The Topology of Statistical Inquiry

Konstantin Genin
Department of Philosophy
Carnegie Mellon University
konstantin.genin@gmail.com

September 14, 2018

Acknowledgements

This dissertation is the culmination of six years of research. Over that time I have acquired a great number of personal debts. I want to acknowledge some of those debts here, without any pretense to completeness — I hope that the omitted will forgive me. Thank you to my parents, Jenny, Joe and Sasha, and my grandparents, Roman and Bronia. Thank you to Matt Evans and Basil Katz, for moving me in. Thank you to Adam Brody and Elizabeth Silver, for support through many long nights. I am especially grateful for the comraderie of our cohort, whom I can't help but find exceptional: Liam Kofi Bright, Aidan Kestigian, Rob Lewis, Dan Malinsky, and Ruben Sanchez. Thank you to Helga Caballero, for your friendship. Thank you to Kerry Filtz, for being there when the tape started rolling again. I am grateful to Adam Bjorndahl, Thomas Icard, Hanti Lin, Alison Springle, Conor Mayo-Wilson, Jonathan Livengood, Sam Fletcher and Jack Parker, for many stimulating discussions. Thank you to my patient and generous committee: Malcolm Forster, Clark Glymour, Thomas Icard, Cosma Shalizi, and Daniel Steel. Finally, none of this would have been possible without Kevin Kelly, whom I am proud to consider a mentor and friend.

Contents

1 Introduction **7**

1.1 The Idea of Progress 7

1.2 Statistical Verifiability and Falsifiability 18

1.3 The Topology of Inquiry 24

1.4 Plan of the Work 27

2 Learning from Propositional Information **29**

2.1 The Propositional Setting 29

2.2 Propositional Verification, Refutation and Decision 32

2.3 Limiting Verification, Refutation and Decision 34

2.4 Problems and Solutions 38

2.5 Simplicity and Ockham's Razor 38

2.6 Simplicity and Progress 40

3 Learning from Statistical Information **43**

3.1 The Statistical Setting 46

3.1.1 Samples and Worlds 46

3.1.2 Statistical Tests 48

3.1.3 The Weak Topology 49

3.1.4 Convergence in Distribution 52

3.2 Statistical Verification, Refutation and Decision 54

3.2.1 Defining the Success Concepts 54

3.2.2 Characterization Theorems 56

3.2.3 Application: Conditional Independence Testing 61

3.2.4 Application: Causal Graphical Models 72

3.3 Monotonic Statistical Verification and Refutation 76

3.3.1 Defining the Success Concept 76

3.3.2 Characterization Theorems 78

3.3.3 Monotonic Verification of Statistical Dependence 90

3.4 Limiting Statistical Verification, Refutation, and Decision 92

3.4.1 Defining the Success Concepts 92

3.4.2 Characterization Theorems 93

3.4.3 Application: Causal Graphical Models 98

3.5 Problems and Solutions 99

3.5.1 A Characterization of Solvable Problems 100
3.5.2 Solving the Markov Class Problem for Causal Graphs . . 101
3.6 Progress, Simplicity and Ockham's Razor 102
3.6.1 Simplicity 102
3.6.2 Progressive Solutions 106
3.6.3 Progress and Ockham's Razor 108
3.6.4 A Progressive Solution to the Markov Class Problem . . . 109
3.7 Related Work 110

Chapter 1

Introduction

1.1 The Idea of Progress

> Scientific progress has often been compared to a mounting tide . . . Whoever casts a brief glance at the waves striking a beach does not see the tide mount, he sees a wave rise, run, uncurl itself, and cover a narrow strip of sand, then withdraw by leaving dry the terrain which it had seemed to conquer; a new wave follows, sometimes going a little farther than the former wave. But under this superficial to-and-fro motion, another movement is produced, deeper, slower, imperceptible to the casual observer; it is a progressive movement continuing steadily in the same direction and by virtue of it the sea constantly rises. [Duhem, 1914]

The justification of inductive method is the problem Hume left to posterity. Two centuries after Hume's *Treatise*, Whitehead [1948] called the problem "the despair of philosophy." If it could not be convincingly demonstrated that "instances of which we have had no experience, must resemble those of which we have had experience," [1739, *Treatise* 1.3.6] then it is altogether implausible that Francis Bacon could derive his axioms from "particular events in a gradual and unbroken ascent" [1645, *Novum Organum*, 1.19] or that Descartes could claim to have deduced, like the geometers he admired, natural laws that "we cannot doubt . . . are strictly adhered to in everything that exists or occurs in the world" [1637, *Discourse*, Part V]. Although the lesson is still only partially digested, Hume taught us that inductive inferences are fallible. If science makes progress, it is not by the monotonic accumulation of certain truths. Hume himself was a believer in scientific progress, but he gave succor to the enemies of Enlightenment, who felt that they had in him an unwitting ally in the camp of their foes [Berlin, 1980]. Methodology has never fully recovered. There remains no convincing explanation of what, if not Cartesian certainty, the methodological advantage of science is supposed to be. This dissertation is aimed at providing such an explanation.

Niiniluoto [2015] divides early twentieth century attempts at justification of inductive method into the *synchronic* and *diachronic* schools.[1] In formal epistemology, the synchronic school remains the dominant paradigm, and has many able defenders. Synchronic theorists have proposed and studied confirmation relations between hypothesis and evidence [Hempel, 1945, Carnap, 1945], the rationality, or coherence, of systems of beliefs [Savage, 1972], and strategies for maximizing epistemic utility [Levi, 1967]. Broadly speaking, the synchronic project is to characterize which systems of belief constitute 'rational' responses to evidence. The trouble is that there is typically no explanation of how synchronic rationality facilitates progress toward the ultimate goal of science, at least as understood by realists: *true* answers to our questions about nature.[2] That makes synchronic-school justifications of inductive method rather attenuated. Carnap admits as much, in a startling passage towards the end of his *On Inductive Logic* [1945]:

> Our system of inductive logic . . . is intended as a rational reconstruction, restricted to a simple language form, of inductive thinking as customarily applied in everyday life and in science. . . . An entirely different question is the problem of the validity of our or any other proposed system of inductive logic, and thereby of the customary methods of inductive thinking. This is the genuinely philosophical problem of induction.

Carnapian reconstruction may systematize our inductive intuitions, but it does not prove that our intuitions are reliable, or that they are any better at arriving at true answers to scientific questions than other methods. The "genuinely philosophical" problem is to prove that our inductive methodology is more reliable than alternatives.

The diachronic school, channeling nineteenth century theorists like Whewell, Peirce, Mach and Duhem, conceived of science as a goal-oriented process, and investigated the dynamics of scientific change. From the diachronic perspective, synchronic norms are justified if their consistent application is conducive toward the goals of science. Popper is perhaps the best-known member of the diachronic school, portraying science as a process aimed at true theories, and driven by bold conjectures followed by dogged attempts at refutation. He was a thoroughgoing fallibilist, believing that the aim of science was not 'highly confirmed' hypotheses, but highly testable conjectures that stand up to the best attempts to falsify them. That compelling story nevertheless raises the alarming possibility that science is *nothing but* an aimless series of bold mistakes, yielding

[1] Jonathan Livengood has suggested (personal communication) that one would do better to distinguish between the internalist and externalist schools. If one were to make a two-by-two table, this work would fall into the diachronic, externalist cell of the partition.

[2] The 'realist' designation is taken here to apply to anyone who takes truth, or at least correctness, to be a constitutive goal of inquiry.

to new, and bolder, mistakes. For that reason, Lakatos [1974] charges that Popper "offers a methodology without an epistemology or a learning theory, and confesses explicitly that his methodology may lead us epistemologically astray, and implicitly, that *ad hoc* stratagems might lead us to Truth." But similar difficulties attended more sophisticated articulations of falsificationism. Lakatos [1970] himself held that a theoretical change was progressive if it made "dramatic, unexpected, stunning" new predictions that were subsequently verified. But we may ask the same question of Lakatos: why think that all this *sturm und drang* approaches, or even arrives eventually, at the truth?

Popper appreciated that difficulty, and attempted to develop a theory of *truthlikeness* to explain how a series of false theories could nevertheless approach closer and closer to the truth. He proposed that one theory is more truthlike than another if it has more true consequences and fewer false consequences. Lakatos, in the final passage of his *Falsification and the methodology of scientific research programmes*, expresses hope that his own sophisticated falsificationism would lead to increased truthlikeness in Popper's sense. Sadly, Popper's definition of truthlikeness was trivialized by Miller [1974] and Tichý [1974] — on his account, no false theory is any more truthlike than any other false theory. The spectacular failure of Popper's definition founded a philosophical specialty in less problematic definitions of truthlikeness. That project reaches a high degree of sophistication in Oddie [1986] and Niiniluoto [1987, 1999].[3] But even if the matter of definition were settled, there is no demonstration that scientific method is guaranteed to produce increasingly truthlike theories. Niiniluoto [1987] addresses the problem of giving "conditions for rationally claiming, on some evidence, that a statement g is truthlike — or at least is more truthlike than some other statement — *even when the truth* h_* *is unknown*". However, "appraisals of the relative distances from the truth presuppose that an epistemic probability distribution . . . is available. In this sense . . . the problem of estimating verisimilitude is neither more nor less difficult than the traditional problem of induction." Therefore, rather than justifying scientific methodology by demonstrating that it produces theories of increasing truthlikeness, the truthlikeness program generates a new problem: demonstrating that a consistent preference for *apparently* truthlike theories is in fact conducive, in some objective sense, toward finding the truth.[4]

Other diachronic theorists, notably Kuhn [1962] and Laudan [1978], abandon the realist project entirely. Kuhn [1962] famously abjures the idea of scientific progress as progress *towards* any goal, opting instead for an analogy with non-teleological, evolutionary progress:

[3] I will not attempt a thorough appraisal of this research program, which I consider a worthy one. Of course, there is the inherent difficulty, of which its proponents are well aware, of applying metrical notions of closeness where metrical structure may be absent.

[4] In fact, the problem may be worse. Why bother with computing expected truthlikeness, when you already have a probability distribution encoding your opinions about what is probably true?

> We are all deeply accustomed to seeing science as the one enterprise that draws constantly nearer to some goal set by nature in advance. But need there be any such goal? Can we not account for both science's existence and its success in terms of evolution from the community's state of knowledge at any given time? Does it really help to imagine that there is some one full, objective, true account of nature and that the proper measure of scientific achievement is the extent to which it brings us closer to that ultimate goal? If we can learn to substitute evolution-from-what-we-do-know for evolution-toward-what-we-wish-to-know, a number of vexing problems may vanish in the process. Somewhere in this maze, for example, must lie the problem of induction.

The introduction to Laudan's *Progress and its Problems* [1978] could serve as a manifesto for the diachronic school. He issues a direct challenge to theorists of synchronic rationality:

> Progress is an unavoidably *temporal* concept; to speak about scientific progress necessarily involves the idea of a process through time. Rationality, on the other hand, has tended to be viewed as an atemporal concept; . . . insofar as rationality and progressiveness have been linked at all, the former has taken priority over the latter—to such a degree that most writers see progress as *nothing more than* the temporal projection of a series of individual rational choices. . . . It will be the assumption here that we may be able to learn something by inverting the presumed dependence of progress on rationality.

Laudan conceives of theoretical progress as increased problem-solving effectiveness, defined by the number of important empirical problems solved minus the number of important anomalies generated. Alas, Laudan does not explain how to identify or count problems or anomalies. But what is more objectionable, on realist grounds, is his willingness to divorce problem-solving effectiveness from truth: "I do not even believe, let alone seek to prove, that problem-solving ability has any direct connection with truth or probabilities." That secures progress at too high a price. It is the goal of this dissertation to demonstrate that Laudan and Kuhn abandon the realist project too easily. More can be said in favor of our best inductive methodology, and designing the right notion of progress is essential to doing so.

A third diachronic tradition descends from Hans Reichenbach [1949] and Hilary Putnam [1965]. Reichenbach held that it was the goal of science to ascertain the probabilities with which various empirical outcomes occur. He advocated the 'straight rule' of induction, which, at every finite stage of inquiry, posits that the probability of an outcome is the relative frequency with which it has so far been observed to occur. Reichenbach justified the straight rule by proving that, on his conception of probability, the posits made by the rule are guaranteed to converge to the true probability, so long as any empirical method can. Putnam [1965] generalized Reichenbach's so-called 'pragmatic' justification, establishing a general framework in which to analyze the reliability of methods for investigating any empirical hypothesis. Suppose that $H_1, H_2, \ldots$ are all the possible answers to an empirical question that are taken to be serious possibilities. Putnam understood a method for investigating the question to be reliable iff, in every possible world consistent with the background assumptions of inquiry, the method converges to H_i iff H_i is true. The investigation of this notion, which came to be known as *logical reliability*, reaches a height of sophistication in Kelly and Glymour [1989], Kelly [1996] and Schulte [1999]. The requirement of logical reliability is non-trivial, but weak enough to be feasible in genuinely inductive inquiries, where Cartesian certainty, or even bounds on the objective chance of error, are impossible to guarantee. Furthermore, the realist goals of science are placed squarely at the center of the action: the goal of reliabilist methodology is to establish a connection between inquiry and the true answer to an empirical question.[5] Steel [2010] argues that logical reliabilism justifies inductive inference by giving a mathematical proof that it is a necessary condition of logical reliability. Since this proof does not rely on any empirical premises, it escapes Hume's circle, and gives a deductive means-ends justification of inductive inference.

Indeed, if a method does not *even* converge to the truth in the limit, it fails to attain realist goals of inquiry. The perennial criticism of logical reliabilism is that the connection between inquiry and the truth is too weak — limiting reliability is consistent with all kinds of arbitrary and irrational behavior in the short run. It is even consistent with the interruption of inquiry by arbitrarily many dark ages, as imagined by Miller Jr [1959] in *A Canticle for Liebowitz*. Therefore the reliabilist can make no significant methodological recommendations. Carnap makes an early version of this criticism when discussing Reichenbach's justification of the straight rule:

> Reichenbach is right in the assertion that any procedure which does not possess the characteristic described above (viz. approximation to the relative frequency in the whole) is inferior to his rule of induction. However, his rule . . . is far from being the only one possessing that characteristic. The same holds for an infinite number of other rules of induction, e.g. for Laplace's rule of succession . . . and likewise for the corresponding rule of our theory of c^* . . . . However, Reichenbach's rule and the other two rules mentioned yield different numerical values for the probability under discussion, although these values converge for an increasing sample towards the same limit. Therefore we need a more general and stronger method for examining and

[5] Note that the goal is not to converge to a true theory of everything, as Kuhn imagines, but piecemeal progress on contextually fixed empirical questions.

> comparing any two given rules of induction in order to find out which of them has more chance of success. [1945, p. 97].

Unfortunately, Carnap never proposes any way of comparing inductive methods that doesn't depend essentially on the choice of a prior probability distribution. A different idea is due to the heirs of Putnam: Schulte [1999] and Kelly and Glymour [2004]. Deductive methods have two desirable properties: they are infallible, given true premises, and they are monotonic, in the sense that conclusions are never withdrawn when more premises are added. Inductive methods, however, are irremediably fallible. They are also non-monotonic: more information often induces the withdrawal of previous conclusions. But even though monotonicity is not feasible in inductive problems, one can still strive to draw inductive conclusions *as monotonically as possible*, given the inherent difficulty of the inductive inference problem one is addressing. If perfect monotonicity is best, more monotonicity is better. Kelly, Glymour, and Schulte have investigating norms of maximal monotonicity that require methods to converge to the truth with as few mind changes, or retractions of opinion, as possible, given the inherent difficult of the problem. In a series of papers Kelly [2004, 2007, 2011] has also shown every method that minimizes mind changes en route to the truth satisfies a version of Ockham's razor. Kelly's results showed that a simplicity-guided strategy of conjectures and refutations is logically reliable and, furthermore, that every logically reliable and mind-change optimal strategy satisfies Ockham's razor. That demonstrates that the canonical biases of inductive methodology are necessary for *maximally monotonic* convergence to the truth. If progressive inquiry converges to the truth without unnecessary U-turns and vacillations along the way, then Kelly's results show that the norms of standard inductive methodology are necessary for progress. Norms of maximally monotonic convergence provide a "middle path" for justifying inductive practice: they are not so strong as to rule out inductive inference altogether, and, unlike mere convergence in the limit, they are strong enough to imply substantive methodological constraints on short-run behavior.

These ideas were expanded and developed in a series of papers [Genin and Kelly, 2015, Kelly et al., 2016, Genin and Kelly, 2018]. In all of these papers, the notion of simplicity receives a precise new topological formulation. In Genin and Kelly [2018], we develop two notions of maximally monotonic convergence that refine mind-change optimality. A sequence of conjectures A_1, A_2, A_3 is a *cycle sequence* if each one is incompatible with its predecessor and the last entails the first.[6] We prove that conjecturing some simplest answer in response to information is necessary for avoiding cycles on the way to convergence [2018, Proposition 9.1]. That shows that Ockham's razor is a necessary condition for avoiding doxastic cycles on the way to the truth. It also shows that every method that violates Ockham's razor must, on some information, conjecture the true answer, only to

[6] Avoidance of cycles is closely related to avoiding U-shaped learning, which has been explored by computational learning theorists cf. Carlucci et al. [2005], Carlucci and Case [2013].

drop it in favor of a false one, upon receiving additional true information. Ockham violators can be forced into these epistemic regressions only if the truth is complex, so the argument motivates preferring simplicity *especially* if the truth is complicated. Furthermore, for a broad class of problems, we construct methods that converge to the truth without cycles. That goes some way toward alleviating methodological pessimism: it is possible to guarantee that inquiry progress without perennial backsliding into darkness.

Reversal optimality is our second notion of maximal monotonicity. A sequence of conjectures $A_1, A_2, \ldots, A_n$ is a *reversal sequence* if for each i, A_{i+1} entails A_i^{c}. A reversal sequence is *forcible* in a problem if every method that converges to the true answer in the limit *must* output that sequence on some sequence of true and increasingly informative information states. The idea is that the forcible sequences, and only the forcible sequences, are the non-monotonicities that a method is justified in performing. For the first time, we give a prior-free topological characterization of all the forcible sequences in a given empirical problem [2018, Proposition 5.1]. That pins down exactly those non-monotonicities that an inductive method is justified in performing. We also show that patiently disjoining equally simple answers is necessary for avoiding retraction sequences that are not forcible [2018, Proposition 9.2].

Intuitively, progress requires that inquiry approach the truth monotonically, or, at least, as monotonically as possible. It is certainly not progressive to conjecture the truth, only to disavow it on receiving additional true information, as every Ockham-violator must do in some circumstances. Norms of maximal monotonicity answer to our intuitive demands on scientific progress, and the results of Kelly, Schulte, Genin et. al. show that the canons of scientific methdology are necessary for achieving these norms. Furthermore, they do so without begging the question, since they do not impose any prior probability distribution on the space of possibilities. That shows that there is some methodological advantage to scientific practice that is not shared by alternative modes of inquiry.

The preceding discussion suggests a systematic approach to questions of inductive methodology. First articulate a spectrum of success concepts of guaranteed efficient convergence to true answers to empirical questions. Then, characterize mathematically the empirical questions for which those success notions are achievable. Finally, justify methodological norms by proving that they are necessary for achieving corresponding success concepts across a wide range of empirical problems. That approach shares with Levi, Laudan and Kuhn a fundamentally *eritetic* stance—the goal is not a true theory of everything, but efficient inquiry guided by a question under discussion. However, it departs from the instrumentalism of Kuhn and Laudan by conceiving of progress as progress toward the *true* answer. It fulfills the realist demands of Popper and Lakatos by showing that progress toward the truth is not only possible, but guaranteed, so long as reliable methods are employed. Finally, it avoids begging the question in favor of canonical methods by demonstrating its results without imposing a potentially question-begging prior probability distribution. Popper wrote that "while we cannot ever have sufficiently good arguments in the empirical sciences for claiming that we have actually reached the truth, we can have strong and reasonably good arguments for claiming that we may have made progress toward the truth" [Popper, 1979, pp. 57-8]. We can do better: we have good arguments for claiming that science *will* make progress toward the truth.

Several characteristic features of the framework are worth calling attention to explicitly.

1. The framework foregrounds key epistemic features of the context of inquiry that are usually ignored.

   A context of inquiry specifies (i) a set of possible worlds (ii) a set of information states, constituting the possible inputs to inquiry, and (iii) a partition of the worlds into a set of answers, constituting the desired outputs of inquiry. Explicit attention to the kind of information that is available in the problem situation exposes the structure of empirical underdetermination, which is obscured by treating all propositions equally as undifferentiated elements of a field of sets on which probabilities are defined.

2. Topology, rather than logic or probability theory, is the fundamental formal tool.

   Verifiability from empirical information is a fundamental notion in methodology and the philosophy of science. The structure of verifiability is fundamentally topological: the verifiable propositions are exactly the open sets of a topological space. Therefore, topology is the most perspicuous setting for studying the structure of empirical inquiry.

3. Results are grounded in objective features of the context of inquiry and are therefore independent of a choice of language or prior probability distribution.

   Topological structure emerges from the kind of information that is available in the context of inquiry. Therefore, it is not imposed by subjective opinion, or linguistic convention.

4. Methodological norms are sensitive to the inherent difficulty of the problem.

   By solving for the inherent topological complexity of an empirical question it is possible to solve for the strongest feasible success criteria. Therefore it is possible to systematically impose strong norms where they are feasible, and relax them when they are not. That enables a principled and systematic right-sizing of epistemic demands to inherent topological complexity. One slogan for such a view is *means-ends epistemology* [Schulte, 1999] and another is *feasibility contextualism*.

5. Successful inquiry consists in efficient convergence to true answers to scientific questions.

   The goal of inquiry is not, as Kuhn imagines, a true theory of everything, but true answers to empirical questions. An answer to a question is not simply a solution to a puzzle, or a new problem-solving technique, but a true theory about the world.

6. Methods, rather than theories, are the locus of epistemic justification.

   A method is justified by demonstrating that it reliably and efficiently converges to the true answers to an empirical question. A methodological norm is justified by demonstrating that all methods violating the norm are less reliable or less efficient for a broad class of empirical questions.

Criticism of the position of Kelly and his students falls broadly into two categories. Previous results were obtained under the idealization that input information is propositional and logically refutes incompatible possibilities. Some critics express skepticism that the theory applies in settings where the data arrives in the form of random samples, and strict logical falsification never occurs. Elliot Sober is representative:

> Jeffreys and Popper both suggest that scientists should start by testing simpler theories; if those simpler theories fail, scientists should move to theories that are more complex. Schulte (1999) and Kelly (2007) also endorses this policy, arguing that it has a desirable sort of *efficiency.* They show that this strategy is optimal with respect to the goal of minimizing the number of times you will need to change your mind. It is important to see that Ockham's razor as a search procedure does not conflict with Ockham's razor as a principle for evaluating the theories at hand. . . . For Kelly and Schulte, the data tell you whether to accept or reject the theory at hand and this decision is made without help from Ockham's razor. According to their picture, theories are tested one at a time. This may be okay if the candidate theories you are considering are deductively related to observations, but when the relationship is probabilistic, I am skeptical of epistemologies that are non-contrastive [Sober, 2015].

Part of Sober's objection seems to be a misunderstanding. Ockham's razor, as defined in Genin and Kelly [2015], Kelly et al. [2016], Genin and Kelly [2018] is not a search procedure, but a normative constraint on theory choice. It is precisely a "principle for evaluating the theories at hand": never choose a theory if a strictly simpler one is compatible with current information. That maxim underdetermines a search procedure, since, in the not atypical situation where there is more than one simplest theory compatible with the data, it does not even determine what theory one should choose at a single stage of inquiry. On the other hand, Sober's worry about whether the justification works when data fail to ever logically falsify theory is well-placed. How can we even state the

Ockham constraint, stated in terms of theories compatible with current information, if all probabilistic theories are logically compatible with every possible random sample? That concern is dealt with decisively in this dissertation by generalizing the reliabilist framework to accomodate statistical theories and random samples. The strategy for executing that generalization is described in the subsequent sections of the introduction, but I foreshadow some of its novel consequences here.

The transition to the statistical framework allows for the articulation of new norms of progress. Previously, philosophers of science have tried to discern some mark by which one could judge whether a change from theory H_1 to H_2 is progressive. If both theories are false, perhaps the second is 'closer' to the truth, in some sense? The hope would be that scientific method is justified because it tends to produce theories that progress toward the truth. I propose a different, more direct, means of investigating the progressiveness of inductive method. Suppose, as before, that $H_1, H_2, \ldots$ are competing probabilistic theories of some phenomena under investigation. Say that a method is reliable iff, in every possible world consistent with the background assumptions of inquiry, the method converges *in probability* to H_i iff H_i is true. Say that a reliable method is *progressive* if, no matter which theory is true, the objective chance that it outputs the true theory is strictly increasing with sample size. In other words, the more data the scientist collects, the higher the chance that her method outputs the true theory. That is a sense in which a method can be progressive, even if it is meaningless to ask whether two of its successive outputs represent progress toward the truth. Progressiveness seems like a straightforwardly desirable property, but it is not always feasible. Nevertheless, it ought to be a regulative ideal that we strive to approximate. Say that a reliable method is *α-progressive* if, no matter which theory is true, the chance that it outputs the true theory never decreases by more than α. That property ensures that collecting more data cannot set your method back too badly. In what follows (Theorem 3.6.3) I prove that for typical problems, there exists a reliable, α-progressive method for every $\alpha > 0$. The method proceeds by a schedule of simplicity-guided conjectures and refutations. That demonstrates that a carefully calibrated Popperian methodology can ensure that the degree of backsliding is arbitrarily low. Furthermore, I prove that every α-progressive method must obey a probabilistic version of Ockham's razor (Theorem 3.6.4). That gives a new, prior-free justification for simplicity bias in statistical inquiry.

