, z) \in R(Y, Z)$. Then

$$\begin{aligned} R(X \mid y, z) &= \{x : x \in R(X \mid y, z)\} \\ &= \{x : (x, y) \in R(X, Y \mid z)\} \\ &= \{x : (x, y) \in R(X \mid z) \times R(Y \mid z)\} \quad \text{by (iii)} \\ &= \{x : x \in R(X \mid z), y \in R(Y \mid z)\} \\ &= \{x : x \in R(X \mid z)\} \\ &= R(X \mid z). \end{aligned}$$

□

Proposition 2.4.3. The following are equivalent.

(i) $W \preceq Y$.

(ii) there exists $f : R(Y) \to R(W)$ such that $W = f(Y)$.

Proof. (i) to (ii): For any $y \in Y(\mathcal{S})$, let σ_y be one element in the preimage $Y^{-1}(\{y\})$. Define $f : Y(\mathcal{S}) \to W(\mathcal{S})$ such that $f(y) = W(\sigma_y)$. To check that $f(Y) = W$, let $\sigma \in \mathcal{S}$ and denote by $y := Y(\sigma)$. Then $f(Y(\sigma)) = f(y) = W(\sigma_y)$. By (i) (since $Y(\sigma) = Y(\sigma_y) = y$), we get that $W(\sigma_y) = W(\sigma)$. Thus $f(Y(\sigma)) = W(\sigma)$ for all $\sigma \in \mathcal{S}$.
(ii) to (i): It follows readily from the definition of $\preceq$. □

Theorem 2.4.4. (Axioms of variation independence.) Let X, Y, Z, W be functions on $\mathcal{S}$. Then the following properties hold.

$P1^v$. $X \perp\!\!\!\perp_v Y \mid Z \Rightarrow Y \perp\!\!\!\perp_v X \mid Z$.

$P2^v$. $X \perp\!\!\!\perp_v Y \mid Y$.

$P3^v$. $X \perp\!\!\!\perp_v Y \mid Z$ and $W \preceq Y \Rightarrow X \perp\!\!\!\perp_v W \mid Z$.

$P4^v$. $X \perp\!\!\!\perp_v Y \mid Z$ and $W \preceq Y \Rightarrow X \perp\!\!\!\perp_v Y \mid (W, Z)$.

$P5^v$. $X \perp\!\!\!\perp_v Y \mid Z$ and $X \perp\!\!\!\perp_v W \mid (Y, Z) \Rightarrow X \perp\!\!\!\perp_v Y \mid (W, Z)$.

Proof.

$P1^v$). Follows directly from Proposition 2.4.2.

$P2^v$). Let $x \in R(X)$. Then

$$\begin{aligned} R(X, Y \mid y) &:= \{(X, Y)(\sigma) : \sigma \in \mathcal{S}, Y(\sigma) = y\} \\ &= \{(X, y)(\sigma)) : \sigma \in \mathcal{S}, Y(\sigma) = y\} \\ &= R(X \mid y) \times y \\ &= R(X \mid y) \times R(Y \mid y) \end{aligned}$$

which proves $P2^v$.

$P3^v$). Let $X \perp\!\!\!\perp_v Y \mid Z$ and $W \preceq Y$. Since $W \preceq Y$, it follows from Proposition 2.4.3 that there exists $f : Y(\mathcal{S}) \to W(\mathcal{S})$ such that $W(\sigma) = f(Y(\sigma))$. For any $(w, z) \in$

$R(W, Z)$,

$$\begin{aligned}
R(X \mid w, z) &:= \{X(\sigma) : \sigma \in \mathcal{S}, (W, Z)(\sigma) = (w, z)\} \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, f(Y(\sigma)) = w, Z(\sigma) = z\} \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, Y(\sigma) \in f^{(-1)}(w), Z(\sigma) = z\} \\
&= \bigcup_{y \in f^{-1}(w)} \{X(\sigma) : \sigma \in \mathcal{S}, Y(\sigma) = y, Z(\sigma) = z\} \\
&= \bigcup_{y \in f^{-1}(w)} R(X \mid y, z) \\
&= \bigcup_{y \in f^{-1}(w)} R(X \mid z) \quad \text{since } X \perp\!\!\!\perp_v Y \mid Z \\
&= R(X \mid z)
\end{aligned}$$

which proves $P3^v$.

$P4^v$). Let $X \perp\!\!\!\perp_v Y \mid Z$ and $W \preceq Y$. Since $W \preceq Y$, it follows from Proposition 2.4.3 that there exists $f : Y(\mathcal{S}) \to W(\mathcal{S})$ such that $W(\sigma) = f(Y(\sigma))$. Now let $(y, z, w) \in R(Y, Z, W)$. Then $f(y) = w$ and we have that:

$$\begin{aligned}
R(X \mid y, z, w) &:= \{X(\sigma) : \sigma \in \mathcal{S}, (Y, Z, W)(\sigma) = (y, z, w)\} \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, Y(\sigma) = y, Z(\sigma) = z, f(Y(\sigma)) = w\} \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, Y(\sigma) = y, Z(\sigma) = z\} \quad \text{since } f(y) = w \\
&= R(X \mid y, z) \\
&= R(X \mid z) \quad \text{since } X \perp\!\!\!\perp_v Y \mid Z \\
&= \bigcup_{y' \in f^{-1}(w)} R(X|z, y') \quad \text{since } X \perp\!\!\!\perp_v Y \mid Z \text{ and } f(y) = w \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, Z(\sigma) = z, Y(\sigma) \in f^{-1}(w)\} \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, Z(\sigma) = z, f(Y(\sigma)) = w\} \\
&= \{X(\sigma) : \sigma \in \mathcal{S}, (Z, W)(\sigma) = (z, w)\} \\
&= R(X \mid z, w)
\end{aligned}$$

which proves $P4^v$.

$P5^v$). Let $X \perp\!\!\!\perp_v Y \mid Z$ and $X \perp\!\!\!\perp_v W \mid (Y, Z)$. Also let $(y, w, z) \in R(Y, W, Z)$. Then

$$R(X \mid y, w, z) = R(X \mid y, z) \quad \text{since } X \perp\!\!\!\perp_v W \mid (Y, Z)$$

$$= R(X \mid z) \quad \text{since } X \perp\!\!\!\perp_v Y \mid Z$$

which proves $P5^v$. □

In the above theorem we have shown that variation independence satisfies the axioms of a separoid, and thus $(V, \preceq, \perp\!\!\!\perp_v)$ is a separoid. In contrast with stochastic conditional independence, variation independence also satisfies the axioms of a strong separoid, and $(V, \preceq, \perp\!\!\!\perp_v)$ is a strong separoid (Dawid, 2001b).

Chapter 3

Conditional independence: Stochastic and non-stochastic variables together

3.1 Extended language for conditional independence

The aim of this chapter is to gradually extend the previous definition of conditional independence to allow it to embrace cases in which some of the variables involved are stochastic and some non-stochastic. We will connect the notion of stochastic conditional independence and variation idependence under a new definition that consistently defines a statement like $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ where X, Y, Z are stochastic variables and K, Θ, Φ are non-stochastic variables. Then we will explore the cases in which, under the extended definition, the axioms of a separoid still hold.

Conditional independence in an extended form that deals simultaneously with stochastic variables as well as parameters was firstly introduced by Dawid (1979a, 1980, 2004). In Dawid (1979a), topics like *sufficiency*, *parameter identification*, *causal inference etc.*, are revisited under the unifying language of conditional independence. A more rigorous approach and justification of the use of the same language can be found in Dawid (1980). In this thesis, we will built on the ideas introduced by the seminal work of Dawid and take a rigorous approach to expand the language and calculus of conditional independence driven by the needs of the DT-framework and in particular the need to establish the axioms under this extended form.

Studying the notions of stochastic conditional independence and variation independence one can identify a very substantial basic similarity. A statement like $X \perp\!\!\!\perp_s Y \mid Z$ for stochastic variables or $X \perp\!\!\!\perp_v Y \mid Z$ for non-stochastic variables, reflects our informal understanding that, upon conditioning on Z, further conditioning on Y will not give us more information than Z alone for making inference on X. Building on this intuitive interpretation, one can extend $X \perp\!\!\!\perp_s Y \mid Z$ to the case that one or both of Y and Z involve non-stochastic variables, such as parameters or regime indicators. An example of this extension is the very fundamental notion of sufficiency. Let $\boldsymbol{X} := X_1, X_2, \ldots, X_n$ be a random sample from a probability distribution with unknown parameter θ and let $T = T(\boldsymbol{X})$ be a statistic. In most textbooks (*e.g.* Casella and Berger (2001), p.272), the definition of sufficiency is given as follows:

Definition 3.1.1. A statistic $T = T(\boldsymbol{X})$ is a *sufficient* statistic for θ if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on the value of θ.

So a sufficient statistic T, although a reduction of the data, still contains all the information about θ contained in $\boldsymbol{X}$. This interpretation becomes clearer when we are concerned with calculating the unconditional distribution of $\boldsymbol{X}$ which does depend on θ. For example, for a discrete sample $\boldsymbol{X}$ and a sufficient statistic $T(\boldsymbol{X})$ we have:

$$\begin{aligned} P_\theta(\boldsymbol{X} = \boldsymbol{x}) &= P_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= P_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x}))P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= P(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x}))P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})) \end{aligned}$$

Since we know that the conditional distribution $P(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x}))$ does not depend on θ, $P_\theta(\boldsymbol{X} = \boldsymbol{x})$ is just proportional (as a function of θ) to $P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))$. That is, information about θ comes only through the statistic function T .

Definition 3.1.1 usually refers to the case where we have a probability mass function or a density with respect to a common measure. To formalize the definition in the general case consider a measurable space $(\Omega, \mathcal{A})$ and a family of $\mathbb{P}$-measures defined on $(\Omega, \mathcal{A})$ indexed by $\theta \in \boldsymbol{\Theta}$ and denoted by $\mathbb{P}_\theta$.

Definition 3.1.2. A statistic $T = T(\boldsymbol{X})$ is a *sufficient* statistic for θ if for any real, bounded and measurable function h, there exist a function $w(T)$ such that for any

$\theta \in \Theta$,

$$\mathbb{E}_\theta[h(X)|T] = w(T) \quad \text{a.s. } [\mathbb{P}_\theta]. \tag{3.1.1}$$

Interpreting carefully the definition, we require that for any real, bounded and measurable $h(X)$, there exists a function $w(T)$ which doesn't depend on the parameter $\theta \in \Theta$ and which serves as a version of the conditional expectation $\mathbb{E}_\theta[h(X) \mid T]$ under $[\mathbb{P}_\theta]$ simultaneously for all $\theta \in \Theta$. We could use conditional independence notation and express this statement as $X \perp\!\!\!\perp \theta \mid T$.

In the DT framework, we will mostly think in terms of different regimes, rather than parameters. So altering the notation to adjust it to the DT framework, we will consider instead of the parameter space the regime space $\mathcal{S} = \{\sigma_i : i \in I\}$, where I is a set of indices. We consider a family of $\mathbb{P}$-measures defined on $(\Omega, \mathcal{A})$ indexed by $\sigma \in \mathcal{S}$ and denoted by $\mathbb{P}_\sigma$. We consider stochastic variables to be random variables defined on Ω (say $X : (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$) which have different distributions under the different regimes $\sigma \in \mathcal{S}$.[1] Thus, we will write "$\mathbb{E}_\sigma(X \mid Y)$" to denote a version of the conditional expectation $\mathbb{E}(X \mid Y)$ under regime σ and "a.s. $[\mathbb{P}_\sigma]$" to differentiate between the different probability measures in the different regimes. We also consider non-stochastic variables to be functions defined on $\mathcal{S}$ (say $\Theta : \mathcal{S} \to \Theta(\mathcal{S})$) and call them *decision variables*. The understanding is that decision variables can give us exact or partial information upon the operating regime. Henceforth, we will denote by Σ the identity (or some injective) function on $\mathcal{S}$ ($\Sigma : \mathcal{S} \to \Sigma(\mathcal{S})$).

We can now extend Definition 3.1.2 to express a statement like $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, where X, Y, Z are random variables and Θ, Φ decision variables. In order to express such a statement, we will first describe what we would like a statement like $X \perp\!\!\!\perp \Theta \mid \Phi$ to reflect intuitively. Writing $X \perp\!\!\!\perp \Theta \mid \Phi$ as a conditional independence statement, we are asking that the distribution of X upon information on Φ no more depends on Θ. Implicitly though, we are assuming that Φ and Θ together give us all the information we need to define a regime $\sigma \in S$ (and thus the distribution of X in this regime). Formally, we will require that (Φ, Θ) defined on $\mathcal{S}$ is a bijection. In this case we say that Φ and Θ are *complementary* (on $\mathcal{S}$) or that Θ *is complementary to* Φ (on $\mathcal{S}$). Thus conditioning on $\phi \in \Phi$ and constraining to all $\sigma \in \mathcal{S}$ that belong to the preimage $\Phi^{-1}(\phi)$, we can get a common version of $\mathbb{E}_\sigma(h(X))$, that is the same in all regimes $\sigma \in \Phi^{-1}(\phi)$.

Definition 3.1.3. Let X, Y and Z be random variables and let Φ and Θ be decision

[1] The regime indicator denoted by $\sigma \in \mathcal{S}$, is not to be confused with the σ-algebra generated by X, denoted by $\sigma(X)$.

variables such that Θ is complementary to Φ. We say that X *is (conditionally) independent of* (Y, Θ) *given* (Z, Φ) and write $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ if for all $\phi \in \Phi(\mathcal{S})$ and all real, bounded and measurable functions h, there exists a function $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[h(X) \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.2}$$

We will refer to this definition of conditional independence as *extended conditional independence* (ECI). Notice that the only important property of Θ in the above definition is that it is complementary to Φ and, although Θ is important in the notation, the actual form of Θ becomes irrelevant. Henceforth, we will write down a conditional independence statement involving two decision variables only when the two variables are complementary.

Remark 3.1.4. Assume that $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and consider $w_\phi(Z)$ as in Definition 3.1.3. Then

$$\begin{aligned}
\mathbb{E}_\sigma[h(X)|Z] &= \mathbb{E}_\sigma\{[\mathbb{E}_\sigma[h(X)|Y, Z]|Z]\} \quad \text{a.s. } \mathbb{P}_\sigma \\
&= \mathbb{E}_\sigma\{[w_\phi(Z)|Z]\} \quad \text{a.s. } \mathbb{P}_\sigma \\
&= w_\phi(Z) \quad \text{a.s. } \mathbb{P}_\sigma
\end{aligned}$$

Thus $w_\phi(Z)$ also serves as a version of $\mathbb{E}_\sigma[h(X)|Z]$ for all $\sigma \in \Phi^{-1}(\phi)$.

The following example will show that even when a property such as (3.1.2) holds, we can't use just any version of the conditional expectation in one regime to serve as a version of this conditional expectation in another regime. Essentially that's because two versions of the same conditional expectation may differ on sets of probability zero and the problem arises because we are allowing sets to have probability zero in some regimes and positive probability in other regimes.

Example 3.1.5. Let $S = \{\sigma_0, \sigma_1\}$ and X, T be random variables. Under the DT framework we can think of X as the variable of interest and T as a binary treatment variable (where $T = 0$ denotes no treatment and $T = 1$ denotes treatment). Also we can think of σ_0 as the interventional regime under control treatment (*i.e.* $\mathbb{P}_{\sigma_0}(T = 0) = 1$) and σ_1 as the interventional regime under active treatment (*i.e.* $\mathbb{P}_{\sigma_1}(T = 1) = 1$). Suppose that $X \perp\!\!\!\perp \Sigma \mid T$ (thus Φ as in Definition 3.1.3 is trivial).

Then there exists $w(T)$ that doesn't depend on σ and can serve simultaneously as a version of both $\mathbb{E}_{\sigma_0}(X \mid T)$ and $\mathbb{E}_{\sigma_1}(X \mid T)$. Assume for illustration purposes

that $w(T)$ is the function defined by

$$w(t) = 1 \text{ for t=0,1.}$$

Thus $\mathbb{E}_{\sigma_0}(X \mid T) = 1$ a.s. $[\mathbb{P}_{\sigma_0}]$ and $\mathbb{E}_{\sigma_1}(X \mid T) = 1$ a.s. $[\mathbb{P}_{\sigma_1}]$. Now for $t = 0, 1$, consider functions

$$k_0(t) = 1 - t \qquad \text{and} \qquad k_1(t) = t.$$

We can see that $k_0(T) = w(T)$ a.s. $[\mathbb{P}_{\sigma_0}]$ and $k_1(T) = w(T)$ a.s. $[\mathbb{P}_{\sigma_1}]$. However almost surely under both $\mathbb{P}_{\sigma_0}$ and $\mathbb{P}_{\sigma_1}$, $k_0(T) \neq k_1(T)$. Hence neither of these variables can replace w in supplying a version of $\mathbb{E}(X \mid T)$ simultaneously in both regimes.

To control sets that have probability zero in some regimes and positive probability in other regimes, consider the following definition.

Definition 3.1.6. (Dominating regime.) Let $\mathcal{S}$ be the regime space that indexes a set of $\mathbb{P}$-measures on $(\Omega, \mathcal{A})$. For $\sigma_1, \sigma_2 \in \mathcal{S}$, we say that $\mathbb{P}_{\sigma_1}$ *dominates* $\mathbb{P}_{\sigma_2}$ or $\mathbb{P}_{\sigma_2}$ *is absolutely continuous with respect to* $\mathbb{P}_{\sigma_1}$ and write $\mathbb{P}_{\sigma_2} \ll \mathbb{P}_{\sigma_1}$ if

$$\mathbb{P}_{\sigma_1}(A) = 0 \Rightarrow \mathbb{P}_{\sigma_2}(A) = 0 \quad \text{for all } A \in \mathcal{A}.$$

For a subset $\mathcal{S}_o \subseteq \mathcal{S}$, we say that $\sigma^* \in \mathcal{S}_o$ is a *dominating regime in* $\mathcal{S}_o$, if $\mathbb{P}_\sigma \ll \mathbb{P}_{\sigma^*}$ for all $\sigma \in \mathcal{S}_o$.

Theorem 3.1.7. Let X, Y, Z be random variables and Φ, Θ be decision variables such that $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$. Suppose that for all $\phi \in \Phi(\mathcal{S})$, there exists a dominating regime $\sigma_\phi \in \Phi^{-1}(\phi)$. Then for all $\phi \in \Phi(\mathcal{S})$ and all real, bounded and measurable functions h,

$$\mathbb{E}_\sigma[h(X) \mid Y, Z] = \mathbb{E}_{\sigma_\phi}[h(X) \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma], \tag{3.1.3}$$

for all $\sigma \in \Phi^{-1}(\phi)$.

Proof. Let $\phi \in \Phi(\mathcal{S})$ and $h(X)$ be a real, bounded and measurable function. Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, there exists a function $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[h(X) \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.4}$$

In particular, for the dominating regime $\sigma_\phi \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_{\sigma_\phi}[h(X) \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_\phi}],$$

and by Remark 3.1.4,

$$\mathbb{E}_{\sigma_\phi}[h(X) \mid Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_\phi}].$$

Since $\mathbb{P}_\sigma \ll \mathbb{P}_{\sigma_\phi}$ for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_{\sigma_\phi}[h(X) \mid Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.5}$$

By (3.1.4) and (3.1.5), it follows that

$$\mathbb{E}_\sigma[h(X) \mid Y, Z] = \mathbb{E}_{\sigma_\phi}[h(X) \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.6}$$

□

The above theorem shows that we can use any version of $\mathbb{E}_{\sigma_\phi}[h(X) \mid Z]$ in the dominating regime σ_ϕ, to serve as the function $w_\phi(Z)$.

Under the DT framework, the form of Definition 3.1.3 comes to represent naturally what we want such a statement to express. However, in most contexts, definition of conditional independence is given in a form that involves indicator functions of measurable sets. Such a form is helpful for the purposes of proving properties of conditional independence. Thus, we will use some standard tools of measure theory to deduce equivalent versions of Definition 3.1.3.

Proposition 3.1.8. Let X, Y, Z be random variables and let Φ, Θ be decision variables. Then the following are equivalent.

(i) $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$.

(ii) For all $\phi \in \Phi(\mathcal{S})$ and all real, bounded and measurable function h_1, there exists a function $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$ and all real, bounded and measurable functions h_2 ,

$$\mathbb{E}_\sigma[h_1(X)h_2(Y) \mid Z] = w_\phi(Z)\mathbb{E}_\sigma[h_2(Y) \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma].$$

(iii) For all $\phi \in \Phi(\mathcal{S})$ and all $A_X \in \sigma(X)$, there exists a function $w_\phi(Z)$ such that

for all $\sigma \in \Phi^{-1}(\phi)$ and all $A_Y \in \sigma(Y)$,

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X \cap A_Y} \mid Z) = w_\phi(Z)\mathbb{E}_\sigma(\mathbb{1}_{A_Y} \mid Z) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

(iv) For all $\phi \in \Phi(\mathcal{S})$ and all $A_X \in \sigma(X)$, there exists a function $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X} \mid Y, Z) = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.7}$$

Proof. (i) $\Rightarrow$ (ii): Let $\phi \in \Phi(\mathcal{S})$ and h_1 to be a real, bounded and measurable function. Then for $\sigma \in \Phi^{-1}(\phi)$ and h_2 a real bounded and measurable function, we have:

$$\begin{aligned}
\mathbb{E}_\sigma[h_1(X)h_2(Y) \mid Z] &= \mathbb{E}_\sigma\{\mathbb{E}_\sigma[h_1(X)h_2(Y) \mid Y, Z] \mid Z\} \quad \text{a.s. } [\mathbb{P}_\sigma] \\
&= \mathbb{E}_\sigma\{h_2(Y)\mathbb{E}_\sigma[h_1(X) \mid Y, Z] \mid Z\} \quad \text{a.s. } [\mathbb{P}_\sigma] \\
&= \mathbb{E}_\sigma\{h_2(Y)w_\phi(Z) \mid Z\} \quad \text{a.s. } [\mathbb{P}_\sigma] \text{ by i)} \\
&= w_\phi(Z)\mathbb{E}_\sigma[h_2(Y) \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma].
\end{aligned}$$

(ii)$\Rightarrow$ (iii): Accrues as a special case of (ii), as the indicator function $\mathbb{1}_{A_X}$ of a measurable set $A_X \in \sigma(X)$ is a real, bounded and $\sigma(X)$-measurable function. Similarly, $\mathbb{1}_{A_Y}$ is a real, bounded and $\sigma(Y)$-measurable function.