The second common objection to reliabilist justification of inductive method is that global features of the path of inquiry can give no good reason to believe the predictions of our current best theories. Here Fitzpatrick [2013a] is representative:

> Logical reliabilism . . . allows that we may have preferences for certain theories over others, given the available data—those selected by efficient logically reliable methods. Unlike Popperian falsificationism,

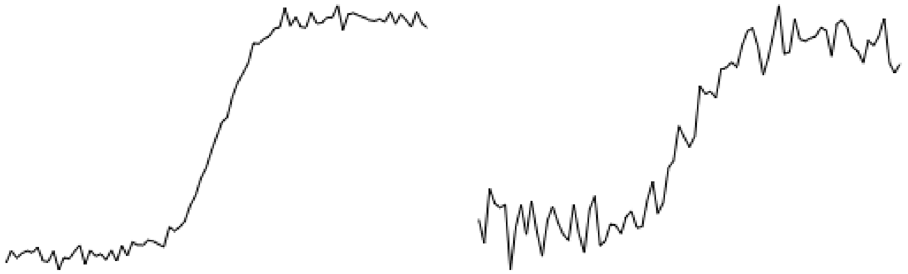


Figure 1.1: The chance of outputting the true theory, as sample size increases, for a .05-progressive and a .25-progressive method.

> it also offers a means-end justification for doxastic commitment to these theories. However, this warrant for belief obtains only relative to the goal of efficient long-run convergence to the truth. Such long-run strategic consideration are completely unrelated to the short-run predictive accuracy of the theories in question . . . Logical reliabilism thus cannot explain why it is rational to use our best theories in practical prediction, and it cannot underwrite the obvious confidence that scientists and ordinary folk have in the predictions of these theories (305).

I can think of two ways of answering such objections. The first way is to simply deny that there is anything more to underwriting our beliefs in the predictions of our best theory than justifying the methods used to arrive at the theory. Reliabilism shifts the locus of justification to method, rather than theory or prediction. Reliabilist analysis demonstrates the sense in which our methods can be made progressive, and proves that violating the canons of methodology makes inquiry vulnerable to unnecessary backsliding into error. That does not prove Hume wrong, or give reason to believe that our best theories must be true, or that their predictions are accurate. Any such reasons must appeal to substantive assumptions, even if they are disguised as subjective probability distributions. Without appealing to such assumptions, logical reliabilism gives a non-trivial, non-question-begging justification of our best inductive methods. The reason to rely on the predictions issued by our best theories is just this: we made them from the theories output by our best methods, which are justified by a reliability analysis.

The second way engages more thoroughly with the eritetic and contextual features of reliabilist analysis. Part of the rhetorical strength of Fitzgerald's objection is due to a subtle shift of the question in context. Settings in which questions of theory choice are salient are different from those in which questions of prediction are salient. A thorough-going contextualism is sensitive to this shift in question, and recommends that one use the most reliable methods to answer the question at hand. That raises the possibility that reliable methods for answering theory-choice questions are not necessarily the most reliable for questions of short-run prediction. In frequentist statistics that observation goes under the slogan that "all models are false," which is taken to mean that one ought not even try to identify true models, and attempt instead only to minimize prediction error.[7] For that reason, frequentists recommend methods like Akaiake's Information Criterion (AIC), which are not guaranteed to converge to the true model even in the limit of infinite data. That defect is supposedly made up for by the fact that the methods guarantee good predictive accuracy on future data [Forster and Sober, 1994]. There are many reasons to doubt this cover story. There is no prior-free argument demonstrating that AIC-type methods are guaranteed to have better predictive accuracy than naive alternatives, even in expectation [Forster, 2002, Leeb and Pötscher, 2005, Kelly, 2011]. The best arguments for these methods suggest that they may minimize "risk-inflation" [Foster and George, 1994], or decision-theoretic regret [Droge, 1998], but fall short of demonstrating that the methods are more predictively accurate. Even if some argument can be made without begging the question, it would apply only to passive (non-interventional) prediction on future samples taken from the same generating distribution used to estimate the prediction methods. That is a very attenuated sense of prediction, which does not capture the robust sense of prediction prevalent in sciences that aim to predict the results of interventions. In fact, there is good reason to believe that it is *impossible* to accurately estimate the effects of interventions without getting the causal model right, especially when inferences are made from non-experimental data [Kelly and Mayo-Wilson, 2010]. In difficult causal inference problems, there is simply no better way to get good predicative accuracy than to get the model right — methods that efficiently identify the correct model are therefore uniquely justified by a reliabilist analysis. For these reasons, causal inference from observational data is the premier example of Chapter 3.

1.2 Statistical Verifiability and Falsifiability

> The relations between probability and experience are also still in need of clarification. In investigating this problem we shall discover what will at first seem an almost insuperable objection to my methodological views. For although probability statements play such a vitally important role in empirical science, they turn out to be in principle impervious to strict falsification. Yet this very stumbling block will become a touchstone upon which to test my theory, in order to find out what it is worth [Popper, 1959, p. 133].

The framework of logical reliabilism, as articulated in Kelly [1996], and more

[7] The slide from "theory" to "model" reflects the instrumentalist perspective that models are not to be *inferred*, but simply *used* for prediction.

recently in Baltag et al. [2016], and Genin and Kelly [2015], Kelly et al. [2016], and Genin and Kelly [2018], relies heavily on the notions of verification and falsification. That suggests that logical reliabilism does not have the conceptual resources to address actual scientific inference, where strict logical falsification hardly ever occurs. That is the core of Sober's [2015] worry about theories that make only probabilistic predictions. As the epigraph to this section shows, Popper was also well aware that probabilistic theories posed an equal challenge for his falsificationism. Nevertheless, there are strong analogies between 'naive' falsificationism, and standard practice in frequentist hypothesis testing. For example, a standard statistical test of a sharp null hypothesis rejects when the data are very improbable under the null hypothesis. Such a procedure has a very low chance of rejecting the null hypothesis in error. That falls slightly short of, but is closely analogous to, the infallibility of rejecting a universal law when it is refuted. Alternatively, a standard statistical test of a simple or sharp statistical hypothesis has an arbitrarily high chance of accepting it in error. That is similar to the fallibility of inferring a universal hypothesis from finitely many instances. Such analogies are natural and sometimes explicit in statistics. For example [Gelman and Shalizi, 2013]:

> . . . the hypothesized model makes certain probabilistic assumptions, from which other probabilistic implications follow deductively. Simulation works out what those implications are, and tests check whether the data conform to them. Extreme p-values indicate that the data violate regularities implied by the model, or approach doing so. If these were strict violations of deterministic implications, we could just apply *modus tollens* to conclude that the model was wrong; as it is, we nonetheless have evidence and probabilities. Our view of model checking, then, is firmly in the long hypothetico-deductive tradition, running from Popper (1934/1959) back through Bernard (1865/1927) and beyond (Laudan, 1981).

Statistical falsification, Gelman and Shalizi suggest, is *all but deductive*.[8] But how extremal, exactly, does a p-value have to be for a test to count as a falsification? Popper was loathe to draw the line at any particular value:

> . . . a physicist is usually quite well able to decide whether he may for the time being accept some particular probability hypothesis as 'empirically confirmed', or whether he ought to reject it as 'practically falsified' . . . It is fairly clear that this 'practical falsification' can be obtained only through a methodological decision to regard

[8] Some frequentists go even further. In their response to the American Statistical Association's controversial statement on p-values, Ionides et al. [2017] distinguish deductive reasoning "based on deducing conclusions from a hypothesis and checking whether they can be falsified" and inductive reasoning "which permits generalization, and therefore allows data to provide direct evidence for the truth of a scientific hypothesis." Furthermore, they write, "it is held widely . . . that only deductive reasoning is appropriate for generating scientific knowledge. Usually, frequentist statistical analysis is associated with deductive reasoning and Bayesian analysis is associated with inductive reasoning."

> highly improbable events as ruled out ... But with what right can they be so regarded? Where are we to draw the line? Where does this 'high improbability' begin? [Popper, 1959, p. 182]

I suggest that such embarrassing questions can be avoided if, instead of asking 'what counts as a statistical falsification?', we ask 'which hypotheses are statistically falsifiable?'. Consider the archetypical examples of falsifiable hypotheses: universal hypotheses like 'all ravens are black', or co-semidecidable formal propositions like 'this program will not halt in a finite number of steps'. Although there is no *a priori* bound on the amount of observation, computation, or proof search required, these hypotheses may be falsified by suspending judgement until the relevant hypothesis is decisively refuted by the provision of a non-black raven, a halting event, or a valid proof. I want to call attention to several properties of paradigmatic falsification methods.

Infallibility: Output conclusions are true.

By suspending judgement until the hypothesis is logically incompatible with the evidence, falsifiers never have to 'stick their neck out' by making a conjecture that might be false.

Monotonicity: Logically stronger inputs yields logically stronger conclusions.

Typical falsifiers never have to retract their previous conclusions; their conjecture at any later time always entails their conjecture at any previous time. In the ornithological context, conjectures made on the basis of more observations always entail conjectures on made on the basis of fewer—once a non-black raven has been observed, the hypothesis is decisively falsified. In the computational context, conjectures made on the basis of more computation always entail conjectures made on the basis of less—once the program has entered a halting state, it will never exit again.

Limiting Convergence: The method converges to $\neg H$ iff H is false.

If all ravens are black, the falsifier may suspend judgement forever; but if there is some non-black raven, diligent observation will turn up a falsifying instance eventually. Similarly, if the program eventually halts, the patient observer will notice.

I propose that statistical verifiability and falsifiability will be found if we look for the minimal weakening of these paradigmatic properties that is feasible in statistical contexts. First, some definitions.[9] *Inference methods* output conjectures on the basis of input information. Here 'information' is understood broadly. One conception of information, articulated explicitly by Bar-Hillel and Carnap [1953] and championed by Floridi [2005, 2011], is true, propositional

[9] The definitions in this section are intentionally rather schematic. Hopefully this will aid, rather than hinder, comprehension. All definitions are formalized in subsequent chapters.

semantic content, logically entailing certain relevant possibilities, and logically refuting others. Call that the *propositional* notion of information. Propositional information is the standard notion in modal and epistemic logic as well as many related formal fields. A second conception, ubiquitous in the natural sciences, is random samples, typically independent and identically distributed, and logically consistent with all relevant possibilites, although more probable under some, and less probable under others. Call that the *statistical* notion of information.

Statistical methods cannot be expected to be infallible. We liberalize that requirement as follows:

> *α-Infallibility*: For every sample size, the objective chance that the output conclusion is false is bounded by α.

The α-infallibility property is closely related to frequentist statistical inference. A confidence interval with coverage probability $1 - \alpha$ is straightforwardly α-infallible: the chance that the interval excludes the true parameter is bounded by α. A hypothesis test with significance level α is also α-infallible, so long as one understands failure to reject the null hypothesis as recommending suspension of judgment, rather than concluding that the null hypothesis is true. The chance of falsely rejecting the null is bounded by α, and failing to reject outputs only the trivially true, or tautological, hypothesis.

We first define a weaker notion. A method *verifies hypothesis H in the limit* by converging, on increasing information, to H, iff H is true. On the propositional conception of information, a method $L(\cdot)$ converges on increasing information to H iff in all possible worlds, there is some true information E, such that on any logically stronger true information F entailing E, $L(F)$ entails H. On the statistical conception of information, a method $L(\cdot)$ converges on increasing information to H if in any possible world w, the chance that $L(\cdot)$ outputs H in w at sample size n converges to 1 as sample size increases. A method *refutes H in the limit* if it verifies not-H in the limit. A method *decides H in the limit* if it verifies H and not-H in the limit. Hypothesis H is verifiable, refutable, or decidable in the limit iff there exists a method that verifies, refutes or decides it in the limit.

Verification *in the limit* is a relatively undemanding concept of success — it is consistent with any finite number of errors and volte-faces prior to convergence. Verification is a stronger success notion. Consider the following cycle of definitions.[10]

[10] The following are only proto-definitions, leaving many things unspecified. The notion of verification in the limit is here a free parameter: compatible notions include convergence in propositional information, convergence in probability and almost sure convergence. The notion of α-infallibility is also parametric: it can mean that the chance of error at any sample size is bounded by α, or that the sum of the chances of error over all sample sizes is bounded by α. Each of these concepts will be developed in detail in the following. These parametric details are omitted here to expose the essential differences between the three concepts.

V1. Hypothesis H is verifiable iff there is a monotonic, infallible method M that verifies H in the limit.

V1.5 Hypothesis H is verifiable iff there is a method M that verifies H in the limit, and for every $\alpha > 0$, M is α-infallible.

V2. Hypothesis H is verifiable iff for every $\alpha > 0$ there is an α-infallible method that verifies H in the limit.

Concept V1 is the familiar one from epistemology, the philosophy of science, and the theory of computation. Our motivating examples are all of this type. These may be falsified by suspending judgement until the relevant hypothesis is logically refuted by information. Although there is no *a priori* bound on the amount of information (or computation) required, the outputs of a verifier are guaranteed to be true, without qualification.

Concept V1.5 weakens concept V1 by requiring only that there exist a method that is infallible with probability one. Hypotheses of type V2 are less frequently encountered in the wild.[11] Concept V1.5 is introduced here to smooth the transition to V2.

V2 weakens V1.5 by requiring only that for every bound on the chance of error, there exists a method that achieves the bound. Hypotheses of this type are ubiquitous in statistical settings. The problem of verifying that a coin is biased by flipping it indefinitely is an archetypical problem of the third kind. For any $\alpha > 0$ there is a consistent hypothesis test with significance level α that verifies, in the third sense, that the coin is biased. Moreover, it is hard to imagine a more stringent notion of verification that could actually be implemented in digital circuitry. Electronics operating outside the protective cover of Earth's atmosphere are often disturbed by space radiation—energetic ions can flip bits or change the state of memory cells and registers [Niranjan and Frenzel, 1996]. Therefore, even routine computations performed in space are subject to non-trivial probabilities of error, although the error rate can be made arbitrarily small by redundant circuitry, error-correcting codes, or simply by repeating the calculation many times and taking the modal result. Electronics operating on Earth are less vulnerable, but are still not immune to these effects.

Concept V2 provides only a partial statistical analogue for V1, since issues of monotonicity are ignored. A statistical analogue of monotonicity is suggested by considerations of replication. Consider the following situation. A group of

[11] For a contrived example, suppose it is known that random samples are distributed uniformly on the interval $(\mu - 1/2, \mu + 1/2)$, for some unknown parameter μ. Although samples may land outside the interval, they only do so with probability zero. Let H be the hypothesis that the true parameter is not μ. Let M be the method that concludes H if some sample lands outside of the interval $(\mu - 1/2, \mu + 1/2)$, and draws no non-trivial conclusion otherwise. Then, M is a deductive verifier of the second type, although not of the first. Clearly, every verifier of the first type is a verifier of the second type.

researchers propose to investigate whether Drug A is better at treating migraine than conventional treatments. Before receiving the funding, the researchers do an analysis of their statistical method and conclude that if Drug A is better than the conventional alternative, the chance that their method rejects the null hypothesis of no effect is greater than 50%. The funding agency is impressed, providing enough funding to perform a pilot study at sample size 100. Elated, the researchers perform the study, and correctly reject the null hypothesis. Now suppose a replication study is proposed at sample size 150, but the objective chance of rejecting has decreased to 40%. That means that the chance of rejecting correctly, and thereby replicating successfully, has gone down, even though the first study was correct, and investigators propose going to the trouble and expense of collecting a larger sample! Such methods are epistemically defective, and more monotonic methods ought to be preferred. Accordingly, consider the following statistical norm

> *Monotonicity in chance*: If H is true, then the objective chance of outputting H is strictly increasing with sample size.

Unfortunately, strict monotonicity is often infeasible. Nevertheless, it should be a regulative ideal that we strive to approximate. The following principle expresses that aspiration:

> *α-Monotonicity in chance*: If H is true, then for any sample sizes $n_1 < n_2$, the objective chance of outputting the true answer does not decrease by more than α.

That property ensures that collecting a larger sample is never a disastrously bad idea. Equipped with a notion of statistical monotonicity, we state the final definition in our cycle:

V3. Hypothesis H is verifiable iff for every $\alpha > 0$ there is an α-infallible and α-monotonic method that verifies H in the limit.

V3 seems like a rather modest strenghtening of V2. Surprisingly, many standard frequentist methods are α-infallible, but not α-monotonic. Chernick and Liu [2002] noticed non-monotonic behavior in the power function of standard hypothesis tests of the binomial proportion, and proposed heuristic software solutions. That defect would have precisely the bad consequences that inspired our statistical notion of monotonicity: attempting replication with a larger sample might actually be a bad idea! That issue has been raised in consumer safety regulation, vaccine studies, and agronomy [Schuette et al., 2012, Musonda, 2006, Schaarschmidt, 2007]. But Chernick and Liu [2002] have only noticed the tip of the iceberg—similar considerations attend all statistical inference methods. One of the results of this dissertation (Theorem 3.3.1) is that V3 is feasible whenever V2 is.

1.3 The Topology of Inquiry

> geometric logic [topology] is the logic of finite observations [Abramsky, 1987].

A central insight of Abramsky [1987], Vickers [1996], Kelly [1996] is that verifiable propositions of type V1 enjoy the following properties:

T1. If H_1, H_2 are verifiable, then so is their conjunction, $H_1 \cap H_2$.

T2. If $\mathcal{H}$ is a (potentially infinite) collection of verifiable propositions, then their union, $\cup\mathcal{H}$, is also verifiable.

Together, T1 and T2 say that verifiable propositions of type V1 are closed under disjunction, and *finite* conjunction. Why the asymmetry? For the same reason that it is possible to verify that bread will continue to nourish for any finite number of days in the future, but not possible to verify that it will nourish forever. It is also important to notice what T1 and T2 *do not* say: if H is verifiable, its logical complement may not be. To convince yourself of this it suffices to notice that it is possible to verify that bread will cease to nourish someday, but not that it will continue to nourish forever. Jointly, T1 and T2 express the fact that the collection of all propositions of type V1 constitute the *open sets* of a *topological space.*[12] Sets of greater topological complexity are formed by set-theoretic operations on open sets. The central point of Kelly [1996] is that degrees of methodological success correspond exactly to increasingly ramified levels of topological complexity, corresponding to elements of the *Borel hierarchy:*

> The Borel hierarchy is a system of mathematical cubbyholes [that] provide a kind of shipshape mathematics, in which there is a place for everything, and everything is put in its place. Each cubbyhole reflects a kind of intrinsic, mathematical complexity of the objects within it. . . . The striking fact is that methodological success can be characterized *exactly* in terms of [these] cubbyholes [Kelly, 1996].

The open sets are exactly the verifiable hypotheses; complements of open sets, called *closed sets*, are exactly the refutable hypotheses; sets that are both open and closed, called *clopen sets*, are exactly the decidable hypotheses. Higher levels of topological complexity correspond to *inductive* notions of methodological success.[13] *Locally closed* sets are sets that can be expressed as an intersection of an open and a closed set.[14] Countable unions of locally closed sets, known as

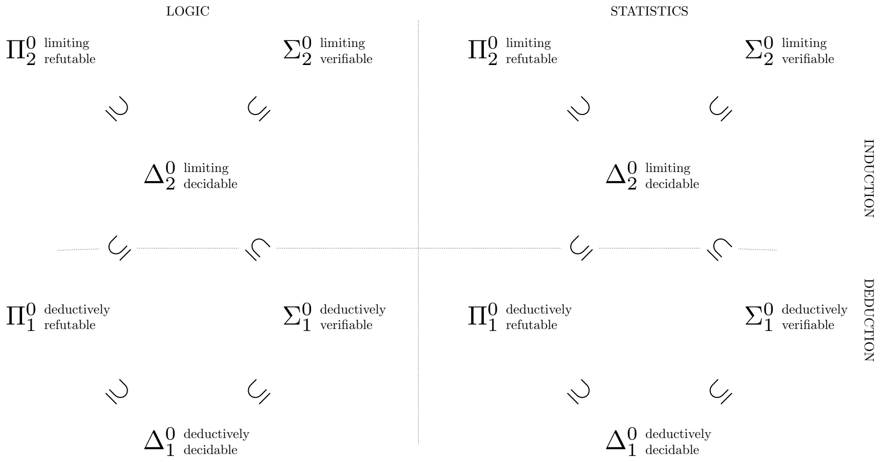
[12] A topological space $\mathcal{T}$ is a structure $\langle W, \mathcal{V}\rangle$ where W is a set, and $\mathcal{V}$ is a collection of subsets of W closed under disjunction and finite conjunction. The elements of $\mathcal{V}$ are called the open sets of $\mathcal{T}$.

[13] By inductive success, I mean any notion where the chance of error is unbounded in the short run.

[14] Although the epistemic interpretation of locally closed sets is significant, we introduce them here only as a building block for sets of higher complexity. Discussion is postponed until Section 2.3.

Σ^0_2 sets, are exactly the hypotheses verifiable in the limit. Their complements, known as Π^0_2 sets, are exactly the hypotheses refutable in the limit. Sets which are both Σ^0_2 and Π^0_2 are decidable in the limit.

Figure 1.2: Pictured below is a hierarchy of topological complexity and corresponding notions of methodological success. The set of all open sets is referred to as Σ^0_1; the set of all closed sets as Π^0_1; and the set of all clopen sets as $\Delta^0_1 = \Sigma^0_1 \cap \Pi^0_1$. Depending on whether the Σ^0_1 sets are propositions of type V1 or V2, we get the logical (left) and statistical (right) hierarchies. Sets of greater complexity are built out of Σ^0_1 sets by logical operations, e.g. Σ^0_2 sets are countable unions of locally closed sets. Inclusion relations between notions of complexity are also indicated.


LOGIC
STATISTICS
Π^0_2 limiting refutable
Σ^0_2 limiting verifiable
Π^0_2 limiting refutable
Σ^0_2 limiting verifiable
Δ^0_2 limiting decidable
Δ^0_2 limiting decidable
INDUCTION
Π^0_1 deductively refutable
Σ^0_1 deductively verifiable
Π^0_1 deductively refutable
Σ^0_1 deductively verifiable
DEDUCTION
Δ^0_1 deductively decidable
Δ^0_1 deductively decidable

The preceding sketches a general complexity theory for empirical inquiry. The theory is extended straightforwardly from individual hypotheses, to empirical questions. Given the topological complexity of the answers, the problem of finding the true answer to the question has an intrinsic difficulty, in that some questions allow one to find the truth in a very strong sense and others only in weaker senses. Deductive questions, in which every answer is clopen, have the property that one can eventually infer the true answer to the question, without ever risking error. Some inductive questions have the property that one can converge to belief in the true answer, after some reversals of opinion along the way, even though one cannot do so with no reversals of opinion. So inductive questions are, in a definite sense, intrinsically harder than deductive questions.

That perspective gives rise to a general view we call *feasibility contextualism*. According to feasibility contextualism, an inferential strategy is epistemically justified insofar as it achieves the best achievable truth-finding performance for the given question, in light of its inherent topological complexity. Weaker connections to the truth are justified when stronger connections are impossible. In particular, inductive methods are justified when deductive methods are impossible. Thus, *pace* philosophical tradition, the infeasibility of deductive solutions to inductive questions does not undermine the justification of inductive inferences. It is, rather, what justifies them. When deductive solutions are available, they should be used. When they aren't, the best one can do is to fall back on inductive methods. Therefore, feasibility contextualism provides a compelling response to inductive skepticism. To reject inductive inference when it is required is to reject the best possible means for finding the truth out of preference for standards that cannot be realized. One might insist that the best possible means are not good enough, but that position is hardly inevitable. It is at least as reasonable to respond that the best is good enough. Feasibility contextualism is not a new idea — it is just business as usual in theoretical computer science, where an algorithm is justified by showing that its efficiency approaches the maximum feasible efficiency for the problem it is intended to solve. As above, it is the negative results showing that the problem cannot be solved in a better way that do the heavy normative lifting.

Taking verifiability of type V1 as the fundamental notion, the perspective sketched above was worked out in its essentials by Kelly [1996] and further generalized by de Brecht and Yamamoto [2009], Genin and Kelly [2015] and Baltag et al. [2016]. But it is not difficult to prove that verifiable propositions of type V2 also satisfy T1 and T2.[15] Therefore, the structure of statistically verifiability is *also* topological. The characteristic asymmetries are all present: while it is possible to verify that the true bias of the coin lies in the interval $(.5 - 1/n, .5 + 1/n)$ for any n, it is not possible to verify that the coin is exactly fair. Nevertheless, it is possible to verify that the coin is *not* fair. Figure 1.3 illustrates how we can build an analogous statistical hierarchy of methodological success by taking verifiability of type V2 (or V3) as the fundamental notion. One of the major achievements of this work is a systematic lifting of the results of Kelly [1996] to the statistical setting.

In Genin and Kelly [2017] we exhibit a topology on probability measures in which the open sets are exactly the propositions verifiable (in the sense of V2) from

[15] It is somewhat more difficult to show the same for propositions of type V3. This is the work of Section 3.3.

random samples. In statistical terminology, our main result provides necessary and sufficient conditions for the existence of a Chernoff consistent hypothesis test. Although there is extensive statistical work on pointwise consistent hypothesis testing, we are unaware of any analogous result. We also characterize topologically the statistical questions for which there exists a pointwise consistent method. We take these results to be novel contributions to mathematical statistics, but the real benefit to statistical methodology is twofold. Firstly, disciplining statistical practice with a complexity theory has the salutory effect of decisively forestalling wishful thinking. Everyone would like methods with a guaranteed bound on the chance of error for their favorite inference problem. Methods that merely converge to the truth in the limit without guaranteed bounds on the chance of error are sometimes scoffed at, although better ones are not exhibited. A topological analysis identifies the best possible sense in which a problem is solvable, and whether guaranteed bounds on error, or even limiting convergence, is feasible. No one should expect solutions to genuinely inductive problems with guaranteed error bounds. Secondly, frequentist statistics as classically formulated by Fisher, Neyman and Pearson, and recently rearticulated by Mayo and Cox [2006], licenses inferences only when the chance of error can be bounded. That is essentially an expression of inductive skepticism, since it is possible to bound the chance of error only for relatively simple problems. Complexity theory allows us to face the problem of induction squarely and honestly, and articulate norms of success that are feasible for the difficult problems scientists face. In light of the fundamental bridge results of Genin and Kelly [2017], we can forge new norms of maximally monotonic convergence in concrete inference problems that arise in statistics and machine learning. That lays the foundation for a new normative program for frequentist statistics, in which considerations of progressiveness play a central role.