(iii)$\Rightarrow$ (iv): Let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$ and consider $w_\phi(Z)$ as in (iii). Note that (3.1.7) is equivalent to:

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}] = \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}] \quad \text{whenever } A_{Y,Z} \in \sigma(Y, Z) \tag{3.1.8}$$

for all $\sigma \in \Phi^{-1}(\phi)$. To show (3.1.8), consider

$$\mathcal{D}_{A_X} = \{A_{Y,Z} \in \sigma(Y, Z) : \mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}] = \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}]\}$$

and

$$\Pi = \{A_{Y,Z} \in \sigma(Y, Z) : A_{Y,Z} = A_X \cap A_Y \text{ for some } A_Y \in \sigma(Y), A_Z \in \sigma(Z)\}.$$

By Proposition 2.3.12, Π is a π-system and $\sigma(\Pi) = \sigma(Y, Z)$. We will show that $\mathcal{D}_{A_X}$ is a d-system that contains Π and apply Dynkin's lemma to conclude that $\mathcal{D}_{A_X}$ contains $\sigma(\Pi) = \sigma(Y, Z)$ and thus prove (3.1.8).

To show that $\mathcal{D}_{A_X}$ contains Π, let $A_{Y,Z} = A_Y \cap A_Z$ such that $A_Y \in \sigma(Y)$ and $A_Z \in \sigma(Z)$. Then

$$\begin{aligned}
\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}] &= \mathbb{E}_\sigma\{\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_Y\cap A_Z} \mid Z]\} \\
&= \mathbb{E}_\sigma\{\mathbb{1}_{A_Z}\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_Y} \mid Z]\} \\
&= \mathbb{E}_\sigma[\mathbb{1}_{A_Z} w_\phi(Z)\mathbb{E}_\sigma(\mathbb{1}_{A_Y} \mid Z)] \quad \text{by iii)} \\
&= \mathbb{E}_\sigma\{\mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_Y\cap A_Z} \mid Z]\} \\
&= \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}].
\end{aligned}$$

To show that $\mathcal{D}_{A_X}$ is a d-system, first notice that $\Omega \in \mathcal{D}_{A_X}$. Also, for $A_1, A_2 \in \mathcal{D}_{A_X}$ such that $A_1 \subseteq A_2$, we have that:

$$\begin{aligned}
\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_2\setminus A_1}] &= \mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_2}] - \mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_1}] \\
&= \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_2}] - \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_1}] \quad \text{since } A_1, A_2 \in \mathcal{D}_{A_X} \\
&= \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_2\setminus A_1}].
\end{aligned}$$

Now consider $(A_n : n \in \mathbb{N})$, an increasing sequence in $\mathcal{D}_{A_X}$. Then $A_n \uparrow \cup_k A_k$ and $\mathbb{1}_{A_X}\mathbb{1}_{A_n} \uparrow \mathbb{1}_{A_X}\mathbb{1}_{\cup_k A_k}$ pointwise. Thus by monotone convergence $\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_n}] \uparrow \mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{\cup_k A_k}]$. Also $w_\phi(Z)\mathbb{1}_{A_n} \uparrow w_\phi(Z)\mathbb{1}_{\cup_k A_k}$ pointwise. Thus by monotone convergence $\mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_n}] \uparrow \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{\cup_k A_k}]$. We can now see that:

$$\begin{aligned}
\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{\cup_n A_n}] &= \lim_n \mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_n}] \\
&= \lim_n \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_n}] \quad \text{since } A_n \in \mathcal{D}_{A_X} \\
&= \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{\cup_n A_n}].
\end{aligned}$$

(iv)$\Rightarrow$ (i): We will prove (iv) using monotone class theorem. Let $\phi \in \Phi(\mathcal{S})$ and consider

$V = \{h$ real, bounded and measurable: there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$, $\mathbb{E}_\sigma[h(X) \mid Y, Z] = w_\phi(Z) \quad$ a.s. $[\mathbb{P}_\sigma]\}$.

By linearity of expectation V is a vector space of real and bounded functions and by (iv) $\mathbb{1}_{A_X} \in V$ for all $A_X \in \sigma(X)$. Now let $h_n \in V$ for all $n \in \mathbb{N}$ and $0 \leq h_n \uparrow h$

for h bounded. Using conditional monotone convergence

$$\begin{aligned}\mathbb{E}_\sigma[h(X) \mid Y, Z] &= \lim_{n\to\infty} \mathbb{E}_\sigma[h_n(X) \mid Y, Z] \\ &= \lim_{n\to\infty} w_\phi^n(Z) \quad \text{since } h_n \in V \\ &=: w_\phi(Z)\end{aligned}$$

which proves that $h \in V$. By monotone class theorem, V contains every bounded measurable function and thus we have shown (i). □

Using Proposition 3.1.8 we can obtain more properties of ECI. For example, we can show that Definition 3.1.3 can be equivalently expressed in two simpler statements of ECI, or that when all the decision variables are confined to the rightmost term symmetry does follow. We will see that even more properties accrue in Section 3.3.

Proposition 3.1.9. Let X, Y, Z be random variables and Φ, Θ decision variables. Then the following are equivalent:

i) $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$

ii) $X \perp\!\!\!\perp Y \mid (Z, \Phi, \Theta)$ and $X \perp\!\!\!\perp \Theta \mid (Z, \Phi)$

Proof. i) ⇒ ii). Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, for all $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[1_{A_X} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

which proves that $X \perp\!\!\!\perp Y \mid (Z, \Phi, \Theta)$. Also, by Remark 3.1.4,

$$\mathbb{E}_\sigma[1_{A_X} \mid Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

which proves that $X \perp\!\!\!\perp \Theta \mid (Z, \Phi)$.

ii) ⇒ i). Since $X \perp\!\!\!\perp Y \mid (Z, \Phi, \Theta)$, for all $\sigma \in \mathcal{S}$ and $A_X \in \sigma(X)$, there exists $w_\sigma(Z)$ such that

$$\mathbb{E}_\sigma[1_{A_X} \mid Y, Z] = w_\sigma(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.9}$$

By Remark 3.1.4,

$$\mathbb{E}_\sigma[1_{A_X} \mid Z] = w_\sigma(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.10}$$

Since $X \perp\!\!\!\perp \Theta \mid (Z, \Phi)$, for all $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[1_{A_X} \mid Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.1.11}$$

By (3.1.10) and (3.1.11)

$$w_\sigma(Z) = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

Thus, by (3.1.9)

$$\mathbb{E}_\sigma[1_{A_X} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

which proves that $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$. □

Proposition 3.1.10. Let X, Y, Z be random variables and let Σ be a decision variable. Then $X \perp\!\!\!\perp Y \mid (Z, \Sigma)$ implies $Y \perp\!\!\!\perp X \mid (Z, \Sigma)$.

Proof. Since $X \perp\!\!\!\perp Y \mid (Z, \Sigma)$, for all $\sigma \in \mathcal{S}$ and all $A_X \in \sigma(X)$, there exists a function $w_\sigma(Z)$ such that for all $A_Y \in \sigma(Y)$,

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X \cap A_Y} \mid Z) = w_\sigma(Z)\mathbb{E}_\sigma(\mathbb{1}_{A_Y} \mid Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

In particular, for $A_Y = \Omega$, we have that

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X} \mid Z) = w_\sigma(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

Thus

$$\mathbb{E}_\sigma(\mathbb{1}_{A_Y \cap A_X} \mid Z) = \mathbb{E}_\sigma(\mathbb{1}_{A_Y} \mid Z)\mathbb{E}_\sigma(\mathbb{1}_{A_X} \mid Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

which concludes the proof. □

3.2 A first approach

At a first instance it might not seem straightforward why ECI follows the same axioms as SCI. However observing Definition 3.1.3 and Definition 2.3.27 one can see that the two concepts are defined in a very similar way. This similar structure allows us to believe that similar properties can also accrue, and motivates the conjecture that the axioms of conditional independence still hold for the extended concept. The approach we will take in this chapter is to extend the original space in order to

conceive both stochastic and non-stochastic variables as measurable functions on the new space and thus create an analogy between ECI and SCI. Similar ideas can be found in a variety of contexts in probability theory and statistics. Examples include Poisson random processes (Kingman (1993), p.82-84) or Bayesian approaches to statistics (Kolmogorov, 1942). We will see that, under the assumption of a discrete regime space, ECI and SCI are equivalent. Thus we can continue to apply all the properties $P1^s$–$P5^s$ of Theorem 2.3.29.

Consider a measurable space $(\Omega, \mathcal{A})$ and a regime space $\mathcal{S}$ and let $\mathcal{F}$ be the σ-algebra of all subsets of $\mathcal{S}$. Further, consider π an arbitrary $\mathbb{P}$-measure on $\mathcal{F}$. We can expand the original space $(\Omega, \mathcal{A})$ and consider the product space $\Omega \times \mathcal{S}$ with its corresponding σ-algebra $\mathcal{A} \otimes \mathcal{F}$, where $\mathcal{A} \otimes \mathcal{F} := \sigma(\mathcal{A} \times \mathcal{F}) := \sigma(\{A \times B : A \in \mathcal{A}, B \in \mathcal{F}\})$. Thus, we can regard all stochastic variables $X, Y, Z, \ldots$ defined on $(\Omega, \mathcal{A})$ also as defined on $(\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F})$ and all decision variables $\Theta, \Phi, \ldots$ defined on $\mathcal{S}$ also as defined on $(\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F})$. To see this, consider any random variable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}_X)$. For any such X we can define $X^* : (\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_X)$ as $X^*(\omega, \sigma) = X(\omega)$. It is readily seen that X^* is $\mathcal{A} \otimes \mathcal{F}$-measurable. Similar justification applies for decision variables. Thus, in the initial space $(\Omega, \mathcal{A})$ we can talk about ECI and in the product space $(\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F})$, after we equip it with a $\mathbb{P}$-measure, we can talk about SCI. To rigorously justify the equivalence of ECI and SCI, we will need some results from measure theory.

Lemma 3.2.1. Let $f : \Omega \times \mathcal{S} \to \mathbb{R}$ be $\mathcal{A} \otimes \mathcal{F}$-measurable. Define for all $\sigma \in \mathcal{S}$, $f_\sigma : \Omega \to \mathbb{R}$ by $f_\sigma(\omega) := f(\omega, \sigma)$. Then f_σ is $\mathcal{A}$-measurable. If further f is bounded, define for all $\sigma \in \mathcal{S}$, $\mathbb{E}_\sigma(f_\sigma) : \mathcal{S} \to \mathbb{R}$ by $\mathbb{E}_\sigma(f_\sigma) := \int_\Omega f_\sigma(\omega)\mathbb{P}_\sigma(d\omega)$. Then $\mathbb{E}_\sigma(f_\sigma)$ is bounded and $\mathcal{F}$-measurable.

Proof. See Billingsley (1995) (p.231, Theorem 18.1. and p.234, Theorem 18.3.). □

Theorem 3.2.2. For $A^* \in \mathcal{A} \otimes \mathcal{F}$, let

$$\mathbb{P}^*(A^*) := \int_{\mathcal{S}} \int_{\Omega} \mathbb{1}_{A^*}(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma). \tag{3.2.1}$$

Then $\mathbb{P}^*$ is the unique $\mathbb{P}$-measure on $\mathcal{A} \otimes \mathcal{F}$ such that

$$\mathbb{P}^*(A \times B) = \int_B \mathbb{P}_\sigma(A)\pi(d\sigma) \tag{3.2.2}$$

for all $A \in \mathcal{A}$ and $B \in \mathcal{F}$.

Proof. By Lemma 3.2.1, $\mathbb{P}^* : \mathcal{A} \otimes \mathcal{F} \to [0,1]$ is a well defined function. To prove that $\mathbb{P}^*$ is a measure, we need to prove countable additivity. So let $(A_n : n \in \mathbb{N})$ be a sequence of disjoint sets in $\mathcal{A} \otimes \mathcal{F}$ and define $B_n := \cup_{k=1}^n A_k$ an increasing sequence of sets. By monotone convergence theorem, for each $\sigma \in \mathcal{S}$, as $n \to \infty$,

$$\int_\Omega \mathbb{1}_{B_n}(\omega, \sigma)\mathbb{P}_\sigma(d\omega) \uparrow \int_\Omega \mathbb{1}_{\cup_k B_k}(\omega, \sigma)\mathbb{P}_\sigma(d\omega),$$

and hence

$$\int_\mathcal{S}\int_\Omega \mathbb{1}_{B_n}(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma) \uparrow \int_\mathcal{S}\int_\Omega \mathbb{1}_{\cup_k B_k}(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma). \tag{3.2.3}$$

Thus

$$\begin{aligned}
\mathbb{P}^*(\bigcup_n A_n) &= \mathbb{P}^*(\bigcup_n B_n) \\
&= \lim_{n\to\infty} \mathbb{P}^*(B_n) \quad \text{by (3.2.3)} \\
&= \lim_{n\to\infty} \int_\mathcal{S}\int_\Omega \sum_{k=1}^n \mathbb{1}_{A_n}(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma) \quad \text{since } A_n \text{ disjoint} \\
&= \sum_n \mathbb{P}^*(A_n).
\end{aligned}$$

We can readily see that $\mathbb{P}^*$ is a $\mathbb{P}$-measure and since $\mathbb{1}_{A\times B} = \mathbb{1}_A\mathbb{1}_B$, property (3.2.2) holds for all $A \in \mathcal{A}$ and $B \in \mathcal{F}$. Since $\mathcal{A} \times \mathcal{F} := \{A \times B : A \in \mathcal{A}, B \in \mathcal{F}\}$ is a π-system generating $\mathcal{A} \otimes \mathcal{F}$ and $\mathbb{P}^*(\Omega \times \mathcal{S}) = 1 < \infty$, $\mathbb{P}^*$ is uniquely determined by its values on $\mathcal{A} \times \mathcal{F}$, by the uniqueness of extension theorem. □

Theorem 3.2.3. Let $f : \Omega \times \mathcal{S} \to \mathbb{R}$ be an $\mathcal{A} \otimes \mathcal{F}$-measurable integrable function. Then

$$\mathbb{E}^*(f) = \int_\mathcal{S}\int_\Omega f(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma) \tag{3.2.4}$$

Proof. Since f is integrable, $\mathbb{E}^*(f) = \mathbb{E}^*(f^+) - \mathbb{E}^*(f^-)$. Thus, it is enough to show (3.2.4) for non-negative $\mathcal{A} \otimes \mathcal{F}$-measurable functions. By definition of $\mathbb{P}^*$ in Theorem 3.2.2, (3.2.4) holds for all $f = \mathbb{1}_A$, where $A \in \mathcal{A} \otimes \mathcal{F}$. By linearity of the integrals, it also holds for functions of the form $f = \sum_{k=1}^m a_k\mathbb{1}_{A_k}$, where $0 \le a_k < \infty$, $A_k \in \mathcal{A} \otimes \mathcal{F}$ for all k and $m \in \mathbb{N}$. We call functions of this form simple functions. For any non-negative $\mathcal{A} \otimes \mathcal{F}$-measurable function f, consider the sequence of non-negative simple functions $f_n = \min\{\frac{\lfloor 2^n f\rfloor}{2^n}, n\}$. Then (3.2.4) holds for f_n and $f_n \uparrow f$.

By monotone convergence theorem, $\mathbb{E}^*(f_n) \uparrow \mathbb{E}^*(f)$ and for each $\sigma \in \mathcal{S}$,

$$\int_\Omega f_n(\omega, \sigma)\mathbb{P}_\sigma(d\omega) \uparrow \int_\Omega f(\omega, \sigma)\mathbb{P}_\sigma(d\omega)$$

and hence

$$\int_\mathcal{S}\int_\Omega f_n(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma) \uparrow \int_\mathcal{S}\int_\Omega f(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(d\sigma).$$

Hence (3.2.4) extends to f. □

In the previous theorems, we have rigorously constructed a new $\mathbb{P}$-measure $\mathbb{P}^*$ on the measurable space $(\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F})$ and also obtained an expression of the integral of a $\mathcal{A} \otimes \mathcal{F}$-measurable function under $\mathbb{P}^*$. We will use this expression to justify the analogy between ECI and SCI in the case of a discrete regime space.

3.2.1 The case of a discrete regime space

We will keep the notation introduced above and for a random variable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}_X)$ we will denote by $X^* : (\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_X)$ the function defined by $X^*(\omega, \sigma) = X(\omega)$. Similarly for a decision variable $\Theta : \mathcal{S} \to \Theta(\mathcal{S})$ we will denote by $\Theta^* : (\Omega \times \mathcal{S}, \mathcal{A} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_X)$ the function defined by $\Theta^*(\omega, \sigma) = \Theta(\sigma)$. We will use similar conventions for all the variables involved in the problem.

In the case where $\mathcal{S}$ is discrete, we can consider an everywhere positive probability mass function $\pi(\sigma)$[2] over $\mathcal{S}$. Then (3.2.4) becomes:

$$\begin{aligned}\mathbb{E}^*(f) &= \sum_{\sigma\in\mathcal{S}}\int_\Omega f(\omega, \sigma)\mathbb{P}_\sigma(d\omega)\pi(\sigma)\\ &= \sum_{\sigma\in\mathcal{S}}\mathbb{E}_\sigma(f_\sigma)\pi(\sigma).\end{aligned}$$

Remark 3.2.4. Notice that for any X^* as above, $\sigma(X^*) = \sigma(X) \times \{\mathcal{S}\}$. Similarly, for any Θ^* as above, $\sigma(\Theta^*) = \{\Omega\} \times \sigma(\Theta)$. Thus

$$\begin{aligned}\sigma(X^*, \Theta^*) &= \sigma(\{A_{X^*} \cap A_{\Theta^*} : A_{X^*} \in \sigma(X^*), A_{\Theta^*} \in \sigma(\Theta^*)\}) \quad \text{by Proposition 2.3.12}\\ &= \sigma(\{A_{X^*} \cap A_{\Theta^*} : A_{X^*} \in \sigma(X) \times \{\mathcal{S}\}, A_{\Theta^*} \in \{\Omega\} \times \sigma(\Theta)\})\\ &= \sigma(\{(A_X \times \mathcal{S}) \cap (\Omega \times A_\Theta) : A_X \in \sigma(X), A_\Theta \in \sigma(\Theta)\})\end{aligned}$$

[2]For a finite regime space $S = \{\sigma_1, \ldots, \sigma_N\}$ consider $\pi(\sigma) = 1/N$ and for a countable regime space $S = \{\sigma_i : i \geq 1\}$ consider $\pi(\sigma_i) = 1/2^i$ (i=1,2,...).

$$= \sigma(\{A_X \times A_\Theta : A_X \in \sigma(X), A_\Theta \in \sigma(\Theta)\})$$
$$=: \sigma(X) \otimes \sigma(\Theta).$$

Thus, for any $\sigma \in \mathcal{S}$ and $A_{X^*} \in \sigma(X^*)$, the function $\mathbb{1}^\sigma_{A_{X^*}} : \Omega \to \{0,1\}$ defined by $\mathbb{1}^\sigma_{A_{X^*}}(\omega) := \mathbb{1}_{A_{X^*}}(\omega, \sigma)$ does not depend on σ. It is equal to $\mathbb{1}_{A_X}$, for $A_X \in \sigma(X)$ such that $A_{X^*} = A_X \times \{\mathcal{S}\}$. Also for $A_{X^*,\Theta^*} \in \sigma(X^*, \Theta^*)$, the function $\mathbb{1}_{A_{X^*,\Theta^*}}$ is $(\sigma(X) \otimes \sigma(\Theta))$-measurable, and by Lemma 3.2.1, for $\sigma \in \mathcal{S}$, the function $\mathbb{1}^\sigma_{A_{X^*,\Theta^*}} : \Omega \to \{0,1\}$ defined by $\mathbb{1}^\sigma_{A_{X^*,\Theta^*}}(\omega) := \mathbb{1}_{A_{X^*,\Theta^*}}(\omega, \sigma)$ is $\sigma(X)$-measurable. $\mathbb{1}^\sigma_{A_{X^*,\Theta^*}}(\omega)$ is equal to $\mathbb{1}_{A^\sigma_X}$ for $A^\sigma_X \in \sigma(X)$ such that A^σ_X is the section of A_{X^*,Θ^*} at σ.

Theorem 3.2.5. Let X, Y, Z be $\mathcal{A}$-measurable functions on Ω, and let Φ, Θ be decision variables on $\mathcal{S}$, where $\mathcal{S}$ is discrete. Suppose that Θ is complementary to Φ. Also, let X^*, Y^*, Z^* and Φ^*, Θ^* be the respective $\mathcal{A} \otimes \mathcal{F}$-measurable functions. Then $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ (ECI) if and only if $X^* \perp\!\!\!\perp_s (Y^*, \Theta^*) \mid (Z^*, \Phi^*)$ (SCI).

Proof. $\Rightarrow$) Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, by Proposition 3.1.8, for all $\phi \in \Phi(\mathcal{S})$ and all $A_X \in \sigma(X)$, there exists a function $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X} \mid Y, Z) = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma], \tag{3.2.5}$$

i.e.,

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X} \mathbb{1}_{A_{Y,Z}}) = \mathbb{E}_\sigma(w_\phi(Z)\mathbb{1}_{A_{Y,Z}}) \quad \text{whenever } A_{Y,Z} \in \sigma(Y, Z). \tag{3.2.6}$$

To show that $X^* \perp\!\!\!\perp (Y^*, \Theta^*) \mid (Z^*, \Phi^*)$, by Proposition 2.3.28, we need to show that for all $A_{X^*} \in \sigma(X^*)$, there exists a function $w(Z^*, \Phi^*)$ such that

$$\mathbb{E}^*[\mathbb{1}_{A_{X^*}} \mid Y^*, \Theta^*, Z^*, \Phi^*] = w(Z^*, \Phi^*) \quad \text{a.s.} \tag{3.2.7}$$

i.e.

$$\mathbb{E}^*(\mathbb{1}_{A_{X^*}} \mathbb{1}_{A_{Y^*,\Theta^*,Z^*,\Phi^*}}) = \mathbb{E}^*(w(Z^*, \Phi^*)\mathbb{1}_{A_{Y^*,\Theta^*,Z^*,\Phi^*}}) \quad \text{whenever } A_{Y^*,\Theta^*,Z^*,\Phi^*} \in \sigma(Y^*, \Theta^*, Z^*, \Phi^*). \tag{3.2.8}$$

Let $A_{X^*} \in \sigma(X^*)$ and define $w(z^*, \phi^*) = w_{\phi^*}(z^*)$ as in (3.2.6). Then for all $A_{Y^*,\Theta^*,Z^*,\Phi^*} \in \sigma(Y^*, \Theta^*, Z^*, \Phi^*)$,

$$\mathbb{E}^*(\mathbb{1}_{A_{X^*}} \mathbb{1}_{A_{Y^*,\Theta^*,Z^*,\Phi^*}}) = \sum_{\sigma \in \mathcal{S}} \mathbb{E}_\sigma(\mathbb{1}_{A_X} \mathbb{1}_{A^\sigma_{Y,Z}})\pi(\sigma)$$

$$
\begin{aligned}
&= \sum_{\sigma \in \mathcal{S}} \mathbb{E}_\sigma(w_\phi(Z) \mathbb{1}_{A^\sigma_{Y,Z}}) \pi(\sigma) \quad \text{by (3.2.6)} \\
&= \mathbb{E}^*(w(Z^*, \Phi^*) \mathbb{1}_{A_{Y^*,\Theta^*,Z^*,\Phi^*}})
\end{aligned}
$$

which proves (3.2.8).
$\Leftarrow$) To show that $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$. Then, for any $\sigma_o = \Phi^{-1}(\phi)$,

$$
\begin{aligned}
\mathbb{E}_{\sigma_o}(\mathbb{1}_{A_X} \mathbb{1}_{A_{Y,Z}}) \pi(\sigma_o) &= \sum_{\sigma \in \mathcal{S}} \mathbb{E}_\sigma[\mathbb{1}_{A_X} \mathbb{1}_{A_{Y,Z}} \mathbb{1}_{\sigma_o}(\sigma)] \pi(\sigma) \\
&= \mathbb{E}^*(\mathbb{1}_{A_{X^*}} \mathbb{1}_{A_{Y,Z} \times \{\sigma_o\}}) \\
&= \mathbb{E}^*(w(Z^*, \Phi^*) \mathbb{1}_{A_{Y,Z} \times \{\sigma_o\}}) \quad \text{by (3.2.8)} \\
&= \sum_{\sigma \in \mathcal{S}} \mathbb{E}_\sigma[w(Z; \Phi(\sigma)) \mathbb{1}_{A_{Y,Z}} \mathbb{1}_{\sigma_o}(\sigma)] \pi(\sigma) \\
&= \mathbb{E}_{\sigma_o}[w(Z; \Phi(\sigma_o)) \mathbb{1}_{A_{Y,Z}}] \pi(\sigma_o).
\end{aligned}
$$

Since $\pi(\sigma_o) > 0$, we have proved (3.2.6) with $w_\phi(z) = w(z, \phi)$. □

Inspecting the proof of Theorem 3.2.5, we see that the assumption of discreteness of the regime space $\mathcal{S}$ is crucial. If we assume an uncountable regime space $\mathcal{S}$ and assign a prior distribution to it, the arguments for the forward direction will still apply but the arguments for the reverse direction will not. Intuitively, this is because (3.2.7) holds almost everywhere and not everywhere. Thus we cannot immediately extend it to hold for all $\sigma \in \mathcal{S}$ as in (3.2.5).

Corollary 3.2.6. Suppose we are given a collection of ECI properties as in the form of Definition 3.1.3. If the regime space $\mathcal{S}$ is discrete, any deduction made using the axioms of conditional independence that produces another ECI property as in the form of Definition 3.1.3 is valid.

The form of Definition 3.1.3 allows us to incorporate decision variables in a conditional independence statement as long as they do not appear in the left-most term of a statement. Thus using the above corollary, we can apply all the axioms of conditional independence as long as in both premisses and conclusions we consider a statement of such form. However, in the intermediate steps of an argument we can violate this condition and even consider some steps that are not so interpretable. For example, we can consider (as an intermediate step) the axiom of symmetry on $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, which induces the statement $(Y, \Theta) \perp\!\!\!\perp X \mid (Z, \Phi)$ that we have not defined when Θ and Φ are decision variables.

3.3 A second approach

Realising the limitations that occur using the previous approach, in this section we will explore a different path. We will examine the axioms of conditional independence using the measure theoretic definition of conditional expectation. As mentioned above, we must exercise some care as we can only be concerned with statements of an appropriate form.

Theorem 3.3.1. Let X, Y, Z, W be random variables and Φ, Θ, Σ be decision variables. Then the following properties hold.

P1′. $X \perp\!\!\!\perp Y \mid (Z, \Sigma) \Rightarrow Y \perp\!\!\!\perp X \mid (Z, \Sigma)$.

P2′. $X \perp\!\!\!\perp (Y, \Sigma) \mid (Y, \Sigma)$.

P3′. $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $W \preceq Y \Rightarrow X \perp\!\!\!\perp (W, \Theta) \mid (Z, \Phi)$.

P4′. $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $W \preceq Y \Rightarrow X \perp\!\!\!\perp (Y, \Theta) \mid (Z, W, \Phi)$.

P5′. $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $X \perp\!\!\!\perp W \mid (Y, Z, \Theta, \Phi) \Rightarrow X \perp\!\!\!\perp (Y, W, \Theta) \mid (Z, \Phi)$.

Proof. *P1′*). Proved in Proposition 3.1.10.

P2′). Let $\sigma \in \mathcal{S}$ and $A_X \in \sigma(X)$. Then for all $A_Y \in \sigma(Y)$,

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X \cap A_Y} \mid Y) = \mathbb{1}_{A_Y} \mathbb{E}_\sigma(\mathbb{1}_{A_X} \mid Y) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

which concludes the proof.

P3′). Let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$. Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

Since $W \preceq Y$, it follows from Proposition 2.3.15 that $\sigma(W) \subseteq \sigma(Y)$ and thus $\sigma(W, Z) \subseteq \sigma(Y, Z)$. Then

$$\begin{aligned} \mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid W, Z] &= \mathbb{E}_\sigma[\mathbb{E}_\sigma(\mathbb{1}_{A_X} \mid Y, Z) \mid W, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= \mathbb{E}_\sigma[w_\phi(Z) \mid W, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma] \end{aligned}$$

which concludes the proof.

$P4'$). Let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$. Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

Since $W \preceq Y$, it follows from Proposition 2.3.15 that $\sigma(W) \subseteq \sigma(Y)$ and thus $\sigma(Y, Z, W) = \sigma(Y, Z)$. Then

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z, W] &= \mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]\end{aligned}$$

which concludes the proof.

$P5'$). Let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$. Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

Since $W \preceq Y$, it follows from Proposition 2.3.15 that $\sigma(W) \subseteq \sigma(Y)$ and thus $\sigma(Y, W, Z) = \sigma(Y, Z)$. Then

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, W, Z] &= \mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]\end{aligned}$$

which concludes the proof. □

In the above theorem we have proven that all the axioms hold apart from symmetry. Lack of symmetry however, introduces some complications as the symmetric equivalents of axioms $P3'$, $P4'$ and $P5'$ cannot be automatically assumed to hold. Consider the following statements.

$P3''$. $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $W \preceq X \Rightarrow W \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$.

$P4''$. $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $W \preceq X \Rightarrow X \perp\!\!\!\perp (Y, \Theta) \mid (Z, W, \Phi)$.

$P5''$. $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $W \perp\!\!\!\perp (Y, \Theta) \mid (X, Z, \Phi) \Rightarrow (X, W) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$.

Although $P3''$ follows straightforwardly, we still need to prove $P4''$ and $P5''$. $P5''$ can be proved to hold (Proposition 3.3.3) but $P4''$ presents some difficulty. We will see that under additional conditions we can obtain $P4''$ (Proposition 3.3.4) but validity in full generality remains an open problem.

Lemma 3.3.2. Let X, Y, Z, W be random variables and Φ, Θ be decision variables. Then

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi) \text{ and } W \preceq X \Rightarrow (W, Z) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi).$$

Proof. Since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, for all $\phi \in \Phi(\mathcal{S})$ and all $A_X \in \sigma(X)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.1}$$

To prove that $(W, Z) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, let $\phi \in \Phi(\mathcal{S})$ and $A_{W,Z} \in \sigma(W, Z)$. We will show that there exists $a_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_{W,Z}} \mid Y, Z] = a_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.2}$$

Consider

$$\mathcal{D} = \{A_{W,Z} \in \sigma(W, Z) : \text{there exists } a_\phi(Z) \text{ such that (3.3.2) holds}\}$$

and

$$\Pi = \{A_{W,Z} \in \sigma(W, Z) : A_{W,Z} = A_W \cap A_Z \text{ for } A_W \in \sigma(W) \text{ and } A_Z \in \sigma(Z)\}.$$

By Proposition 2.3.12, Π is a π-system and $\sigma(\Pi) = \sigma(W, Z)$. We will show that $\mathcal{D}$ is a d-system that contains Π and apply Dynkin's lemma to conclude that $\mathcal{D}$ contains $\sigma(\Pi) = \sigma(W, Z)$.

To show that $\mathcal{D}$ contains Π, let $A_{W,Z} = A_W \cap A_Z$ such that $A_W \in \sigma(W)$ and $A_Z \in \sigma(Z)$. Then

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{A_W}\mathbb{1}_{A_Z} \mid Y, Z] &= \mathbb{1}_{A_Z}\mathbb{E}_\sigma[\mathbb{1}_{A_W} \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= \mathbb{1}_{A_Z} w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma] \text{ by (3.3.1).}\end{aligned}$$

We define $a_\phi(Z) := \mathbb{1}_{A_Z} w_\phi(Z)$ and we are done.

To show that $\mathcal{D}$ is a d-system, notice first that $\Omega \in \mathcal{D}$. Also, for $A_1, A_2 \in \mathcal{D}$ such that $A_1 \subseteq A_2$, we can readily see that $A_2 \setminus A_1 \in \mathcal{D}$. Now consider $(A_n : n \in \mathbb{N})$, an increasing sequence in $\mathcal{D}$ and denote by $a_\phi^{A_n}(Z)$ the corresponding function such that

(3.3.2) holds. Then $A_n \uparrow \cup_n A_n$ and $\mathbb{1}_{A_n} \uparrow \mathbb{1}_{\cup_n A_n}$ pointwise. Thus, by conditional monotone convergence

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{\cup_n A_n} \mid Y, Z] &= \lim_{n\to\infty} \mathbb{E}_\sigma[\mathbb{1}_{A_n} \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= \lim_{n\to\infty} a_\phi^{A_n}(Z) \quad \text{a.s. } [\mathbb{P}_\sigma].\end{aligned}$$

We define $a_\phi(Z) := \lim_{n\to\infty} a_\phi^{A_n}(Z)$ and we are done. □

Proposition 3.3.3. Let X, Y, Z, W be random variables and Φ, Θ be decision variables. Then

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi) \text{ and } W \perp\!\!\!\perp (Y, \Theta) \mid (X, Z, \Phi) \Rightarrow (X, W) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi).$$

Proof. Following the same approach as in the proof of Lemma 3.3.2, to prove that $(X, W) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, it's enough to show that for all $\phi \in \Phi(\mathcal{S})$ and all $A_{X,W} = A_X \cap A_W$ where $A_X \in \sigma(X)$ and $A_W \in \sigma(W)$, there exists $w_\phi(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_{X,W}} \mid Y, Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.3}$$

Since $W \perp\!\!\!\perp (Y, \Theta) \mid (X, Z, \Phi)$, for all $\phi \in \Phi(\mathcal{S})$ and all $A_W \in \sigma(W)$, there exists $w_\phi^1(X, Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_W} \mid X, Y, Z] = w_\phi^1(X, Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.4}$$

Also by Lemma 3.3.2,

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi) \Rightarrow (X, Z) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi).$$

Thus, for all $\phi \in \Phi(\mathcal{S})$ and all $h(X, Z)$, there exists $w_\phi^2(Z)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[h(X, Z) \mid Y, Z] = w_\phi^2(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.5}$$

So let $\phi \in \Phi(\mathcal{S})$ and $A_{X,W} = A_X \cap A_W$, where $A_X \in \sigma(X)$ and $A_W \in \sigma(W)$. Then

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{A_X \cap A_W} \mid Y, Z] &= \mathbb{E}_\sigma[\mathbb{E}_\sigma(\mathbb{1}_{A_X \cap A_W} \mid X, Y, Z) \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= \mathbb{E}_\sigma[\mathbb{1}_{A_X} \mathbb{E}_\sigma(\mathbb{1}_{A_W} \mid X, Y, Z) \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \\ &= \mathbb{E}_\sigma[\mathbb{1}_{A_X} w_\phi^1(X, Z) \mid Y, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \text{ by (3.3.4)} \\ &= w_\phi^2(Z) \quad \text{a.s. } [\mathbb{P}_\sigma] \text{ by (3.3.5)}\end{aligned}$$

which proves (3.3.3). □

Proposition 3.3.4. Let X, Y, Z, W be random variables, Φ, Θ be decision variables and consider the following conditions:

(i) For all $\phi \in \Phi(\mathcal{S})$, there exists a dominating regime $\sigma_\phi \in \Phi^{-1}(\phi)$.

(ii) X, Y, Z and W are discrete random variables.

If either (i) or (ii) holds, then

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi), W \preceq X \Rightarrow X \perp\!\!\!\perp (Y, \Theta) \mid (Z, W, \Phi).$$

Proof. Suppose that (i) holds. To prove that $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, W, \Phi)$, by Proposition 3.1.9, it suffices to prove the following two statements:

$$X \perp\!\!\!\perp Y \mid (Z, W, \Phi, \Theta) \tag{3.3.6}$$

and

$$X \perp\!\!\!\perp \Theta \mid (Z, W, \Phi). \tag{3.3.7}$$

To prove (3.3.6), we will use Proposition 3.1.10 and prove equivalently that $Y \perp\!\!\!\perp X \mid (Z, W, \Phi, \Theta)$. Notice first that since $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, by Proposition 3.1.9, it follows that $X \perp\!\!\!\perp Y \mid (Z, \Phi, \Theta)$ and by Proposition 3.1.10, it follows that $Y \perp\!\!\!\perp X \mid (Z, \Phi, \Theta)$. Also, since $W \preceq X$, by Proposition 2.3.15 it follows that $\sigma(W) \subseteq \sigma(X)$. Thus, let $(\phi, \theta) \in (\Phi(\mathcal{S}), \Theta(\mathcal{S}))$, $\sigma = (\Phi, \Theta)^{-1}(\phi, \theta)$ and $A_Y \in \sigma(Y)$. Then

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{A_Y} \mid X, Z, W] &= \mathbb{E}_\sigma[\mathbb{1}_{A_Y} \mid X, Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \text{ since } \sigma(W) \subseteq \sigma(X) \\ &= \mathbb{E}_\sigma[\mathbb{1}_{A_Y} \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma] \text{ since } Y \perp\!\!\!\perp X \mid (Z, \Phi, \Theta),\end{aligned}$$

which proves that $Y \perp\!\!\!\perp X \mid (Z, W, \Phi, \Theta)$.
To prove (3.3.7), let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$. We will show that there exists $w_\phi(Z, W)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Z, W] = w_\phi(Z, W) \quad \text{a.s. } [\mathbb{P}_\sigma], \tag{3.3.8}$$

i.e.,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mathbb{1}_{A_{Z,W}}] = \mathbb{E}_\sigma[w_\phi(Z, W) \mathbb{1}_{A_{Z,W}}] \quad \text{whenever } A_{Z,W} \in \sigma(Z, W). \tag{3.3.9}$$

Let $A_{Z,W} \in \sigma(Z,W)$ and notice that

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}}] = \mathbb{E}_\sigma[\mathbb{E}_\sigma(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z)]. \tag{3.3.10}$$

Since $X \perp\!\!\!\perp (Y,\Theta) \mid (Z,\Phi)$, by Lemma 3.3.2, it follows that $(X,Z) \perp\!\!\!\perp (Y,\Theta) \mid (Z,\Phi)$ and by Proposition 3.1.9 that $(X,Z) \perp\!\!\!\perp \Theta \mid (Z,\Phi)$. Also, since $W \preceq X$, there exists $a_\phi(Z)$ such that

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z) = a_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.11}$$

In particular, for the dominating regime $\sigma_\phi \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z) = a_\phi(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_\phi}]$$

and thus

$$\mathbb{E}_{\sigma_\phi}[\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z,W) \mid Z] = a_\phi(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_\phi}].$$

Since $\mathbb{P}_\sigma \ll \mathbb{P}_{\sigma_\phi}$ for all $\sigma \in \Phi^{-1}(\phi)$, it follows that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_{\sigma_\phi}[\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z,W) \mid Z] = a_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.12}$$

Thus, by (3.3.11) and (3.3.12), we get that

$$\mathbb{E}_\sigma(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z) = \mathbb{E}_{\sigma_\phi}[\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z,W) \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.13}$$

Similarly,

$$\mathbb{E}_{\sigma_\phi}[\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z,W) \mid Z] = \mathbb{E}_\sigma[\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}} \mid Z,W) \mid Z] \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.3.14}$$

Returning back to (3.3.10), it follows that

$$\begin{aligned}
\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Z,W}}] &= \mathbb{E}_\sigma\{\mathbb{E}_{\sigma_\phi}[\mathbb{1}_{A_{Z,W}}\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X} \mid Z,W) \mid Z]\} \quad \text{by (3.3.13)} \\
&= \mathbb{E}_\sigma\{\mathbb{E}_\sigma[\mathbb{1}_{A_{Z,W}}\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X} \mid Z,W) \mid Z]\} \quad \text{by (3.3.14)} \\
&= \mathbb{E}_\sigma[\mathbb{1}_{A_{Z,W}}\mathbb{E}_{\sigma_\phi}(\mathbb{1}_{A_X} \mid Z,W)].
\end{aligned}$$

Now suppose that (ii) holds. To show that $X \perp\!\!\!\perp (Y,\Theta) \mid (Z,W,\Phi)$, we need to show that for all $\phi \in \Phi(\mathcal{S})$ and all $A_X \in \sigma(X)$, there exists $w_\phi(Z,W)$ such that for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y,Z,W] = w_\phi(Z,W) \quad \text{a.s. } [\mathbb{P}_\sigma], \tag{3.3.15}$$

i.e.,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z,W}}] = \mathbb{E}_\sigma[w_\phi(Z,W)\mathbb{1}_{A_{Y,Z,W}}] \quad \text{whenever } A_{Y,Z,W} \in \sigma(Y,Z,W). \quad (3.3.16)$$

Observe that it's enough to show (3.3.16) for $A_{Y,Z,W} \in \sigma(Y,Z,W)$ such that $\mathbb{P}_\sigma(A_{Y,Z,W}) \neq 0$. Also since X, Y, Z and W are discrete we need to show (3.3.16) only for sets of the form $\{X = x\}$ and $\{Y = y, Z = z, W = w\}$.[3] Thus, it's enough to show that for all $\phi \in \Phi(\mathcal{S})$ and $\{X = x\} \in \sigma(X)$, there exists $w_\phi(Z,W)$ such that, for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{\{X=x\}}\mathbb{1}_{\{Y=y,Z=z,W=w\}}] = \mathbb{E}_\sigma[w_\phi(Z,W)\mathbb{1}_{\{Y=y,Z=z,W=w\}}] \quad (3.3.17)$$

whenever $\mathbb{P}_\sigma(Y = y, Z = z, W = w) \neq 0$.
So, let $\phi \in \Phi(\mathcal{S})$ and $\{X = x\} \in \sigma(X)$. Then for $\sigma \in \Phi^{-1}(\phi)$ and $\{Y = y, Z = z, W = w\}$ such that $\mathbb{P}_\sigma(Y = y, Z = z, W = w) \neq 0$,

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{\{X=x\}}\mathbb{1}_{\{Y=y,Z=z,W=w\}}] &= \mathbb{P}_\sigma[X = x, Y = y, Z = z, W = w] \\ &= \mathbb{P}_\sigma[X = x \mid Y = y, Z = z, W = w]\mathbb{P}_\sigma[Y = y, Z = z, W = w] \\ &= \frac{\mathbb{P}_\sigma[X = x, W = w \mid Y = y, Z = z]}{\mathbb{P}_\sigma[W = w \mid Y = y, Z = z]}\mathbb{P}_\sigma[Y = y, Z = z, W = w].\end{aligned} \quad (3.3.18)$$

Since $X \perp\!\!\!\perp (Y,\Theta) \mid (Z,\Phi)$ and $W \preceq X$, there exist $w^1_\phi(Z)$ and $w^2_\phi(Z)$ such that

$$\mathbb{E}_\sigma[\mathbb{1}_{\{X=x,W=w\}} \mid Y, Z] = w^1_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

and

$$\mathbb{E}_\sigma[\mathbb{1}_{\{W=w\}} \mid Y, Z] = w^2_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]$$

where $w^1_\phi(Z) = 0$ unless $w = W(x)$.
Define

$$w_\phi(z) = \begin{cases} \frac{w^1_\phi(z)}{w^2_\phi(z)} & \text{if } w^2_\phi(z) \neq 0, \\ 0 & \text{if } w^2_\phi(z) = 0 \end{cases}$$

and notice that $w^2_\phi(z) \neq 0$ when $\mathbb{P}_\sigma(Y = y, Z = z, W = w) \neq 0$. Also notice that since $w_\phi(Z)$ is $\sigma(Z)$-measurable it is also $\sigma(W,Z)$-measurable. Returning back to

[3] More precisely sets of the form $\{\omega \in \Omega : X(\omega) = x\}$ and $\{\omega \in \Omega : Y(\omega) = y, Z(\omega) = z, W(\omega) = w\}$.

(3.3.18) we get

$$\begin{aligned}\mathbb{E}_\sigma[\mathbb{1}_{\{X=x\}}\mathbb{1}_{\{Y=y,Z=z,W=w\}}] &= w_\phi(z)\mathbb{P}_\sigma[Y=y,Z=z,W=w] \\ &= \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{\{Y=y,Z=z,W=w\}}]\end{aligned}$$

which concludes the proof. □

In Section 3.2.1, we have also shown that $P4''$ holds when the regime space is discrete. Thus the result of Proposition 3.3.4, should also hold if we remove conditions (i) and (ii) and instead impose the condition that the regime space is discrete. An alternative proof of $P4''$ is given below.