1.4 Plan of the Work

The remainder of this dissertation is divided into two chapters. Chapter 2 presents an overview of learning from propositional information. This chapter attempts to summarize quickly and painlessly the relevant results from previous work in the propositional framework [Genin and Kelly, 2015, Kelly et al., 2016, Genin and Kelly, 2018]. Since many issues are set into sharper relief when abstracting from statistical complications, the reader who is interested in a systematic development is invited to begin here. The reader familiar with the basic outlines of the learning theoretic approach, as well as those that are in a hurry, can safely skip to Chapter 3. Chapter 3 contains all of the substantive original contributions made in this document. Advancing from the developments in Genin and Kelly [2017], it develops a point-by-point statistical analogue of the framework outlined in Chapter 2. For a thorough outline of what I take to be its novel contribution, see the beginning of Chapter 3.

Chapter 2

Learning from Propositional Information

2.1 The Propositional Setting

To clarify the intended analogy with statistics, we briefly introduce the setting of propositional information. The impatient reader can proceed to the next section. More expansive developments of the results in this section appear in de Brecht and Yamamoto [2009], Genin and Kelly [2015], and Baltag et al. [2016].

Let W be a set of possible worlds, or epistemic possibilities one takes seriously, consistent with the background assumptions of inquiry. A proposition is identified with the set of worlds in which it is true, so propositions are subsets of W. Let P, Q be arbitrary propositions. P is true in w iff $w \in P$. Logical operations correspond to set-theoretic operations in the usual way: $P \cap Q$ is conjunction, $P \cup Q$ is disjunction, $P^{\mathsf{c}} = W \setminus P$ is negation, and $P \subseteq Q$ is deductive entailment of Q by P. Finally, P is deductively valid iff $P = W$ and is deductively contradictory iff $P = \varnothing$.

In the setting of propositional information, *information states* are propositions that rule out relevant possibilities. For every w in W, let $\mathcal{I}(w)$ be the set of all information states true in w (i.e. which contain w as an element). It is assumed that $\mathcal{I}(w)$ is non-empty — at worst, one receives the trivial information W. That motivates the following axiom:

Axiom I.1 $\mathcal{I}(w) \neq \varnothing$.

The information states in $\mathcal{I}(w)$ are interpreted not merely as information that might, with luck, be afforded in w. Rather, they reflect what *will* be afforded, eventually, to a diligent inquirer, in the sense that for each information state $E \in \mathcal{I}(w)$, there is a stage of inquiry after which the total information entails E. That assumption reflects the normative requirement that all scientific information can be replicated with sufficient diligence. Let E, F be two information states in $\mathcal{I}(w)$. If the total information in w eventually entails each of E and F, then it must eventually entail them both. That motivates the following:

Axiom I.2 For each E, F in $\mathcal{I}(w)$, there exists G in $\mathcal{I}(w)$ such that $G \subseteq E \cap F$.

Finally, we assume that $\mathcal{I}$ is countable, since any language in which the data are recorded is at most countably infinite.

Axiom I.3 $\mathcal{I}$ is countable.

The following examples help to fix ideas.

Example 2.1.1. *Let W be the set of all infinite binary sequences. Each world w determines an infinite sequence of observable outcomes. Let $w|_n$ be the initial segment of w of length n. Let $[w|_n]$ be the set of all worlds having $w|_n$ as an initial segment. Let $\mathcal{I}(w)$ be the set of all $[w|_n]$ for every n. Think of the length of the initial segment observed as the "stage" of inquiry. There is exactly one such information state in w at every stage, and $[w|_n]$ is entailed by $[w|_m]$ for every $m \geq n$.*

Example 2.1.2. *Let W be the set of all real numbers. Think of the possible "stage-n" information states in w as the open intervals of width $1/2^n$ that contain w. Then $\mathcal{I}(w)$ is the set of all intervals containing w of width $1/2^n$, for some natural number n. It follows that for every $E \in \mathcal{I}(w)$ there is a stage n such that every stage-n information state in $\mathcal{I}(w)$ entails E.*

Define $\mathcal{I} = \cup_w \mathcal{I}(w)$. We call the structure $(W, \mathcal{I})$ an *information basis.* The set

$$\mathcal{I}(w|E) = \{F \in \mathcal{I}(w) : F \subseteq E\}$$

is the set of information states that might be presented in w from E onwards. It is straightforward to check that the restriction of $\mathcal{I}$ to E, $\mathcal{I}|_E = \cup_w \mathcal{I}(w|E)$, satisfies Axioms I.1-3 relative to the background space of possibilities E. Therefore, $(E, \mathcal{I}|_E)$ determines the information basis relative to E.

Information state E *verifies* proposition H iff E entails H. Information state E *refutes* H iff it verifies H^{c}. Information state E *decides* H iff it either verifies or refutes H. We now introducte some topological operators, and their epistemological interpretations. The *interior* of a proposition H, denoted $\mathsf{int}(H)$, is the set of all worlds w, such that there is $E \in \mathcal{I}(w)$ verifying H. Hence, $\mathsf{int}(H)$ is the set of worlds in which H is eventually verified by information. The *exterior* of a proposition H, denoted $\mathsf{ext}(H)$, is the set of all worlds in which H will be refuted, i.e. $\mathsf{int}(H^{\mathsf{c}})$. The *closure* of H, denoted $\mathsf{cl}(H)$, is the set of all worlds in which H will never be refuted by information, defined by $(\mathsf{ext}H)^{\mathsf{c}}$. Of course, H is never refuted if H is true. The worrisome possibility is if H is never refuted in w even though H is false. Then, a Popperian might say that H poses the *problem of metaphysics* in w. The proposition that H poses the problem of metaphysics is called the *frontier* of H, defined by: $\mathsf{frnt}(H) = \mathsf{cl}(H) \cap H^{\mathsf{c}}$. Again, following Popper, hypothesis H poses the *problem of induction* in w iff H is true, but will never be verified. So the proposition that H poses the problem of induction is just $\mathsf{frnt}(H^{\mathsf{c}})$. The proposition that H will never be decided is called the *boundary* of H, defined by: $\mathsf{bdry}\, H \;=\; \mathsf{frnt}\, H \cup \mathsf{frnt}\, H^{\mathsf{c}}$. The frontier of H and the frontier of H^{c} partition the boundary into the problem of metaphysics and the problem of induction, respectively—the two fundamental problems with which Popper begins *The Logic of Scientific Discovery.* The following key translates between the topological operators and the propositions to which they correspond.

$$\begin{aligned}
\mathsf{int}(H) &\equiv \; H \text{ will be verified;}\\
\mathsf{ext}(H) &\equiv \; H \text{ will be refuted;}\\
\mathsf{cl}(H) &\equiv \; H \text{ will not be refuted;}\\
\mathsf{bdry}(H) &\equiv \; H \text{ will not be decided;}\\
\mathsf{frnt}(H) &\equiv \; H \text{ is false and will not be refuted;}\\
\mathsf{frnt}(H^{\mathsf{c}}) &\equiv \; H \text{ is true and will not be verified.}
\end{aligned}$$

Proposition H is *open* iff $H \subseteq \mathsf{int}H$, i.e., if H entails that H will be verified. Proposition H is *closed* iff $\mathsf{cl}H \subseteq H$, i.e. if H^{c} entails that H^{c} will be verified. It is clear from the definition that H is closed iff H^{c} is open. Proposition H is *clopen* iff H is both open and closed. The following summarizes the above correspondences.

$$\begin{aligned}
H \text{ is open } &\equiv \; H \text{ entails that } H \text{ will be verified;}\\
H \text{ is closed} &\equiv \; H^{\mathsf{c}} \text{ entails that } H \text{ will be refuted;}\\
H \text{ is clopen} &\equiv \; H \text{ will be decided.}
\end{aligned}$$

Proposition H is closed iff it does not pose the problem of metaphysics. Dually, H is open iff it does not pose the problem of induction.

Lemma 2.1.1. *H is closed iff* $\mathsf{frnt}(H) = \varnothing$.

Proof. H is closed iff $\mathsf{cl}(H) \subseteq H$ iff $\mathsf{cl}(H) \cap H^{\mathsf{c}} = \varnothing$ iff $\mathsf{frnt}(H) = \varnothing$. □

It is not difficult to show that if H is open, it is a countable union of information states in $\mathcal{I}$. Let $\mathcal{T}$ be the closure of $\mathcal{I}$ under arbitrary unions. The elements of $\mathcal{T}$ are exactly the open sets. Readers familiar with topology will have already noticed that Axioms I.1-3 guarantee that $\mathcal{I}$ is a topological basis, and therefore, that $(W, \mathcal{T})$ is a topological space. We call $\mathcal{T}$ the *information topology* on W. The open sets of $\mathcal{T}$ are governed by the following axioms:

Axiom T.1 Any (finite or infinite) union of elements of $\mathcal{T}$ belongs to $\mathcal{T}$;

Axiom T.2 The intersection of finitely many members of $\mathcal{T}$ belongs to $\mathcal{T}$.

By the axioms it follows straightforwardly that the closed sets are closed under (finite or infinite) intersections, and finite unions.

2.2 Propositional Verification, Refutation and Decision

A *method* $L : \mathcal{I} \to \mathcal{P}(W)$ is a function from information states to propositions. Method L is *non-ampliative* iff it never leaps beyond the information, i.e. if $E \subseteq L(E)$ for all E in $\mathcal{I}$. Method L is *infallible* iff its output is always true, i.e. iff $w \in L(E)$, for all $E \in \mathcal{I}(w)$. Infallibility and non-ampliativity are equivalent.

Lemma 2.2.1. *Method L is infallible iff L is non-ampliative.*

Proof. Left to right. Suppose that L is not non-ampliative. Then there is $E \in \mathcal{I}$, such that $w \in E \setminus L(E)$. So $L(E)$ is false in w, although $E \in \mathcal{I}(w)$. Therefore, L is not infallible. Right to left. Suppose that L is non-ampliative. Then, for all $E \in \mathcal{I}(w)$, we have that $w \in E \subseteq L(E)$. Therefore, L is infallible. □

Say that method L is *monotonic* iff $L(F) \subseteq L(E)$ whenever $F \subseteq E$. Monotonicity and infallibility are logically independent. To see that infallibility does not imply monotonicity, notice that an infallible method with $L(E) = E$ can always retract to the trivial information state W on further information. To see that monotonicity does not imply infallibility, notice that the method that always outputs the incoherent proposition $\varnothing$ is monotonic.

A method *converges to proposition* A in w, iff there is $E \in \mathcal{I}(w)$ such that $L(F) \subseteq A$ for all $F \in \mathcal{I}(w|E)$. Say that method L is a *verification method* for H iff it is an infallible method that converges to H in the limit, if H is true. That is, L is a verifier of H iff

INFAL. L is infallible;

LIMCON. L converges to H, if H is true.

Say that H is *verifiable* iff there exists a verifier for H. A method L is a refutation method for H iff it is a verification method for H^{c}. Say that H is *refutable* iff there exists a refutation method for H. A method L is a decision method for H iff it is a verification method for H and H^{c}. A proposition is *decidable* iff there exists a decision procedure for H. A monotonic strengthening of these concepts immediately suggests itself. Say that H is *monotonically verifiable/refutable/decidable* iff there exists a monotonic verification/refutation/decision method for H.

For an illustration, suppose you are observing a computation by an unknown program; it is verifiable that that the program will halt at some point, but it is not verifiable that it will never halt. In the setting of Example 2.1.1, it is verifiable that a 0 will be observed at some stage, but not that 0 will be observed at every stage. In that setting, it is refutable that no 1s will ever be observed. It is decidable whether a 0 appears at stage n.

Theorem 2.2.1 provides a topological characterization of the verifiable hypotheses.

Theorem 2.2.1. *For $H \subseteq W$, the following are equivalent:*

1. *H is verifiable;*

2. *H is monotonically verifiable;*

3. *H is open in the information topology.*

Proof of Theorem 2.2.1. It is clear from the defintions that 2 implies 1. First, we show that 1 entails 3. Suppose that H is not open. Then H is true in some w, such that for all information E true in w, E does not entail H, i.e. there is $w \in H \setminus \mathsf{int}(H)$. Suppose, for contradiction, that L verifies H. Then $L(F) \subseteq H$, for some F true in w. But, by assumption, there is $v \in F \cap H^{\mathsf{c}}$. So L does not avoid error in v, and fails to satisfy Infal. We show that 3 entails 1. Suppose that H is open. Let $L(E) = H$ if E entails H, and let $L(E) = W$ otherwise. It is clear that L is monotonic and non-ampliative, and therefore infallible. Suppose that $w \in H$. Since $w \in \mathsf{int}H$, there is an information state F true in w that entails H. Therefore, $L(F) = H$. Furthermore, for any information state G true in w, we have that $L(G \cap F) = H$. So L satisfies LimCon. □

By the theorem, and Axioms T.1 and T.2, it follows that verifiable propositions are closed under finite conjunctions and arbitrary disjunctions. Theorem 2.2.1 implies that if H is not open, then there is in general no error-avoiding method that arrives at true belief in H. Every method that converges to true belief in worlds in which H is never verified must leap beyond the information available, and expose itself to error thereby. The characterization of refutable propositions follows immediately.

Theorem 2.2.2. *For $H \subseteq W$, the following are equivalent:*

1. *H is refutable;*

2. *H is monotonically refutable;*

3. *H is closed in the information topology.*

Finally, the decidable propositions are exactly the clopen sets.

Theorem 2.2.3. *For $H \subseteq W$, the following are equivalent:*

1. *H is decidable;*

2. *H is monotonically decidable;*

3. *H is clopen in the information topology.*

Proof of Theorem 2.2.3. It is clear from the definitions that 2 implies 1. If H is decidable, it is both verifiable and refutable. By Theorems 2.2.1, and 2.2.2, H is both open and closed, and therefore, clopen. We have shown that 1 implies 3. Finally, we show that 3 implies 1. Suppose that H is clopen in the information topology. Let

$$L(E) = \begin{cases} H, & \text{if } E \subseteq H; \\ H^{\mathsf{c}}, & \text{if } E \subseteq H^{\mathsf{c}}; \\ W, & \text{otherwise.} \end{cases}$$

It is clear that L is both monotonic and non-ampliative, and therefore infallible. Suppose that $w \in H$. Since H is open, $w \in \mathsf{int}H$. That implies that there is $E \in \mathcal{I}(w)$ such that $E \subseteq H$. Therefore, $L(E) \subseteq H$ and $L(F) \subseteq H$ for all $F \subseteq E$, and L verifies H. Suppose that $w \in H^{\mathsf{c}}$. Since H^{c} is also open, there is $E \in \mathcal{I}(w)$ such that $E \subseteq H^{\mathsf{c}}$. Therefore, $L(E) \subseteq H^{\mathsf{c}}$ and $L(F) \subseteq H^{\mathsf{c}}$ for all $F \in \mathcal{I}(w|E)$, which implies that L verifies H^{c}. □

2.3 Limiting Verification, Refutation and Decision

The requirement of infallibility is too strict to allow for ampliative, inductive inferences that draw conclusions beyond the information provided. But induction is necessary for inquiry, since most scientific hypotheses are neither verifiable nor refutable, but have a less familiar topological property of significant methodological importance. Consider the hypothesis H, which says:

$$Y = \alpha X^2 + \beta X.$$

Suppose that the truth is:

$$Y = \beta X.$$

Hypothesis H is not verifiable, because finitely many inexact observations along a parabola are compatible with a cubic function with a very small cubic term. Hypothesis H is not refutable, because the truth might be linear, in which case inexact measurements would never rule out arbitrarily flat parabolas. But H does have this important property: however H is true, one receives, eventually, information ruling out all simpler laws, after which H would be refuted if H were false. That is the characteristic epistemological property of concrete, scientific hypotheses and models. In general, say that H is *verifutable* iff H entails that $\mathsf{frnt}(H)$, the worlds in which H is false but will not be refuted, will be refuted, i.e. if $H \subseteq \mathsf{ext}(\mathsf{frnt}H)$. In topology, verifutable propositions are said to be *locally closed.*

Theorem 2.3.1. *The following are all equivalent*

1. *H is verifutable;*

2. *$\mathsf{frnt}(H)$ is refutable;*

3. $H \subseteq \mathsf{ext}(\mathsf{frnt}H)$*;*

4. $H = V \cap R$*, for verifiable* V *and refutable* R*.*

Proof. The proof relies on several standard facts about the closure operator (1) it is extensive, i.e. $H \subseteq \mathsf{cl}H$, (2) it is increasing, i.e. that if $A \subseteq B$, then $\mathsf{cl}A \subseteq \mathsf{cl}B$, and (3) it is idempotent, i.e. that $\mathsf{clcl}A = \mathsf{cl}A$. It is an immediate consequence of idempotence that $\mathsf{cl}A$ is closed. It is an immediate consequence of extensivity that if A is closed, then $\mathsf{cl}A = A$.

2 implies 3. Suppose the $\mathsf{frnt}(H)$ is refutable. By Theorem 2.2.2, $\mathsf{frnt}H$ is closed, and therefore $\mathsf{cl}(\mathsf{frnt}H) \subseteq \mathsf{frnt}(H)$. Since the closure operator is extensive, $\mathsf{cl}(\mathsf{frnt}H) = \mathsf{frnt}(H)$. For all propositions, $H \subseteq (\mathsf{frnt}H)^{\mathsf{c}}$. Substituting equals for equals, $H \subseteq (\mathsf{clfrnt}H)^{\mathsf{c}} = \mathsf{extfrnt}H$.

3 implies 2. By definition, $\mathsf{frnt}H \subseteq \mathsf{cl}H$. Since the closure operator is increasing and idempotent, $\mathsf{clfrnt}H \subseteq \mathsf{clcl}H = \mathsf{cl}H$. Suppose that (3) holds. Then, $\mathsf{clfrnt}H \subseteq H^{\mathsf{c}}$, and therefore $\mathsf{clfrnt}H \subseteq H^{\mathsf{c}} \cap \mathsf{cl}H = \mathsf{frnt}H$. So $\mathsf{frnt}H$ is closed and, by Theorem 2.2.2, refutable.

2 implies 4. For all propositions $H = \mathsf{cl}H \cap H = \mathsf{cl}H \cap (\mathsf{frnt}H)^{\mathsf{c}}$. Suppose that (2) holds. Then, $\mathsf{frnt}H$ is closed and therefore, $\mathsf{frnt}H = \mathsf{clfrnt}H$. So $H = \mathsf{cl}H \cap (\mathsf{clfrnt}H)^{\mathsf{c}}$. Furthermore, $\mathsf{cl}H$ is closed, and $(\mathsf{clfrnt}H)^{\mathsf{c}}$ is open. By Theorems 2.2.1 and 2.2.2, we have exhibited H as the conjunction of a verifiable and refutable proposition.

4 implies 2. Suppose that $H = V \cap R$ for verifiable V and refutable R. Then,

$$\begin{aligned} \mathsf{frnt}H &= \mathsf{cl}(V \cap R) \cap (V \cap R)^{\mathsf{c}} \\ &= \mathsf{cl}(V \cap R) \cap V^{\mathsf{c}} \cup \mathsf{cl}(V \cap R) \cap R^{\mathsf{c}} \end{aligned}$$

Since the closure operator is increasing, $\mathsf{cl}(V \cap R) \subseteq \mathsf{cl}R = R$. Therefore, the right-hand disjunct is empty, and $\mathsf{frnt}(H) = \mathsf{cl}(V \cap R) \cap V^{\mathsf{c}}$, a conjunction of closed sets. □

Scientific paradigms, or research programs, are typically not even verifutable—they must be articulated with auxiliary assumptions of increasing complexity to make them verifutable. That familiar idea motivates the concept of a *limiting open* proposition, which is a countable union (disjunction) of locally closed propositions that may be viewed as its possible, concrete articulations.

Limiting open propositions have an important methodological character. Say that method L is a *limiting decision procedure* for H iff L converges to H in the limit, if H is true, and converges to H^{c} in the limit, if H^{c} is true. That is, L is a limiting decision procedure for H iff

LimDec L converges to H in w, if $w \in H$, and converges to H^{c} in w, if $w \notin H$.

Say that H is *decidable in the limit* iff there is a limiting decision procedure for H. It is a basic result, proved independently by a number of authors in philosophy and informatics (de Brecht and Yamamoto [2009], Genin and Kelly [2015], and Baltag et al. [2016]), that H is decidable in the limit iff H and H^{c} are both countable unions of locally closed sets—i.e., iff H and H^{c} are both research programs. That implies that the catch-all hypothesis "neither H nor H^{c}" is off the table, either by presupposition or by assumption, and explains why science typically focuses on competitions between two salient research programs.

Theorem 2.3.2. *H is decidable in the limit iff H and H^{c} are limiting open.*

Proof of Theorem 2.3.2. Left to right. Suppose that H is decidable in the limit. Then H and H^{c} are verifiable in the limit. By Theorem 2.3.3, both H and H^{c} are limiting open.

Right to left. Suppose that H and H^{c} are limiting open. Then, $H = \cup_{i=1} A_{1,i}$ and $H^{\mathsf{c}} = \cup_{i=1} A_{2,i}$, where each $A_{j,i}$ is verifutable and $A_{1,k} \cap A_{2,j} = \varnothing$ for all j, k. Let $f : \mathbb{N} \to \{1, 2\} \times \mathbb{N}$ be a bijection. For $E \in \mathcal{I}$, let

$$\sigma(E) = f \circ \min\{i : E \subseteq \mathsf{extfrnt}(A_{f(i)}) \text{ and } E \nsubseteq \mathsf{ext}(A_{f(i)})\}.$$

Define

$$L(E) = \begin{cases} A_{\sigma(E)}, & \text{if } \sigma(E) < \infty; \\ W, & \text{otherwise.} \end{cases}$$

We show that L satisfies LimDec. Suppose, without loss of generality, that $w \in H$. Let $k = \min\{i : w \in A_{f(i)}\}$. For all $j < k$, $w \notin A_{f(j)}$ and therefore, either $w \in \mathsf{frnt}(A_{f(j)})$ or $w \in \mathsf{ext}(A_{f(j)})$. Let $Y = \{j < k : w \in \mathsf{frnt}(A_{f(j)})\}$, and $X = \{j < k : w \in \mathsf{ext}(A_{f(j)})\}$. Let

$$O' = \cap_{j \in X} \mathsf{ext} A_{f(j)};$$
$$O'' = \mathsf{extfrnt}(A_{f(k)}).$$

Let $O = O' \cap O''$. O' is a finite conjunction of open sets, and therefore open. Since $A_{f(k)}$ is verifutable, O'' is open, by Theorem 2.3.1. Therefore O is open. By construction, $w \in O$. Let $E \in \mathcal{I}(w)$ such that $E \subseteq O$. Suppose that $F \in \mathcal{I}(w|E)$. We claim that $A_{\sigma(F)} = A_{f(k)}$. For $j \in X$, $F \subseteq E \subseteq \mathsf{ext}(A_{f(j)})$. For $j \in Y$, $w \in \mathsf{frnt} A_{f(j)}$ and therefore, $F \nsubseteq \mathsf{extfrnt}(A_{f(j)})$. Since $F \in \mathcal{I}(w)$, $F \nsubseteq \mathsf{ext} A_{f(k)}$. Furthermore, $F \subseteq E \subseteq \mathsf{extfrnt}(A_{f(k)})$. Therefore $L(F) = L(E) = A_{f(k)} \subseteq H$, as required.

□

If the catch-all hypothesis is taken seriously, one can still verify research program H in the limit, in the sense that method L converges to an articulation of H iff H is true. Otherwise, L may cycle forever through alternative articulations of H. The converse is also true—verification in the limit is demonstrably possible only for research programs. Concretely, say that L *converges to an articulation* of H in w iff there $E \in \mathcal{I}(w)$ such that $L(F) \subseteq L(E) \subseteq H$, for all

$F \in \mathcal{I}(w|E)$. Say that L *verifies* H *in the limit* iff L converges to an articulation of H in w iff $w \in H$. Say that H is verifiable in the limit iff there is a method that verifies H in the limit.

Theorem 2.3.3. *H is verifiable in the limit iff H is limiting open.*

Proof of Theorem 2.3.3. Left to right. Suppose that L is a limiting verifier of H. Let

$$\mathcal{S} = \{E \in \mathcal{I} : L(E) \subseteq H\}.$$

For each $E \in \mathcal{S}$, let $\mathcal{D}_E = \{F \in \mathcal{I} : F \subseteq E \text{ and } L(F) \not\subseteq L(E)\}$, and let $E' = \bigcup \mathcal{D}_E$. We claim that:

$$H = \bigcup_{E \in \mathcal{T}} E \setminus E'.$$

To prove the claim, $w \in H$ iff there is $E \in \mathcal{I}(w)$ such that for all information states $F \in \mathcal{I}(w|E)$, $L(F) \subseteq L(E) \subseteq H$ iff there is $E \in \mathcal{S}$ such that $w \in E \setminus E'$. Since $\mathcal{S} \subseteq \mathcal{I}$, and $\mathcal{I}$ is countable, H is expressed as a countable union of locally closed sets.