Proposition 3.3.5. Let X, Y, Z, W be random variables and Φ, Θ be decision variables on $\mathcal{S}$ where $\mathcal{S}$ is discrete. Then

$$X \perp\!\!\!\perp (Y,\Theta) \mid (Z,\Phi), W \preceq X \Rightarrow X \perp\!\!\!\perp (Y,\Theta) \mid (Z,W,\Phi).$$

Proof. To prove that $X \perp\!\!\!\perp (Y,\Theta) \mid (Z,W,\Phi)$, let $\phi \in \Phi(\mathcal{S})$ and $A_X \in \sigma(X)$. Since $\mathcal{S}$ is discrete, the set $\Phi^{-1}(\phi)$ will be finite ($\Phi^{-1}(\phi) = \{\sigma_{\phi,n} : n = 1,\ldots,k\}$) or countable ($\Phi^{-1}(\phi) = \{\sigma_{\phi,n} : n = 1,2,\ldots\}$). If $\Phi^{-1}(\phi)$ is finite, define $\mathbb{P}_{\sigma^*_\phi} := \frac{1}{k}\sum_{n=1}^{k}\mathbb{P}_{\sigma_{\phi,n}}$ and if $\Phi^{-1}(\phi)$ is countable, define $\mathbb{P}_{\sigma^*_\phi} := \sum_{n=1}^{\infty}\frac{1}{2^n}\mathbb{P}_{\sigma_{\phi,n}}$. In both cases, we can readily see that $\mathbb{P}_{\sigma^*_\phi}$ is a probability measure and that $\mathbb{P}_\sigma \ll \mathbb{P}_{\sigma^*_\phi}$ for all $\sigma \in \Phi^{-1}(\phi)$. Also, since $X \perp\!\!\!\perp (Y,\Theta) \mid (Z,\Phi)$, let $\phi \in \Phi$ and $A_X \in \sigma(X)$. Then, there exists $w_\phi(Z)$ such that, for all $\sigma \in \Phi^{-1}(\phi)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y,Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_\sigma], \tag{3.3.19}$$

i.e.,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}] = \mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}] \quad \text{whenever } A_{Y,Z} \in \sigma(Y,Z).$$

We will show that, for the $\mathbb{P}$-measure $\mathbb{P}_{\sigma^*_\phi}$,

$$\mathbb{E}_{\sigma^*_\phi}[\mathbb{1}_{A_X} \mid Y,Z] = w_\phi(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma^*_\phi}]$$

i.e.

$$\mathbb{E}_{\sigma^*_\phi}[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}] = \mathbb{E}_{\sigma^*_\phi}[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}] \quad \text{whenever } A_Z \in \sigma(Y,Z).$$

Indeed, denoting the respective integral by $\mathbb{E}^*_\phi$ for the finite case we get that

$$\begin{aligned}\mathbb{E}^*_\phi[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}] &= \frac{1}{k}\sum_{\sigma\in\Phi^{-1}(\phi)}\mathbb{E}_\sigma[\mathbb{1}_{A_X}\mathbb{1}_{A_{Y,Z}}]\\ &= \frac{1}{k}\sum_{\sigma\in\Phi^{-1}(\phi)}\mathbb{E}_\sigma[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}]\\ &= \mathbb{E}^*_\phi[w_\phi(Z)\mathbb{1}_{A_{Y,Z}}] \quad \text{whenever } A_{Y,Z}\in\sigma(Y,Z).\end{aligned}$$

Similarly, the above equation holds for the countable case. So, for all $\phi\in\Phi(\mathcal{S})$, we have constructed a new $\mathbb{P}$-measure, indexed by σ^*_ϕ, that satisfies (3.3.19) and dominates $\mathbb{P}_\sigma$, for all $\sigma\in\Phi^{-1}(\phi)$. Essentially, we have constructed the dominating regime required by (i) in Proposition 3.3.4. Thus, the rest follows as in the corresponding proof. □

3.3.1 The case of discrete random variables

In this section we will explore a more general definition of conditional independence when we are concerned with discrete random variables. We will define a statement of the form $(X,K) \perp\!\!\!\perp (Y,\Theta) \mid (Z,\Phi)$, where X,Y,Z are discrete stochastic variables and K,Θ,Φ are decision variables, and show that the ternary operation conforms with the axioms.

Before considering the general definition, we shall firstly discuss what we might mean when we call two decision variables independent upon conditioning on a random variable. Writing down $K \perp\!\!\!\perp \Theta \mid Z$, essentially we require that conditioning on the stochastic variable Z, we get information upon which we render the decision variables K and Θ variation independent. Seeking to define $K \perp\!\!\!\perp \Theta \mid Z$, we need to consider the link that connects the three functions and in particular what information Z can provide to make K and Θ variation independent. Recall that Z is a discrete random variable that has a different distribution in the different regimes $\sigma\in\mathcal{S}$ and K and Θ are functions defined on $\mathcal{S}$. One way to define such a statement is to consider for each value z of Z, the set of regimes that assign to this value a positive probability and render the decision variables independent on this set. We formalise this interpretation in the following definitions.

Definition 3.3.6. Let Z be a discrete random variable and K,Θ be decision variables. For each outcome z of Z we define:

(i) the *restriction of* z *in* $\mathcal{S}$ to be the set

$$\mathcal{S}|_z := \{\sigma \in \mathcal{S} : \mathbb{P}_\sigma(Z = z) > 0\},$$

(ii) the *conditional image of* Θ*, given* $Z = z$, to be the set

$$R(\Theta \mid Z = z) := \{\Theta(\sigma) : \sigma \in \mathcal{S}|_z\},$$

(iii) the *conditional image of* Θ*, given* $(Z = z, K = k)$, to be the set

$$R(\Theta \mid Z = z, K = k) := \{\Theta(\sigma) : \sigma \in \mathcal{S}|_z, K(\sigma) = k\}.$$

We can now give suitable definitions for $\Theta \perp\!\!\!\perp K \mid Z$ and $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$.

Definition 3.3.7. Let Z be a discrete random variable and K, Θ, Φ be decision variables. We say that Θ *is (conditionally) independent of* K *given* (Z, Φ) and write $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$ if for all $z \in Z(\Omega)$ and all $(k, \phi) \in (K, \Phi)(\mathcal{S})$, $R(\Theta \mid Z = z, K = k, \Phi = \phi) = R(\Theta \mid Z = z, \Phi = \phi)$.

Notice that if $\{\sigma \in \mathcal{S}|_z : \Phi(\sigma) = \phi\} = \emptyset$ then $R(\Theta \mid Z = z, \Phi = \phi) = \emptyset = R(\Theta \mid Z = z, K = k, \Phi = \phi)$. Also notice that we can equivalently define $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$ in terms of variation independence and write instead that for all $z \in Z(\Omega)$, $\Theta \perp\!\!\!\perp_v K \mid \Phi\ [\mathcal{S}|_z]$. Then for all $z \in Z(\Omega)$, we can apply the axioms of variation independence to $\Theta \perp\!\!\!\perp_v K \mid \Phi\ [\mathcal{S}|_z]$.

To introduce further definitions where the random variables will appear in the left-most term of a conditional independence statement, we need to ascertain that the decision variables involved form a bijection on $\mathcal{S}$. Recall that we have defined $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ only when Θ and Φ are complementary on $\mathcal{S}$. Similarly, we will define $(Y, \Theta) \perp\!\!\!\perp K \mid (Z, \Phi)$ and consequently $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ only when K, Θ and Φ are *complementary* (on $\mathcal{S}$), *i.e.*, when (K, Θ, Φ) forms a bijection on $\mathcal{S}$.

Definition 3.3.8. Let Y, Z be discrete random variables and K, Θ, Φ be decision variables such that K, Θ and Φ are complementary on $\mathcal{S}$. Then we say that:

(i) (Y, Θ) *is (conditionally) independent of* K *given* (Z, Φ), and write $(Y, \Theta) \perp\!\!\!\perp K \mid (Z, \Phi)$, if $Y \perp\!\!\!\perp K \mid (Z, \Phi, \Theta)$ and $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$.

(ii) K *is (conditionally) independent of* (Y, Θ) *given* (Z, Φ), and write $K \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, if $(Y, \Theta) \perp\!\!\!\perp K \mid (Z, \Phi)$.

Now, we have all the elements we need to define the most general form.

Definition 3.3.9. Let X, Y, Z be discrete random variables and K, Θ, Φ be decision variables such that K, Θ and Φ are complementary on $\mathcal{S}$. We say that (X, K) *is (conditionally) independent of* (Y, Θ) *given* (Z, Φ), and write $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, if $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi, K)$ and $K \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$.

Defining $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ as a ternary relation where each term is a pair of a stochastic and a non-stochastic variable, allows us to introduce the corresponding quasiorder that is consistent with the theory developed separately for stochastic and non-stochastic variables. Let V be the set of functions of the form (Y, Θ), where Y is a random variable defined on Ω and Θ is a decision variable defined on $\mathcal{S}$. For $(Y, \Theta), (W, \Lambda) \in V$, we write $(W, \Lambda) \preceq (Y, \Theta)$ to denote that $W = f(Y)$ for some measurable function f (henceforth denoted by $W \preceq Y$) and $\Lambda = h(\Theta)$ for some function h (henceforth denoted by $\Lambda \preceq \Theta$). [4] As we can only write $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ when K, Θ and Φ are complementary on $\mathcal{S}$, any statement that involves K, Θ and Φ in the same context cannot involve further decision variables. As a consequence, introducing in the same statement an additional decision variable Λ such that $\Lambda \preceq \Theta$, can only mean that $\Lambda = h(\Theta)$ for some bijective function h. This understanding changes the form of the axioms. Consider for example $P3$ that should state that $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $(W, \Lambda) \preceq (Y, \Theta)$ implies $(X, K) \perp\!\!\!\perp (W, \Lambda) \mid (Z, \Phi)$. As the exact form of Λ is unneeded in the definition of $(X, K) \perp\!\!\!\perp (W, \Lambda) \mid (Z, \Phi)$, Λ can be replaced with Θ and $(W, \Lambda) \preceq (Y, \Theta)$ can be replaced with $W \preceq Y$. Thus $P3$ can be rephrased as $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $W \preceq Y$ implies $(X, K) \perp\!\!\!\perp (W, \Theta) \mid (Z, \Phi)$. We will prove this and the corresponding forms of all the axioms in the following theorem.

Theorem 3.3.10. (Axioms of Conditional Independence.) Let X, Y, Z, W be random variables and K, Θ, Φ be decision variables. Then the following properties hold:

$P1$. $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi) \Rightarrow (Y, \Theta) \perp\!\!\!\perp (X, K) \mid (Z, \Phi)$.

$P2$. $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Y, \Theta)$.

$P3$. $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi), W \preceq Y \Rightarrow (X, K) \perp\!\!\!\perp (W, \Theta) \mid (Z, \Phi)$.

$P4$. $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi), W \preceq Y \Rightarrow (X, K) \perp\!\!\!\perp Y \mid (Z, W, \Phi, \Theta)$.

[4] Related to Section 2.2, $\preceq$ is a quasiorder with join $(Y, \Theta) \vee (W, \Lambda) \approx ((Y, W), (\Theta, \Lambda))$.

$$P5.\quad \left.\begin{array}{c}(X,K)\ \perp\!\!\!\perp\ (Y,\Theta)\mid(Z,\Phi)\\ \text{and}\\ (X,K)\ \perp\!\!\!\perp\ (Y,W,\Theta)\mid(Z,\Phi)\end{array}\right\}\ \Rightarrow\quad (X,K)\ \perp\!\!\!\perp\ W\mid(Y,Z,\Theta,\Phi).$$

Proof. $P1$). By Definition 3.3.9 and Proposition 3.1.9, we need to show that

$$Y\ \perp\!\!\!\perp\ X\mid(Z,\Phi,\Theta,K) \tag{3.3.20}$$
$$Y\ \perp\!\!\!\perp\ K\mid(Z,\Phi,\Theta) \tag{3.3.21}$$
$$X\ \perp\!\!\!\perp\ \Theta\mid(Z,\Phi,K) \tag{3.3.22}$$
$$K\ \perp\!\!\!\perp\ \Theta\mid(Z,\Phi). \tag{3.3.23}$$

Since $(X,K)\ \perp\!\!\!\perp\ (Y,\Theta)\mid(Z,\Phi)$, we have that

$$X\ \perp\!\!\!\perp\ Y\mid(Z,\Phi,K,\Theta) \tag{3.3.24}$$
$$X\ \perp\!\!\!\perp\ \Theta\mid(Z,\Phi,K) \tag{3.3.25}$$
$$Y\ \perp\!\!\!\perp\ K\mid(Z,\Phi,\Theta) \tag{3.3.26}$$
$$\Theta\ \perp\!\!\!\perp\ K\mid(Z,\Phi). \tag{3.3.27}$$

It follows that (3.3.21) and (3.3.22) hold automatically. Also applying $P1'$ to (3.3.24) we deduce (3.3.20). Rephrasing (3.3.27) in terms of variation independence, we have that, for all $z\in Z(\Omega)$, $\Theta\ \perp\!\!\!\perp_v\ K\mid\Phi\ [\mathcal{S}|_z]$. Thus applying $P1^v$ to (3.3.27) we deduce that, for all $z\in Z(\Omega)$, $K\ \perp\!\!\!\perp_v\ \Theta\mid\Phi\ [\mathcal{S}|_z]$, *i.e.* (3.3.23).

$P2$). By Definition 3.3.9, we need to show that

$$X\ \perp\!\!\!\perp\ (Y,\Theta)\mid(Y,\Theta,K) \tag{3.3.28}$$
$$Y\ \perp\!\!\!\perp\ K\mid(Y,\Theta) \tag{3.3.29}$$
$$\Theta\ \perp\!\!\!\perp\ K\mid(Y,\Theta). \tag{3.3.30}$$

By $P2'$ we have that $X\ \perp\!\!\!\perp\ (Y,\Theta,K)\mid(Y,\Theta,K)$ which is identical to (3.3.28). To show (3.3.29), let $\theta\in\Theta(\mathcal{S})$ and $A_Y\in\sigma(Y)$. We seek to find $w_\theta(Y)$ such that, for all $\sigma\in\Theta^{-1}(\theta)$,

$$\mathbb{E}_\sigma[\mathbb{1}_{A_Y}\mid Y]=w_\theta(Y)\quad\text{a.s. }[\mathbb{P}_\sigma].$$

But notice that

$$\mathbb{E}_\sigma[\mathbb{1}_{A_Y}\mid Y]=\mathbb{1}_{A_Y}\quad\text{a.s. }[\mathbb{P}_\sigma].$$

To show (3.3.30), let $y \in Y(\Omega)$. By $P2^v$, we have that

$$K \perp\!\!\!\perp_v \Theta \mid \Theta \quad [\mathcal{S}|_y]. \tag{3.3.31}$$

Applying $P1^v$ to (3.3.31) we deduce that $\Theta \perp\!\!\!\perp_v K \mid \Theta\ [\mathcal{S}|_y]$, *i.e.* (3.3.30).

$P3$). By Definition 3.3.9, we need to show that

$$X \perp\!\!\!\perp (W, \Theta) \mid (Z, \Phi, K) \tag{3.3.32}$$
$$W \perp\!\!\!\perp K \mid (Z, \Phi, \Theta) \tag{3.3.33}$$
$$\Theta \perp\!\!\!\perp K \mid (Z, \Phi). \tag{3.3.34}$$

Since $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, we have that

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi, K) \tag{3.3.35}$$
$$Y \perp\!\!\!\perp K \mid (Z, \Phi, \Theta) \tag{3.3.36}$$
$$\Theta \perp\!\!\!\perp K \mid (Z, \Phi). \tag{3.3.37}$$

Since $W \preceq Y$, applying $P3'$ to (3.3.35), we deduce (3.3.32). Also, applying $P3''$ to (3.3.36) we deduce (3.3.33).

$P4$). By Definition 3.3.9, we need to show that

$$X \perp\!\!\!\perp Y \mid (Z, W, \Phi, \Theta, K) \tag{3.3.38}$$
$$Y \perp\!\!\!\perp K \mid (Z, W, \Phi, \Theta). \tag{3.3.39}$$

Since $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$, we have that

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi, K) \tag{3.3.40}$$
$$Y \perp\!\!\!\perp K \mid (Z, \Phi, \Theta) \tag{3.3.41}$$
$$\Theta \perp\!\!\!\perp K \mid (Z, \Phi). \tag{3.3.42}$$

Since $W \preceq Y$, applying $P4'$ to (3.3.40) we deduce that $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, W, \Phi, K)$ which implies (3.3.38). Also, since all the random variables are discrete, by (3.3.41) and Proposition 3.3.4 we deduce (3.3.39).

$P5$). By Definition 3.3.9, we need to show that

$$X \perp\!\!\!\perp (Y, W, \Theta) \mid (Z, \Phi, K) \tag{3.3.43}$$

$$(Y, W) \perp\!\!\!\perp K \mid (Z, \Phi, \Theta) \tag{3.3.44}$$
$$\Theta \perp\!\!\!\perp K \mid (Z, \Phi). \tag{3.3.45}$$

Since $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ and $(X, K) \perp\!\!\!\perp W \mid (Y, Z, \Theta, \Phi)$, we have that

$$X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi, K) \tag{3.3.46}$$
$$Y \perp\!\!\!\perp K \mid (Z, \Phi, \Theta) \tag{3.3.47}$$
$$\Theta \perp\!\!\!\perp K \mid (Z, \Phi) \tag{3.3.48}$$
$$X \perp\!\!\!\perp W \mid (Y, Z, \Theta, \Phi, K) \tag{3.3.49}$$
$$W \perp\!\!\!\perp K \mid (Y, Z, \Theta, \Phi). \tag{3.3.50}$$

It follows that (3.3.45) holds automatically. Also applying $P5'$ to (3.3.46) and (3.3.49) we deduce (3.3.43) and applying $P5''$ to (3.3.47) and (3.3.50) we deduce (3.3.44). □

3.3.2 Further extensions

Following the approach we took in the previous section to define the statement $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ for X, Y, Z discrete random variables and K, Θ, Φ decision variables, we will now explore possible generalisations that do not confine to discrete random variables.

The pivotal step that allowed us to extend the definition of extended conditional independence from $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ to $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ was the interpretation of a statement of the form $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$ as a variation independence statement. We defined $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$, for a discrete random variable Z, to mean that for all $z \in Z(\Omega)$, $\Theta \perp\!\!\!\perp_v K \mid \Phi\ [\mathcal{S}|_z]$. Adopting the same definition for continuous random variables would only induce a vacuous statement as for a continuous random variable Z, for all $z \in Z(\Omega)$, $\mathcal{S}|_z := \{\sigma \in \mathcal{S} : \mathbb{P}_\sigma(Z = z) > 0\} = \emptyset$. Seeking to find a more general definition for the restriction of z in $\mathcal{S}$ that will incorporate the definition we already gave for discrete random variables, we will need to introduce the concept of the *support* of a measure. This concept can be defined on measurable spaces that are also *separable metric spaces*. The required definitions follow.

Definition 3.3.11. Let (Δ, d) be a metric space and $A \subseteq \Delta$. We denote by $\bar{A}$ the intersection of all closed subsets of Δ containing A and call $\bar{A}$ the *closure* of A. We say that A is *dense* in Δ if $\bar{A} = \Delta$.

Notice that if $A \neq \emptyset$ then $\bar{A} \neq \emptyset$ as Δ is a closed set containing A. Also $A \subseteq \bar{A}$.

Definition 3.3.12. A metric space (Δ, d) is called *separable* if it contains a countable dense subset.

We can readily see that any subset of a separable metric space is also separable.

Definition 3.3.13. A collection $U_{i\in I}$ of open subsets of a metric space (Δ, d) is called an *open base* if every open set in Δ can be written as a union of $U_{i\in I_0}$ for some $I_0 \subseteq I$.

Theorem 3.3.14. Let (Δ, d) be a metric space. Then the following are equivalent:

(i) (Δ, d) is separable.

(ii) (Δ, d) has a countable open base.

Proof. See Kaplansky (2001) (p.95, Theorem 59.). □

Using the above elements we can prove the following theorem (Parthasarathy, 1967) that essentially formulates the concept of the support of a measure.

Theorem 3.3.15. Let (Δ, d) be a separable metric space equipped with a measure μ. Then there exists a unique closed set C_μ satisfying:

(i) $\mu(C_\mu) = 1$

(ii) If D is any closed set such that $\mu(D) = 1$, then $C_\mu \subseteq D$.

Moreover, C_μ is the set of all points $x \in \Delta$ having the property that $\mu(U) > 0$ for each open set U containing x.

Proof. Let $\mathcal{U} = \{U : U \text{ open}, \mu(U) = 0\}$. Since Δ is separable, there are countably many open sets $U_1, U_2, \ldots$ such that $\bigcup_{n\in\mathbb{N}} U_n = \bigcup\{U : U \in \mathcal{U}\}$. Denote this union by U_μ and set $C_\mu = \Delta \setminus U_\mu$. Since $\mu(U_\mu) = \mu(\bigcup_{n\in\mathbb{N}} U_n) = \sum_{n\in\mathbb{N}} \mu(U_n) = 0$, $\mu(C_\mu) = 1$. Further, if D is any closed set with $\mu(D) = 1$, $\Delta \setminus D \in \mathcal{U}$ and hence $\Delta \setminus D \subseteq U_\mu$, *i.e.*, $C_\mu \subseteq D$. The uniqueness of C_μ is obvious. To prove the last assertion notice that, for any $x \in \Delta \setminus C_\mu$, U_μ is an open set containing x and $\mu(U_\mu) = 0$, whereas if $x \in C_\mu$ and U is an open set containing x, $\mu(U)$ must be positive, as otherwise $U \subseteq U_\mu$ by the definition of U_μ. □

The following definition accrues.

Definition 3.3.16. Let (Δ, d) be a separable metric space equipped with a measure μ. The *support* or *spectrum* of the measure μ is defined to be the unique closed set $C_\mu := \{x \in \Delta : \mu(U) > 0 \text{ for each open set } U \text{ containing } x\}$.

To give an example of this concept, consider μ to be the Lebesgue measure on the real line $\mathbb{R}$. For any arbitrary $x \in \mathbb{R}$, any open set U containing x is such that $(x-\epsilon, x+\epsilon) \subseteq U$ for some $\epsilon > 0$. But $\mu(U) \geq \mu(x-\epsilon, x+\epsilon) = 2\epsilon > 0$. Thus $x \in C_\mu$ and since x was arbitrary, $C_\mu = \mathbb{R}$.

Remark 3.3.17. For a discrete random variable Z that takes values on a separable metric space, we can readily see that $\{z : \mathbb{P}(Z = z) > 0\} = \{z : \mathbb{P}(U) > 0 \text{ for each open set } U \text{ containing } z\}$.

For the following section, we assume that the condition of a separable metric space is met. Also for $\sigma \in \mathcal{S}$, we denote by $C_{\mathbb{P}_\sigma}$ the support of the $\mathbb{P}$-measure $\mathbb{P}_\sigma$. Thus, we can extend the definition of the restriction of z in $\mathcal{S}$ as follows.

Definition 3.3.18. Let Z be a random variable on $(\Omega, \mathcal{A})$. For each outcome z of Z we define the *restriction of z in $\mathcal{S}$* to be the set $\mathcal{S}|_z := \{\sigma \in \mathcal{S} : z \in C_{\mathbb{P}_\sigma}\}$.