Right to left. Let $H = \cup_{i=1}^{\infty} A_i$, for A_i verifutable. Let

$$\sigma(E) = \min\{i : E \subseteq \mathsf{extfrnt} A_i \text{ and } E \not\subseteq \mathsf{ext} A_i\}.$$

Define

$$L(E) = \begin{cases} A_{\sigma(E)}, & \text{if } \sigma(E) < \infty; \\ W, & \text{otherwise.} \end{cases}$$

Suppose that $w \in H$. Let k be the least integer such that $w \in A_k$. Then for $j < k$, either $w \in \mathsf{ext}(A_j)$ or $w \in \mathsf{frnt}(A_j)$. Let $Y = \{j < k : w \in \mathsf{frnt}(A_j)\}$, and $X = \{j < k : w \in \mathsf{ext}(A_j)\}$. Let

$$O' = \cap_{j \in X} \mathsf{ext} A_j;$$
$$O'' = \mathsf{extfrnt}(A_k).$$

Let $O = O' \cap O''$. O' is a finite conjunction of open sets, and therefore open. Since A_k is verifutable, O'' is open, by Theorem 2.3.1. Therefore O is open. By construction, $w \in O$. Let $E \in \mathcal{I}(w)$ such that $E \subseteq O$. Suppose that $F \in \mathcal{I}(w|E)$. We claim that $A_{\sigma(F)} = A_k$. For $j \in X$, $F \subseteq E \subseteq \mathsf{ext}(A_j)$. For $j \in Y$, $w \in \mathsf{frnt} A_j$ and therefore, $F \not\subseteq \mathsf{extfrnt}(A_j)$. Since $F \in \mathcal{I}(w)$, $F \not\subseteq \mathsf{ext} A_k$. Furthermore, $F \subseteq E \subseteq \mathsf{extfrnt}(A_k)$. Therefore $L(F) = L(E) = A_k \subseteq H$, as required. Suppose that $w \notin H$, and that for $E \in \mathcal{I}(w)$, $L(E) = A_i$. Then $w \in E \subseteq \mathsf{extfrnt} A_i$. Since $w \notin A_i$, and $w \notin \mathsf{frnt} A_i$, it must be that $w \in \mathsf{ext} A_i$. Let $F \in \mathcal{I}(w|E)$ be such that $F \subseteq \mathsf{ext} A_i$. Then $L(F) \subseteq A_i^{\mathsf{c}}$, and therefore $L(F) \not\subseteq L(E)$, as required. □

2.4 Problems and Solutions

An *empirical problem* is a triple $(W, \mathcal{I}, \mathcal{Q})$ countable partition $\mathcal{Q}$ of the worlds in W into a set of *answers*. For $w \in W$, write $\mathcal{Q}(w)$ for the answer true in w. A *relevant response* is any disjunction of answers to $\mathcal{Q}$. Let $\mathcal{Q}^*$ be the set of all relevant responses. For any proposition A, let $\mathcal{Q}(A)$ be the strongest relevant response entailed by A, i.e. $\mathcal{Q}(A) = \bigcap\{R \in \mathcal{Q}^* : A \subseteq R\}$. A method is a *solution* to $\mathcal{Q}$ iff it converges, on increasing information, to the true answer in $\mathcal{Q}$, i.e. iff for every $w \in W$, there exists $E \in \mathcal{I}(w)$ such that $L(F) \subseteq \mathcal{Q}(w)$ for all $F \in \mathcal{I}(w)$ entailing E. A problem is *solvable* iff it has a solution.

Theorem 2.4.1. *Problem $\mathcal{Q}$ is solvable iff every answer is limiting open.*

Proof. Left to right. Suppose that $\mathcal{Q}$ is solvable. Let L be a solution to $\mathcal{Q}$. Then, L is a limiting decision procedure for each $A \in \mathcal{Q}$. By Theorem 2.3.2, A is limiting open. Right to left. Let $A_1, A_2, \ldots$ enumerate the answers to $\mathcal{Q}$. Suppose that each A_i is limiting open. Then, by Theorem 2.3.2, each answer is decidable in the limit. Let L_i be a limiting decision prodecure for A_i. Let $\sigma(E) = \min\{i : L_i(E) \subseteq A_i\}$. Define

$$L(E) = \begin{cases} A_{\sigma(E)}, & \text{if } \sigma(E) < \infty; \\ W, & \text{otherwise.} \end{cases}$$

Suppose that $w \in A_i$. Since, $w \notin A_j$ for each $j < i$, there is $E_j \in \mathcal{I}(w)$ such that for all $F \in \mathcal{I}(w|E_j)$, $L_j(F) \subseteq A_j^{\mathsf{c}}$. Furthermore, since $w \in A_i$, there is $E_i \in \mathcal{I}(w)$ such that for all $F \in \mathcal{I}(w|E_i)$, $L_i(F) \subseteq A_i$. Let $O' = \cap_{j<i} E_j$. Let $O'' = E_i$. Let $O = O' \cap O''$. By construction, O is open and $w \in O$. Let $E \in \mathcal{I}(w)$ such that $E \subseteq O$. Let $F \in \mathcal{I}(w|E)$. Then, $L_i(F) \subseteq A_i$ and, for $j < i$, $L_j(F) \subseteq A_j^{\mathsf{c}}$. Therefore, $L(F) \subseteq A_i$, as required. □

Theorems like 2.2.1, 2.3.3, and 2.4.1 constitute an exact correspondence between topology and learnability.

2.5 Simplicity and Ockham's Razor

Popper [1959] proposed that A is as simple as B, which we abbreviate with $A \preceq B$, iff A is at least as falsifiable as B, i.e. if every information state that refutes B also refutes A. In the equivalent, contrapositive formulation: $A \preceq B$ iff any information state consistent with A is consistent with B. Therefore, Popper's thesis is that $A \preceq B$ iff $A \subseteq \mathsf{cl}B$, i.e. A entails that B will never be refuted. That elegant notion captures many of our intuitions about simplicity. Since any collection of points consistent with a linear polynomial is consistent with a quadratic, *linear* is simpler than *quadratic*. In the ornithological context, it means the sharp hypothesis 'all ravens are black' is simpler than its negation. However, Popper's proposal has the unintuitive consequence that the result of "tacking-on" an irrelevant conjunct to A is simpler than A itself [Glymour, 1980]. That is undesirable feature is a consequence of the elementary fact that if B entails A, then B entails $\mathsf{cl}A$ as well. A particularly striking consequence is that the unopinionated hypothesis W, expressing a suspension of belief, is maximally complex. By demanding greater logical content, Popper's version of Ockham's razor has the paradoxical effect of preventing you from suspending judgment, even before you have seen any data.

In Genin and Kelly [2018], we amend Popper's notion to answer that type of objection. According to Popper, A is simpler than B iff A entails that B will never be refuted. But there is nothing wrong with a *true* theory never being refuted. The worrisome possibilities are those in which B is *false*, but never refuted — in such worlds, mistaken belief in B would never be detected. Those are exactly the worlds in $\mathsf{cl}(B) \setminus B$, or $\mathsf{frnt}B$. That suggests the following explication of simplicity: $A \preceq B$ iff $A \subseteq \mathsf{frnt}B$. That is better than Popper's original formulation, because it no longer confounds underdetermination by information with logical strength. Since $\mathsf{frnt}W = \varnothing$, suspension of belief is always a simplest response. We still get the intuitive verdict that *linear* is simpler than *quadratic.* But there are tricky cases where it is still difficult to decide what to say. Ought we to say that *linear or cubic* is also simpler than *quadratic*? The proposed simplicity notion says that the disjunctive hypothesis is not simpler than *quadratic*, since the cubic worlds are not in the frontier of the quadratic worlds. That means that simplicity relations can be obscured by disjoining irrelevant possibilities, e.g. if $A \preceq B$, and $w \notin \mathsf{frnt}B$, then $A \cup \{w\} \not\preceq B$. Those sorts of considerations suggest the following defintion: say that A is as simple as B, written $A \lhd B$ iff $A \cap \mathsf{frntr}(B) \neq \varnothing$, which says that A is compatible with the possibility that B is false, but irrefutable. Say that A is *simplest* iff there is no B such that $B \lhd A$. This latest definition is also motivated by a suggestive correspondence with refutability.

Theorem 2.5.1. *A is simplest iff A is refutable.*

Proof. By Lemma 2.1.1, A is closed iff $\mathsf{frnt}A = \varnothing$, iff there is no $B \lhd A$. □

The simplicity relation may change in light of new information. Say that A is as simple as B *in light of* E, written $A \lhd_E B$, iff $A \cap E \lhd B \cap E$. Say that A is *simplest in light of* E iff there is no B such that $B \lhd_E A$.

Defining simplicity is preliminary to defining Ockham's razor, and other methodological norms. In the context of question $\mathcal{Q}$, let $\widehat{L}(E) = \mathcal{Q}(L(E))$. Say that a method L is *Popperian* iff $\widehat{L}(E)$ is closed (refutable) in E, i.e. $E \cap \mathsf{cl}\widehat{L}(E) \subseteq E \cap \widehat{L}(E)$. Say that a method L is *Ockham* iff $\widehat{L}(E)$ is simplest in light of E. Say that a method L *never errs on the side of complexity* iff L never outputs anything more complex than the truth, i.e. $\mathcal{Q}(w) \not\lhd_E \widehat{L}(E)$, for all $E \in \mathcal{I}(w)$. The first two concepts, are global, and refer to what is simplest, or refutable, when the possibilities are restricted to current information. The last concept is local, and refers to the unknown truth: "in each world, never conjecture any relevant response more complex than the truth." Surprisingly, all three principles are equivalent.

Theorem 2.5.2. *The following are equivalent:*

1. *L is Popperian;*
2. *L is Ockham;*
3. *L never errs on the side of complexity.*

Proof of Theorem 2.5.2. 1 and 2 are equivalent by Theorem 3.6.2. We show that 1 and 3 are equivalent. L errs on the side of comlpexity iff there is $E \in \mathcal{I}(w)$ such that $E \cap \mathcal{Q}(w) \lhd E \cap \mathcal{Q}(L(E))$ iff there is $w \in E \cap \mathsf{frnt}\mathcal{Q}(L(E))$ iff $\mathcal{Q}(L(E))$ is not closed in E. □

Thus, if any of the above hold, we say that L *satisfies Ockham's razor*, otherwise say that L is an *Ockham violator*.[1] Say that L is an *opinionated* Ockham violator iff L ever conjectures an *answer* more complex than the truth, i.e. iff there is $E \in \mathcal{I}(w)$ such that $\mathcal{Q}(w) \lhd_E \mathcal{Q}(v) \supseteq L(E)$. Clearly, every opinionated Ockham violator is an Ockham violator, though the converse is not true.

2.6 Simplicity and Progress

It is one thing to define simplicity, and another to provide an epistemic justification for preferring it. In this section, we provide what Lakatos laments is missing from Popperian methodology: some reason to believe that preferring simplicity, or equivalently, falisifiability, is any better than some other *ad hoc* stratagem. It seems like a minimal requirement of progressive inquiry that truth not be forfeited once it is in our grasp. Say that a solution L to a problem $(W, \mathcal{I}, \mathcal{Q})$ is *progressive* iff for $E \in \mathcal{I}(w)$, if $L(E) \subseteq \mathcal{Q}(w)$, then $L(F) \subseteq \mathcal{Q}(w)$ for $F \in \mathcal{I}(w|E)$. That is to say that the true answer to $\mathcal{Q}$ is a *fixed point* of inquiry in w: once L has conjectured the true answer, no further information can dislodge it from the truth. The following provides a simple sufficient condition for progressive solvability.

Theorem 2.6.1. *Suppose that there is an enumeration $A_1, A_2, \ldots,$ of the answers to $\mathcal{Q}$, in agreement with the simplicity relation, i.e. that if $i < j$ then $A_j \ntriangleleft A_i$. Then there is a progressive solution for $(W, \mathcal{I}, \mathcal{Q})$.*

Proof of Theorem 2.6.1. Suppose that the preconditions of the theorem hold. First we argue that, for all i, $\cup_{j \leq i} A_j$ is refutable. Suppose that $\cup_{j \leq i} A_j$ is not refutable. Then, by Lemma 2.1.1 there is $w \in \mathsf{frnt} \cup_{j \leq i} A_j$. Then, it must be

[1] That maximal simplicity corresponds to refutability is a very Popperian result, but Popper failed to obtain it. Recall that on Popper's proposal, W is (trivially) refutable, but maximally complex.

that $w \in A_k$ for $k > i$. But then, there is $j < k$ such that $A_k \lhd A_j$. Contradiction. Therefore, each $\cup_{j \leq i} A_j$ is refutable. Let $\sigma(E) = \min\{i : E \nsubseteq \cap_{j \leq i} A_j^{\mathsf{c}}\}$. Define

$$L(E) = \begin{cases} A_{\sigma(E)}, & \text{if } \sigma(E) < \infty; \\ W, & \text{otherwise.} \end{cases}$$

Suppose that $w \in A_i$. First, we show that L is a solution. Let $w \in A_i$. Then w is in the open set $O = \cap_{j<i} A_j^{\mathsf{c}}$. Let $E \in \mathcal{I}(w)$, such that $E \subseteq O$. Suppose that $F \in \mathcal{I}(w|E)$. Then, for $j < i$, $F \subseteq E \subseteq \cap_{k \leq j} A_k^{\mathsf{c}}$. Since $F \in \mathcal{I}(w)$, $F \nsubseteq \cap_{j \leq i} A_j^{\mathsf{c}}$. Therefore, $L(F) \subseteq A_i$, as required. It remains to show that L is progressive. Suppose that $E \in \mathcal{I}(w)$ and $L(E) \subseteq \mathcal{Q}(w)$. Let $F \in \mathcal{I}(w|E)$. Then, $F \subseteq E \subseteq \cap_{j<i} A_j^{\mathsf{c}}$ and, since $F \in \mathcal{I}(w)$, $F \nsubseteq \cap_{j \leq i} A_j^{\mathsf{c}}$. Therefore, $L(F) = A_i$, as required. □

So long as it is possible to enumerate the answers to a question in agreement with the simplicity relation, it is possible to solve it in a progressive way. In the propositional setting this result is not terribly exciting, nor difficult to prove.[2] Proving an analogous result in the statistical setting (Theorem 3.6.3) is not as easy, but of commensurately greater interest.

In Genin and Kelly [2018], we justify Ockham's razor by showing that it is a necessary condition of avoiding unnecessary cycles of opinion on the way to the truth. Here, I give a justification of a somewhat weaker principle, proving that every progressive solution must not be an opinionated violator of Ockham's razor. That means that the rather minimal requirement of progressiveness mandates a rather strong methodological preference for simple answers. This result speaks to intuition, and inspires an analogous result in the statistical setting (Theorem 3.6.4).

Theorem 2.6.2. *Suppose that L is a solution for $(W, \mathcal{I}, \mathcal{Q})$. L is progressive only if L is not an opinionated Ockham violator.*

Proof of Theorem 2.6.2. Suppose that L is a solution for $(W, \mathcal{I}, \mathcal{Q})$. Suppose that L is an opinionated Ockham violator. By Theorem 2.5.2, there is w and $E \in \mathcal{I}(w)$, such that $\mathcal{Q}(w) \lhd_E \mathcal{Q}(v) \supseteq L(E)$. Therefore, there is $w' \in E \cap \mathcal{Q}(w) \cap \mathsf{frnt}\mathcal{Q}(v)$. Since L is a solution, there is $F \in \mathcal{I}(w'|E)$ such that $L(F) \subseteq \mathcal{Q}(w') = \mathcal{Q}(w)$. Since $w' \in \mathsf{frnt}\mathcal{Q}(v)$, there is $w'' \in F \cap \mathcal{Q}(v)$. Therefore, $E, F \in \mathcal{I}(w'')$, $F \subseteq E$, but $L(E) \subseteq \mathcal{Q}(w'')$ and $L(F) \nsubseteq \mathcal{Q}(w'')$. Therefore, the true answer is not a fixed point of inquiry in w''. □

[2] For more results about when, exactly, progressive solutions are feasible, see Genin and Kelly [2018]

Chapter 3

Learning from Statistical Information

There seems to be a gulf between propositional and statistical information. Propositional information rules out relevant possibilities. In contrast, a typical random sample is logically compatible with *every* possible generating distribution. That well-worn observation has convinced many to abandon theories of scientific method in which refutation occupies a central role. If scientific data typically fail to refute *any* relevant possibility, then the theory developed in Chapter 2, which relies on decisive refutation, has no real subject matter. In this chapter, we address that fundamental difficulty by solving for the *unique* topology in which the open sets are precisely the *statistically verifiable* propositions, and the closed sets are precisely the *statistically refutable* propositions. That allows for a systematic translation out of the idiom of falsificationist philosophy of science and into the language of the data-driven sciences.

In Section 3.1, we set the stage on which statistical inquiry occurs. The unstructured point-worlds of Chapter 2 are replaced with probability measures on a sample space. The task of inquiry is to infer from random samples something about the probability measure governing their statistical behavior. In Section 3.1, we introduce the *weak topology*, a common topology on probability measures, and prove some simple results about it. Most of the theorems in this section can be found in Billingsley [1999]. The point of the section is firstly, to quickly and painlessly develop some necessary tools, and secondly, to introduce an epistemic interpretation of the weak topology. Roughly speaking: a sequence of measures μ_n converges to μ in the weak topology iff a statistical test would have difficulty refuting the (μ_n) on the basis of data from μ.

In Section 3.2, we introduce several candidates for the definition of statistical verifiability. In the setting of Part I, an hypothesis H is said to be verifiable iff there exists a monotonic and infallible method that converges on increasing information to H iff H is true. That condition implies that there is a method that achieves *every* bound on *chance* of error, and converges to H iff H is true. In statistical settings, one typically cannot insist on such a high standard of infallibility. Instead, say that H is *statistically verifiable* iff for every bound on error, there is a method that achieves it, and that converges on increasing samples to H iff H is true. The reversal of quantifiers expresses the fundamental difference between statistical and propositional verifiability. Of course, this rough definition leaves many things unspecified. Section 3.2 considers several different ways of rendering it precise.

Theorem 3.2.1 shows that no matter which plausible definition one prefers, a hypothesis is statistically verifiable if, and only if, it is open in the weak topology. In statistical terminology, our Theorem 3.2.1 provides a topological criterion for the existence of a consistent hypothesis test, with genuine finite-sample bounds on the chance of Type I error. To the best of my knowledge, there is no precedent for this in the statistical literature.

In Section 3.2.3, we illustrate the power and utility of the framework by applying it to conditional independence testing. We show that under a weak, non-parametric condition, hypotheses of conditional dependence between random variables are statistically verifiable. That proves the existence of a general, consistent, non-parametric test of conditional independence with finite-sample bounds on the chance of Type I error. That improves on previous non-parametric results given by Gretton and Györfi [2010], Györfi and Walk [2012] which, while guaranteeing the existence of tests that are consistent in the limit of infinite data, do not guarantee finite-sample bounds on the chance of error. In Section 3.2.4 we apply these results on conditional independence testing to learning causal Bayes nets from observational data.

In the preceding, we said that statistical verifier of H converges to H if H is true, and otherwise has a small chance of drawing an erroneous conclusion. But that standard is consistent with a wild see-sawing between the chance of producing the informative conclusion H and the uninformative conclusiosn W as sample sizes increase, even if H is true. Of course, it is desirable that the chance of correctly producing H increases with the sample size. If that is not feasible, then we can at least expect that there are no two sample sizes $n_1 < n_2$ such that the chance of correctly producing H is much smaller at n_2 than at n_1. That property ensures that collecting a larger sample is never a disastrously bad idea. Surprisingly, standard hypothesis tests fail to satisfy even that weak requirement. Chernick and Liu [2002] noticed non-monotonic behavior in the power function of textbook tests of the binomial proportion, and proposed heuristic software solutions. The test exhibited in the proof of Theorem 3.2.1 also displays occasionally dramatic non-monotonicity. For that reason, Theorem 3.2.1 provides only a partial statistical analogue of propositional verifiability, since issues of monotonicity are ignored.

In Section 3.3 we pay closer attention to monotonicity. Theorem 3.3.1 states that every hypothesis which is open in the weak topology is also *monotonically verifiable*, in the following sense: for every $\alpha > 0$, there is a statistical verifier of H such that (1) the chance of erroneously concluding H is bounded by α and (2) as sample sizes increase, the probability of correctly producing the informative concluison H never decreases by more than α. The import of the Theorem is that, with care, non-monotonicities can be rendered abitrarily small. Section 3.3.3 applies the results of the previous section to prove the existence of monotonic methods for verifying conditional dependencies and related causal hypotheses. That feature turns out to be crucial in Section 3.6.2.

In Section 3.4.1, we leave behind deductive standards of statistical success. In other words: we begin to consider success concepts that do not require finite-sample bounds on the chance of error. The central theorems of Section 3.4.1 give topological characterizations of hypotheses that are statistically verifiable, refutable and decidable in the limit. The last of these results gives a generalization of the topological criterion of decidability in the limit given by Dembo and Peres [1994].[1] As an illustration of these concepts, we give a simple proof that, in the framework of graphical causal models that, under the causal Markov and faithfulness assumption, it is always possible to decide in the limit whether the true causal structure belongs to a particular Markov equivalence class.

In Section 3.5, we continue with a study of inductive statistical problems. A *statistical problem* is a partition of the set of probability measures compatible with background knowledge into an exhaustive set of competing *answers*. A method is a *solution* to a statistical problem if, on increasing samples, it converges to the true answer to the statistical problem. Theorem 3.5.1 gives a topological characterization of solvable statistical problems. This defines the outer limits of problems that are tractable by statistical means. As an illustration, we prove that, under the usual assumptions, it is always possible to converge in the limit to the Markov equivalence class of the true causal graph. Spirtes et al. [2000], give similar results, exhibiting several algorithms that solve this problem provided that one “plugs in” reliable procedures for making the requisite decisions about conditional independence. Spirtes et al. [2000] cite appropriate tests for the linear Gaussian case, and the discrete case. One may still wonder, however, whether appropriate procedures exist in general. In Section 3.5.2, we give general results that hold for discrete variables, random variables with density functions, and any mixture of the two.

A method may count as a solution to a statistical problem even if its chance of producing the true conclusion at sample size 100 decreases dramatically from the chance at sample size 20. Of course one would prefer if, no matter which answer is the true one, the chance of getting the right answer increases monotonically with sample size. Even if that standard is infeasible, it should be our regulative

[1] Dembo and Peres refer to this decidability in the limit as *discernability*.

ideal. Say that a method is *α-progressive* if, no matter which answer is the true one, the chance that the method outputs the true answer never decreases by more than α as the sample size grows. That property ensures that collecting more data cannot set the method back too badly. Theorem 3.6.3 demonstrates that, for typical problems, there exists an α-progressive method for every $\alpha > 0$. In Section 3.6.5 we show that there exist progressive solutions to the problem of inferring Markov equivalence classes of causal graphs. That provides a stronger justification for standard methods of causal discovery from observational data than previous arguments, which only demonstrated their pointwise consistency.

Finally, Theorem 3.6.4 shows that every α-progressive method must obey a probabilistic version of Ockham's razor. That provides a non-circular, prior-free justification for simplicity bias in statistical methodology. If progressiveness strikes the reader as a weak property, then we have given a *strong* justification of Ockham's razor, since it is necessary for achieving even this very weak standard of success.

3.1 The Statistical Setting

3.1.1 Samples and Worlds

A *sample space* $\mathfrak{S} = (\Omega, \mathcal{I})$ is a set of possible random samples Ω equipped with a topological basis $\mathcal{I}$. The topology $\mathcal{T}$ is formed by closing the basis $\mathcal{I}$ under unions. The topology on the sample space reflects what is verifiable about the sample itself. As in the purely propositional setting, it is *verifiable* that sample ω lands in A iff A is open, and it is *decidable* whether sample ω falls into region A iff A is clopen. No one can tell whether a real-valued sample point is rational or irrational, or if it is exactly π, because these regions are not open in the usual topology on $\mathbb{R}$. For another example, suppose that region A is the closed interval $[1/2, \infty]$, and that the sample ω happens to land right on the end-point $1/2$ of A. Suppose, furthermore, that given enough time and computational power, the sample ω can be specified to arbitrary, finite precision. No finite degree of precision: $\omega \approx .50$; $\omega \approx .500$; $\omega \approx .5000$; ... suffices to determine that ω is truly in A. But the mere possibility of a sample hitting the boundary of A does not matter statistically, if the chance of obtaining such a sample is zero, as it typically is, unless there is discrete probability mass on the geometrical point $\frac{1}{2}$.

The worlds in W are probability measures on the measurable space $(\Omega, \mathcal{B})$, where $\mathcal{B}$ is the smallest σ-algebra generated by $\mathcal{T}$. The elements of $\mathcal{B}$ are called *Borel* sets. Call the triple $(W, \Omega, \mathcal{I})$, consisting of a set of probability measures W, and sample space $(\Omega, \mathcal{I})$, a *chance setup.*[2]

[2] The term is borrowed from Ian Hacking [1965], although my usage is not exactly the same.

A Borel set A for which $\mu(\mathsf{bdry}A) = 0$ is said to be *almost surely clopen (decidable) in* μ.[3] Say that a collection of Borel sets $\mathcal{S}$ is almost surely clopen in μ iff every element of $\mathcal{S}$ is almost surely clopen in μ. We say that a Borel set A is almost surely decidable iff it is almost surely decidable in every μ in W. Similarly, we say that a collection of Borel sets $\mathcal{S}$ is almost surely clopen iff every element of $\mathcal{S}$ is almost surely clopen.

In the following, we will often assume that the basis $\mathcal{I}$ is almost surely clopen. That assumption is satisfied, for example, in the standard case in which the worlds in W are Borel measures on $\mathbb{R}^n$, and all measures are absolutely continuous with respect to Lebesgue measure, i.e. when all measures have probability density functions, which includes normal, chi-square, exponential, Poisson, and beta distributions. It is also satisfied for discrete distributions like the binomial, for which the topology on the sample space is the discrete (power set) topology, so every region in the sample space is clopen. It is satisfied in the particular cases of Examples 3.1.1 and 3.1.2.