The subsequent definitions of $\Theta \perp\!\!\!\perp K \mid Z$ and $\Theta \perp\!\!\!\perp K \mid (Z, \Phi)$ in terms of variation independence and the most general definition $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ remain the same. Thus using the same arguments as in the discrete case we can prove that $(X, K) \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ satisfies the axioms of conditional independence with the exception of axiom $P4$. However, making the additional assumption of the existence of a dominating regime as required in Proposition 3.3.4 we can also deduce axiom $P4$. Examination of the necessity to impose existence of a dominating regime on a separable metric space has not been attempted in this thesis.

3.4 Pairwise Conditional Independence

In this section we will consider a more relaxed notion of ECI that we will term *pairwise extended conditional independence* (PECI).

Definition 3.4.1. Let X, Y and Z be random variables and let Θ and Φ be decision variables. We say that X *is pairwise (conditionally) independent of* (Y, Θ) *given* (Z, Φ), and write $X \perp\!\!\!\perp_{pw} (Y, \Theta) \mid (Z, \Phi)$, if for all $\phi \in \Phi(\mathcal{S})$, all real, bounded and measurable functions h, and all pairs $\{\sigma_1, \sigma_2\} \in \Phi^{-1}(\phi)$, there exists a function

$w_\phi^{\sigma_1,\sigma_2}(Z)$ such that

$$\mathbb{E}_{\sigma_1}[h(X) \mid Y, Z] = w_\phi^{\sigma_1,\sigma_2}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_1}]$$

and

$$\mathbb{E}_{\sigma_2}[h(X) \mid Y, Z] = w_\phi^{\sigma_1,\sigma_2}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_2}].$$

It is readily seen that ECI implies PECI but not the other way round. In Definition 3.4.1, for all $\phi \in \Phi(\mathcal{S})$, we only require a common version for the corresponding conditional expectation for every pair of regimes $\{\sigma_1, \sigma_2\} \in \Phi^{-1}(\phi)$. We do not require that these versions agree on one function that can serve as a version for the corresponding conditional expectation simultaneously in all regimes $\sigma \in \Phi^{-1}(\phi)$. However, such a conclusion can follow if we wish to bring in the additional assumption of the existence of a dominating regime.

Proposition 3.4.2. Let $\mathcal{S}$ be the regime space and Φ be a decision variable. Assume that for all $\phi \in \Phi(\mathcal{S})$ there exists a dominating regime $\sigma_\phi \in \Phi^{-1}(\phi)$. Then $X \perp\!\!\!\perp_{pw} (Y, \Theta) \mid (Z, \Phi)$ implies $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$.

Proof. Since $X \perp\!\!\!\perp_{pw} (Y, \Theta) \mid (Z, \Phi)$, for all $\phi \in \Phi(\mathcal{S})$, all $A_X \in \sigma(X)$ and all pairs $\{\sigma_\phi, \sigma\} \in \Phi^{-1}(\phi)$, there exists a function $w_\phi^{\sigma_\phi,\sigma}(Z)$ such that

$$\mathbb{E}_{\sigma_\phi}[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi^{\sigma_\phi,\sigma}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_\phi}] \tag{3.4.1}$$

and

$$\mathbb{E}_{\sigma}[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi^{\sigma_\phi,\sigma}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma}]. \tag{3.4.2}$$

Consider one of these functions for a fixed $\sigma_o \in \Phi^{-1}(\phi)$ such that

$$\mathbb{E}_{\sigma_\phi}[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi^{\sigma_\phi,\sigma_o}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma_\phi}]. \tag{3.4.3}$$

Since $\mathbb{P}_\sigma \ll \mathbb{P}_{\sigma_\phi}$ for all $\sigma \in \Phi^{-1}(\phi)$, by (3.4.1) and (3.4.3)

$$\mathbb{E}_{\sigma_\phi}[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi^{\sigma_\phi,\sigma}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma}] \tag{3.4.4}$$

and

$$\mathbb{E}_{\sigma_\phi}[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi^{\sigma_\phi,\sigma_o}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma}]. \tag{3.4.5}$$

Thus, by (3.4.4) and (3.4.5)

$$w_\phi^{\sigma_\phi,\sigma}(Z) = w_\phi^{\sigma_\phi,\sigma_o}(Z) \quad \text{a.s. } [\mathbb{P}_{\sigma}] \tag{3.4.6}$$

and by (3.4.2)

$$\mathbb{E}_\sigma[\mathbb{1}_{A_X} \mid Y, Z] = w_\phi^{\sigma_\phi, \sigma_o}(Z) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{3.4.7}$$

So, we have shown that any pairwise version of $\mathbb{E}_{\sigma_\phi}[h(X) \mid Y, Z]$ can serve as a version of $\mathbb{E}_\sigma[h(X) \mid Y, Z]$ in any regime $\sigma \in \Phi^{-1}(\phi)$. □

Replacing $X \perp\!\!\!\perp (Y, \Theta) \mid (Z, \Phi)$ with $X \perp\!\!\!\perp_{pw} (Y, \Theta) \mid (Z, \Phi)$ in Theorem 3.3.1 we can still deduce properties $P1'$ to $P5'$. Essentially we use the same arguments as in the corresponding proof just confined to the comparison of two regimes at a time. Similar arguments can be used to deduce not only the symmetric equivalents $P3''$ and $P5''$ but also $P4''$ that in the case of ECI requires additional assumptions. In the pairwise case, since we are only comparing two regimes at a time, we can use similar arguments to the ones we used in the case of a discrete regime space (Proposition 3.3.5). Thus additional assumptions are not required.

Chapter 4

Conditional independence in the DT framework

4.1 Expressing causal quantities

In this chapter we return to the DT framework where we apply the language and calculus of ECI to express and identify causal quantities. We consider the random variables Y and T that denote the response and treatment variable respectively. We also consider the decision variable Σ that denotes the regime indicator and takes values on the regime space $\mathcal{S}$. We express causal quantities in interventional terms but, acknowledging that we might only be able to obtain observational data, we seek to find conditions that will allow us to transfer information across regimes, and, in particular, allow us to identify interventional terms from the observational regime.

In this context T denotes a binary treatment variable, that takes values $T = 0$ for control treatment and $T = 1$ for active treatment. Y denotes a discrete or continuous variable of interest, for example, a disease indicator, an income counter, *etc.*. We think of T as the potential *cause* variable and Y as the *effect* variable and we aim to compare the distribution of Y under the two interventional regimes. Thus the regime indicator Σ takes values in $\mathcal{S} = \{\emptyset, 0, 1\}$, where $\sigma = \emptyset$ denotes the observational regime, $\sigma = 0$ denotes the interventional regime under control treatment and $\sigma = 1$ denotes the interventional regime under active treatment.

Example 4.1.1 (Average Causal Effect)**.** As discussed in Chapter 1, the simplest quantity we usually seek to calculate is the Average Causal Effect (ACE) defined by

$$ACE := \mathbb{E}_1(Y) - \mathbb{E}_0(Y). \tag{4.1.1}$$

When the data is gathered from a Randomised Control Trial (RCT), where the sample is randomly chosen and the treatment is randomly allocated, we can use the observational regime directly to identify interventional quantities. In this case we are able to impose the condition,

$$Y \perp\!\!\!\perp \Sigma \mid T. \tag{4.1.2}$$

This condition expresses the belief that the distribution of Y is independent of the regime given information on the treatment T. When (4.1.2) holds, by definition, for all functions $h(Y)$ there exists $w(T)$ that doesn't depend on the regime such that, for all $\sigma \in \{\emptyset, 0, 1\}$,

$$\mathbb{E}_\sigma[h(Y) \mid T] = w(T) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

In particular, for $h(Y) = Y$ there exists $w(T)$ such that for all $\sigma \in \{\emptyset, 0, 1\}$,

$$\mathbb{E}_\sigma(Y \mid T) = w(T) \quad \text{a.s. } [\mathbb{P}_\sigma]. \tag{4.1.3}$$

Recall that, in the interventional regimes, for $t = 0, 1$, $\mathbb{P}_t(T = t) = 1$. Thus for $t = 0, 1$,

$$\mathbb{E}_t(Y \mid T) = \mathbb{E}_t(Y \mid T = t) \quad \text{a.s. } [\mathbb{P}_t]. \tag{4.1.4}$$

Assuming that in the observational regime both treatments are allocated, by (4.1.3), we obtain for $t = 0, 1$,

$$\mathbb{E}_\emptyset(Y \mid T = t) = \mathbb{E}_t(Y \mid T = t) = w(t). \tag{4.1.5}$$

Thus

$$\begin{aligned}
ACE &= \mathbb{E}_1(Y) - \mathbb{E}_0(Y) \\
&= \mathbb{E}_1[\mathbb{E}_1(Y \mid T)] - \mathbb{E}_0[\mathbb{E}_0(Y \mid T)] \\
&= \mathbb{E}_1[\mathbb{E}_1(Y \mid T = 1)] - \mathbb{E}_0[\mathbb{E}_0(Y \mid T = 0)] \quad \text{by (4.1.4)} \\
&= \mathbb{E}_1(Y \mid T = 1) - \mathbb{E}_0(Y \mid T = 0) \\
&= \mathbb{E}_\emptyset(Y \mid T = 1) - \mathbb{E}_\emptyset(Y \mid T = 0) \quad \text{by (4.1.5)}
\end{aligned}$$

and can be estimated from observational data.

In most practical applications, the observational regime will not represent a RCT and thus the condition $Y \perp\!\!\!\perp \Sigma \mid T$ will not be justified. Consequently, com-

paring the response variable of the observed treatment groups directly can not be interpreted as the treatment effect. In the following, we will investigate different conditions that take into account the problem of confounding and still allow us to identify causal quantities.

4.2 Sufficient covariates

In some contexts, we could think of additional unobserved variables that can influence the treatment choice in the observational regime: for example, a variable U that indicates the individual's preference for treatment. This variable can be assumed to belong to a set of pre-treatment variables, meaning that it is fully determined before deciding on the treatment and therefore can be regarded as independent of the regime in which we could observe it. Considering this variable, one could argue that individuals who themselves choose to receive active treatment respond differently to it than when forced to take it. Correspondingly, individuals who choose themselves not to receive the treatment may respond differently than when forced not to take it. But upon information on both the treatment and the preference for treatment, the response can be considered as independent of the regime. We can mathematically express these properties using the language of conditional independence.

Definition 4.2.1. For a random variable U, consider the following ECI properties:

$$U \perp\!\!\!\perp \Sigma \tag{4.2.1}$$

$$Y \perp\!\!\!\perp \Sigma \mid (U, T) \tag{4.2.2}$$

$$\text{For } t = 0, 1 \quad \mathbb{P}_\emptyset(T = t \mid U) > 0 \quad \text{a.s. } [\mathbb{P}_\emptyset] \tag{4.2.3}$$

We call U a *covariate* (with respect to treatment T) if it satisfies (4.2.1), a *sufficient covariate* (with respect to treatment T) if it satisfies (4.2.1) and (4.2.2) and a *strongly sufficient covariate* (with respect to treatment T) if it satisfies (4.2.1), (4.2.2) and (4.2.3).

Property (4.2.1) states that U is independent of the regime. This condition is satisfied when we consider U to represent attributes of the individual that are decided prior to the treatment: for example, treatment preference, genotype of the individual, *etc.*. Property (4.2.2) states that Y is independent of the regime given (U, T).

Thus upon information on U and T, the regime doesn't provide additional information for making probabilistic inference on Y. Property (4.2.2) has been described as "strongly ignorable treatment assignment, given U" (Rosenbaum and Rubin, 1983). Considering a specific sufficient covariate will prove useful for identifying causal quantities, but, depending on the problem, there might be several or no variables with these properties. Property (4.2.3) states that, given any observable value of U, we should be able to observe both treatments.

Graphical models in the form of Directed Acyclic Graphs (DAGs) (Pearl, 1997; Cowell et al., 2007) or Influence Diagrams (IDs) (Howard and Matheson, 1984; Shachter, 1986; Lauritzen et al., 1990; Dawid, 2002) can sometimes be used to represent collections of conditional independence properties. We can then use graphical techniques (in particular, the *d-separation*, or the equivalent *moralization*, criterion) to derive, in a visual and transparent way, further conditional independence properties that are implied by our assumptions. As we will see in Section 5.6.3, a graphical representation is not always possible and can sometimes be misleading. But even when it is possible, it is never essential: all that can be achieved through the graph-theoretic properties of DAGs or IDs, and more, can be achieved using the calculus of conditional independence. In the following passage, we will briefly describe the semantics of DAGs and IDs and show that properties (4.2.1) and (4.2.2) can be graphically represented by means of the ID of Figure 4.1. For a more detailed account of DAGs and IDs, the reader is referred to the above references.

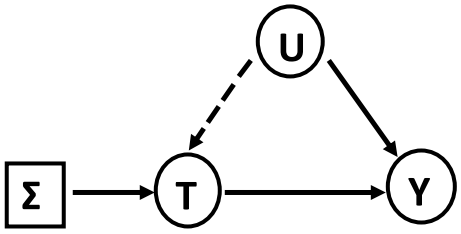

Figure 4.1: Influence Diagram representing (4.2.1) and (4.2.2)

A graph $\mathcal{D} = (V, E)$ is defined by means of a set V of vertices (or nodes) and a set $E \subseteq V \times V$ of edges. We say that a graph is *directed* when the edges have directions (given by arrows) and *acyclic* when the directions of the arrows do not create a cycle from one node back to itself. Each node of a DAG will represent a random variable that is under consideration and will be marked by a circle or oval. In Chapter 3,

we have seen that stochastic conditional independence can be extended to allow for some variables to be non-stochastic (parameters, regimes, *etc.*). Similarly, DAGs can be extended to allow for some nodes to represent non-stochastic variables. What results is an ID, where the nodes which represent stochastic variables are marked by a circle or oval, and the nodes which represent non-stochastic variables are marked by a square or rectangle.

The connections between nodes can be expressed using a special language. For nodes labeled A, B, C, when we see that $A \to B$, we say that A is a parent of B and B is a child of A. We denote the set of parents of node A by $pa(A)$ and the set of children of node A by $ch(A)$. When $A \to \ldots \to B$, we say that A is an ancestor of B and B is a descendant of A, and when $A \to B$ and $C \to B$, we say that A and C are co-parents of B. For example, in the ID presented in Figure 4.1, Y is a child of T, T is a parent of Y, σ is an ancestor of Y, Y is a descendant of σ and σ and U are co-parents of T. When a node is not a descendant of its own then no cycle occurs.

In general, for the ordered sequence of variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ with joint distribution P, we can construct a DAG representing P in the following way. Let $V = \{v : 1, \ldots, N\}$ denote the set of nodes corresponding to the variables $X_1, \ldots, X_N$. We define $pre(v) = (1, \ldots, v-1)$ (where $pre(1) = \emptyset$) and $X_{pre(v)} = (X_1, \ldots, X_{v-1})$. Then we introduce nodes in order 1 to N, and for each node v we consider the conditional distribution of X_v given $X_{pre(v)}$. Then the set of parents of v can be defined as the subset of $pre(v)$, such that the conditional distribution of X_v given $X_{pre(v)}$ depends in fact only on $X_{pa(v)}$. The defining property of the set $pa(v)$ is expressed in terms of conditional independence as $X_v \perp\!\!\!\perp X_{pre(v)} \mid X_{pa(v)}$. Then we draw a directed arrow from each $w \in pa(v)$ to v. In the resulting DAG, $pa(v)$ is indeed the set of parents of v.

Similarly, a collection of joint distributions (one for each regime) can be represented by an ID following the above steps. Consider the variables of interest in the order (Σ, U, T, Y) and name the nodes correspondingly. Then $pre(\Sigma) = \emptyset$, $pre(U) = \Sigma$, $pre(T) = (\Sigma, U)$ and $pre(Y) = (\Sigma, U, T)$, and taking into account (4.2.1) and (4.2.2), we conclude that $pa(\Sigma) = \emptyset$, $pa(U) = \emptyset$, $pa(T) = (\Sigma, U)$ and $pa(Y) = (U, T)$. Thus we draw directed arrows from U and Σ to T, and from U and T to Y, to obtain the ID shown in Figure 4.1. The dotted arrow from U to T, indicates a link that disappears under an interventional regime: for $t = 0, 1$, when $\Sigma = t$, T will have the 1-point distribution at t, independently of U.

Whenever a given collection of joint distributions is represented by an ID $\mathcal{D}$,

we can use graphical techniques (the d-separation, or the equivalent moralization, criterion), not only to read off the initial conditional independence properties from the graph, but also to derive further conditional independence properties that are implied by our assumptions (see Section 5.6.2). Here we briefly describe the moralization criterion and apply it to the ID of Figure 4.1, to read off properties (4.2.1) and (4.2.2).

For any DAG or ID $\mathcal{D}$, its moral graph denoted by $mo(\mathcal{D})$, is the undirected graph obtained in the following way: first we insert an undirected edge between two parents of a common child which are not already linked (we marry unmarried parents), and then we ignore the directions of the edges. A set S of nodes of $\mathcal{D}$ is termed *ancestral* if, whenever $v \in S$ and $u \to v$, then $u \in S$. Then S must contain all the ancestors of any of its members. For any such set S, we denote the smallest ancestral subgraph of $\mathcal{D}$ containing S by $an_{\mathcal{D}}(S)$, and its moral graph by $man_{\mathcal{D}}(S)$ (when $\mathcal{D}$ is implied, we may omit the specification and just write $man(S)$). For sets A, B, C of nodes of $\mathcal{D}$ we write $A \perp\!\!\!\perp_{\mathcal{D}} B \mid C$, and say C separates A from B (with respect to $\mathcal{D}$) to mean that, in $man(A \cup B \cup C)$, every path joining A to B intersects C. Then it can be shown (Lauritzen et al., 1990; Dawid, 2002) that, whenever a collection of joint distributions is represented by D, we have

$$A \perp\!\!\!\perp_{\mathcal{D}} B \mid C \Rightarrow A \perp\!\!\!\perp B \mid C. \tag{4.2.4}$$

Observing the ID in Figure 4.1 we can easily derive (4.2.1), as in $man(\Sigma, U)$ trivially $U \perp\!\!\!\perp_{\mathcal{D}} \Sigma$. Thus (4.2.1) follows from (4.2.4). To derive (4.2.2), we observe that in $man(\mathcal{D})$, presented in Figure 4.2, every path from Y to Σ intersects (U, T) and thus $Y \perp\!\!\!\perp_{\mathcal{D}} \Sigma \mid (U, T)$. Then (4.2.2) follows from (4.2.4).

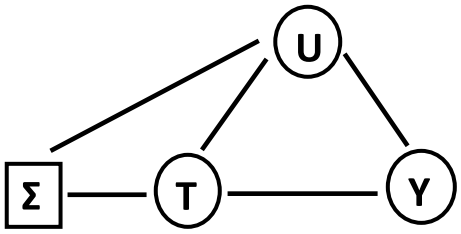

Figure 4.2: Moral graph of the Influence Diagram of Figure 4.1

4.3 Average Causal Effect

Exploring the properties of a strongly sufficient covariate, in this section, we seek to identify the Average Causal Effect.

Remark 4.3.1. Notice that for $t = 0, 1$,

$$\begin{aligned}\mathbb{E}_t(Y \mid U) &= \mathbb{E}_t[\mathbb{E}_t(Y \mid U, T) \mid U] \quad \text{a.s. } [\mathbb{P}_t] \\ &= \mathbb{E}_t[\mathbb{E}_t(Y \mid U, T = t) \mid U] \quad \text{a.s. } [\mathbb{P}_t] \quad \text{since } \mathbb{P}_t(T = t) = 1 \\ &= \mathbb{E}_t(Y \mid U, T = t) \quad \text{a.s. } [\mathbb{P}_t].\end{aligned}$$

Thus in the interventional regime $\sigma = t$, any version of $\mathbb{E}_t(Y \mid U)$ can serve as a version of $\mathbb{E}_t(Y \mid U, T = t)$ and, since $\mathbb{P}_t(T = t) = 1$, as a version of $\mathbb{E}_t(Y \mid U, T)$.

Lemma 4.3.2. Let U be a covariate (with respect to treatment T). Then (4.2.3) holds if and only if $\mathbb{P}_\emptyset$ dominates $\mathbb{P}_t$ on $\sigma(U, T)$.