Example 3.1.1. *Consider the outcome of a single coin flip. The set Ω of possible outcomes is $\{H, T\}$. Since every outcome is decidable, the appropriate topology on the sample space is $\mathcal{T} = \{\varnothing, \{H\}, \{T\}, \{H, T\}\}$, the discrete topology on Ω. Let W be the set of all probability measures assigning a bias to the coin. Since every element of $\mathcal{T}$ is clopen, every element is also almost surely clopen.*

Example 3.1.2. *Consider the outcome of a continuous measurement. Then the sample space Ω is the set of real numbers. Let the basis $\mathcal{I}$ of the sample space topology be the usual interval basis on the reals. That captures the intuition that it is verifiable that the sample landed in some open interval, but it is not verifiable that it landed exactly on the boundary of an open interval. There are no nontrivial decidable (clopen) propositions in that topology. However, in typical statistical applications, W contains only probability measures μ that assign zero probability to the boundary of an arbitrary open interval. Therefore, every open interval E is almost surely decidable, i.e. $\mu(\mathsf{bdry}(E)) = 0$.*

The following Lemma, given in Parthasarathy [1967] states that the almost surely clopen sets are closed under finitary set-theoretic operations.

Lemma 3.1.1 (Lemma 6.4 [Parthasarathy, 1967])**.** *The almost surely clopen sets in μ, denoted $\mathcal{C}(\mu)$, form an algebra.*

Proof of Lemma 3.1.1. One has that $\Omega \in \mathcal{C}(\mu)$, since $\mathsf{bdry}(\Omega) = \varnothing$. Moreover, $\mathcal{C}(\mu)$ is closed under complement, since $\mathsf{bdry}(A) = \mathsf{bdry}(A^{\mathsf{c}})$. Furthermore, since $\mathsf{bdry}(A \cup B) \subseteq \mathsf{bdry}(A) \cup \mathsf{bdry}(B)$, it follows that if $A, B \in \mathcal{C}(\mu)$, then $\mu(\mathsf{bdry}(A \cup B)) \leq \mu(\mathsf{bdry}(A) \cup \mathsf{bdry}(B)) \leq \mu(\mathsf{bdry}(A)) + \mu(\mathsf{bdry}(B)) = 0$. Therefore, $\mathcal{C}(\mu)$ is closed under finite union as well. □

[3] A set that is almost surely clopen in μ is sometimes called a *continuity set* of μ.

As an immediate consequence, we have the following:

Corollary 3.1.1. *If $\mathcal{S}$ is almost surely clopen in μ, then $\mathcal{A}(\mathcal{S})$, the smallest algebra generated by $\mathcal{S}$, is almost surely clopen in μ.*

Proof of Corollary 3.1.1. Recall that for arbitrary index set I, if every element of the collection $(\mathcal{A}_i, i \in I)$ is an algebra, then $\cap_{i\in I}\mathcal{A}_i$ is an algebra. Since $\mathcal{A}(\mathcal{S})$ is the intersection of all algebras containing $\mathcal{S}$, and $\mathcal{C}(\mu)$ is an algebra containing $\mathcal{S}$, $\mathcal{A}(\mathcal{S}) \subseteq \mathcal{C}(\mu)$. □

Product spaces represent the outcomes of repeated sampling. Let I be an index set, possibly infinite. Let $(\Omega_i, \mathcal{T}_i)_{i\in I}$ be sample spaces, each with basis $\mathcal{I}_i$. Define the *product* $(\Omega, \mathcal{T})$ of the $(\Omega_i, \mathcal{T}_i)$ as follows: let Ω be the Cartesian product of the Ω_i; let $\mathcal{T}$ be the product topology, i.e. the topology in which the open sets are unions of Cartesian products $\times_i O_i$, where each O_i is an element of $\mathcal{T}_i$, and all but finitely many O_i are equal to Ω_i. When I is finite, the products of basis elements in $\mathcal{I}_i$ are the intended basis for $\mathcal{T}$. Let $\mathcal{B}$ be the σ-algebra generated by $\mathcal{T}$. Let μ_i be a probability measure on $\mathcal{B}_i$, the Borel σ-algebra generated by the $\mathcal{T}_i$. The *product measure* $\mu = \times_i \mu_i$ is the unique measure on $\mathcal{B}$ such that, for each $B \in \mathcal{B}$ expressible as a Cartesian product of $B_i \in \mathcal{B}_i$, where all but finitely many of the B_i are equal to Ω_i, $\mu(B) = \prod \mu_i(B_i)$. (For a simple proof of the existence of the infinite product measure, see Saeki [1996].) Let $\mu^{|I|}$ denote the $|I|$-fold product of μ with itself.

The following Lemma shows that almost sure decidability is preserved by products.

Lemma 3.1.2. *Suppose that A', A'' are almost surely decidable in μ', μ'' respectively. Then $A' \times A''$ is almost surely decidable in $\mu = \mu' \times \mu''$.*

Proof of Lemma 3.1.2. Note that $\mathsf{bdry}(A' \times A'') \subset (\mathsf{bdry}A' \times \Omega'') \cup (\Omega' \times \mathsf{bdry}A'')$. Therefore,

$$\begin{aligned}\mu(\mathsf{bdry}(A' \times A'')) &\le \mu(\mathsf{bdry}A' \times \Omega'') + \mu(\Omega' \times \mathsf{bdry}A'')\\ &= \mu'(\mathsf{bdry}A') \cdot \mu''(\Omega'') + \mu'(\Omega') \cdot \mu''(\mathsf{bdry}A'')\\ &= 0.\end{aligned}$$

□

3.1.2 Statistical Tests

A statistical *method* is a measurable function from random samples to propositions over W.[4] A *test* of a statistical hypothesis $H \subseteq W$ is a statistical method $\psi : \Omega \to \{W, H^{\mathrm{c}}\}$. Call $\psi^{-1}(W)$ the *acceptance region*, and $\psi^{-1}(H^{\mathrm{c}})$ the *rejection region* of the test.[5] The *power* of test $\psi(\cdot)$ is the worst-case probability

[4] The σ-algebra on the range of the method is assumed to be the power set.

[5] The acceptance region is $\psi^{-1}(W)$, rather than $\psi^{-1}(H)$, because failing to reject H licenses only the trivial inference W.

that it rejects truly, i.e. $\inf_{\mu \in H^{\mathsf{c}}} \mu[\psi^{-1}(H^{\mathsf{c}})]$. The *significance level* of a test is the worst-case probability that it rejects falsely, i.e. $\sup_{\mu \in H} \mu[\psi^{-1}(H^{\mathsf{c}})]$.

A test is *feasible in* μ iff its acceptance region is almost surely decidable in μ. Say that a test is *feasible* iff it is feasible in every world in W. More generally, say that a method is *feasible* iff the preimage of every element of its range is almost surely decidable in every world in W. Tests that are not feasible in μ are impossible to implement — as described above, if the acceptance region is not almost surely clopen in μ, then with non-zero probability, the sample lands on the boundary of the acceptance region, where one cannot decide whether to accept or reject. If one were to draw a conclusion at some finite stage, that conclusion might be reversed in light of further computation. Tests are supposed to *solve* inductive problems, not to generate new ones.[6] Therefore we consider only feasible methods in the following development.

Hypothesis tests are often constructed to reject if the number of samples landing in a particular region exceeds some threshold. The following lemma states that such a test is μ-feasible, if the region is almost surely clopen in μ.

Lemma 3.1.3. *Suppose that A is almost surely clopen in μ. Then:*

$$\left\{(\omega_1, \ldots, \omega_n) : \sum_{i=1}^{n} \mathbb{1}[\omega_i \in A] \geq k\right\}$$

is almost surely clopen in μ^n, for $n \geq 1$, and $k \geq 0$.[7]

Proof of Lemma 3.1.3. Let $L_1, L_2, \ldots, L_{n\mathrm{C}\lceil k \rceil}$ enumerate all $\lceil k \rceil$-element subsets of $\{1, 2, \ldots, n\}$. Then

$$\{(\omega_1, \ldots, \omega_n) : \sum_{i=1}^{n} \mathbb{1}[\omega_i \in A] \geq k\} = \bigcup_{i=1}^{n\mathrm{C}\lceil k \rceil} \times_{j=1}^{n} B_{ij},$$

where $B_{ij} = A$ if $j \in L_i$, and $B_{ij} = \Omega$ otherwise. By Lemma 3.1.1, the almost surely clopen sets in μ^n are closed under finite disjunctions. Therefore it suffices to show that $\times_{j=1}^{n} B_{ij}$ is an almost surely clopen set in μ^n, which follows immediately from Lemma 3.1.2. □

3.1.3 The Weak Topology

A sequence of measures $(\mu_n)_n$ *converges weakly* to μ, written $\mu_n \Rightarrow \mu$, iff $\mu_n(A) \to \mu(A)$, for every A almost surely clopen in μ. It is immediate that

[6] Considerations of feasibility provide a new perspective on the assumption that appears throughout this work: that the basis $\mathcal{I}$ is almost surely clopen. If that assumption fails, then it is not an *a priori* matter whether geometrically simple zones are suitable acceptance zones for statistical methods. But if that is not determined *a priori*, then presumably it must be investigated by statistical means. That suggests a methodological regress in which we must use statistical methods to decide which statistical methods are feasible to use.

[7] The indicator function $\mathbb{1}[\omega \in A]$ is defined to take the value 1, if $\omega \in A$, and 0, otherwise. In the following we will write $\mathbb{1}_A(\cdot)$ for that indicator function.

$\mu_n \Rightarrow \mu$ iff for every μ-feasible test $\psi(\cdot)$, $\mu_n(\psi \text{ rejects}) \to \mu(\psi \text{ rejects})$. It follows that no feasible test of $H = \{\mu\}$ achieves power strictly greater than its significance level. Furthermore, every feasible method that correctly infers H with high chance in μ, exposes itself to a high chance of error in "nearby" μ_n. It is a standard fact that one can topologize W in such a way that weak convergence is exactly convergence in the topology: the usual sub-basis is given by sets of the form $\{\nu : |\mu(A) - \nu(A)| < \epsilon\}$, where A is almost surely clopen in μ.[8] That topology is called the *weak topology*, n.b.: the weak topology is a topology on *probability measures*, whereas all previously mentioned topologies were topologies on *random samples*. The following theorem provides useful sufficient conditions for weak convergence. It is essentially Theorem 2.1 in Billingsley [1999], omitting conditions irrelevant to the current development.

Theorem 3.1.1. *(Billingsley [1999, Theorem 2.1]) 1 and 2 are equivalent. 2 implies 3.*

1. $\limsup_n \mu_n(F) \leq \mu(F)$ *for all closed* F*;*

2. $\liminf_n \mu_n(G) \geq \mu(G)$ *for all open* G*;*

3. $\mu_n \Rightarrow \mu$.

Proof of Theorem 3.1.1. The fact that 1 and 2 are equivalent is immediate by duality. To see that 1 and 2 imply 3, note that:

$$\begin{aligned}\mu(\mathsf{cl}A) &\geq \limsup_n \mu_n(\mathsf{cl}A) \geq \limsup_n \mu_n(A)\\ &\geq \liminf_n \mu_n(A) \geq \liminf_n \mu_n(\mathsf{int}A) \geq \mu(\mathsf{int}A).\end{aligned}$$

If A is almost surely clopen, then $\mu(\mathsf{int}A) = \mu(A) = \mu(\mathsf{cl}A)$ and $\liminf_n \mu_n(A) = \limsup_n \mu_n(A) = \lim_n \mu_n(A) = \mu(A)$. □

As a consequence of Theorem 3.1.1, Billingsley proves the following, which provides a way to demonstrate weak convergence by showing that $\mu_n(A) \to \mu(A)$ holds for a convenient class of events.

Theorem 3.1.2. *(Billingsley [1999, Theorem 2.2]) Suppose (1) that* $\mathcal{S}$ *is closed under finite conjunction and (2) that every open set is a countable union of* $\mathcal{S}$ *sets. If* $\mu_n(A) \to \mu(A)$ *for every* A *in* $\mathcal{S}$*, then* $\mu_n \Rightarrow \mu$*.*

Proof of theorem 3.1.2. Following Billingsley: if $A_1, \ldots, A_r$ are in $\mathcal{S}$, then so

[8] Recall that a sequence (μ_n) converges to μ in a topology iff for every open set E containing μ, there is n_0 such that $\mu_n \in E$ for all $n \geq n_0$. If a topology is first countable, (μ_n) converges to μ in the topology iff μ is in the topological closure of the μ_n.

are their intersections. Therefore, by the inclusion-exclusion principle:

$$\begin{aligned}
&\mu_n(\cup_{i=1}^r A_i) = \\
&= \sum_i \mu_n(A_i) - \sum_{i<j} \mu_n(A_i \cap A_j) + \sum_{i<j<k} \mu_n(A_i \cap A_j \cap A_k) - \cdots \\
&\to \sum_i \mu(A_i) - \sum_{i<j} \mu(A_i \cap A_j) + \sum_{i<j<k} \mu(A_i \cap A_j \cap A_k) - \cdots \\
&= \mu(\cup_{i=1}^r A_i).
\end{aligned}$$

If G is open, then $G = \cup_i A_i$ for some sequence $\{A_i\} \subseteq \mathcal{S}$. Let $\epsilon > 0$. Since μ is countable additive, there is r such that $\mu(\cup_{i \leq r} A_i) > \mu(G) - \epsilon$. By the above,

$$\mu(G) - \epsilon \leq \mu(\cup_{i \leq r} A_i) = \lim_n \mu_n(\cup_{i=1}^r A_i) = \liminf_n \mu_n(\cup_{i=1}^r A_i) \leq \liminf_n \mu_n(G).$$

Since ϵ was arbitrary, $\mu(G) \leq \liminf_n \mu_n(G)$. So by Theorem 3.1.1, $\mu_n \Rightarrow \mu$. □

The preceding theorem allows us to exhibit a very tractable sub-basis for the weak topology. That sub-basis for the weak topology has two fundamental advantages over the standard sub-basis. First, its closure under finite intersection is evidently a countable basis. Second, it is easy to show that all the elements of the sub-basis are statistically verifiable, which we demonstrate in Lemma 3.2.1.

Theorem 3.1.3. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely decidable in every $\mu \in W$. Let $\mathcal{A}$ be the algebra generated by $\mathcal{I}$. Then, the collection*

$$\{\{\mu : \mu(A) > r\} : r \in \mathbb{Q} \text{ and } A \in \mathcal{A}\}$$

is a sub-basis for the weak topology on W.

Proof of Theorem 3.1.3. It is sufficient to show that $\mu_n \Rightarrow \mu$ iff the μ_n converge to μ in the topology generated by the sub-basis. *Left to right.* Suppose $\mu_n \Rightarrow \mu$. Let E be open in the topology generated by the sub-basis. Suppose μ lies in E. Then there is a basic open set:

$$B = \bigcap_{i=1}^k \{\mu : \mu(A_i) > b_i\},$$

such that $\mu \in B \subseteq E$. Since $\mathcal{I}$ is feasible for W, $\mu_n(B_i) \to \mu(B_i)$ for each i. Therefore, there exists n_i such that $\mu_n \in \{\mu : \mu(A_i) > b_i\}$ for all $n \geq n_i$. Letting $m = \max\{n_1, \ldots, n_k\}$, it follows that $\mu_n \in B \subseteq E$ for all $n \geq m$. Therefore, the μ_n converge to μ in the topology generated by the sub-basis. *Right to left.* Suppose that the μ_n converge to μ in the topology generated by the sub-basis. Notice that $\{\mu : \mu(A) \in (a, b)\} = \{\mu : \mu(A) > a\} \cap \{\mu : \mu(A^c) > 1 - b\}$. Therefore, since $\mathcal{A}$ is an algebra, the collection $\{\mu : \mu(A) \in (a, b)\}$ for $A \in \mathcal{A}$, and $a, b \in \mathbb{Q}$ generates the same topology. Let $A_1, A_2, \ldots$ enumerate the elements of $\mathcal{A}(\mathcal{I})$. Let $l_{ij} < \mu(A_i) < r_{ij}$ be rationals lying in $(\mu(A_i) - 1/j,\, \mu(A_i) + 1/j)$. Let

Z_{ij} denote the sub-basis element $\{\nu : \nu(A_i) \in (l_{ij}, r_{ij})\}$. Let f be a surjective function from $\mathbb{N}$ to $\mathbb{N} \times \mathbb{N}$. Let $U_k = Z_{f(k)}$. By assumption, for every $m \geq 1$, there is n_0 such that the μ_n lie in $\cap_{k=1}^{m} U_k$, for all $n \geq n_0$. So $\mu_n(A) \to \mu(A)$, for every $A \in \mathcal{A}$. Since $\mathcal{A}$ is closed under conjunction, and every open set in $\mathcal{T}$ is a countable union of elements of $\mathcal{A}$, it follows from Theorem 3.1.2 that $\mu_n \Rightarrow \mu$. □

The following observations are easy consequences of the Lemma. In the setting of Example 3.1.1, the set of all $\{\mu : \mu(\{H\}) \in (a, b)\}$ for $a, b \in \mathbb{Q}$, assigning open intervals of biases for the coin, forms a sub-basis for the weak topology on W. In fact, it forms a basis. If μ is the world in which the bias of the coin is exactly .5 and μ_n is the world in which the bias is exactly $.5 + 1/2^n$, then the μ_n converge to μ in the weak topology.

The preceding results enable us to characterize how convergence in the weak topology interacts with product spaces.

Theorem 3.1.4. *Suppose (1) that $\mathcal{I}', \mathcal{I}''$ are countable bases, (2) that μ'_n, μ' and μ''_n, μ'' are Borel measures on $(\Omega', \mathcal{I}')$, $(\Omega'', \mathcal{I}'')$ respectively, and (3) that $\mathcal{I}', \mathcal{I}''$ are almost surely decidable in μ', μ'' respectively. Then $\mu'_n \times \mu''_n \Rightarrow \mu' \times \mu''$ iff $\mu'_n \Rightarrow \mu'$ and $\mu''_n \Rightarrow \mu''$.*

Proof. Right to left. Suppose that $\mu'_n \Rightarrow \mu'$ and that $\mu''_n \Rightarrow \mu''$. Let $\mathcal{A}', \mathcal{A}''$ be the smallest algebras generated by $\mathcal{I}', \mathcal{I}''$. By Corollary 3.1.1, $\mathcal{A}', \mathcal{A}''$ are almost surely decidable in μ', μ''. Consider the class of sets

$$\mathcal{A}' \times \mathcal{A}'' = \{A' \times A'' : A' \in \mathcal{A}', A'' \in \mathcal{A}''\}.$$

Since $(A_1 \times B_1) \cap (A_2 \times B_2) = (A_1 \cap A_2) \times (B_1 \cap B_2)$, $\mathcal{A}' \times \mathcal{A}''$ is closed under finite intersection. Furthermore, since $\mathcal{I}' \times \mathcal{I}''$ is a countable basis for $\mathcal{T}' \times \mathcal{T}''$, and $\mathcal{A}' \times \mathcal{A}'' \supseteq \mathcal{I}' \times \mathcal{I}''$, every open set in $\mathcal{T}' \times \mathcal{T}''$ is a countable union of $\mathcal{A}' \times \mathcal{A}''$ sets. Therefore, by Theorem 3.1.2 it suffices to show that $\mu'_n \times \mu''_n(A) \to \mu' \times \mu''(A)$ for every $A \in \mathcal{A}' \times \mathcal{A}''$. Let $A = A' \times A''$ be in $\mathcal{A}' \times \mathcal{A}''$. Then $\mu'_n \times \mu''_n(A) = \mu'_n(A') \cdot \mu''_n(A'') \to \mu'(A') \cdot \mu''(A'') = \mu' \times \mu''(A)$. Left to right. Suppose that $\mu'_n \times \mu''_n \Rightarrow \mu' \times \mu''$. By Lemma 3.1.2, every element of $\mathcal{A}' \times \mathcal{A}''$ is almost surely decidable in $\mu' \times \mu''$. Therefore, for every $A' \in \mathcal{A}$, $\mu'_n(A') = \mu'_n(A') \cdot \mu''_n(\Omega'') = \mu'_n \times \mu''_n(A' \times \Omega'') \to \mu' \times \mu''(A' \times \Omega) = \mu'(A')$. Similarly, for every $A'' \in \mathcal{A}''$, $\mu''_n \Rightarrow \mu''$. By Theorem 3.1.2, $\mu'_n \Rightarrow \mu'$ and $\mu''_n \Rightarrow \mu''$.

□

3.1.4 Convergence in Distribution

In this section, we paraphrase the theory of weak convergence in terms of convergence in distribution. When stated in these terms, the theory may look more familiar to those accustomed to working with random variables.

A *probability space* is a sample space $(\Omega, \mathcal{I})$ endowed with a measure μ defined on $\mathcal{B}$, the Borel σ-algebra generated by $\mathcal{I}$. Let X be a function from a probability space $(\Omega, \mathcal{B}, \mu)$ to a metric space $(S, \mathcal{S})$. We say that X is a *random variable* if it is measurable $\mathcal{B}/\mathcal{S}$. If $\mathcal{S}$ is $\mathbb{R}$, we say that X is a random scalar. If $\mathcal{S}$ is $\mathbb{R}^k$, we say that X is a *random vector*. If $\mathcal{S}$ is $\mathbb{R}^{m\times n}$, then we say that X is a *random matrix*. The *distribution* of X is the probability measure P_μ on $(S, \mathcal{S})$ defined by $P_\mu(A) = \mu(X^{-1}A)$. The distribution of X is also called the *law* of X and denoted $\mathcal{L}(X)$. When X is a random vector, there is also the associated *distribution function* of $X = (X_1, \ldots, X_k)$, defined by $F(x_1, \ldots, x_k) = P_\mu[y \in \mathbb{R}^k : y_i \leq x_i, i \leq k]$.

We say that a sequence $\{X_n\}$ of random variables *converges in distribution* to the random variable X iff $\mathcal{L}(X_n) \Rightarrow \mathcal{L}(X)$. Then write $X_n \Rightarrow X$. The definition makes sense only if the range and the topology on it $(S, \mathcal{S})$ are the same for all $X, X_1, X_2 \ldots$. The domains may be distinct, and their structure is largely irrelevant, since they enter into the definition only by way of the distribution on $(S, \mathcal{S})$ that they induce. If $(S, \mathcal{S})$ is $\mathbb{R}^k$ with the usual topology, then $X_n \Rightarrow X$ iff $F_n(x_1, \ldots, x_k) \to F(x_1, \ldots, x_k)$ for all $(x_1, \ldots, x_k)$ such that $\{y : y_i \leq x_i, i \leq k\}$ is almost surely clopen in $\mathcal{L}(X)$.[9] That connects weak convergence with the more familiar definition of convergence in distribution.

We also introduce the notion of convergence in probability. Let $X, X_1, X_2, \ldots$ be a sequence of random vectors defined on the same space. Let $||X_n - X||$ be the Euclidean distance between the random variables, i.e.

$$||X_n - X|| = \sqrt{[X_{n,1}(\omega) - X_1(\omega)]^2 + \cdots + [X_{n,k}(\omega) - X_k(\omega)]^2}.$$

We say that the X_n converge to X iff

$$\lim_{n\to\infty} P(||X_n - X|| > \epsilon) = 0,$$

for all $\epsilon > 0$. Note that, for each n, $||X_n - X||$ is a real-valued random variable, so the relevant distributions are the $\mathcal{L}(||X_n - X||)$. Convergence in probability is indicated by writing $X_n \xrightarrow{P} X$. The definition is generalized to matrices by reading $||X_n - X||$ as the Euclidean distance between matrices.

We also state, without proof, Slutsky's well-known theorem. It will be invoked in Sections 3.6.1 and 3.6.1.

Theorem 3.1.5 (Slutsky's Theorem). *Let $\{X_n\}, \{Y_n\}$ be sequences of random scalars/vectors/matrices. If $X_n \Rightarrow X$ and $Y_n \xrightarrow{P} c$, where c is a constant, then*

- $X_n + Y_n \Rightarrow X + c$;
- $Y_n X_n \Rightarrow cX$,

provided that multiplication and addition are defined.

[9] For a proof of this fact, see Example 2.3 in Billingsley [1999].

3.2 Statistical Verification, Refutation and Decision

3.2.1 Defining the Success Concepts

In the setting of propositional information, hypothesis H was said to be verifiable iff there is an infallible method that converges on increasing information to H iff H is true. That condition implies that there is a method that achieves *every* bound on *chance* of error, and converges to H iff H is true.[10] In statistical settings, one cannot insist on such a high standard of infallibility. Instead, say that H is *verifiable in chance* iff for every bound on error, there is a method that achieves it, and that converges in probability to H iff H is true. The reversal of quantifiers expresses the fundamental difference between statistical and propositional verifiability and, hence, between statistical and propositional information.

Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible tests of H^{c} is an *α-verifier in chance* of $H \subseteq W$ iff for all $n \in \mathbb{N}$:

BndErr. $\mu^n[\lambda_n^{-1}(H)] \leq \alpha$, if $\mu \in H^{\mathsf{c}}$;

LimCon. $\mu^n[\lambda_n^{-1}(H)] \xrightarrow{n} 1$, if $\mu \in H$.

Say that H is *α-verifiable in chance* iff there is an α-verifier in chance of H. Say that H is *verifiable in chance* iff H is α-verifiable in chance for every $\alpha > 0$.

Several strengthenings of verifiability in chance immediately suggest themselves. One could demand that, in addition to BndErr, the chance of error vanishes to zero:

VanErr. $\mu^n[\lambda_n^{-1}(H)] \xrightarrow{n} 0$, if $\mu \in H^{\mathsf{c}}$.

If we wanted to be even more demanding, we could strengthen BndErr to the requirement that the total chance of error is bounded:

σ-BndErr. $\sum_{n=1}^{\infty} \mu^n[\lambda_n^{-1}(H)] \leq \alpha$, if $\mu \in H^{\mathsf{c}}$.

That implies both BndErr and VanErr, but also, by the Borel-Cantelli lemma, that with probability one, a verifier makes only finitely many errors on an infinite sample path, whenever H is false:

SVanErr. $\mu^{\infty}[\liminf \lambda_n^{-1}(W)] = 1$, if $\mu \in H^{\mathsf{c}}$.