Proof. $\Rightarrow$) Let $A \in \sigma(U, T)$ such that $\mathbb{P}_\emptyset(A) = 0$. Since T takes values 0 or 1 and $U : (\Omega, \sigma(U)) \to (F_U, \mathcal{F}_U)$, $A = (U, T)^{-1}((C^0 \times \{0\}) \cup (C^1 \times \{1\}))$ for some $C^0, C^1 \in \mathcal{F}_U$. Notice that for $t = 0, 1$, $U^{-1}(C^t) \in \sigma(U)$. For simplicity of notation we will just write $A = (C^0 \times \{0\}) \cup (C^1 \times \{1\})$. Then for $t = 0, 1$, $\mathbb{P}_\emptyset(C^t \times \{t\}) = 0$. It follows that

$$\begin{aligned}\mathbb{E}_\emptyset[\mathbb{1}_{C^t}\mathbb{P}_\emptyset(T = t \mid U)] &= \mathbb{E}_\emptyset[\mathbb{1}_{C^t}\mathbb{E}_\emptyset(\mathbb{1}_{\{T=t\}} \mid U)] \\ &= \mathbb{E}_\emptyset[\mathbb{E}_\emptyset(\mathbb{1}_{C^t}\mathbb{1}_{\{T=t\}} \mid U)] \\ &= \mathbb{E}_\emptyset(\mathbb{1}_{C^t}\mathbb{1}_{\{T=t\}}) \\ &= \mathbb{P}_\emptyset(C^t \times \{t\}) \\ &= 0.\end{aligned}$$

By (4.2.3), it follows that $\mathbb{P}_\emptyset(C^t) = 0$ and by (4.2.1), it follows that $\mathbb{P}_0(C^t) = 0$ and $\mathbb{P}_1(C^t) = 0$. Consequently, for t=0,1, $\mathbb{P}_t(A) = \mathbb{P}_t(\{(C^0 \times \{0\}) \cup (C^1 \times \{1\})\}) = 0$.
$\Leftarrow$) Let $t = 0, 1$ and consider

$$C = \{u : \mathbb{P}(T = t \mid U = u) = 0\}$$

We aim to show that $\mathbb{P}_\emptyset(C) = 0$. We have that

$$\mathbb{E}_\emptyset[\mathbb{1}_C\mathbb{1}_{T=t}] = \mathbb{E}_\emptyset\{\mathbb{E}_\emptyset[\mathbb{1}_C\mathbb{1}_{T=t} \mid U]\}$$

$$
\begin{aligned}
&= \mathbb{E}_\emptyset\{\mathbb{1}_C \mathbb{E}_\emptyset[\mathbb{1}_{T=t} \mid U]\} \\
&= \mathbb{E}_\emptyset\{\mathbb{1}_C \mathbb{P}_\emptyset[T = t \mid U]\} \\
&= 0
\end{aligned}
$$

This implies that $\mathbb{P}_\emptyset(C \cap \{T = t\}) = 0$, which by hypothesis implies that $\mathbb{P}_t(C \cap \{T = t\}) = 0$. Since $\mathbb{P}_t(T = t) = 1$, it follows that $\mathbb{P}_t(C) = 0$ which by (4.2.1) implies that $\mathbb{P}_\emptyset(C) = 0$. □

Theorem 4.3.3. Let Y, T, U be random variables and Σ be the regime indicator. Suppose that U is a strongly sufficient covariate. Then, for any versions of the conditional expectations, for $t = 0, 1$,

$$\mathbb{E}_t(Y \mid U) = \mathbb{E}_\emptyset(Y \mid U, T = t) \quad \text{a.s. in any regime.} \tag{4.3.1}$$

Proof. Since U is a strongly sufficient covariate, by Lemma 4.3.2 and Theorem 3.1.7, for all $\sigma \in \{\emptyset, 0, 1\}$,

$$\mathbb{E}_\sigma(Y \mid U, T) = \mathbb{E}_\emptyset(Y \mid U, T) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

In particular, for $t = 0, 1$,

$$
\begin{aligned}
\mathbb{E}_t(Y \mid U, T) &= \mathbb{E}_\emptyset(Y \mid U, T) \quad \text{a.s. } [\mathbb{P}_t] \\
&= \mathbb{E}_\emptyset(Y \mid U, T = t) \quad \text{a.s. } [\mathbb{P}_t] \quad \text{since } \mathbb{P}_t(T = t) = 1.
\end{aligned}
$$

By Remark 4.3.1, it follows that

$$\mathbb{E}_t(Y \mid U) = \mathbb{E}_\emptyset(Y \mid U, T = t) \quad \text{a.s. } [\mathbb{P}_t]$$

and by (4.2.1), it follows that

$$\mathbb{E}_t(Y \mid U) = \mathbb{E}_\emptyset(Y \mid U, T = t) \quad \text{a.s. in any regime.}$$

□

Theorem 4.3.4. Let Y, T, U be random variables and Σ be the regime indicator. Suppose that U is a strongly sufficient covariate. Then the ACE can be identified purely from observational data. In particular,

$$ACE = \mathbb{E}_\emptyset(Y \mid T = 1) - \mathbb{E}_\emptyset(Y \mid T = 0).$$

Proof.

$$\begin{aligned}
ACE &:= \mathbb{E}_1(Y) - \mathbb{E}_0(Y) \\
&= \mathbb{E}_1[\mathbb{E}_1(Y \mid U)] - \mathbb{E}_0[\mathbb{E}_0(Y \mid U)] \\
&= \mathbb{E}_\emptyset[\mathbb{E}_1(Y \mid U)] - \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U)] \quad \text{by (4.2.1)} \\
&= \mathbb{E}_\emptyset[\mathbb{E}_\emptyset(Y \mid U, T = 1)] - \mathbb{E}_\emptyset[\mathbb{E}_\emptyset(Y \mid U, T = 0)] \quad \text{by Theorem 4.3.3} \\
&= \mathbb{E}_\emptyset(Y \mid T = 1) - \mathbb{E}_\emptyset(Y \mid T = 0).
\end{aligned}$$

□

4.4 Effect of Treatment on the Treated

In some cases, we might want to confine our attention to the individuals that in fact receive the active treatment and calculate the average of the ACE for a certain subgroup of the population in the observational regime relative to some covariate U. In this case we want to identify what is called the *Effect of Treatment on the Treated* (ETT). While more on the ETT can be found in Geneletti and Dawid (2011), here we will confine to giving the definition and showing that we can identify the ETT only from observational data and interventional data under control treatment.

Definition 4.4.1. The *effect of treatment on the treated* (ETT) for a specified sufficient covariate U is defined by

$$ETT_U := \mathbb{E}_\emptyset\{[\mathbb{E}_1(Y \mid U) - \mathbb{E}_0(Y \mid U)] \mid T = 1\}. \tag{4.4.1}$$

The definition of the ETT involves a specified sufficient covariate U. However, when such a covariate exists it does not need to be unique. Nonetheless, in Theorem 4.4.4 we will see that the ETT doesn't depend on the choice of U but only on the distribution of Y in regimes $\sigma = \emptyset$ and $\sigma = 0$ and the distribution of T in regime $\sigma = \emptyset$.

Lemma 4.4.2. Let $\sigma \in \{\emptyset, 0, 1\}$ and suppose that for $t = 0, 1$, $\mathbb{P}_\sigma(T = t) > 0$. Then $\mathbb{P}_\sigma(\cdot \mid T = t) \ll \mathbb{P}_\sigma$.

Theorem 4.4.3. Let Y, T, U be random variables and let Σ be the regime indicator. Suppose that U is a sufficient covariate. Then, for $t = 0, 1$, when $\mathbb{P}_\emptyset(T = t) > 0$, for

any versions of the conditional expectations,

$$\mathbb{E}_t(Y \mid U) = \mathbb{E}_\emptyset(Y \mid U, T = t) \quad \text{a.s. } [\mathbb{P}_\emptyset(\cdot \mid T = t)].$$

Proof. By (4.2.2), there exists $w(U, T)$ such that for all $\sigma \in \{\emptyset, 0, 1\}$,

$$\mathbb{E}_\sigma(Y \mid U, T) = w(U, T) \quad \text{a.s. } [\mathbb{P}_\sigma].$$

In particular,

$$\mathbb{E}_\emptyset(Y \mid U, T) = w(U, T) \quad \text{a.s. } [\mathbb{P}_\emptyset]$$

and thus,

$$\mathbb{E}_\emptyset(Y \mid U, T = t) = w(U, t) \quad \text{a.s. } [\mathbb{P}_\emptyset]. \tag{4.4.2}$$

Also,

$$\mathbb{E}_t(Y \mid U, T) = w(U, T) \quad \text{a.s. } [\mathbb{P}_t] \quad (t = 0, 1).$$

By Remark 4.3.1,

$$\mathbb{E}_t(Y \mid U) = w(U, t) \quad \text{a.s. } [\mathbb{P}_t]$$

and by (4.2.1),

$$\mathbb{E}_t(Y \mid U) = w(U, t) \quad \text{a.s. } [\mathbb{P}_\emptyset]. \tag{4.4.3}$$

By (4.4.2) and (4.4.3), it follows that

$$\mathbb{P}_\emptyset(\mathbb{E}_t(Y \mid U) \neq \mathbb{E}_\emptyset(Y \mid U, T = t)) = 0.$$

Since $\mathbb{P}_\emptyset(T = t) > 0$, it follows that

$$\mathbb{P}_\emptyset(\mathbb{E}_t(Y \mid U) \neq \mathbb{E}_\emptyset(Y \mid U, T = t) \mid T = t) = 0,$$

which concludes the proof. □

The following theorem that proves identification of the ETT from the observational regime and interventional regime under control treatment has been first proved by Geneletti and Dawid (2011) for a strongly sufficient covariate. Then Guo (2010) showed that the same result can be obtained only by considering a sufficient covariate.

Theorem 4.4.4. Let Y, T, U be random variables and Σ be the regime indicator.

Suppose that U is a sufficient covariate and that $\mathbb{P}_\emptyset(T=1)>0$. Then

$$ETT_U = \frac{\mathbb{E}_\emptyset(Y) - \mathbb{E}_0(Y)}{\mathbb{P}_\emptyset(T=1)}. \tag{4.4.4}$$

Proof. We will consider separately the cases $\mathbb{P}_\emptyset(T=0)>0$ and $\mathbb{P}_\emptyset(T=0)=0$. First suppose that $\mathbb{P}_\emptyset(T=0)>0$. Then

$$\begin{aligned}\mathbb{E}_0(Y) &= \mathbb{E}_0[\mathbb{E}_0(Y \mid U)] \\ &= \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U)] \quad \text{by (4.2.1)} \\ &= \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T=0]\mathbb{P}_\emptyset(T=0) + \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T=1]\mathbb{P}_\emptyset(T=1)\end{aligned} \tag{4.4.5}$$

By Theorem 4.4.3, for $t=0,1$,

$$\mathbb{E}_t(Y \mid U) = \mathbb{E}_\emptyset(Y \mid U, T=t) \quad \text{a.s. } [\mathbb{P}_\emptyset(\cdot \mid T=t)].$$

Thus,

$$\begin{aligned}\mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T=0] &= \mathbb{E}_\emptyset[\mathbb{E}_\emptyset(Y \mid U, T=0) \mid T=0] \\ &= \mathbb{E}_\emptyset(Y \mid T=0)\end{aligned} \tag{4.4.6}$$

and

$$\begin{aligned}\mathbb{E}_\emptyset[\mathbb{E}_1(Y \mid U) \mid T=1] &= \mathbb{E}_\emptyset[\mathbb{E}_\emptyset(Y \mid U, T=1) \mid T=1] \\ &= \mathbb{E}_\emptyset(Y \mid T=1).\end{aligned} \tag{4.4.7}$$

Combining (4.4.5) and (4.4.6), we obtain

$$\mathbb{E}_0(Y) = \mathbb{E}_\emptyset(Y \mid T=0)\mathbb{P}_\emptyset(T=0) + \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T=1]\mathbb{P}_\emptyset(T=1).$$

Rearranging we get

$$\mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T=1] = \frac{\mathbb{E}_0(Y) - \mathbb{E}_\emptyset(Y \mid T=0)\mathbb{P}_\emptyset(T=0)}{\mathbb{P}_\emptyset(T=1)}. \tag{4.4.8}$$

Therefore

$$ETT_U := \mathbb{E}_\emptyset[\mathbb{E}_1(Y \mid U) - \mathbb{E}_0(Y \mid U) \mid T=1]$$

$$
\begin{aligned}
&= \mathbb{E}_\emptyset[\mathbb{E}_1(Y \mid U) \mid T = 1\} - \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T = 1] \\
&= \mathbb{E}_\emptyset(Y \mid T = 1) - \frac{\mathbb{E}_0(Y) - \mathbb{E}_\emptyset(Y \mid T = 0)\mathbb{P}_\emptyset(T = 0)}{P_\emptyset(T = 1)} \quad \text{by (4.4.7) and (4.4.8)} \\
&= \frac{\mathbb{E}_\emptyset(Y) - \mathbb{E}_0(Y)}{\mathbb{P}_\emptyset(T = 1)}
\end{aligned}
$$

Now suppose that $\mathbb{P}_\emptyset(T = 0) = 0$. Then $\mathbb{P}_\emptyset(T = 1) = 1$ and by Theorem 4.4.3,

$$
\begin{aligned}
\mathbb{E}_\emptyset[\mathbb{E}_1(Y \mid U) \mid T = 1] &= \mathbb{E}_\emptyset\{\mathbb{E}_\emptyset(Y \mid U, T = 1) \mid T = 1\} \\
&= \mathbb{E}_\emptyset(Y \mid T = 1) \\
&= \mathbb{E}_\emptyset(Y) \quad \text{since } \mathbb{P}_\emptyset(T = 1) = 1. \qquad (4.4.9)
\end{aligned}
$$

Also

$$
\begin{aligned}
\mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U) \mid T = 1] &= \mathbb{E}_\emptyset[\mathbb{E}_0(Y \mid U)] \quad \text{since } \mathbb{P}_\emptyset(T = 1) = 1 \\
&= \mathbb{E}_0[\mathbb{E}_0(Y \mid U)] \quad \text{by (4.2.2)} \\
&= \mathbb{E}_0(Y) \qquad (4.4.10)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
ETT_U &:= \mathbb{E}_\emptyset[\mathbb{E}_1(Y \mid U) - \mathbb{E}_0(Y \mid U) \mid T = 1] \\
&= \mathbb{E}_\emptyset(Y) - \mathbb{E}_0(Y) \quad \text{by (4.4.9) and (4.4.10)} \\
&= \frac{\mathbb{E}_\emptyset(Y) - \mathbb{E}_0(Y)}{P_\emptyset(T = 1)} \quad \text{since } \mathbb{P}_\emptyset(T = 1) = 1.
\end{aligned}
$$

□

4.5 Further extensions

Examples of more general cases, where the regime is determined by two complementary functions on the regime space (as defined in Section 3.1), can also be considered. We will describe three such cases here and graphically represent them by means of IDs.

We can extend the setting of sufficient covariates, to consider cases where the regime $\sigma \in \mathcal{S}$, under which we observe the stochastic variables U, T and Y, is determined by the values of two complementary functions, $\Sigma_1 : \mathcal{S} \to \Sigma_1(\mathcal{S})$ and $\Sigma_2 : \mathcal{S} \to \Sigma_2(\mathcal{S})$. To emphasize this we write $\sigma = (\sigma_1, \sigma_2)$, for $\sigma_1 \in \Sigma_1(\mathcal{S})$ and

$\sigma_2 \in \Sigma_2(\mathcal{S})$.

As a simple example, consider an educational experiment in which we can select certain pupils to undergo additional tutoring. The form of tutoring consists of two decisions: adding or not adding one extra hour of tutoring at school and adding or not adding one extra hour of tutoring at home. In this setting, the regimes consist of information on the way the two different forms of tutoring (at school or at home) were assigned to pupils. We can thus consider two decision variables Σ_1 and Σ_2, each at three levels. In particular, we have that:

$$\sigma_1 = \begin{cases} \emptyset, & \text{for observational regime with regards to school tutoring} \\ 0, & \text{for intervention of no extra tutoring at school} \\ 1, & \text{for intervention of one extra hour of tutoring at school} \end{cases}$$

and

$$\sigma_2 = \begin{cases} \emptyset, & \text{for observational regime with regards to home tutoring} \\ 0, & \text{for intervention of no extra tutoring at home} \\ 1, & \text{for intervention of one extra hour of tutoring at home.} \end{cases}$$

Then the treatment variable T is no longer binary as, in this example, treatment consists of more than one level. In particular, we have that:

$$T = \begin{cases} 0, & \text{for no extra tutoring at school and no extra tutoring at home} \\ 1, & \text{for no extra tutoring at school and one extra hour of tutoring at home} \\ 2, & \text{for one extra hour of tutoring at school and no extra tutoring at home} \\ 3, & \text{for one extra hour of tutoring at school and one extra hour of tutoring at home.} \end{cases}$$

The response variable Y, represents some suitable measurement on the progress of the pupils and our aim is to compare the distribution of Y under the different interventional regimes. Similarly to Section 4.2, variable U represents some unobserved confounding (*e.g.* pupils' preference to treatment) which influences the distribution of Y in the different regimes.

The joint distributions of the three stochastic variables U, T and Y, will be different in the different regimes. In this example, under purely observational settings (*i.e.* $\sigma = (\emptyset, \emptyset)$), the distribution of T will be completely determined by Nature. Conversely, under purely interventional settings (*e.g.* $\sigma = (0, 0)$ or $\sigma = (0, 1)$, *etc.*),

the distribution of T will be fully determined by the regime. For example, for $\sigma = (0, 0)$, $\mathbb{P}_{0,0}(T = 0) = 1$, $\mathbb{P}_{0,0}(T = 1) = 0$, $\mathbb{P}_{0,0}(T = 2) = 0$ and $\mathbb{P}_{0,0}(T = 3) = 0$. But when part of the decision is due to Nature and part of the decision is due to intervention (*e.g.* $\sigma = (\emptyset, 0)$ or $\sigma = (1, \emptyset)$, *etc.*), the distribution of T is partly determined by the regime. For example, $\mathbb{P}_{\emptyset,0}(T = 1) = 0$, $\mathbb{P}_{\emptyset,0}(T = 3) = 0$ and $\mathbb{P}_{\emptyset,0}(T = 0) + \mathbb{P}_{\emptyset,0}(T = 2) = 1$.

Thinking of U as the individual's preference for treatment, we can express analogous statements to (4.2.1) and (4.2.2) of Section 4.2. Further, we can express the assumption that the two decision variables are (variation) independent as the form of school tutoring can be assigned independently of the form of home tutoring. More specifically, we can consider properties:

$$\Sigma_1 \perp\!\!\!\perp \Sigma_2 \tag{4.5.1}$$

$$U \perp\!\!\!\perp (\Sigma_1, \Sigma_2) \tag{4.5.2}$$

$$Y \perp\!\!\!\perp (\Sigma_1, \Sigma_2) \mid (U, T) \tag{4.5.3}$$

and graphically represent them by means of the ID of Figure 4.3.

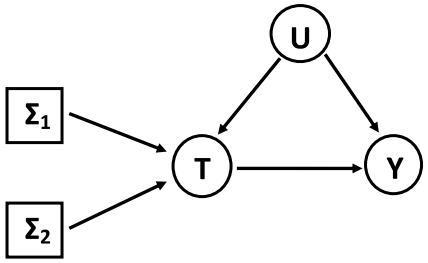

Figure 4.3: Example 1

A different formulation of the above example, would be to represent the treatment variable with two binary variables T_1 and T_2, instead of one variable T, at four different levels. Then T_1 will correspond to extra tutoring at school and T_2 will correspond to extra tutoring at home. In particular, we have that:

$$T_1 = \begin{cases} 0, & \text{for no extra tutoring at school} \\ 1, & \text{for one hour of extra tutoring at school} \end{cases}$$

and

$$T_2 = \begin{cases} 0, & \text{for no extra tutoring at home} \\ 1, & \text{for one hour of extra tutoring at home.} \end{cases}$$

Under this formulation, the corresponding properties (4.5.1), (4.5.2), and (4.5.3), become:

$$\Sigma_1 \perp\!\!\!\perp \Sigma_2 \tag{4.5.4}$$

$$U \perp\!\!\!\perp (\Sigma_1, \Sigma_2) \tag{4.5.5}$$

$$Y \perp\!\!\!\perp (\Sigma_1, \Sigma_2) \mid (U, T_1, T_2). \tag{4.5.6}$$

In addition to the above properties, we also consider:

$$T_1 \perp\!\!\!\perp \Sigma_2 \mid (\Sigma_1, U) \tag{4.5.7}$$

$$T_2 \perp\!\!\!\perp (\Sigma_1, T_1) \mid (\Sigma_2, U). \tag{4.5.8}$$

Property (4.5.7) reflects the understanding that upon knowledge on the outcome of decision variable Σ_1 and the individual's preference for treatment U, the outcome of decision variable Σ_2 adds no further knowledge regarding the distribution of T_1. Similarly, property (4.5.8) reflects the understanding that upon knowledge on the outcome of decision variable Σ_2 and the individual's preference for treatment U, the outcome of decision variable Σ_1 and corresponding treatment T_1, add no further knowledge regarding the distribution of T_2. We can graphically represent (4.5.4), (4.5.5), (4.5.6), (4.5.7) and (4.5.8) by means of the ID of Figure 4.4.

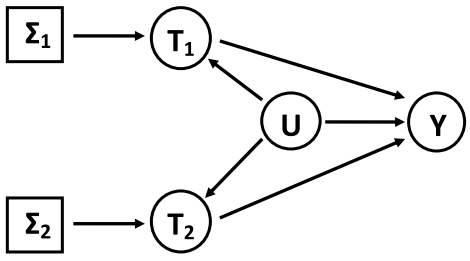

Figure 4.4: Example 2

A more complex example where stochastic variables U, T_1, T_2, T_3 and Y are considered together with three complementary functions, $\Sigma_1 : \mathcal{S} \to \Sigma_1(\mathcal{S})$, $\Sigma_2 : \mathcal{S} \to$

$\Sigma_2(\mathcal{S})$ and $\Sigma_3 : \mathcal{S} \to \Sigma_3(\mathcal{S})$, which together determine the regime, is considered in Figure 4.5.

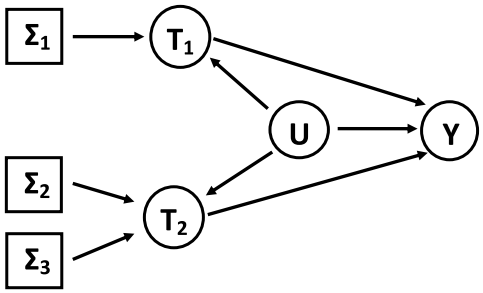

Figure 4.5: Example 3

Chapter 5

Dynamic treatment strategies

5.1 A Sequential Decision Problem

In this chapter we will use the language and calculus of extended conditional independence, as described in Chapter 3, to express and explore more complex problems. We remain in the DT framework where we lose the formality of the underlying spaces to facilitate more intuitive understanding of the problem under study.

More specifically, the problem we will be concerned with, is that of controlling some variable of interest through a sequence of consecutive actions. An example in a medical context is maintaining a critical variable, such as blood pressure, within an appropriate risk-free range. To achieve such control, the doctor will administer treatments over a number of stages, taking into account, at each stage, a record of the patient's history, which provides him with information on the level of the critical variable, and possibly other related measurements, as well as the patient's reactions to the treatments applied in preceding stages. Consider, for instance, practices followed after events such as stroke, pulmonary embolism or deep vein thrombosis (Rosthøj et al., 2006; Sterne et al., 2009). The aim of such practices is to keep the patient's prothrombin time (international normalized ratio, INR) within a recommended range. Such efforts are not confined to a single decision and instant allocation of treatment, marking the end of medical care. Rather, they are effected over a period of time, with actions being decided and applied at various stages within this period, based on information available at each stage. So the patient's INR and related factors will be recorded throughout this period, along with previous actions taken, and at each stage all the information so far recorded, as well, possibly, as other, unrecorded information, will form the basis upon which the doctor will decide

on allocation of the subsequent treatment.

A well-specified algorithm that takes as input the recorded history of a patient at each stage and gives as output the choice of the next treatment to be allocated constitutes a *dynamic decision strategy*. Such a strategy gives guidance to the doctor on how to take into account the earlier history of the patient, including reactions to previous treatments, in allocating the next treatment. There can be an enormous number of such strategies, having differing impacts on the variable of interest. We should like to have criteria to evaluate these strategies, and so allow us to choose the one that is optimal for our problem (Murphy, 2003).

5.2 Notation and terminology

We consider two sets of variables: $\mathcal{L}$, a set of *observable* variables, and $\mathcal{A}$, a set of *action* variables. We term the variables in $\mathcal{L} \cup \mathcal{A}$ *domain variables*. An alternating ordered sequence $\mathcal{I} := (L_1, A_1, \ldots, L_n, A_n, L_{n+1} \equiv Y)$ with $L_i \subseteq \mathcal{L}$ and $A_i \in \mathcal{A}$ defines an *information base*, the interpretation being that the specified variables are observed in this time order. We will adopt notation conventions such as (L_1, L_2) for $L_1 \cup L_2$, $\overline{L}_i$ for $(L_1, \ldots, L_i)$, *etc.*.

The observable variables $\mathcal{L}$ represent initial or intermediate symptoms, reactions, personal information, *etc.*, observable between consecutive treatments, over which we have no direct control; they are perceived as generated and revealed by Nature. The action variables $\mathcal{A}$ represent the treatments, which we could either control by external intervention, or else leave to Nature to determine. Thus at each stage i we will have a realization of the random variable (or set of random variables) $L_i \subseteq \mathcal{L}$, followed by a value for the variable $A_i \in \mathcal{A}$. After the realization of the final $A_n \in \mathcal{A}$, we will observe the outcome variable $L_{n+1} \in \mathcal{L}$, which we also denote by Y.

For any stage i, a configuration $h_i := (l_1, a_1, \ldots, a_{i-1}, l_i)$ of the variables $(L_1, A_1, \ldots, A_{i-1}, L_i)$ constitutes a *partial history*. A clearly described way of specifying, for each action A_i, its value a_i as a function of the partial history h_i to date defines a *strategy*: the values $(\overline{l}_i, \overline{a}_{i-1})$ of the earlier domain variables $(\overline{L}_i, \overline{A}_{i-1})$ can thus be taken into account in determining the current and subsequent actions.

In a *static*, or *atomic*, strategy, the sequence of actions is predetermined, entirely unaffected by the information provided by the L_i's. In a *non-randomized dynamic strategy* we specify, for each stage i and each partial history h_i, a fixed value a_i of A_i, that is then to be applied. We can also consider *randomized strategies*, where, for each stage i and associated partial history h_i we specify a probability

distribution for A_i, so allowing randomization of the decision for the next action. In this chapter we consider general randomized strategies, since we can regard static and non-randomized strategies as special cases of these. Then all the L_i's and A_i's have the formal status of random variables.

Under the decision theoretic approach, we proceed by making suitable assumptions relating the probabilistic behaviours of stochastic variables across a variety of different regimes. We denote the set of all regimes under consideration by $\mathcal{S}$ and the regime indicator by σ. Thus σ specifies which (known or unknown) joint distribution is operating over the domain variables $\mathcal{L} \cup \mathcal{A}$. Information about the operating regime comes through decision variables defined on $\mathcal{S}$. Recall that Σ is the identity function. Any probabilistic statement about the domain variables must, explicitly or implicitly, be conditional on some specified value $\sigma \in \mathcal{S}$.

We focus here on the case that we want to make inference about one or more interventional regimes on the basis of data generated under an observational regime. So we take $\mathcal{S} = \{\emptyset\} \cup \mathcal{S}^*$, where $\emptyset$ is the observational regime under which data have been gathered, and $\mathcal{S}^*$ is the collection of contemplated interventional strategies with respect to a given information base $(L_1, A_1, \ldots, L_N, A_N, Y)$. We will follow closely the approach taken by Dawid and Didelez (2010) and study rigorously conditions that allow identification of a *control strategy*, a type of strategy that will be defined later. To adjust to the notation used in Dawid and Didelez (2010), we write $s \in \mathcal{S}$ (instead of $\sigma \in \mathcal{S}$) and *e.g.* $\mathbb{E}(L_i \mid \overline{A}_{i-1}, \overline{L}_{i-1}\,; s)$ to denote any version of the conditional expectation $\mathbb{E}(L_i \mid \overline{A}_{i-1}, \overline{L}_{i-1})$ under the joint distribution $\mathbb{P}_s$ generated by following strategy s (instead of $\mathbb{E}_s(L_i \mid \overline{A}_{i-1}, \overline{L}_{i-1})$).

5.3 Evaluating a strategy

Suppose we want to identify the effect of some strategy s on the outcome variable Y: we then need to be able to assess the overall effect that the action variables have on the distribution of Y. An important application is where we have a loss $L(y)$ associated with each outcome y of Y, and want to compute the expected loss $\mathbb{E}\{L(Y)\}$ under the distribution for Y induced by following strategy s. We shall see in Section 5.5 below that, if we know or can estimate the conditional distribution, under this strategy, of each observable variable L_i $(i = 1, \ldots, n+1)$ given the preceding variables in the information base, then we would be able to compute $\mathbb{E}\{L(Y)\}$. Following this procedure for each contemplated strategy, we could compare the various strategies, and so choose that minimizing expected loss.

In order to evaluate a particular strategy of interest, we need to be able to mimic the experimental settings that would give us the data we need to estimate the probabilistic structure of the domain variables. Thus suppose that we wish to evaluate a specified non-randomized strategy for a certain patient P, and consider obtaining data under two different scenarios.

The first scenario corresponds to precisely the strategy that we wish to evaluate: that is, the doctor knows the prespecified plan defined by the strategy, and at each stage i, taking into account the partial history h_i, he allocates to patient P the treatment that the strategy recommends. The expected loss $\mathbb{E}\{L(Y)\}$ computed under the distribution of Y generated by following this strategy is exactly what we need to evaluate it.

Now consider a second scenario. Patient P does not take part in the experiment described above, but it so happens he has received exactly the same sequence of treatments that would be prescribed by that strategy. However, the doctor did not decide on the treatments using the strategy, but based on a combination of criteria, that might have involved variables beyond the domain variables $\mathcal{L} \cup \mathcal{A}$. For example, the doctor might have taken into account, at each stage, possible allergies or personal preferences for certain treatments of patient P, variables that the strategy did not encompass.

Because these extra variables are not recorded in the data, the analyst does not know them. Superficially, both scenarios appear to be the same, since the variables recorded in each scenario are the same. However, without further assumptions there is no reason to believe that they have arisen from the same distribution.

The regime described in the first scenario above is one of the interventional regimes, where the doctor was intervening in a specified fashion (which we assume known to the analyst), according to a given strategy for allocating treatment. The regime described in the second scenario is the observational regime, where the analyst has just been observing the sequence of domain variables, but does not know just how the doctor has been allocating treatments.

Data actually generated under this interventional regime would provide exactly the information required to evaluate the strategy. However, typically the data available will not have been generated this way—and in any case there are so many possible strategies to consider that it would not be humanly possible to obtain such experimental data for all of them. Instead, the analyst may have observed how patients (and doctors) respond, in a single, purely observational, regime. Direct use of such observational data as if generated by intervention, though tempting, can be

very misleading. For example, suppose the analyst wants to estimate, at each stage i, the conditional distribution of L_i given $(\overline{L}_{i-1}, \overline{A}_{i-1})$ in an interventional regime (which he has not observed), using data from the observational regime (which he has). Since all the variables in this conditional distribution have been recorded in the observational regime, he might instead estimate (as he can) the conditional distribution of L_i given $(\overline{L}_{i-1}, \overline{A}_{i-1})$ in the observational regime, and consider this as a proxy for its interventional counterpart. However, since the doctor may have been taking account of other variables, that the analyst has not recorded and so can not adjust for, this estimate will typically be biased, often seriously so. One of the main aims of this chapter is to consider conditions under which this bias disappears.

For simplicity, we assume that all the domain variables under consideration can be observed for every patient. However, the context in which we observe these variables will determine if and how we can use the information we collect. In order to tackle issues such as the bias, or otherwise, introduced by making computations under a "wrong" regime, we will need to make assumptions relating the probabilistic behaviours under the differing regimes. Armed with such understanding of the way the regimes interconnect, we can then investigate whether, and if so how, we can transfer information from one regime to another.

5.3.1 Conditional Independence

In order to address the problem of making inference from observational data we need to assume (and justify) some relationships between the probabilistic behaviours of the variables in the differing regimes, interventional and observational. These assumptions will typically relate certain conditional distributions across different regimes, thus be expressed in the language of ECI.

We have seen in Chapter 3 that if we make the assumption of countably many regimes we can show that properties $P1^s$–$P5^s$ continue to hold when some of the variables involved are non-stochastic. For the purposes of this problem, we will only ever need to compare two regimes at a time: the observational regime $\emptyset$ and one particular interventional regime s of interest. Then the properties $P1^s$–$P5^s$ of conditional independence can be applied, and help us to axiomatically pursue identification of interventional quantities from observational data.

Graphical models in the form of Influence Diagrams (IDs) will also be used (where possible) to represent collections of conditional independence properties and help us derive further conditional independence properties that are implied by our

assumptions.

5.4 Consequence of a strategy

We seek to calculate the expectation $\mathbb{E}\{k(Y)\,;\,s\}$ of some given function $k(\cdot)$ of Y in a particular interventional regime s; for example, $k(\cdot)$ could be a loss function, $k(y) \equiv L(y)$, associated with the outcome of Y. We shall use the term *consequence* of s to denote the expectation $\mathbb{E}\{k(Y)\,;\,s\}$ of $k(Y)$ under the contemplated interventional regime s.

We can factorize the overall joint density of $(L_1, A_1, \ldots, L_N, A_N, Y)$ in interventional regime s as:

$$p(y, \overline{l}, \overline{a}\,;\,s) = \left\{\prod_{i=1}^{n+1} p(l_i \mid \overline{l}_{i-1}, \overline{a}_{i-1}\,;\,s)\right\} \times \left\{\prod_{i=1}^{n} p(a_i \mid \overline{l}_i, \overline{a}_{i-1}\,;\,s)\right\} \tag{5.4.1}$$

with $l_{n+1} \equiv y$.

5.4.1 G-recursion

If we knew all the terms on the right-hand side of (5.4.1), we could in principle compute the joint density for $(Y, \overline{L}, \overline{A})$ under strategy s, hence, by marginalization, the density of Y, and finally the desired consequence $\mathbb{E}\{k(Y); s\}$. However, a more efficient way to compute this is by means of the *G-computation* formula introduced by Robins (1986). Here we describe the recursive formulation of this formula, *G-recursion*, as presented in Dawid and Didelez (2010).

Let h denote a partial history of the form $(\overline{l}_i, \overline{a}_{i-1})$ or $(\overline{l}_i, \overline{a}_i)$ $(0 \le i \le n+1)$. We denote the set of all partial histories by $\mathcal{H}$. Fixing a regime $s \in \mathcal{S}$, define a function f on $\mathcal{H}$ by:

$$f(h) := \mathbb{E}\{k(Y) \mid h\,;\,s\}. \tag{5.4.2}$$

Note: When we are dealing with non-discrete distributions (and also in the discrete case when there are non-trivial events of $\mathbb{P}_s$-probability 0), the conditional expectation on the right-hand side of (5.4.2) will not be uniquely defined, but can be altered on a set of histories that has $\mathbb{P}_s$-probability 0. Thus we are in fact requiring, for each i:

$$f(\overline{L}_i, \overline{A}_i) := \mathbb{E}\{k(Y) \mid \overline{L}_i, \overline{A}_i\,;\,s\} \quad \text{a.s. } [\mathbb{P}_s]$$

(and similarly when the argument is $(\overline{L}_i, \overline{A}_{i-1})$). And we allow the left-hand side of

(5.4.2) to denote any selected version of the conditional expectation on the right-hand side.

For any versions of these conditional expectations, applying the law of repeated expectations yields:

$$f(\overline{L}_i, \overline{A}_{i-1}) = \mathbb{E}\{f(\overline{L}_i, \overline{A}_i) \mid \overline{L}_i, \overline{A}_{i-1}\,;\, s\} \quad \text{a.s. } [\mathbb{P}_s] \tag{5.4.3}$$

$$f(\overline{L}_{i-1}, \overline{A}_{i-1}) = \mathbb{E}\{f(\overline{L}_i, \overline{A}_{i-1}) \mid \overline{L}_{i-1}, \overline{A}_{i-1}\,;\, s\} \quad \text{a.s. } [\mathbb{P}_s]. \tag{5.4.4}$$

For h a full history $(\bar{l}_n, \overline{a}_n, y)$, we have $f(h) = k(y)$. Using these as starting values, by successively implementing (5.4.3) and (5.4.4) in turn, starting with (5.4.4) for $i = n + 1$ and ending with (5.4.4) for $i = 1$, we step down through ever shorter histories until we have computed $f(\emptyset) = \mathbb{E}\{k(Y)\,;\, s\}$, the consequence of regime s. Note that this equality is only guaranteed to hold almost surely, but since both sides are constants they must be the same constant. In particular, it can not matter which version of the conditional expectations we have chosen in conducting the above recursion: in all cases we will exit with the desired consequence $\mathbb{E}\{k(Y)\,;\, s\}$.

5.4.2 Using observational data

In order to compute $\mathbb{E}\{k(Y)\,;\, s\}$, whether directly from (5.4.1) or using G-recursion, (5.4.3) and (5.4.4), we need (versions of) the following conditional distributions under $\mathbb{P}_s$:

(i) $A_i \mid \overline{L}_i, \overline{A}_{i-1}\,;\, s$ for $i = 1, \ldots, n$.

(ii) $L_i \mid \overline{L}_{i-1}, \overline{A}_{i-1}\,;\, s$ for $i = 1, \ldots, n+1$.

Since s is an interventional regime, corresponding to a well-defined (possibly randomized) treatment strategy, the conditional distributions in (i) are fully specified by the treatment protocol. So we only need to get a handle on each term of the form (ii). However, since we have not implemented the strategy s, we do not have data directly relevant to this task. Instead, we have observational data, arising from a joint distribution we shall denote by $\mathbb{P}_\emptyset$. We might then be tempted to replace the desired but not directly accessible conditional distribution, under $\mathbb{P}_s$, of $L_i|L_{i-1}, A_{i-1}$, by its observational counterpart, computed under $\mathbb{P}_\emptyset$, which is (in principle) estimable from observational data. This will generally be a dangerous ploy, since we are dealing with two quite distinct regimes, with strong possibilities for confounding and other biases in the observational regime; however, it can

be justifiable if we can impose suitable extra conditions, relating the probabilistic behaviours of the different regimes.

5.5 Simple Stability

We now use ECI to express and explore some conditions that will allow us to perform G-recursion for the strategy of interest on the basis of observational data.

Consider first the term $p(a_i \mid \overline{l}_i, \overline{a}_{i-1}\,;\, s)$ as needed for (5.4.3). This term requires knowledge of the mechanism that allocates the treatment at stage i in the light of the preceding variables in the information base. We assume that, for an interventional regime $s \in \mathcal{S}^*$, this distribution (degenerate for a non-randomized strategy) will be known *a priori* to the analyst, as it will be encoded in the strategy. In such a case we call $s \in \mathcal{S}^*$ a *control strategy* (with respect to the information base $\mathcal{I} = (L_1, A_1, \ldots, L_N, A_N, Y)$).

Next we consider how we might gain knowledge of the conditional density $p(l_i \mid \overline{l}_{i-1}, \overline{a}_{i-1}\,;\, s)$, as required for (5.4.4). This distribution is unknown, and we need to explore conditions that will enable us to identify it from observational data.

Definition 5.5.1. We say that the problem exhibits *simple stability*[1] with respect to the information base $\mathcal{I} = (L_1, A_1, \ldots, L_n, A_n, Y)$ if, for each $s \in \mathcal{S}^*$, with $\Sigma_{\emptyset,s}$ denoting the identity function on $\{\emptyset, s\}$:

$$L_i \;\perp\!\!\!\perp\; \Sigma_{\emptyset,s} \mid (\overline{L}_{i-1}, \overline{A}_{i-1}) \quad (i = 1, \ldots, n+1). \tag{5.5.1}$$

Formally, simple stability requires that, for any bounded measurable function h, there exists a common version of the conditional expectations $\mathbb{E}\{h(L_i) \mid \overline{L}_{i-1}, \overline{A}_{i-1}\,;\, \emptyset\}$ and $\mathbb{E}\{h(L_i) \mid \overline{L}_{i-1}, \overline{A}_{i-1}\,;\, s\}$. In particular, it apparently[2] supports the identification of $p(l_i \mid \overline{l}_{i-1}, \overline{a}_{i-1}\,;\, s)$ with $p(l_i \mid \overline{l}_{i-1}, \overline{a}_{i-1}\,;\, \emptyset)$, thus allowing observational estimation of the former term, which is what is required for (5.4.4).

The ID describing simple stability (5.5.1) for $i = 1, 2, 3$ is shown in Figure 5.1. The specific property (5.5.1) is represented by the absence of arrows from σ to L_1, L_2, and $L_3 \equiv Y$.

[1] This definition is slightly weaker than that of Dawid and Didelez (2010), as we are only requiring a common version of the corresponding conditional expectations between each single control strategy and the observational regime. We do not require that there exists one function that can serve as common version across all regimes simultaneously.

[2] but see Section 5.5.1 below

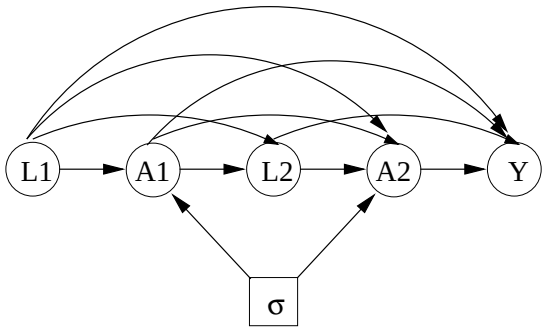

Figure 5.1: Stability

5.5.1 Positivity

We have indicated that simple stability might allow us to identify the consequence of a control strategy s on the basis of data from the observational regime $\emptyset$. However, while this condition ensures the existence of a common version of the relevant conditional expectation valid for both regimes, deriving this function from the observational regime alone might be problematic, because versions of the same conditional expectation can differ on events of probability 0, and we have not ruled out that an event having probability 0 in one regime might have positive probability in another. Thus we can only obtain the desired function from the observational regime on a set that has probability 1 in the observational regime; and this might not have probability 1 in the interventional regime (recall Example 3.1.5).

To evade this problem, we can, where appropriate, impose a condition that requires an event to have zero probability in the interventional regime whenever it has zero probability in the observational regime:

Definition 5.5.2. We say the problem exhibits *positivity* or *absolute continuity* if, for any interventional regime $s \in \mathcal{S}^*$, the joint distribution of $(\overline{L_n}, \overline{A_n}, Y)$ under $\mathbb{P}_s$ is absolutely continuous with respect to that under $\mathbb{P}_\emptyset$, *i.e.*:

$$\mathbb{P}_s(E) > 0 \Rightarrow \mathbb{P}_\emptyset(E) > 0 \tag{5.5.2}$$

for any event E defined in terms of $(\overline{L_n}, \overline{A_n}, Y)$.

Notice that the above definition is equivalent to asking that $\mathbb{P}_\emptyset$ dominates $\mathbb{P}_s$ for any event E defined in terms of $(\overline{L_n}, \overline{A_n}, Y)$.

5.6 Sequential Ignorability

Simple stability might be a tenable assumption when, in the observational regime, the treatments are physically (sequentially) randomized; or when we know that we have included in the information base all the information that has been taken into account in assigning the treatments in the observational regime. However in many cases—for example because of the suspected presence of confounding variables—we would not be willing to accept, at any rate without further justification, this assumption of simple stability. Here we consider conditions that might seem more acceptable, and investigate when these will, after all, imply simple stability—thus supporting the application of G-recursion.

5.6.1 Extended stability and extended positivity

Let $\mathcal{U}$ denote a set of variables that, while they might potentially influence actions taken under the observational regime, are not available to the decision maker, and so are not included in his information base $\mathcal{I} := (L_1, A_1, \ldots, L_n, A_n, L_{n+1} \equiv Y)$. We define the *extended information base* $\mathcal{I}' := (L_1, U_1, A_1, \ldots, L_n, U_n, A_n, L_{n+1})$, with U_i denoting the variables in $\mathcal{U}$ realized just before action A_i is taken. However, while thus allowing U_i to influence A_i in the observational regime, we still only consider interventional strategies where there is no such influence—since the decision maker does not have access to the (U_i). This motivates an extended formal definition of "control strategy" in this context:

Definition 5.6.1 (Control strategy)**.** A regime s is a *control strategy* if

$$A_i \perp\!\!\!\perp \overline{U}_i \mid (\overline{L}_i, \overline{A}_{i-1}\,;\, s) \quad (i = 1, \ldots, n) \tag{5.6.1}$$

and in addition, the conditional distribution of A_i, given $(\overline{L}_i, \overline{A}_{i-1})$, under regime s, is known to the analyst.

We again denote the set of interventional regimes corresponding to the control strategies under consideration by $\mathcal{S}^*$.

Definition 5.6.2. We say that the problem exhibits *extended stability* (with respect to the extended information base $\mathcal{I}'$) if, for any $s \in \mathcal{S}^*$, with $\Sigma_{\emptyset,s}$ denoting the identity function on $\{\emptyset, s\}$:

$$(L_i, U_i) \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_{i-1}, \overline{U}_{i-1}, \overline{A}_{i-1}) \quad (i = 1, \ldots, n+1). \tag{5.6.2}$$

Extended stability is formally the same as simple stability, but using a different information base, where L_i is expanded to (L_i, U_i). The real difference is that the extended information base is not available to the decision maker in the interventional regime, so that his decisions can not take account of the (U_i). An ID faithfully representing property (5.6.2) for $i = 1, 2, 3$ is shown in Figure 5.2. The property (5.6.2) is represented by the absence of arrows from σ to L_1, U_1, L_2, U_2 and Y. However, the diagram does not explicitly represent the additional property (5.6.1), which implies that, *when* $\sigma = s$, the arrows into A_1 from U_1 and into A_2 from U_1 and U_2 can be dropped.

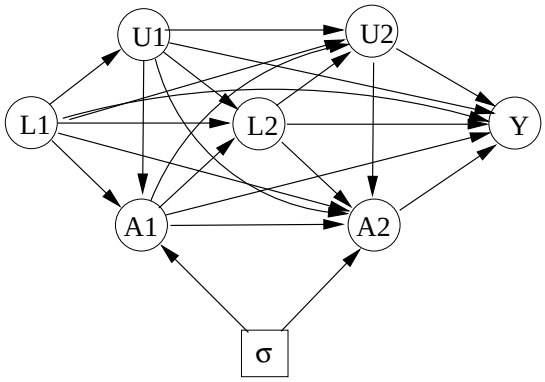

Figure 5.2: Extended stability

To evade problems with events of zero probability, we can extend Definition 5.5.2:

Definition 5.6.3. We say the problem exhibits *extended positivity* if, for any $s \in \mathcal{S}^*$, the joint distribution of $(\overline{U_n}, \overline{L_n}, \overline{A_n}, Y)$ under $\mathbb{P}_s$ is absolutely continuous with respect to that under $\mathbb{P}_\emptyset$, *i.e.*

$$\mathbb{P}_s(E) > 0 \Rightarrow \mathbb{P}_\emptyset(E) > 0 \tag{5.6.3}$$

for any event E defined in terms of $(\overline{L_n}, \overline{U_n}, \overline{A_n}, Y)$.