We might demand similar asymptotic behavior when H is true. Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible tests of $H^{\mathsf{c}} \subseteq W$ is an *almost sure α-verifier* of H iff

σ-BndErr. $\sum_{n=1}^{\infty} \mu^n[\lambda_n^{-1}(H)] \leq \alpha$, if $\mu \in H^{\mathsf{c}}$, and

[10] If for every $\epsilon > 0$ the chance of error is less than ϵ, then the chance of error is zero: the method is *almost surely* infallible.

SLimCon. $\mu^\infty\left[\liminf \lambda_n^{-1}(H)\right] = 1$, if $\mu \in H$.

Say that H is *almost surely α-verifiable* iff there is an almost sure α-verifer of H. Say that H is *almost surely verifiable* iff H is almost surely α-verifiable, for every $\alpha > 0$.

Since almost sure convergence entails convergence in chance, almost surely verifiability entails verifiability in chance.[11] At this point, the reader may wonder why we introduce so many alternative definitions of statistical verification. There are two principal reasons. Firstly, α-verification in chance seems like the weakest success notion worthy of the name. Considering this weak success concept strengthens the necessity side of Theorem 3.2.1. Secondly, almost sure verification in chance, although it is logically stronger, is not thereby a "better" version of statistical verification. Almost sure verification is a reasonable goal for a single inquirer, or a single laboratory, that will collect a larger and larger sample. The in-chance notions are more natural for the situation in which different laboratories coordinate their methods in order to replicate an effect from larger and larger *independent* samples. From the perspective of an individual inquirer, almost sure verifiability is the natural notion. From the perspective of a science planner interested in patterns of independent replication, the in-chance notion is the natural one.

Defining statistical refutability requires no new ideas. Say that H is *α-refutable in chance* iff there is an α-verifier in chance of H^c. Say that H is *refutable in chance* iff H^c is α-verifiable in chance for every $\alpha > 0$. Say that H is *almost surely α-refutable* iff there is an almost sure α-verifer of H^c. Say that H is *almost surely refutable* iff H^c is almost surely α-verifiable, for every $\alpha > 0$.

Verification and refutation are asymmetrical concepts. If H is false, a verifier of H must bound its chance of error, but is excused from drawing any non-trivial conclusions. The two-sided notion of *statistical decision* removes that epistemic asymmetry. Say that a statistical method is a *two-sided test* of $H \subseteq W$ iff it is a function $\lambda : \Omega \to \{W, H, H^\mathsf{c}\}$. Let $H_\mu = H$ if $\mu \in H$ and $H_\mu = H^\mathsf{c}$ if $\mu \notin H$. Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible, two-sided tests of H is an *α-decision procedure in chance* for $H \subseteq W$ iff

BndErr. $\mu^n[\lambda_n^{-1}(H_\mu^c)] \leq \alpha$;

LimCon. $\mu^n[\lambda_n^{-1}(H_\mu)] \xrightarrow{n} 1$.

Say that H is *α-decidable in chance* iff there is an α-decision procedure in chance for H. Say that H is *decidable in chance* iff H is α-decidable in chance for every $\alpha > 0$.

Similarly, say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible, two-sided tests of H is an *almost sure α-decision procedure* for $H \subseteq W$ iff

[11] This entailment holds only for countably additive measures, to which we restrict attention.

σ-BndErr. $\sum_{n=1}^{\infty} \mu^n[\lambda_n^{-1}(H_\mu^{\mathsf{c}})] \leq \alpha$;

SLimCon. $\mu^\infty \left[\liminf \lambda_n^{-1}(H_\mu)\right] = 1$.

Say that H is *almost surely α-decidable* iff there is an almost sure α-decision procedure for H. Say that H is *almost surely decidable* iff H is almost surely α-decidable for every $\alpha > 0$.

3.2.2 Characterization Theorems

The central theorem of this work states that, for sample spaces with countable, almost surely clopen bases, verifiability in chance and almost sure verifiability are both equivalent to being open in the weak topology. As promised in the introduction, that fundamental result lifts the topological perspective to inferential statistics.

Theorem 3.2.1 (Fundamental Characterization Theorem)**.** *Suppose (1) that $\mathcal{I}$ is a countable base, (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-verifiable in chance for some $\alpha > 0$;*
2. *H is almost surely verifiable;*
3. *H is open in the weak topology on W.*

The characterization of statistical refutability follows immediately.

Theorem 3.2.2. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-refutable in chance for some $\alpha > 0$;*
2. *H is almost surely refutable;*
3. *H is closed in the weak topology on W.*

Finally, we give a characterization of statistical decidability.

Theorem 3.2.3. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-decidable in chance for some $\alpha > 0$;*
2. *H is almost surely decidable;*
3. *H is clopen in the weak topology on W.*

In light of Theorems 3.2.1, 3.2.2 and 3.2.3 we will say that a hypothesis is simply *statistically verifiable/refutable/decidable* when the precise sense of verifiability/refutability/decidability is not relevant.

Since a topological space is determined uniquely by its open sets, Theorem 3.2.1 implies that the weak topology is the *unique* topology that characterizes statistical verifiability under the weak conditions stated in the antecedent of the theorem. Thus, under those conditions, the weak topology is not merely a convenient formal tool—it is *the* topology of statistical information.

For an elementary application of Theorem 3.2.1, consider, in the setting of Example 3.1.1, the sharp hypothesis, $F := \{\mu_0\}$, that the bias of the coin is exactly .5. By Theorem 3.1.3, the collection $\{\{\mu : \mu(A) > r\} : A \in \{H, T\}\}$ is a sub-basis for the weak topology on W. Therefore, the hypotheses of head-bias, $B_H := \{\mu : \mu(H) > .5\}$, and tail-bias, $B_T := \{\mu : \mu(T) > .5\}$, are open in the weak topology. It follows that the hypothesis that the coin is biased, $B := B_H \cup B_T$, is open, and therefore, statistically verifiable, and the hypothesis that it is fair, $F = B^{\mathsf{c}}$, is closed, and therefore statistically refutable. However, since the collection (μ_n), defined by $\mu_n(H) = .5 + \epsilon/n$, converges weakly to μ_0, the hypothesis that the coin is fair is not open, and therefore, not statistically verifiable.

The proof of Theorem 3.2.1 proceeds in three steps. It is immediate from the definitions (and the assumption of countable additivity) that 2 implies 1. Given what we have developed so far, it is easy to show that 1 implies 3. The idea is simple, but fundamental. If a hypothesis H is not open, then it must contain a boundary point μ_0. Then, there must be a sequence (μ_n) contained in H^{c} converging weakly to μ_0 and, therefore, the probability of falsely inferring H in the μ_n is converging to the probability of correctly inferring H in μ_0. That is the fundamental expression of the statistical problem of induction. We use that feature to derive a contradiction from the assumption of the existence of a statistical verifier — any such ‘verifier’ would have to violate BndErr.

Proof of Theorem 3.2.1. 1 implies 3. Suppose, for contradiction, that H is not open, but that $(\lambda_n)_{n\in\mathbb{N}}$ is an α-verifier in chance for H. Let $\mu \in H \cap \mathsf{bdry} H$. Then there is a sequence of μ_n in H^{c} such that $\mu_n \Rightarrow \mu$. Since $(\lambda_n)_{n\in\mathbb{N}}$ is a verifier for H, there is a sample size k such that $\mu^k(\lambda_k^{-1}(H)) > \alpha + \epsilon$. By Theorem 3.1.4, $\mu_n^k \Rightarrow \mu^k$, and therefore $\mu_n^k(\lambda_k^{-1}(H)) \to \mu^k(\lambda_k^{-1}(H))$. So there is a $\mu_m \in H^{\mathsf{c}}$ such that $\mu_m^k(\lambda_k^{-1}(H)) > \alpha$. Contradiction. □

To show that 3 implies 2, we first prove that for almost surely decidable A, the hypothesis $\{\mu : \mu(A) > r\}$ is almost surely verifiable. That entails that every element of the subbasis for the weak topology exhibited in Theorem 3.1.3 is almost surely verifiable. Finally, we show that almost sure verifiability is preserved by finite conjunctions and countable disjunctions, which completes the proof. The proof of Theorem 3.2.3 is provided at the end of the section.

Lemma 3.2.1. *Suppose that B is almost surely decidable for every $\mu \in W$. Then, for all $b \in \mathbb{R}$, the hypothesis $H = \{\mu : \mu(B) > b\}$ is almost surely verifiable.*

Proof of Lemma 3.2.1. Define the indicator random variable $\mathbb{1}_B : \Omega \to \{0,1\}$ by $\mathbb{1}_B(\omega) = 1$ if $\omega \in B$, otherwise $\mathbb{1}_B = 0$. By Hoeffding's inequality,

$$\mu^n \left[\left\{ (\omega_1, \dots, \omega_n) : \sum_{i=1}^{n} \mathbb{1}_B(\omega_i) \geq n\left(\mu(B) + t_n\right) \right\} \right] \leq \exp(-2nt_n^2).$$

Letting $t_n = \sqrt{\frac{1}{2n} \ln(\pi^2 n^2 / 6\alpha)}$, it follows from Hoeffding's inequality that:

$$\mu^n \left[\left\{ (\omega_1, \dots, \omega_n) : \sum_{i=1}^{n} \mathbb{1}_B(\omega_i) \geq n\left(\mu(B) + t_n\right) \right\} \right] \leq \frac{6\alpha}{\pi^2 n^2}.$$

We argue that $t_n \overset{n}{\to} 0$. Rewriting, we have that $t_n = \sqrt{\frac{\ln(\pi n)}{n} - \frac{\ln(\sqrt{6\alpha})}{n}}$. Clearly, $\frac{\ln(\sqrt{6\alpha})}{n} \to 0$. By L'Hopital's rule, it follows that $\frac{\ln(\pi n)}{n} \to 0$. Therefore, by limit algebra, $t_n^2 \to 0$. Next, we notice that $\frac{\ln(\pi n)}{n} > \frac{\ln(\sqrt{6\alpha})}{n}$, and therefore that $t_n^2 \geq 0$. Finally, we appeal to the standard fact that, if $a_n \geq 0$ and $a_n \to t$, then $\sqrt{a_n} \to \sqrt{t}$.

Let

$$\lambda_n(\omega_1, \dots, \omega_n) = \begin{cases} H, & \text{if } \sum_{i=1}^{n} \mathbb{1}_B(\omega_i) \geq n(b + t_n) \\ W, & \text{otherwise.} \end{cases}$$

By Lemma 3.1.3, λ^n is a feasible method. First, we show that (λ_n) satisfies σ-BndErr. Suppose that $\mu \notin H$. Then, $b \geq \mu(B)$ and:

$$\begin{aligned} \sum_{n=1}^{\infty} \mu^n \left[\lambda_n^{-1}(H)\right] &= \sum_{n=1}^{\infty} \mu^n \left[\left\{ (\omega_1, \dots, \omega_n) : \sum_{i=1}^{n} \mathbb{1}_B(\omega_i) \geq n(b + t_n) \right\} \right] \\ &\leq \sum_{n=1}^{\infty} \mu^n \left[\left\{ (\omega_1, \dots, \omega_n) : \sum_{i=1}^{n} \mathbb{1}_B(\omega_i) \geq n(\mu(B) + t_n) \right\} \right] \\ &\leq \sum_{n=1}^{\infty} \frac{6\alpha}{\pi^2 n^2} = \alpha, \end{aligned}$$

where the final equality follows from the fact that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \pi^2/6$.

It remains to show that (λ_n) satisfies SLimCon. Suppose that $\mu \in H$. Then, $\mu(B) > b$. Then, since $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_B(\omega_i) \overset{a.s.}{\to} \mathbb{E}[\mathbb{1}_B] = \mu(B)$, by the strong law of large numbers, and $t_n \to 0$, we have that $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_B(\omega_i) - t_n \overset{a.s.}{\to} \mu(B) > b$. Therefore, $\mu^\infty \left[\liminf_{n\to\infty} \lambda_n^{-1}(H)\right] = 1$, as required. □

Lemma 3.2.2. *The almost surely verifiable propositions are closed under finite conjunctions, and countable disjunctions.*

Proof of Lemma 3.2.2. Suppose that A_1, A_2 are a.s. verifiable. Let $\alpha > 0$. Let $\{\lambda_n^i\}_{n \in \mathbb{N}}$ be an a.s. α-verifier for A_i. Let $\lambda_n(\vec{\omega}) = A_1 \cap A_2$ if $\lambda_n^i(\vec{\omega}) = A_i$, for $i \in \{1, 2\}$. By Lemma 3.1.1, λ_n is feasible, for each $\mu \in W$, $n \in \mathbb{N}$. Suppose that $\mu \in A_1 \cap A_2$. Then:

$$\begin{aligned}
&\mu^\infty \left[\liminf_{n\to\infty} \lambda_n^{-1}(A_1 \cap A_2)\right] = \\
&= \mu^\infty \left[\liminf_{n\to\infty} (\lambda_n^1)^{-1}(A_1) \cap (\lambda_n^2)^{-1}(A_2)\right] \\
&= 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda_n^1)^{-1}(W) \cup (\lambda_n^2)^{-1}(W)\right] \\
&= 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda_n^1)^{-1}(W) \cup \limsup_{n\to\infty} (\lambda_n^2)^{-1}(W)\right] \\
&\geq 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda_n^1)^{-1}(W)\right] - \mu^\infty \left[\limsup_{n\to\infty} (\lambda_n^2)^{-1}(W)\right] \\
&= -1 + \mu^\infty \left[\liminf_{n\to\infty} (\lambda_n^1)^{-1}(A_1)\right] + \mu^\infty \left[\liminf_{n\to\infty} (\lambda_n^2)^{-1}(A_2)\right] \\
&= 1.
\end{aligned}$$

Suppose that $\mu \notin A_1 \cap A_2$. Without loss of generality, suppose $\mu \notin A_1$. Then:

$$\begin{aligned}
&\sum_{n=1}^{\infty} \mu^\infty \left[\lambda_n^{-1}(A_1 \cap A_2)\right] = \\
&= \sum_{n=1}^{\infty} \mu^\infty \left[(\lambda_n^1)^{-1}(A_1) \cap (\lambda_n^2)^{-1}(A_2)\right] \\
&\leq \sum_{n=1}^{\infty} \mu^\infty \left[(\lambda_n^1)^{-1}(A_1)\right] \leq \alpha.
\end{aligned}$$

To show that the a.s. verifiable propositions are closed under countable union, suppose that $A_1, A_2, \ldots$ are a.s. verifiable. For $i \in \mathbb{N}$, let $\{\lambda_n^i\}_{n\in\mathbb{N}}$ be an a.s. α_i-verifier for A_i with $\alpha_i = \alpha/2^i$. Let $\lambda_n(\vec{\omega}) = \bigcup_{i=1}^{\infty} A_i$ if $\lambda_n^i(\vec{\omega}) = A_i$ for some $i \in \{1, \ldots, n\}$, and let $\lambda_n(\vec{\omega}) = W$ otherwise. By Lemma 3.1.1, λ_n is feasible for each $\mu \in W, n \in \mathbb{N}$. Suppose that $\mu \in \bigcup_{i=1}^{\infty} A_i$. Then there exists $j \in \mathbb{N}$ such that $\mu \in A_j$. Furthermore:

$$\begin{aligned}
\mu^\infty \left[\liminf_{n\to\infty} \lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)\right] &= \mu^\infty \left[\liminf_{n\to\infty} \cup_{k\leq n} (\lambda_n^k)^{-1}(A_k)\right] \\
&\geq \mu^\infty \left[\liminf_{n\to\infty} (\lambda_n^j)^{-1}(A_j)\right] = 1.
\end{aligned}$$

Suppose that $\mu \notin \cup_{i=1}^{\infty} A_i$. Then:

$$\begin{aligned}\sum_{n=1}^{\infty} \mu^{\infty}\left[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)\right] &= \sum_{n=1}^{\infty} \mu^{\infty}\left[\cup_{k=1}^{n} (\lambda_n^k)^{-1}(A_k)\right] \\ &\leq \sum_{n=1}^{\infty} \mu^{\infty}\left[\cup_{k=1}^{\infty} (\lambda_n^k)^{-1}(A_k)\right] \\ &\leq \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \mu^{\infty}\left[(\lambda_n^k)^{-1}(A_k)\right] \\ &\leq \sum_{k=1}^{\infty} \alpha/2^k = \alpha.\end{aligned}$$

□

Proof of Theorem 3.2.3. It is immediate from the definitions that 2 implies 1. To see that 1 implies 3, suppose that (λ_n) is an α-decision procedure in chance for H. For $H \subseteq W$, define the set function $\chi_H : \mathcal{P}(W) \to \{H, W\}$ as:

$$\chi_H(A) = \begin{cases} H, & \text{if } A \subseteq H, \\ W, & \text{otherwise.} \end{cases}$$

Then, $(\chi_H \circ \lambda_n)$ is a family of (one-sided) tests of H^{c}. We claim that $(\chi_H \circ \lambda_n)$ is an α-verifier in chance of H. Clearly, $\mu^n[\chi_H \circ \lambda_n^{-1}(H)] = \mu^n[\lambda_n^{-1}(H)]$. First, we verify that $(\chi_H \circ \lambda_n)$ satisfies BndErr. Suppose that $\mu \in H^{\mathsf{c}}$. Then, $\mu^n[\chi_H \circ \lambda_n^{-1}(H)] = \mu^n[\lambda_n^{-1}(H)] \leq \alpha$, as required. To see that $(\chi_H \circ \lambda_n)$ satisfies LimCon, suppose that $\mu \in H$. Then, $\lim_{n\to\infty} \mu^n[\chi_H \circ \lambda_n^{-1}(H)] = \lim_{n\to\infty} \mu^n[\lambda_n^{-1}(H)] = 1$, as required. Therefore, H is α-verifiable in chance and, by Theorem 3.2.1, open in the weak topology. An identical argument establishes that $(\chi_{H^{\mathsf{c}}} \circ \lambda_n)$ is an α-verifier in chance of H^{c}, which is therefore open in the weak topology.

It remains to show that 3 implies 2. Suppose that H is clopen in the weak topology. Let (λ_n^H) be an almost sure α-verifier of H, and let $(\lambda_n^{H^{\mathsf{c}}})$ be an almost sure α-verifier of H^{c}. Let

$$\lambda_n(\vec{\omega}) = \begin{cases} H^{\mathsf{c}}, & \text{if } \lambda_n^H(\vec{\omega}) = W \text{ and } \lambda_n^{H^{\mathsf{c}}}(\vec{\omega}) = H^{\mathsf{c}}, \\ H, & \text{if } \lambda_n^H(\vec{\omega}) = H \text{ and } \lambda_n^{H^{\mathsf{c}}}(\vec{\omega}) = W, \\ W, & \text{otherwise.} \end{cases}$$

First we show that (λ_n) satisfies σ-BndErr:

$$\begin{aligned}\sum_{i=1}^{\infty}\mu^n[\lambda_n^{-1}(H_\mu^{\mathsf{c}})] &= \sum_{i=1}^{\infty}\mu^n\left[(\lambda_n^{H_\mu})^{-1}(W)\cap(\lambda_n^{H_\mu^{\mathsf{c}}})^{-1}(H_\mu^{\mathsf{c}})\right]\\ &\leq \sum_{i=1}^{\infty}\mu^n\left[(\lambda_n^{H_\mu^{\mathsf{c}}})^{-1}(H_\mu^{\mathsf{c}})\right]\leq\alpha.\end{aligned}$$

It remains to show that (λ_n) satistifes SLimCon:

$$\begin{aligned}\mu^\infty[\liminf_{n\to\infty}\lambda_n^{-1}(H_\mu)] &= \mu^\infty[\liminf_{n\to\infty}(\lambda_n^{H_\mu})^{-1}(H_\mu)\cap(\lambda_n^{H_\mu^{\mathsf{c}}})^{-1}(W)]\\ &= 1-\mu^\infty[\limsup_{n\to\infty}(\lambda_n^{H_\mu})^{-1}(W)\cup(\lambda_n^{H_\mu^{\mathsf{c}}})^{-1}(H_\mu^{\mathsf{c}})]\\ &= 1-\mu^\infty[\limsup_{n\to\infty}(\lambda_n^{H_\mu})^{-1}(W)\cup\limsup_{n\to\infty}(\lambda_n^{H_\mu^{\mathsf{c}}})^{-1}(H_\mu^{\mathsf{c}})]\\ &\geq 1-\mu^\infty[\limsup_{n\to\infty}(\lambda_n^{H_\mu})^{-1}(W)]-\mu^\infty[\limsup_{n\to\infty}(\lambda_n^{H_\mu^{\mathsf{c}}})^{-1}(H_\mu^{\mathsf{c}})]\\ &= 1.\end{aligned}$$

□

3.2.3 Application: Conditional Independence Testing

The concepts of marginal and conditional independence are central to statistics, machine learning and neighboring fields [Dawid, 1979, Zhang et al., 2011]. Conditional independence plays an especially crucial role in causal discovery and Bayesian network learning [Spirtes et al., 2000, Pearl, 2000]. In this section we consider the verifiability of hypotheses about conditional dependence of random variables. The central result of this section is Theorem 3.2.7, which shows that, under a weak condition, conditional dependencies are statistically verifiable. That result is crucial for learning causal hypotheses in the framework of causal Bayes nets. First, we develop some basic results.

Theorem 3.2.4. *For almost surely clopen events A, B, C and $r \in [0,1]$, the following hypotheses are statistically verifiable:*

1. $\{\mu : \mu(A) > r\}$;
2. $\{\mu : \mu(A) < r\}$;
3. $\{\mu : \mu(A)\mu(B) > r\}$;
4. $\{\mu : 0 < \mu(A)\mu(B) < r\}$;
5. $\{\mu : \mu(A\cap B) \neq \mu(A)\mu(B)\}$;
6. $\{\mu : \mu(C) > 0, \frac{\mu(A)}{\mu(C)} > r\}$;
7. $\{\mu : \mu(C) > 0, \frac{\mu(A)}{\mu(C)} < r\}$;
8. $\{\mu : \mu(C) > 0, \frac{\mu(A)\mu(B)}{\mu(C)} > r\}$;
9. $\{\mu : \mu(C) > 0, 0 < \frac{\mu A \mu B}{\mu C} < r\}$;

10. $\{\mu : \mu(C) > 0, \frac{\mu(A\cap B\cap C)}{\mu(C)} \neq \frac{\mu(A\cap C)\mu(B\cap C)}{\mu(C)}\}$.

Proof of Theorem 3.2.4.

1. By Lemma 3.2.1.

2. Follows from (1) and the fact that

$$\{\mu : \mu(A) < r\} = \{\mu : 1 - \mu(A) > 1 - r\} = \{\mu : \mu(A^c) > 1 - r\}.$$

Since A^c is almost surely clopen iff A is almost surely clopen, we are done by part (1).

3. We argue that:

$$\{\mu : \mu(A)\mu(B) > r\} = \bigcup_{s\in(0,1]\cap\mathbb{Q}} \left(\{\mu : \mu(A) > r/s\} \cap \{\mu : \mu(B) > s\}\right).$$

Suppose that ν is contained in the right-hand side. Then, there is $s \in (0,1]\cap\mathbb{Q}$ such that $\nu(A)s > r$ and $\nu(B) > s$. It follows that $\nu(A)\nu(B) > \nu(A)s > r$. Suppose that ν is contained in the left-hand side. Since $\nu(A)\nu(B) > r \geq 0$, $0 \leq \frac{r}{\nu(A)} < \nu(B) \leq 1$. Since the rationals are dense in the reals, there is $s' \in (\frac{r}{\nu(A)}, \nu(B)) \cap (0,1] \cap \mathbb{Q}$. Furthermore, $\nu(A)s' > \nu(A)\frac{r}{\nu(A)} = r$, and $\nu(B) > s'$. By (1), we have expressed $\{\mu : \mu(A)\mu(B) > r\}$ as a countable union of finite intersections of statistically verifiable hypotheses. By Lemma 3.2.2, the statistically verifiable hypotheses are closed under finite conjunctions and countable disjunctions, therefore $\{\mu : \mu(A)\mu(B) > r\}$ is statistically verifiable.

4. We argue that:

$$\{\mu : 0 < \mu(A)\mu(B) < r\} = \bigcup_{s\in(0,1]\cap\mathbb{Q}} \left(\{\mu : 0 < \mu(A) < r/s\} \cap \{\mu : 0 < \mu(B) < s\}\right) \cup \{\mu : 0 < \mu(A) < r, \mu(B) > 0\}$$

Suppose that ν is an element of the rhs. Then, $\nu(B) > 0$ and either $0 < \nu(A) < r$, or there is $s \in (0,1]\cap\mathbb{Q}$ such that $0 < \nu(A)s < r$ and $0 < \nu(B) < s$. In the first case, $0 < \nu(A)\nu(B) \leq \nu(A) < r$, and therefore ν is an element of the lhs. In the second case, $0 < \nu(A)\nu(B) < \nu(A)s < r$. We have shown that the rhs is included in the lhs. Suppose that ν is an element of the lhs. Then either $\nu(B) = 1$ or $\nu(B) \in (0,1)$. In the first case, $\mu(B) > 0$ and $0 < \mu(A) < r$, and therefore ν is an element of the rhs. In the second case, $\nu(B) \in (0,1)$. Then, $0 < \nu(A)\nu(B) < r$, and $\frac{r}{\nu(A)} > \nu(B)$. Since the rationals are dense in the reals, there is $s' \in (\nu(B), \frac{r}{\nu(A)}) \cap \mathbb{Q} \cap (0,1]$. Furthermore, $0 < \nu(A)s' < \nu(A)\frac{r}{\nu(A)} = r$, and $0 < \nu(B) < s'$. By parts (1) and (2), each disjunct is a finite intersection of statistically verifiable hypotheses. Since the disjunction is countable, $\{\mu : 0 < \mu(A)\mu(B) < r\}$ is statistically verifiable by Lemma 3.2.2.