Notice that the above definition is equivalent to asking that $\mathbb{P}_\emptyset$ dominates $\mathbb{P}_s$ for any event E defined in terms of $(\overline{L_n}, \overline{U_n}, \overline{A_n}, Y)$.

5.6.2 Sequential randomization

Extended stability represents the belief that, for each i, the conditional distribution of (L_i, U_i), given all the earlier variables $(\overline{L}_{i-1}, \overline{U}_{i-1}, \overline{A}_{i-1})$ in the extended information base, is the same in the observational regime as in the interventional regime.

This will typically be defensible if we can argue that we have included in $\mathcal{L} \cup \mathcal{U}$ all the variables influencing the actions in the observational regime.

However, extended stability, while generally more defensible than simple stability, typically does not imply simple stability, which is what is required to support G-recursion. But it may do so if we impose additional conditions. Here and in Section 5.6.3 below we explore two such conditions.

Our first is the following:

Condition 5.6.4 (Sequential randomization)**.**

$$A_i \perp\!\!\!\perp \overline{U}_i \mid (\overline{L}_i, \overline{A}_{i-1}\,; \emptyset) \quad (i = 1, \ldots, n). \tag{5.6.4}$$

Taking account of (5.6.1), we see that (5.6.4) is equivalent to:

$$A_i \perp\!\!\!\perp \overline{U}_i \mid (\overline{L}_i, \overline{A}_{i-1}, \Sigma) \quad (i = 1, \ldots, n) \tag{5.6.5}$$

where Σ in the identity function on $\mathcal{S} = \{\emptyset\} \cup \mathcal{S}^*$. In particular, for any $s \in \mathcal{S}^*$

$$A_i \perp\!\!\!\perp \overline{U}_i \mid (\overline{L}_i, \overline{A}_{i-1}, \Sigma_{\emptyset,s}) \quad (i = 1, \ldots, n) \tag{5.6.6}$$

where $\Sigma_{\emptyset,s}$ in the identity function on $\{\emptyset, s\}$.

Under sequential randomization, the observational distribution of A_i, given the earlier variables in the information base, would be unaffected by further conditioning on the earlier unobservable variables, $\overline{U}_i$. Hence the (U_i) are redundant for explaining the way in which actions are determined in the observational regime. The following result is therefore unsurprising.

Theorem 5.6.5 (Dawid and Didelez (2010))**.** Suppose we have both extended stability, (5.6.2) and sequential randomization, (5.6.4). Then we have simple stability, (5.5.1).

Proof. Let $s \in \mathcal{S}^*$ and let E_i, R_i, H_i denote, respectively, the following assertions:

$$\begin{aligned}
E_i &: (L_i, U_i) \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_{i-1}, \overline{U}_{i-1}, \overline{A}_{i-1}) \\
R_i &: A_i \perp\!\!\!\perp \overline{U}_i \mid (\overline{L}_i, \overline{A}_{i-1}, \Sigma_{\emptyset,s}) \\
H_i &: (L_i, \overline{U}_i) \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_{i-1}, \overline{A}_{i-1})
\end{aligned}$$

Extended stability (5.6.2) is equivalent to E_i holding for all i and sequential randomisation implies R_i for all i. We shall show that these assumptions imply H_i for

all i, which in turn implies $L_i \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_{i-1}, \overline{A}_{i-1})$, *i.e.*, simple stability.

We proceed by induction. Since E_1 and H_1 are both equivalent to $(L_1, U_1) \perp\!\!\!\perp \Sigma_{\emptyset,s}$, H_1 holds.

Suppose now H_i holds. Conditioning on L_i yields

$$\overline{U}_i \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_i, \overline{A}_{i-1}), \tag{5.6.7}$$

and this together with R_i is equivalent to $\overline{U}_i \perp\!\!\!\perp (A_i, \Sigma_{\emptyset,s}) \mid (\overline{L}_i, \overline{A}_{i-1})$, which on conditioning on A_i then yields

$$\overline{U}_i \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_i, \overline{A}_i). \tag{5.6.8}$$

Also, by E_{i+1} we have

$$(L_{i+1}, U_{i+1}) \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (\overline{L}_i, \overline{U}_i, \overline{A}_i). \tag{5.6.9}$$

Taken together, (5.6.8) and (5.6.9) are equivalent to H_{i+1}, so the induction is established. □

An ID faithfully representing the conditional independence relationships assumed in Theorem 5.6.5, for $i = 1, 2, 3$, is shown in Figure 5.3. Figure 5.3 can be obtained

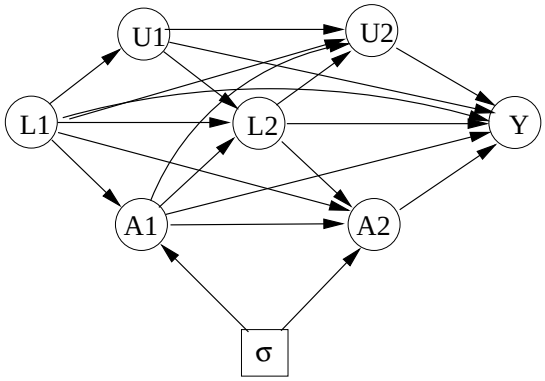

Figure 5.3: Sequential randomization

from Figure 5.2 on deleting the arrows into A_1 from U_1 and into A_2 from U_1 and U_2, so representing (5.6.6). (However, as we shall see below in Section 5.6.3, in general such "surgery" on IDs can be hazardous.)

The conditional independence properties (5.5.1) characterising simple stability can now be read off from Figure 5.3, by applying the *d*-separation or moralization

criteria. These criteria will not be expounded in this thesis but are only given as a reference for an alternative approach. Details can be found in Dawid (2002). [3]

Corollary 5.6.6. Suppose we have extended stability, sequential randomization, and extended positivity. Then we can apply G-recursion to compute the consequence of a strategy $s \in \mathcal{S}^*$.

5.6.3 Sequential irrelevance

Consider now the following alternative condition:

Condition 5.6.7 (Sequential Irrelevance)**.**

$$L_i \perp\!\!\!\perp \overline{U}_{i-1} \mid (\overline{L}_{i-1}, \overline{A}_{i-1}, \Sigma) \quad (i = 1, \ldots, n+1), \tag{5.6.10}$$

where Σ in the identity function on $\mathcal{S} = \{\emptyset\} \cup \mathcal{S}^*$.

In contrast to (5.6.5), (5.6.10) permits the unobserved variables that appear in earlier stages to influence the next action A_i—but not the development of the subsequent observable variables (including the ultimate response variable Y).

By analogy with the passage from Figure 5.2 to Figure 5.3, we might attempt to represent the additional assumption (5.6.10) by removing from Figure 5.2 all arrows from U_j to L_i $(j < i)$. This would yield Figure 5.4. On applying d-separation

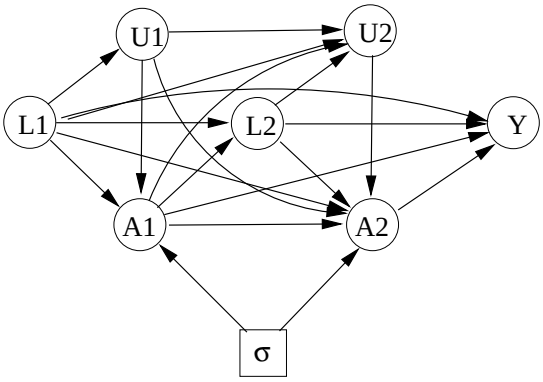

Figure 5.4: Sequential irrelevance?

or moralization to Figure 5.4 we could then deduce the simple stability property (5.5.1). However, this approach is not valid, since Figure 5.4 encodes the property

[3]Note that, in either of these approaches, we can restrict σ to the two values $\emptyset$ and s, so fully justifying treating the non-stochastic variable σ as if it were stochastic.

$L_2 \perp\!\!\!\perp \Sigma \mid (L_1, A_1)$, which can not be derived from (5.6.2) and (5.6.10) using only the axioms. In fact there is no ID that faithfully represents the combination of the properties (5.6.2) and (5.6.10), since these do not form a recursive system. And indeed, in full generality, simple stability is not implied by extended stability, (5.6.2), together with sequential irrelevance, (5.6.10), as the following counter-example demonstrates.

Counterexample 5.6.8. Take $n = 1$, $\mathcal{L} = \emptyset$ and $\mathcal{U} = \{U\}$. The extended information base is $\mathcal{I}' = (U, A, Y)$. Assume that we are only concerned with two regimes, the observational regime $\emptyset$ and the interventional regime s. Thus in (5.6.10) above we can replace Σ with $\Sigma_{\emptyset,s}$. We suppose that in both regimes, $Y = 1$ if $A = U$, else $Y = 0$. Also, in each regime, the marginal distribution of U is uniform on $[0, 1]$. It remains to specify the distribution of A, given U: we assume that, in regime $\emptyset$, $A = U$, while in regime s, A is uniform on $[0, 1]$, independently of U.

It is readily seen that $U \perp\!\!\!\perp \Sigma_{\emptyset,s}$ and $Y \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid (U, A)$. Thus we have extended stability, (5.6.2), as represented by the ID of Figure 5.5.

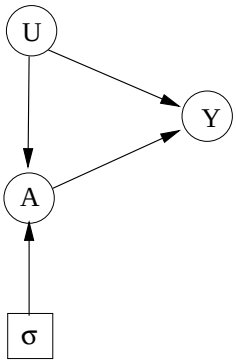

Figure 5.5: Counter-example

Also, since $U \perp\!\!\!\perp A$ in regime s, (5.6.1) holds, so s is a control strategy. Finally, in regime $\emptyset$, $Y = 1$ a.s. $[\mathbb{P}_\emptyset]$, while in regime s, $Y = 0$ a.s. $[\mathbb{P}_\emptyset]$. Because these are both degenerate distributions, trivially $Y \perp\!\!\!\perp U \mid (A, \Sigma_{\emptyset,s})$, and we have sequential irrelevance. However, because they are different distributions, $Y \not\perp\!\!\!\perp \Sigma_{\emptyset,s} \mid A$: so we do *not* have simple stability, (5.5.1). In particular, we can not remove the arrow from U to Y in Figure 5.5, since this would encode the false property $Y \perp\!\!\!\perp \Sigma_{\emptyset,s} \mid A$.

So, if we wish to deduce simple stability from extended stability and sequential irrelevance, further conditions, and a different approach, will be required.

In Theorem 6.2 of Dawid and Didelez (2010) it is shown that this result does follow if we additionally impose the extended positivity condition of Definition 5.6.3;

and then we need only require sequential irrelevance, (5.6.10), to hold for the observational regime $\sigma = \emptyset$.

However, in Section 5.6.4 below we show that, if we restrict all variables to be discrete, no further conditions are required for this result to hold; moreover, in this case we only require sequential irrelevance to hold for the interventional regime $\sigma = s$.

5.6.4 Discrete case

To control null events, we need the following lemma:

Lemma 5.6.9. Let all variables be discrete. Suppose that we have extended stability, (5.6.2), and let s be a control strategy, so that (5.6.1) holds. Then, for any $(\overline{u}_k, \overline{l}_k, \overline{a}_k)$ such that

$\mathbf{A_k}$: $P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) > 0$, and

$\mathbf{B_k}$: $P(\overline{U}_k = \overline{u}_k, \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, \emptyset) > 0$, we have

$\mathbf{C_k}$: $P(\overline{U}_k = \overline{u}_k, \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) > 0$.

Proof. Let H_k denote the assertion that $\mathbf{A_k}$ and $\mathbf{B_k}$ imply $\mathbf{C_k}$. We establish H_k by induction.

To start, we note that H_0 holds vacuously.

Now suppose H_{k-1} holds. Assume further $\mathbf{A_k}$ and $\mathbf{B_k}$. Together these conditions imply that all terms appearing throughout the following argument are positive.

We have

$$
\begin{aligned}
&P(\overline{U}_k = \overline{u}_k, \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) \\
&= P(\overline{U}_k = \overline{u}_k \mid \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) && (5.6.11) \\
&= P(\overline{U}_k = \overline{u}_k \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s) P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) && (5.6.12) \\
&= \frac{P(\overline{U}_k = \overline{u}_k, \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s)}{P(\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s)} P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s) \\
&= P(U_k = u_k, L_k = l_k \mid \overline{U}_{k-1} = \overline{u}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s) \times \\
&\frac{P(\overline{U}_{k-1} = \overline{u}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s) P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s)}{P(\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s)} && (5.6.13) \\
&= P(U_k = u_k, L_k = l_k \mid \overline{U}_{k-1} = \overline{u}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, \emptyset) \times \\
&\frac{P(\overline{U}_{k-1} = \overline{u}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s) P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k \,;\, s)}{P(\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \,;\, s)} && (5.6.14)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{P(\overline{U}_k = \overline{u}_k, \overline{L}_k = l_k, \overline{A}_{k-1} = \overline{a}_{k-1}\,;\,\emptyset)}{P(\overline{U}_{k-1} = \overline{u}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}\,;\,\emptyset)} \times \\
&\frac{P(\overline{U}_{k-1} = \overline{u}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}\,;\,s) P(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k\,;\,s)}{P(\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}\,;\,s)} \qquad (5.6.15) \\
&> 0.
\end{aligned}
$$

Here (5.6.12) holds by (5.6.1) and (5.6.14) holds by (5.6.2). The induction is established. □

Theorem 5.6.10. Suppose the conditions of Lemma 5.6.9 apply, and, further, that we have sequential irrelevance in the interventional regime s:

$$L_i \perp\!\!\!\perp \overline{U}_{i-1} \mid (\overline{L}_{i-1}, \overline{A}_{i-1}\,;\,s) \quad (i = 1, \ldots, n+1). \qquad (5.6.16)$$

Then the simple stability property (5.5.1) holds.

Proof. The result will be established if we can show that, for any l_i, there exists a function $w(\overline{L}_{i-1}, \overline{A}_{i-1})$ such that, for both $\sigma = \emptyset$ and $\sigma = s$,

$$P(L_i = l_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\sigma) = w(\overline{l}_{i-1}, \overline{a}_{i-1})$$

whenever $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\sigma) > 0$.

If $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\sigma) = 0$ for both regimes, we have nothing to show.

If, say, $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) > 0$ while $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s) = 0$, we can take

$$w(\overline{l}_{i-1}, \overline{a}_{i-1}) := P(\overline{L}_i = \overline{l}_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset),$$

and similarly, *mutatist mutandis*, if $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s) > 0$ while $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) = 0$.

Otherwise, $P(\overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\sigma) > 0$ for both regimes. Then

$$
\begin{aligned}
&P(L_i = l_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) \\
&= \sum_{\overline{u}_{i-1}} P(L_i = l_i \mid \overline{U}_{i-1} = \overline{u}_{i-1}, \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) \times \\
&\qquad P(\overline{U}_{i-1} = \overline{u}_{i-1} \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset). \qquad (5.6.17)
\end{aligned}
$$

The non-zero summands in (5.6.17) have $P(\overline{U}_{i-1} = \overline{u}_{i-1}, \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} =$

$\overline{a}_{i-1}\,;\,\emptyset) > 0$, and so, by Lemma 5.6.9, $P(\overline{U}_{i-1} = \overline{u}_{i-1}, \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s) > 0$. Then by (5.6.2),

$$\begin{aligned}
&P(L_i = l_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) \\
&= \sum_{\overline{u}_{i-1}} P(L_i = l_i \mid \overline{U}_{i-1} = \overline{u}_{i-1}, \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s) \times \\
&\qquad\qquad P(\overline{U}_{i-1} = \overline{u}_{i-1} \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) \\
&= \sum_{\overline{u}_{i-1}} P(L_i = l_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s) \times \\
&\qquad\qquad P(\overline{U}_{i-1} = \overline{u}_{i-1} \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,\emptyset) \qquad (5.6.18) \\
&= P(L_i = l_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s)
\end{aligned}$$

where (5.6.18) holds by (5.6.16). Thus we can take

$$w(\overline{l}_{i-1}, \overline{a}_{i-1}) := P(L_i = l_i \mid \overline{L}_{i-1} = \overline{l}_{i-1}, \overline{A}_{i-1} = \overline{a}_{i-1}\,;\,s)$$

to conclude the proof. □

Chapter 6

Conclusions

The need to express causal concepts necessitates the use of an appropriate framework which differentiates between seemingly related observations and causally related observations. Working under the Decision Theoretic framework we use an index variable to denote the different regimes that generate data. In particular, we differentiate between the observational regime where variables of interest are merely observed under natural conditions and the interventional regimes where variables of interest are observed under intervention.

In practice, we are usually not able to obtain data from the interventional regimes of interest and hence compare them (a comparison with a directly causal interpretation). Thus, we want to explore under what conditions we can deduce information for the interventional regimes using the observational regime.

The conditions that allow transfer of probabilistic information between different regimes can be expressed in the language of conditional independence. In this thesis, we have developed a formal underpinning for this approach, based on an extension of the axiomatic theory of conditional independence to include non-stochastic variables.

In Chapter 2 we have studied the axioms of conditional independence separately for stochastic and non-stochastic variables, presenting all the mathematical tools that are needed to extend the language. In Chapter 3, we have combined stochastic and non-stochastic variables in a single language and explored the validity of the axioms. We have shown that when the regime space is discrete or the random variables are discrete the axioms follow. Generalising to the case of a non-discrete space and non-discrete random variables, we saw that additional conditions were required. This formal foundation now supplies a rigorous justification for various more informal arguments that have previously been presented (Dawid, 1979a, 2002;

Dawid and Didelez, 2010).

As applications of the extended analysis of conditional independence, we considered in Chapter 4 the identification of the Average Causal Effect (ACE) and the Effect of Treatment on the Treated (ETT). We introduced sufficient covariates as variables that satisfy conditions that allow identification of the ACE from observational data, and the ETT from observational data and interventional data under control treatment. Also in Chapter 5 we applied this theory to the problem of dynamic treatment assignment. We have shown how, and under what conditions, the assumptions of sequential randomization or sequential irrelevance can support observational identification of the consequence of some treatment strategy under consideration.

Bibliography

Patrick Billingsley. *Probability & Measure*. Wiley, 3rd edition, 1995. ISBN 0-471-0071-02.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2001.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 0387718230, 9780387718231.

A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):pp. 1–31, 1979a. URL `http://www.jstor.org/stable/2984718`.

A. Philip Dawid. Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):249–252, 1979b. URL `http://www.jstor.org/stable/2985039`.

A. Philip Dawid. Conditional independence for statistical operations. *The Annals of Statistics*, 8(3):598–617, 1980. URL `http://www.jstor.org/stable/2240595`.

A. Philip Dawid. Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*, 95:407 – 424, 2000.

A. Philip Dawid. Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32:335–372, 2001a. ISSN 1012-2443. URL `http://dx.doi.org/10.1023/A:1016734104787`.

A. Philip Dawid. Some variations on variation independence. In *Proceedings of Artificial Intelligence and Statistics 2001*, pages pp. 187 – 191. Morgan Kaufmann Publishers, 2001b.

A. Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002. doi: 10.1111/j.1751-5823.2002.tb00354.x. URL `http://dx.doi.org/10.1111/j.1751-5823.2002.tb00354.x`.

A. Philip Dawid. *Conditional Independence*. John Wiley & Sons, Inc., 2004. ISBN 9780471667193. doi: 10.1002/0471667196.ess0618.pub2. URL `http://dx.doi.org/10.1002/0471667196.ess0618.pub2`.

A. Philip Dawid. Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. *In Causality and Probability in the Sciences, edited by F. Russo and J. Williamson. London: College Publications*, Texts In Philosophy Series Vol. 5:503–532, 2007a.

A. Philip Dawid. Fundamentals of statistical causality. (279), 2007b.

A. Philip Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010. ISSN 1935-7516. doi: 10.1214/10-SS081. URL `arXiv:1010.3425`.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4 edition, 2010. ISBN 978-0521765398.

S. Geneletti and A. P. Dawid. Defining and identifying the effect of treatment on the treated. *In Causality in the Sciences (P. McKay Illari, F. Russo and J. Williamson, Eds.). Oxford University Press (to appear)*, 2011.

H. Guo. *Statistical Causal Inference and Propensity Analysis*. PhD thesis, University of Cambridge, October 2010.

R. A. Howard and J. E. Matheson. Influence diagrams. 1984.

Irving Kaplansky. *Set theory and metric spaces*. AMS Chelsea Publishing, 2nd edition, 2001.

John Frank Charles Kingman. *Poisson processes*. Oxford studies in probability ;3. Oxford University Press., 1993. ISBN 9780198536932.

Andrei Nikolaevich Kolmogorov. Sur l'éstimation statistique des paramètres be la loi de gauss. *Izv. Akad. Nauk SSSR Ser. Mat.*, 6:3–32, 1942.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990. ISSN 1097-0037. doi: 10.1002/net.3230200503. URL `http://dx.doi.org/10.1002/net.3230200503`.

E.L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 2nd edition, 1998.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003. doi: 10.1111/1467-9868.00389. URL `http://dx.doi.org/10.1111/1467-9868.00389`.

K. R. Parthasarathy. *Probability Measures on Metric Spaces*, volume 352. AMS Chelsea Publishing, 1967. ISBN 978-0-8218-3889-1.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1997. ISBN 1558604790.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393 – 1512, 1986. ISSN 0270-0255. doi: 10.1016/0270-0255(86)90088-6. URL `http://www.sciencedirect.com/science/article/pii/0270025586900886`.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41•55, 1983.

Susanne Rosthøj, Catherine Fullwood, Robin Henderson, and Syd Stewart. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Statistics in Medicine*, 25(24):4197–4215, 2006. ISSN 1097-0258. doi: 10.1002/sim.2694. URL `http://dx.doi.org/10.1002/sim.2694`.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688•701, 1974.

Donald B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2:1•26, 1977.

Donald B. Rubin. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6:34•68, 1978.

R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871•82, 1986.

Satish Shirali and Harkrishan Lal Vasudeva. *Metric Spaces*. Springer, 2006. ISBN 1852339225.

J. A. C. Sterne, M. May, D. Costagliola, F. de Wolf, A. N. Phillips, R. Harris, M. J. Funk, R. B. Geskus, J. Gill, F. Dabis, J. M. Miro, A. C. Justice, B. Ledergerber, G. Fatkenheuer, R. S. Hogg, A. D'Arminio-Monforte, M. Saag, C. Smith, S. Staszewski, M. Egger, and S. R. Cole. Timing of initiation of antiretroviral therapy in aids-free hiv-1-infected patients: a collaborative analysis of 18 hiv cohort studies. *The Lancet*, 373:1352–1363, 2009. URL `http://linkinghub.elsevier.com/retrieve/pii/S0140673609606127`.