5. Follows from (10), taking Ω for C.

6. Follows from (8), taking Ω for C.

7. In the case where $r = 0$,

$$\{\mu : \mu(C) > 0, \frac{\mu(A)}{\mu(C)} < 0\} = \varnothing,$$

which is trivially verifiable. Suppose that $r \in (0, 1]$. We argue that:

$$\{\mu : \mu(C) > 0, \frac{\mu(A)}{\mu(C)} < r\} = \bigcup_{s \in (0,1] \cap \mathbb{Q}} \{\mu : \mu(A) < rs\} \cap \{\mu : \mu(C) > s\}.$$

Suppose that ν is an element of the rhs. Then, there is $s \in (0, 1]$ such that $0 < s < \nu(C)$, and therefore $0 < \frac{1}{\nu(C)} < \frac{1}{s}$. It follows that $\frac{\nu(A)}{\nu(C)} < \frac{\nu(A)}{s} < r$. We have shown that the rhs is contained in the lhs. Suppose that ν is an element of the lhs. Then, $0 \leq \frac{\nu(A)}{r} < \nu(C) \leq 1$. Since the rationals are dense in the reals, there is $s' \in (\frac{\nu(A)}{r}, \nu(C)) \cap (0, 1] \cap \mathbb{Q}$. Clearly, $\nu(C) > s'$. Furthermore, since $\frac{\nu(A)}{r} < s'$, it follows that $\nu(A) < rs'$. We have shown that the lhs is contained in the rhs. By (1) and (2), we have expressed the hypothesis as a countable union of finite intersections of statistically verifiable hypotheses. Therefore, it is statistically verifiable by Lemma 3.2.2.

8. In the case where $r = 0$,

$$\begin{aligned}\{\mu : \mu(C) > 0 \text{ and } \frac{\mu(A)\mu(B)}{\mu(C)} > 0\} = \\ = \{\mu : \mu(C) > 0\} \cap \{\mu : \mu(A) > 0\} \cap \{\mu : \mu(B) > 0\}.\end{aligned}$$

By part (1), this is a finite conjunction of statistically verifiable hypotheses. By Lemma 3.2.2, it is statistically verifiable.

Suppose that $r \in (0, 1]$. We argue that

$$
\begin{aligned}
&\{\mu : \mu(C) > 0 \text{ and } \frac{\mu(A)\mu(B)}{\mu(C)} > r\} = \\
&\qquad = \bigcup_{s \in (0,1] \cap \mathbb{Q}} (\{\mu : \mu(A)\mu(B) > rs\} \cap \{\mu : 0 < \mu(C) < s\}) \cup \\
&\qquad \cup \{\mu : \mu(C) > 0 \text{ and } \mu(A)\mu(B) > r\}.
\end{aligned}
$$

Suppose that ν is an element of the rhs. Then, either $\nu(C) > 0$ and $\nu(A)\nu(B) > r$, or there is $s \in (0,1] \cap \mathbb{Q}$ such that $rs < \nu(A)\nu(B)$ and $0 < \nu(C) < s$. In the first case, $\frac{\nu(A)\nu(B)}{\nu(C)} \geq \nu(A)\nu(B) > r$, and therefore ν is an element of the lhs. In the second case, $0 < \frac{1}{s} < \frac{1}{\nu(C)}$, and $r < \frac{\nu(A)\nu(B)}{s} < \frac{\nu(A)\nu(B)}{\nu(C)}$. We have shown that the rhs is contained in the lhs.

Suppose that ν is an element of the lhs. Then, either $\nu(C) = 1$ or $\nu(C) \in (0,1)$. In the first case, $\nu \in \{\mu : \mu(C) > 0 \text{ and } \mu(A)\mu(B) > r\}$, so ν is an element of the rhs. Suppose that $\nu(C) \in (0,1)$. Then, $\frac{\mu(A)\mu(B)}{r} > \mu(C) > 0$. Since $\mu(C) \in (0,1)$, and the rationals are dense in the reals, there is $s' \in (\mu(C), \frac{\mu(A)\mu(B)}{r}) \cap \mathbb{Q} \cap (0,1]$. Furthermore, $\mu(A)\mu(B) > r\mu(C) > rs'$ and $0 < \mu(C) < s'$. We have shown the lhs is contained in the rhs. By (1), (2) and (3), we have expressed the hypothesis as a countable union of finite intersections of statistically verifiable hypotheses. Therefore, it is statistically verifiable by Lemma 3.2.2.

9. We argue that

$$
\begin{aligned}
&\{\mu : \mu(C) > 0 \text{ and } 0 < \frac{\mu(A)\mu(B)}{\mu(C)} < r\} \\
&\qquad = \bigcup_{s \in (0,1] \cap \mathbb{Q}} (\{\mu : 0 < \mu(A)\mu(B) < rs\} \cap \{\mu : \mu(C) > s\}).
\end{aligned}
$$

Suppose that ν is an element of the rhs. Then, there is $s \in (0,1]$ such that $0 < \nu(A)\nu(B) < rs$ and $0 < s < \nu(C)$. That entails that $0 < \frac{1}{\nu(C)} < \frac{1}{s}$, and that $0 < \frac{\nu(A)\nu(B)}{\nu(C)} < \frac{\nu(A)\nu(B)}{s} < r$. We have shown the the rhs is contained in the lhs. Suppose that ν is an element of the lhs. Then, $r > 0$ and $0 < \frac{\nu(A)\nu(B)}{r} < \nu(C) \leq 1$. Since the rationals are dense in the reals, there is $s' \in (\frac{\nu(A)\nu(B)}{r}, \nu(C)) \cap (0,1] \cap \mathbb{Q}$. Clearly, $\nu(C) > s'$ and furthermore, $\frac{\nu(A)\nu(B)}{r} < s'$, which implies that $0 < \nu(A)\nu(B) < rs'$. We have shown the lhs is contained in the rhs. By (1) and (4), we have expressed the hypothesis as a countable union of finite intersections of statistically verifiable hypotheses. Therefore, it is statistically verifiable.

10. We argue that

$$\{\mu : \mu(C) > 0, \frac{\mu(A \cap B \cap C)}{\mu(C)} \neq \frac{\mu(A \cap C)\mu(B \cap C)}{\mu(C)}\} =$$

$$= \bigcup_{r \in (0,1) \cap \mathbb{Q}} \left(\{\mu : \mu(C) > 0, \frac{\mu(ABC)}{\mu(C)} < r\} \cap \{\mu : \mu(C) > 0, \frac{\mu(AC)\mu(BC)}{\mu(C)} > r\} \right) \cup$$
$$\left(\{\mu : \mu(C) > 0, \frac{\mu(ABC)}{\mu(C)} > r\} \cap \{\mu : \mu(C) > 0, 0 < \frac{\mu(AC)\mu(BC)}{\mu(C)} < r \ \} \right).$$

It is clear that the rhs is contained in the lhs. Suppose that ν is an element of the lhs. Then, it must be that $\nu(AC), \nu(BC) > 0, \nu(C) > 0$. Furthermore, either $\frac{\nu(ABC)}{\nu(C)} > \frac{\nu(AC)\nu(BC)}{\nu(C)}$ or $\frac{\nu(ABC)}{\nu(C)} < \frac{\nu(AC)\nu(BC)}{\nu(C)}$. In the first case, since, $0 < \frac{\nu(AC)\nu(BC)}{\nu(C)} < \frac{\nu(ABC)}{\nu(C)} \leq 1$, and the rationals are dense in the reals, there is $r \in (\frac{\nu(AC)\nu(BC)}{\nu(C)}, \frac{\nu(ABC)}{\nu(C)}) \cap (0,1) \cap \mathbb{Q}$. Therefore ν is an element of the rhs. In the second case, $0 < \frac{\nu(ABC)}{\nu(C)} < \frac{\nu(AC)\nu(BC)}{\nu(C)} \leq 1$. Since the rationals are dense in the reals, there is $r \in (\frac{\nu(ABC)}{\nu(C)}, \frac{\nu(AC)\nu(BC)}{\nu(C)}) \cap (0,1) \cap \mathbb{Q}$. By (1), and (6-9) we have expressed the hypothesis as a union of finite intersections of statistically verifiable hypotheses. Therefore, it is statistically verifiable.

□

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, and let $\mathcal{A}, \mathcal{B}$ be two subsets of $\mathcal{F}$. We say that $\mathcal{A}, \mathcal{B}$ are μ-independent if, whenever $A \in \mathcal{A}$ and $B \in \mathcal{B}$, $\mu(A \cap B) = \mu(A)\mu(B)$. Say that a collection of subsets of $\mathcal{F}$, $(\mathcal{A}_i, i \in I)$, for arbitrary index set I, are mutually μ-independent iff for any finite $J \subseteq I$ and any $A_i \in \mathcal{A}_i$, $\mu(\cap_{i \in J} A_i) = \prod_{i \in J} \mu(A_i)$.

The σ-algebra, $\sigma(X)$, generated by a random variable X taking values in some measurable space $(S, \mathcal{S})$, is defined as the collection

$$\{X^{-1}(U) : U \in \mathcal{S}\}.^{12}$$

Two random variables X, Y defined over Ω are said to be *μ-independent* iff the σ-algebras they generate, $\sigma(X), \sigma(Y)$ are μ-independent, i.e. if, whenever $A \in \sigma(X)$ and $B \in \sigma(Y)$, $\mu(A \cap B) = \mu(A)\mu(B)$. We say that a collection of random variables $X_i : (\Omega, \mathcal{F}, \mu) \to (S_i, \mathcal{S}_i)$, $i \in I$ for some index set I, are mutually μ-independent if for any finite $J \subseteq I$, the collection of σ-algebras $(\sigma(X_i), i \in J)$ are mutually μ-independent.

It is a well known consequence of the $\pi - \lambda$ theorem that μ-independence of random variables X, Y is equivalent to independence of any π-systems, $\pi(X), \pi(Y)$,

[12] The fact that this is a σ-algebra is a consequence of the general fact that $f^{-1}(\sigma(\mathcal{C})) = \sigma(f^{-1}(\mathcal{C}))$.

generating $\sigma(X), \sigma(Y)$. As a warm-up, we prove this reduction. Recall that a λ-system is a collection $\mathcal{L}$ of subsets of Ω such that:

1. $\Omega \in \mathcal{L}$,

2. if $A \in \mathcal{L}$, then $A^c \in \mathcal{L}$,

3. if $(A_i)_{i \in \mathbb{N}}$ is a sequence of disjoint subsets in $\mathcal{L}$ then, $\cup_{i=1}^{\infty} A_i$ is in $\mathcal{L}$.

Theorem 3.2.5 ($\pi - \lambda$ Theorem)**.** *Let $\mathcal{L}$ be a λ-system, and let $\mathcal{P} \subseteq \mathcal{L}$ be a π-system contained in $\mathcal{L}$. Then the σ-algebra generated by $\mathcal{P}$ is contained in $\mathcal{L}$.*

Proof. For a proof see Theorem 3.2 in Billingsley [1986], or any other book on measure theory. □

Lemma 3.2.3. *If $\mathcal{G}_i \subseteq \mathcal{F}$, $i \in I$ is a mutually μ-independent collection of π-systems, then $\sigma(\mathcal{G}_i)$, $i \in I$, is a mutually μ-independent collection of σ-algebras.*

Proof. Suppose that $\mathcal{G}_i \subseteq \mathcal{F}$, $i \in I$ is a mutually μ-independent collection of π-systems. Let J be a finite subset of I. Let $j \in J$. Let

$$\mathcal{E} = \{E \in \mathcal{G}_j : \mu\left(\cap_{i \in J \setminus \{j\}} B_i \cap E\right) = \mu(E) \prod_{i \in J \setminus \{i\}} \mu(B_i), \quad \forall (B_i, i \in J \setminus \{j\}) \in \times_{i \in J \setminus \{j\}} \mathcal{G}_i\}.$$

We show that $\mathcal{E}$ is a λ-system. Since the $\mathcal{G}_i$ are mutually μ-independent, it is clear that $\Omega \in \mathcal{E}$. Next, we show that $\mathcal{E}$ is closed under complements. Suppose that $A \in \mathcal{E}$. Let $(B_i, i \in J \setminus \{j\}) \in \times_{i \in J \setminus \{j\}} \mathcal{G}_i\}$. Then, $\mu(\cap_{i \in J \setminus \{j\}} B_i) = \mu(\cap_{i \in J \setminus \{j\}} B_i \cap A) + \mu(\cap_{i \in J \setminus \{j\}} B_i \cap A^{\mathsf{c}})$. But $\mu(\cap_{i \in J \setminus \{j\}} B_i \cap A) = \mu(A) \prod_{i \in J \setminus \{i\}} \mu(B_i)$. Rearranging, $\mu(\cap_{i \in J \setminus \{j\}} B_i \cap A^{\mathsf{c}}) = \prod_{i \in J \setminus \{i\}} \mu(B_i)(1 - \mu(A)) = \mu(A^{\mathsf{c}}) \prod_{i \in J \setminus \{i\}} \mu(B_i)$. Finally, we show that $\mathcal{E}$ is closed under disjoint unions. Suppose the collection $(A_i, i \in \mathbb{N})$ is disjoint and each A_i is in $\mathcal{E}$. Then, $\mu(\cap_{i \in J \setminus \{j\}} B_i \cap \cup_i A_i) = \sum_{i=1}^{\infty} \mu(\cap_{i \in J \setminus \{j\}} B_i \cap A_i) = \mu(\cap_{i \in J \setminus \{j\}} B_i) \sum_{i=1}^{\infty} \mu(A_i) = \mu(\cap_{i \in J \setminus \{j\}} B_i) \mu(\cup_i A_i)$. Now, by the $\pi - \lambda$ theorem, if the $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n$ are mutually μ-independent, then $\sigma(\mathcal{G}_1), \mathcal{G}_2, \ldots, \mathcal{G}_n$ are mutually μ-independent. Iterating application of the $\pi - \lambda$ theorem, we have that $\sigma(\mathcal{G}_1), \sigma(\mathcal{G}_2), \ldots, \sigma(\mathcal{G}_n)$ are mutually μ-independent. □

It follows immediately that a collection of random variables $X_i : (\Omega, \mathcal{F}, \mu) \to (S_i, \mathcal{S}_i)$, $i \in I$ are mutually μ-independent if for any finite $J \subseteq I$, there is a collection of mutually μ-independent π-systems, $(\pi(X_i), i \in J)$, that generate the σ-algebras $(\sigma(X_i), i \in J)$.

Corollary 3.2.1. *Suppose that X, Y are random variables and that the sigma algebras $\sigma(X), \sigma(Y)$ are generated by countable almost surely clopen bases $\mathcal{I}(X)$, $\mathcal{I}(Y)$. Then the hypothesis of dependence $\{\mu : X \not\perp\!\!\!\perp_{\mu} Y\}$ is statistically verifiable.*

Proof. By Lemma 3.2.3, random variables X, Y are μ-independent iff $\mu(A \cap B) = \mu(A)\mu(B)$ for all A, B in a π-system generating $\sigma(X), \sigma(Y)$ respectively. Let $\mathcal{A}(X), \mathcal{A}(Y)$ be the algebras generated by $\mathcal{I}(X), \mathcal{I}(Y)$. Obviously, $\mathcal{A}(X), \mathcal{A}(Y)$

are π-systems. Furthermore, they are almost surely clopen by Corollary 3.1.1. Therefore,

$$\{\mu : X \not\perp\!\!\!\perp_\mu Y\} = \bigcup_{A \in \mathcal{A}(X), B \in \mathcal{A}(Y)} \{\mu : \mu(A \cap B) \neq \mu(A)\mu(B)\}.$$

By Theorem 3.2.4, we have expressed the hypothesis of dependence as a countable union of statistically verifiable hypotheses. Therefore, $\{\mu : X \not\perp\!\!\!\perp_\mu Y\}$ is statistically verifiable by Lemma 3.2.2. □

Proving that conditional dependence is in general statistically verifiable is somewhat more involved. Let $(\Omega, \mathcal{B}, \mu)$ be a probability space. A *conditional expectation* of a random variable X given a sub σ-algebra $\mathcal{F}$ of $\mathcal{B}$, $E[X|\mathcal{F}]$, is any real-valued random variable Y such that

1. Y is $\mathcal{F}$ measurable;

2. for all $A \in \mathcal{F}$, $\int_A X d\mu = \int_A Y d\mu$.

Any Y satisfying (1-2) is called a *version* of the conditional expectation. A version of the conditional expectation is guaranteed to exist by the Radon-Nikodym theorem. Furthermore, if Y, Y' are versions of the conditional expectation, then $Y \overset{\text{a.s.}}{=} Y'$. As a way to shake hands with the conditional expectation, we prove the following simple facts:

Lemma 3.2.4.

1. $E[\mathbb{1}_\Omega|\mathcal{F}] = 1$;

2. $1 - E[\mathbb{1}_A|\mathcal{F}] = E[\mathbb{1}_{A^c}|\mathcal{F}]$;

3. *For disjoint countable A_i, $E[\mathbb{1}_{\cup A_i}|\mathcal{F}] = \sum_{i=1}^{\infty} E[\mathbb{1}_{A_i}|\sigma(Z)]$.*

Proof of Lemma 3.2.4.
1. Constant functions are obviously measurable, so the first condition is satisfied. Furthermore, for all $A \in \mathcal{F}$, $\int_A \mathbb{1}_\Omega d\mu = \int \mathbb{1}_A d\mu = \int_A 1 \cdot d\mu$.

2. Let $Y = E[\mathbb{1}_A|\mathcal{F}]$. Since pointwise sums and differences of measurable functions are measurable, $X = 1 - Y$ is measurable. Suppose $B \in \mathcal{F}$.

$$\begin{aligned}
\int_B \mathbb{1}_{A^c} d\mu &= \mu(A^c \cap B) \\
&= \mu(B) - \mu(A \cap B) \\
&= \int_B d\mu - \int_B \mathbb{1}_A d\mu \\
&= \int_B d\mu - \int_B Y d\mu \\
&= \int_B X d\mu.
\end{aligned}$$

3. Let $Y_k = \sum_{i=1}^k E[\mathbb{1}_{A_i}|\mathcal{F}]$. Let $B \in \mathcal{F}$. There is a null set N such that for every $\omega \in B \setminus N$,

$$0 \leq Y_k(\omega) \leq Y_{k+1}(\omega) \leq \infty.$$

(This is by almost sure positivity of the expectation of positive random variables.) Then, by the Monotone Convergence Theorem, $Y = \lim_{k\to\infty} Y_k$ is $\mathcal{F}$-measurable, and

$$\lim_{k\to\infty} \int_B Y_k d\mu = \int_B Y d\mu.$$

Similarly, by the Monotone Convergence Theorem, we have that

$$\lim_{k\to\infty} \int_B \sum_{i=1}^k \mathbb{1}_{A_i} d\mu = \int_B \lim_{k\to\infty} \sum_{i=1}^k \mathbb{1}_{A_i} d\mu.$$

Therefore:

$$\begin{aligned}\int_B E[\mathbb{1}_{\cup A_i}|\mathcal{F}]d\mu &= \int_B \mathbb{1}_{\cup_{i=1}^\infty A_i} d\mu = \int_B \lim_{k\to\infty} \sum_{i=1}^k \mathbb{1}_{A_i} d\mu \\ &= \lim_{k\to\infty} \int_B \sum_{i=1}^k \mathbb{1}_{A_i} d\mu = \lim_{k\to\infty} \int_B \sum_{i=1}^k E[\mathbb{1}_{A_i}|\mathcal{F}]d\mu \\ &= \int_B \lim_{k\to\infty} \sum_{i=1}^k E[\mathbb{1}_{A_i}|\mathcal{F}]d\mu = \int_B Y d\mu.\end{aligned}$$

□

The following Lemma connects the conditional expectation with familiar ideas from basic probability theory.

Lemma 3.2.5. *In the case that $\mathcal{F}$ is generated by a countable partition $\{F_i\}$, then*

$$E(\mathbb{1}_B|\mathcal{F}) \overset{a.s.}{=} \sum_{i:\mu(F_i)>0} \frac{\mu(B\cap F_i)}{\mu(F_i)} \mathbb{1}_{F_i} \qquad \forall B \in \mathcal{B}.$$

Proof of Lemma 3.2.5. Let

$$Y = \sum_{i:\mu(F_i)>0} \frac{\mu(B\cap F_i)}{\mu(F_i)} \mathbb{1}_{F_i}.$$

Since Y is constant on the F_i, it is $\mathcal{F}$-measurable. Furthermore,

$$\begin{aligned}\int_{F_j} Y d\mu &= \int_{F_j} \sum_{i:\mu(F_i)>0} \frac{\mu(B\cap F_i)}{\mu(F_i)} \mathbb{1}_{F_i} d\mu \\ &= \int \sum_{i:\mu(F_i)>0} \frac{\mu(B\cap F_i)}{\mu(F_i)} \mathbb{1}_{F_i\cap F_j} d\mu\end{aligned}$$

Therefore,

$$\int_{F_j} Y d\mu = \begin{cases} \int \frac{\mu(B\cap F_j)}{\mu(F_j)} \mathbb{1}_{F_j} d\mu = \mu(B \cap F_j), & \text{if } \mu(F_j) > 0 \\ 0, & \text{if } \mu(F_j) = 0, \end{cases}$$

and $\int_{F_j} Y d\mu = \int_{F_j} \mathbb{1}_B d\mu = \mu(B \cap F_j)$. Since the elements of $\mathcal{F}$ are countable, disjoint unions of the F_i, we are done. □

We also state without proof the following theorem, sometimes called Lévy's Upward Theorem, which helps to illuminate the conditional expectation:

Theorem 3.2.6 (Lévy's Zero-One Law)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and let X be a random variable with finite mean. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$ be a filtration of $\mathcal{F}$. Then*

$$E[X|\mathcal{F}_k] \quad \xrightarrow{a.s.} \quad E[X|\mathcal{F}].$$

Proof of Theorem 3.2.6. See, for example, Theorem 5.6 in Ladd [2011]. □

As a consequence of Theorem 3.2.6 and Lemma 3.2.5, if the $\mathcal{F}_i$ are generated by partitions $\{F_1^i, \ldots, F_{n_i}^i\}$, then

$$E(\mathbb{1}_B|\mathcal{F}_i) = \sum_{k\,:\,\mu(F_k^i)>0} \frac{\mu(B \cap F_k^i)}{\mu(F_k^i)} \mathbb{1}_{F_k^i} \quad \xrightarrow{a.s.} \quad E(\mathbb{1}_B|\mathcal{F}) \qquad \forall B \in \mathcal{B}.$$

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, and let $\mathcal{A}, \mathcal{B}$ be two subsets of $\mathcal{F}$. We say that $\mathcal{A}, \mathcal{B}$ are μ-independent given a sub σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, iff for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$,

$$E[\mathbb{1}_{A\cap B}|\mathcal{G}] = E[\mathbb{1}_A|\mathcal{G}]E[\mathbb{1}_B|\mathcal{G}].$$

Two random variables X, Y defined over Ω are *conditionally independent* given a third Z, written $X \perp\!\!\!\perp_\mu Y|Z$, iff $\sigma(X)$ and $\sigma(Y)$ are conditionally independent given $\sigma(Z)$. The following Lemma states a result for conditional independence analogous to Lemma 3.2.3 for unconditional independence.

Lemma 3.2.6. *Two random variables X, Y are μ-conditionally independent given Z iff for any π-systems, $\pi(X), \pi(Y)$, generating $\sigma(X), \sigma(Y)$, $\pi(X)$ is independent of $\pi(Y)$ given $\sigma(Z)$.*

Proof of Lemma 3.2.6. Let X, Y be random variables taking values in measurable spaces $(S, \mathcal{S}), (S', \mathcal{S}')$. Let $\mathcal{G} \subseteq \sigma(Y)$, and let

$$\mathcal{E} = \{E : E \in \sigma(X) \text{ and } E[\mathbb{1}_{E\cap B}|\sigma(Z)] = E[\mathbb{1}_E|\sigma(Z)]E[\mathbb{1}_B|\sigma(Z)] \text{ for all } B \in \mathcal{G}\}.$$

We show that $\mathcal{E}$ is a λ-system. First, we show that $\Omega \in \mathcal{E}$. Since X is total, $X^{-1}(S) = \Omega \in \sigma(X)$. By Lemma 3.2.4, $E[\mathbb{1}_\Omega|\sigma(Z)] = 1$. Therefore, for $B \in \mathcal{G}$, $E[\mathbb{1}_B|\sigma(Z)] = E[\mathbb{1}_{\Omega\cap B}|\sigma(Z)] = E[\mathbb{1}_\Omega|\sigma(X)]E[\mathbb{1}_B|\sigma(Z)]$, as required. Now we show that $\mathcal{E}$ is closed under complements. Suppose that $A \in \mathcal{E}$ and $B \in \mathcal{G}$. By the third part of Lemma 3.2.4,

$$\begin{aligned} E[\mathbb{1}_B|\sigma(Z)] &= E[\mathbb{1}_{A\cap B}|\sigma(Z)] + E[\mathbb{1}_{A^c\cap B}|\sigma(Z)] \\ &= E[\mathbb{1}_A|\sigma(Z)]E[\mathbb{1}_B|\sigma(Z)] + E[\mathbb{1}_{A^c\cap B}|\sigma(Z)], \end{aligned}$$

and therefore:

$$E[\mathbb{1}_B|\sigma(Z)](1 - E[\mathbb{1}_A|\sigma(Z)]) = E[\mathbb{1}_{A^c\cap B}|\sigma(Z)].$$

By the second part of Lemma 3.2.4:

$$E[\mathbb{1}_{A^c\cap B}|\sigma(Z)] = E[\mathbb{1}_B|\sigma(Z)]E[\mathbb{1}_{A^c}|\sigma(Z)].$$

Finally, we show that $\mathcal{E}$ is closed under disjoint unions. Suppose the collection $(A_i)_{i\in\mathbb{N}}$ is disjoint and that each A_i is in $\mathcal{E}$. By the third part of Lemma 3.2.4:

$$\begin{aligned} E[\mathbb{1}_{B\cap\cup_i A_i}|\sigma(Z)] &= \sum_{i=1}^{\infty} E[\mathbb{1}_{B\cap A_i}|\sigma(Z)] \\ &= \sum_{i=1}^{\infty} E[\mathbb{1}_B|\sigma(Z)]E[\mathbb{1}_{A_i}|\sigma(Z)] \\ &= E[\mathbb{1}_B|\sigma(Z)]\sum_{i=1}^{\infty} E[\mathbb{1}_{A_i}|\sigma(Z)] \\ &= E[\mathbb{1}_B|\sigma(Z)]E[\mathbb{1}_{\cup_{i=1}^{\infty} A_i}|\sigma(Z)]. \end{aligned}$$

Now suppose that $\pi(X)$ is μ-independent of $\pi(Y)$ given $\sigma(Z)$. Then, by the $\pi - \lambda$ Theorem, $\sigma(X)$ is μ-independent of $\pi(Y)$ given $\sigma(Z)$. Applying the $\pi - \lambda$ theorem again, $\sigma(X)$ is μ-independent of $\sigma(Y)$ given $\sigma(Z)$. □

The following theorem collects these facts to show that conditional dependence is statistically verifiable.

Theorem 3.2.7. *Suppose that X, Y, Z are random variables taking values in an arbitrary measurable space, and that the σ-algebras $\sigma(X), \sigma(Y), \sigma(Z)$ are generated by countable, almost surely clopen bases $\mathcal{I}(X), \mathcal{I}(Y), \mathcal{I}(Z)$. Then the hypothesis of conditional dependence, $\{\mu : X \not\perp\!\!\!\perp_\mu Y|Z\}$, is statistically verifiable.*

Proof of Theorem 3.2.7. By Corollary 3.1.1, if $\mathcal{I}(Z)$ is almost surely clopen, then the algebra that it generates, $\mathcal{A}(Z)$, is also almost surely clopen. Since $\mathcal{I}(Z)$ is countable, it is possible to form a filtration of $\mathcal{I}(Z)$ by larger and larger finite subsets. Let $\mathcal{I}_1(Z) \subset \mathcal{I}_2(Z) \subset \mathcal{I}_3(Z) \subset \cdots \subset \mathcal{I}(Z)$ be such a filtration of $\mathcal{I}(Z)$ such that for each i, $|\mathcal{I}_i(Z)| < \omega$. Let $\mathcal{A}_i(Z)$ be the finite algebra generated by $\mathcal{I}_i(Z)$. Each $\mathcal{A}_i(Z)$ is generated by a finite, almost surely clopen partition

$\{A^i_1, \dots, A^i_{n_i}\}$.[13]

Suppose that $X \not\perp\!\!\!\perp_\mu Y|Z$. By Lemma 3.2.6, there is $R \in \mathcal{A}(X)$, and $S \in \mathcal{A}(Y)$ such that

$$E[\mathbb{1}_{R\cap S}|\sigma(Z)] \quad \overset{\mu-a.s.}{\neq} \quad E[\mathbb{1}_R|\sigma(Z)]E[\mathbb{1}_S|\sigma(Z)].$$

Let $X = E[\mathbb{1}_{R\cap S}|\sigma(Z)]$ and $Y = E[\mathbb{1}_R|\sigma(Z)]E[\mathbb{1}_S|\sigma(Z)]$, and

$$C = \{\omega : X(\omega) \neq Y(\omega)\}.$$

Then $\mu(C) > 0$. Furthermore, let

$$\begin{aligned} X_i &= E[\mathbb{1}_{R\cap S}|\mathcal{A}_i(Z)] \\ Y_i &= E[\mathbb{1}_R|\mathcal{A}_i(Z)]E[\mathbb{1}_S|\mathcal{A}_i(Z)]. \end{aligned}$$

Note that, since the $\mathcal{A}_i(Z)$ are generated by finite partitions $\{A^i_1, \dots, A^i_{n_i}\}$,

$$\begin{aligned} X_i &= \sum_{k:\mu(A^i_k)>0} \frac{\mu(R\cap S\cap A^i_k)}{\mu(A^i_k)} \mathbb{1}_{A^i_k}, \\ Y_i &= \sum_{k:\mu(A^i_k)>0} \frac{\mu(R\cap A^i_k)\mu(S\cap A^i_k)}{\mu(A^i_k)} \mathbb{1}_{A^i_k}. \end{aligned}$$

Letting $B = \{\omega : X_i(\omega) \to X(\omega)\}$ and $C = \{\omega : Y_i(\omega) \to Y(\omega)\}$, we have by Levy's Zero-One Law that $\mu(A) = \mu(B) = 1$. Furthermore, $\mu(A \cap B \cap C) \geq \mu(C) > 0$. Therefore, there is $\omega \in A\cap B\cap C$ and N such that $X_N(\omega) \neq Y_N(\omega)$. That implies that there is an A^N_k such that $\mu(A^N_k) > 0$ and

$$\frac{\mu(R\cap S\cap A^N_k)}{\mu(A^N_k)} \neq \frac{\mu(R\cap A^N_k)\mu(S\cap A^N_k)}{\mu(A^N_k)}.$$

That demonstrates that

$$\begin{aligned} &\{\mu : X \not\perp\!\!\!\perp_\mu Y|Z\} = \\ &= \bigcup_{A\in\mathcal{A}(X), B\in\mathcal{A}(Y), C\in\mathcal{A}(Z)} \left\{\mu : \mu(C) > 0, \frac{\mu(A\cap B\cap C)}{\mu(C)} \neq \frac{\mu(A\cap C)\mu(B\cap C)}{\mu(C)}\right\}, \end{aligned}$$

which, by part (10) of Theorem 3.2.4, is a countable union of statistically verifiable hypothesis. By Lemma 3.2.2, $\{\mu : X \not\perp\!\!\!\perp_\mu Y|Z\}$ is statistically verifiable.

□

[13] To see that, note that if $\mathcal{I}_i(Z) = \{I^i_1, \dots, I^i_{n_i}\}$, then $\mathcal{A}_i(Z)$ is generated by events of the form $\bigcap_{j\neq i} I^c_j \cap I_i$.

3.2.4 Application: Causal Graphical Models

In the preceding we have taken a set of probability measures to represent all the relevant epistemic possibilities, or "possible worlds". If one were to pin down the chances of all the events in the background algebra, one would know all there is to know about the world. The framework of causal Bayes nets begins with an enriched set of epistemic possibilities. The possible worlds are pairs $(\mu, \mathcal{G})$ where μ is a probability measure that determines a probability distribution over a set of observable variables $\mathcal{V}$, and $\mathcal{G}$ is a causal graph over the variables in $\mathcal{V}$. Crucially, μ is taken to be the *observational* distribution, governing the chances of events only when the world is passively observed. In turn, the graph $\mathcal{G}$ determines the causal relations between the variables, which fixes the results of *interventions* on variables in $\mathcal{V}$.

Conventional statistical wisdom warns that it is foolhardy to infer anything about the causal graph $\mathcal{G}$ from observational, or non-experimental, data. A statistical dependency between X and Y is compatible with X being a cause of Y, Y being a cause of X, or with their being some common cause of both X and Y. One of the crucial insights of the literature inaugurated by Pearl and Verma [1995], Spirtes et al. [2000] is that, given certain bridge principles between causation and probability, some interesting causal conclusions are determined by patterns of conditional independence. In this section, we introduce some of the basic notions of the theory of causal discovery from observational data, and demonstrate how it is illuminated by the preceding theory.

Let M be a set of probability measures on a measurable space $(\Omega, \mathcal{B})$. Let $\mathcal{V}$ be a fixed, finite set of random variables $X_1, X_2, \ldots, X_n$ taking values in measurable spaces $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n)$. Assume that each X_i is $(\mathcal{B}, \mathcal{S}_i)$ measurable.

Let DAG be the set of all directed, acyclic graphs on the fixed variable set $\mathcal{V}$. The presence of a directed edge from X_i to X_j in $\mathcal{G}$ is understood to mean that X_i is a *direct cause* of X_j. Let $\mathcal{G} \in \mathsf{DAG}$. We say that X_i is a *parent* of X_j in $\mathcal{G}$ iff there is a directed edge in $\mathcal{G}$ out of X_i and into X_j. We say that X_j is a *child* of X_i in $\mathcal{G}$ iff there is a directed edge in $\mathcal{G}$ out of X_i and into X_j. We say that X_i, X_j are *adjacent* in $\mathcal{G}$ iff X_i is either a parent or chid of X_j in $\mathcal{G}$. We say that X_i is an *ancestor* of X_j in $\mathcal{G}$ iff there is a directed path in $\mathcal{G}$ from X_i to X_j. We say that X_j is a *descendant* of X_i in $\mathcal{G}$ iff there is a directed path in $\mathcal{G}$ from X_i to X_j. Let $\mathsf{parents}(X_i, \mathcal{G}), \mathsf{children}(X_i, \mathcal{G}), \mathsf{ancestors}(X_i, \mathcal{G}), \mathsf{descendants}(X_i, \mathcal{G})$ be the set of parents, children, ancestors and descendants of X_i in $\mathcal{G}$.

One of the most important notions in the causal framework is *d-separation.* Several preliminary notions are required. If $\mathcal{U} \subseteq \mathcal{V}$, Let $\mathcal{G}(\mathcal{U})$ be the subgraph of $\mathcal{G}$ that contains only vertices in $\mathcal{U}$. The *moralization* of $\mathcal{G} \in \mathsf{DAG}$ is the undirected graph $\mathcal{G}^M$, with the same vertices as $\mathcal{G}$, where a pair of vertices X_i, X_j are connected iff X_i and X_j are adjacent in $\mathcal{G}$, or they have a common child in $\mathcal{G}$. If $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are disjoint subsets of $\mathcal{V}$, we say that $\mathcal{X}$ is *separated* from $\mathcal{Y}$ given $\mathcal{Z}$ in

$\mathcal{G}^M$ iff every undirected path in $\mathcal{G}^M$ from $\mathcal{X}$ to $\mathcal{Y}$ contains a member of $\mathcal{Z}$. If $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are disjoint subsets of $\mathcal{V}$, we say that $\mathcal{X}$ and $\mathcal{Y}$ are *d-separated given* $\mathcal{Z}$ in $\mathcal{G}$ iff $\mathcal{X}$ and $\mathcal{Y}$ are separated given $\mathcal{Z}$ in $\mathcal{G}^M(\mathsf{ancestors}(\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}, \mathcal{G}))$. Write $\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})$ as shorthand for $\mathcal{X}$ and $\mathcal{Y}$ are d-separated given $\mathcal{Z}$ in $\mathcal{G}$. Write $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})$ as shorthand for $\mathcal{X}$ and $\mathcal{Y}$ are not d-separated given $\mathcal{Z}$ in $\mathcal{G}$.

The I-map partial order on DAGs is defined by setting $\mathcal{G}' \preceq \mathcal{G}''$ iff $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G}')$ implies $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G}'')$. In other words, $\mathcal{G}' \preceq \mathcal{G}''$ if $\mathcal{G}'$ entails all the same d-separations that $\mathcal{G}''$ does. If $\mathcal{G}' \preceq \mathcal{G}''$ and $\mathcal{G}'' \preceq \mathcal{G}'$, we say that $\mathcal{G}', \mathcal{G}''$ are *Markov equivalent*, i.e. they entail all the same d-separation relations. Write $\mathcal{G}' \sim \mathcal{G}''$ iff $\mathcal{G}'$ and $\mathcal{G}''$ are Markov equivalent, and let

$$[\mathcal{G}] = \{\mathcal{G}' \in \mathsf{DAG} : \mathcal{G}' \sim \mathcal{G}\},$$

be the *Markov equivalence class* of $\mathcal{G}$. The I-map order is lifted to Markov equivalence classes in the natural way. Say that $\mathcal{G}' \preceq [\mathcal{G}'']$ iff $\mathcal{G}' \preceq \mathcal{G}''$. Say that $[\mathcal{G}'] \preceq [\mathcal{G}'']$ iff $\mathcal{G}' \preceq \mathcal{G}''$. Verma and Pearl [1992] prove the following Theorem:

Theorem 3.2.8. *$\mathcal{G}, \mathcal{G}'' \in \mathsf{DAG}$ are Markov equivalent iff*

1. *$\mathcal{G}'$ and $\mathcal{G}''$ have the same adjacencies, and*
2. *X_k is a common child of non-adjacent X_i, X_j in $\mathcal{G}'$ iff X_k is a common child of non-adjacent X_i, X_j in $\mathcal{G}''$.*

The following concepts connect graphical concepts with probabilistic ones. A measure $\mu \in M$ is *Markov* for $\mathcal{G} \in \mathsf{DAG}$ iff for any three disjoint subsets of $\mathcal{V}$, $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, if $\mathcal{X}$ is d-separated from $\mathcal{Y}$ given $\mathcal{Z}$, then $\mathcal{X} \perp\!\!\!\perp_\mu \mathcal{Y} | \mathcal{Z}$. A measure $\mu \in M$ is *faithful* to $\mathcal{G} \in \mathsf{DAG}$ iff for any three disjoint subsets of $\mathcal{V}$, $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, if $\mathcal{X}$ is not d-separated from $\mathcal{Y}$ given $\mathcal{Z}$, then $\mathcal{X} \not\perp\!\!\!\perp_\mu \mathcal{Y} | \mathcal{Z}$. It is clear that a measure $\mu \in W$ is Markov (or faithful) to $\mathcal{G}$ iff it is Markov (or faithful) to every $\mathcal{G}'$ in the Markov equivalence class $[\mathcal{G}]$.

The set of possible worlds W is a subset of the cross product $M \times \mathsf{DAG}$. The *causal Markov assumption* says that for all $(\mu, \mathcal{G}) \in W$, μ is Markov for $\mathcal{G}$. The *causal faithfulness assumption* is that for all $(\mu, \mathcal{G}) \in W$, μ is faithful to $\mathcal{G}$. Crucially, the causal Markov/faithfulness assumptions are not saying that the true probability measure is Markov/faithful to *some* DAG, but rather that the true probability measure is Markov and faithful to the *true* DAG. In particular, faithfulness rules out certain perfect observational illusions, where the conditional independences are drastically misleading about the causal truth. So misleading, in fact, that even in the limit of infinite data, one would not be able to identify the true graph, not even up to Markov equivalence.[14] Although the causal Markov and faithfulness assumptions are hotly contested, together they ensure that patterns of conditional dependence and independence are reliable

[14] For a good discussion of the causal faithfulness assumption, with examples, see Lin and Zhang [2018].

guides to the causal truth, at least up to Markov equivalence. Therefore, it is crucial to investigate how hard it is to learn about conditional dependencies and independencies from observational data. Given what we have developed in the previous section, this is a straightforward task.

Theorem 3.2.9. *Suppose that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$.*[15] *Then, the following hypotheses are statistically refutable:*

1. *The true measure is Markov to $\mathcal{G} \in \mathsf{DAG}$;*
2. *The true measure is Markov to $[\mathcal{G}]$, for $\mathcal{G} \in \mathsf{DAG}$;*
3. *The true measure is Markov to some $\mathcal{G} \in \mathsf{DAG}$.*

Proof of Theorem 3.2.9.

1. The set of measures Markov to $\mathcal{G} \in \mathsf{DAG}$ can be expressed as

$$\{\mu \in M : \mu \text{ is Markov to } \mathcal{G}\} = \bigcap_{\mathsf{dsep}(\mathcal{X},\mathcal{Y},\mathcal{Z},\mathcal{G})} \{\mu : \mathcal{X} \perp\!\!\!\perp_\mu \mathcal{Y} | \mathcal{Z}\}.$$

By Theorem 3.2.7, each element of the finite conjunction is statistically refutable. By Lemma 3.2.2, the conjunction is statistically refutable.

2. The set of measures Markov to $[\mathcal{G}]$ is exactly the set of measures Markov to $\mathcal{G}$. Therefore, the second part is an immediate consequence of the first.

3. The set of measures Markov to some $\mathcal{G} \in \mathsf{DAG}$ can be expressed as

$$\bigcup_{\mathcal{G} \in \mathsf{DAG}} \{\mu \in M : \mu \text{ is Markov to } \mathcal{G}\}.$$

By Part 2, each member of the union is statistically refutable. Since the union is finite, the disjunction is statistically refutable by Lemma 3.2.2. □

Corollary 3.2.2. *Suppose that the causal Markov and faithfulness assumptions hold, and that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the causal hypothesis*

$$\{(\mu, \mathcal{G}') : [\mathcal{G}'] \preceq [\mathcal{G}]\}$$

is statistically refutable.

Proof of Corollary 3.2.2. It suffices to show that, under the causal Markov and faithfulness assumptions,

$$\{(\mu, \mathcal{G}') : \mathcal{G}' \preceq \mathcal{G}\} = \{(\mu, \mathcal{G}') : \mu \text{ is Markov to } \mathcal{G}\},$$

[15] Note that this entails, by Lemma 3.1.2, that for any $I \subseteq \{1, \dots, n\}$, $\prod_{i \in I} \mathcal{I}(X_i)$ is a countable, almost surely clopen basis for $\bigotimes_{i \in I} \sigma(X_i)$.

which is statistically refutable by Theorem 3.2.9. Suppose $(\nu, \mathcal{G}'')$ is an element of the lhs. Suppose that $\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})$. Then, since $\mathcal{G}' \preceq \mathcal{G}$, $\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G}')$. By the causal Markov assumption, $\mathcal{X} \perp\!\!\!\perp_\nu \mathcal{Y} | \mathcal{Z}$. Therefore ν is Markov to $\mathcal{G}$. We have shown that the lhs is contained in the rhs. Suppose that $(\nu, \mathcal{G}')$ is an element of the rhs. Then, ν is Markov to $\mathcal{G}$. Suppose that $\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})$. Then $\mathcal{X} \perp\!\!\!\perp_\nu \mathcal{Y} | \mathcal{Z}$. By the causal faithfulness assumption, ν is faithful to $\mathcal{G}'$. Therefore, $\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G}')$. So $\mathcal{G}' \preceq \mathcal{G}$. Since the rhs is statistically refutable by Theorem 3.2.9, we are done. □

Theorem 3.2.10. *Suppose that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the following hypotheses are statistically verifiable:*

1. *The true measure is faithful to $\mathcal{G} \in \mathsf{DAG}$;*
2. *The true measure is faithful to $[\mathcal{G}]$, for $\mathcal{G} \in \mathsf{DAG}$;*
3. *The true measure is faithful to some $\mathcal{G} \in \mathsf{DAG}$.*

Proof of Theorem 3.2.10.

1. The set of measures faithful to $\mathcal{G} \in \mathsf{DAG}$ can be expressed as

$$\{\mu \in M : \mu \text{ is faithful to } \mathcal{G}\} = \bigcap_{\neg \mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})} \{\mu : \mathcal{X} \not\!\perp\!\!\!\perp_\mu \mathcal{Y} | \mathcal{Z}\}.$$

By Theorem 3.2.7, each element of the finite conjunction is statistically verifiable. Since the conjunction is finite, it is statistically verifiable by Lemma 3.2.2.

2. The set of measures faithful to $[\mathcal{G}]$ is exactly the set of measures faithful to $\mathcal{G}$. Therefore, the second part is an immediate consequence of the first.

3. The set of measures faithful to some $\mathcal{G} \in \mathsf{DAG}$ can be expressed as

$$\bigcup_{\mathcal{G} \in \mathsf{DAG}} \{\mu \in M : \mu \text{ is faithful to } \mathcal{G}\}.$$

By Part 1, each member of the union is statistically verifiable. By Lemma 3.2.2, the union is statistically verifiable. □

Corollary 3.2.3. *Suppose that the causal Markov and faithfulness assumptions hold, and that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the causal hypothesis*

$$\{(\mu, \mathcal{G}') : [\mathcal{G}] \preceq [\mathcal{G}']\}$$

is statistically verifiable.

Proof of Corollary 3.2.3. It suffices to show that, under the causal Markov and faithfulness assumptions,

$$\{(\mu, \mathcal{G}') : \mathcal{G} \preceq \mathcal{G}'\} = \{(\mu, \mathcal{G}') : \mu \text{ is faithful to } \mathcal{G}\},$$

which is statistically verifiable by Theorem 3.2.10. Suppose $(\nu, \mathcal{G}')$ is an element of the lhs. Suppose that $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})$. Then, since $\mathcal{G} \preceq \mathcal{G}'$, $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G}')$. By the causal faithfulness assumption, ν is faithful to $\mathcal{G}'$, and therefore $\mathcal{X} \not\perp\!\!\!\perp_\nu \mathcal{Y} | \mathcal{Z}$. Therefore ν is faithful to $\mathcal{G}$. We have shown that the lhs is contained in the rhs. Suppose that $(\nu, \mathcal{G}')$ is an element of the rhs. Then, ν is faithful to $\mathcal{G}$. Suppose that $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G})$. Then $\mathcal{X} \not\perp\!\!\!\perp_\nu \mathcal{Y} | \mathcal{Z}$. By the causal Markov assumption, ν is Markov to $\mathcal{G}'$. Therefore, $\neg\mathsf{dsep}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{G}')$. So $\mathcal{G} \preceq \mathcal{G}'$, and the rhs is contained in the lhs. Since the rhs is statistically verifiable by Theorem 3.2.9, we are done. □

3.3 Monotonic Statistical Verification and Refutation

3.3.1 Defining the Success Concept

In section 3.2, we said that (λ_n) is a statistical verifier of H if it converges to H if H is true, and that otherwise has a small chance of drawing an erroneous conclusion. But that standard is consistent with a wild see-sawing in the chance of producing the informative conclusion H as sample sizes increase, even if H is true. Of course, it is desirable that the chance of correctly producing H increases with the sample size, i.e. that for all $\mu \in H$ and $n_1 < n_2$,

Mon. $\mu^{n_2}[\lambda_{n_2}^{-1}(H)] > \mu^{n_1}[\lambda_{n_1}^{-1}(H)]$.

Failing to satisfy Mon has the perverse consequence that collecting a larger sample might be a bad idea! Researchers would have to worry whether a failure of replication was due merely to a clumsily designed statistical method that converges to the truth along a needlessly circuitous route. Unfortunately, Mon is infeasible in typical cases, so long as we demand that verifiers satisfy VanErr. Lemma 3.3.1 expresses that misfortune.

Say that a chance setup $(W, \Omega, \mathcal{I})$ is *purely statistical* iff for all $\mu \in W$ and all events $B \in \mathcal{B}^{\otimes n}$ such that $\mu^n(\mathsf{bdry} B) = 0$, $\mu^n(B) > 0$. This is the formal expression of the idea that almost surely clopen sample events have no logical bearing on statistical hypotheses.

Lemma 3.3.1. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Suppose furthermore that $(W, \Omega, \mathcal{I})$ is purely statistical. Let H be open, but not closed in the weak topology. If (λ_n) is an α-verifier in chance of H, then, (λ_n) satisfies* VanErr *only if it does not satisfy* Mon.

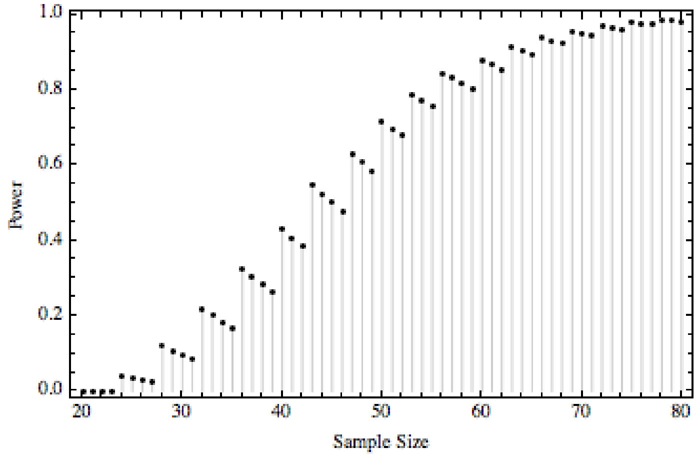


Figure 3.1: Diachronic plot of power of the test exhibited in the proof of Lemma 3.2.1. The plot exhibits the characteristic "saw-tooth" shape. The null hypothesis is that $\mu(B) \leq .5$. At the true world $\mu(B) = .875$. Note that the power drops by $> .06$ between sample sizes 43 and 46.

Proof of Lemma 3.3.1. Suppose that H is open, but not closed in the weak topology. Let $\nu \in \mathsf{cl}H \setminus H$. Suppose that (λ_n) is an α-verifier in chance of H and that λ_n satistfies VanErr. Since the chance setup is purely statistical, $\alpha_{n_1} = \mu^{n_1}[\lambda_{n_1}^{-1}(H)] > 0$. Let $O' = \{\mu : \alpha_{n_1} - \epsilon < \mu^{n_1}[\lambda_{n_1}^{-1}(H)]\}$. By construction $\nu \in O'$. Since (λ_n) satisfies VanErr, there is $n_2 > n_1$ such that $\mu^{n_2}[\lambda_{n_2}^{-1}(H)] < \alpha_{n_1} - \epsilon$. Let $O'' = \{\mu : \alpha_{n_1} - \epsilon > \mu^{n_1}[\lambda_{n_1}^{-1}(H)]\}$. By construction $\nu \in O''$. Since (λ_n) is feasible, both O', O'' are open in the weak topology. Therefore, $O = O' \cap O''$ is open in the weak topology and $\nu \in O$. But since $\nu \in \mathsf{cl}(H)$, there is $\mu \in H \cap O$. But then $\mu^{n_1}[\lambda_{n_1}^{-1}(H)] > \alpha_{n_1} - \epsilon$, whereas $\mu^{n_2}[\lambda_{n_2}^{-1}(H)] < \alpha_{n_1} - \epsilon$. Therefore, (λ_n) does not satisfy Mon. □

But even if strict monotonicity of power is infeasible, it ought to be our regulative ideal. Say that an α-verifier $(\lambda_n)_{n \in \mathbb{N}}$ of H, whether in chance, or almost sure, is *α-monotonic* iff for all $\mu \in H$ and $n_1 < n_2$:

α-Mon $\mu^{n_2}[\lambda_{n_2}^{-1}(H)] + \alpha > \mu^{n_1}[\lambda_{n_1}^{-1}(H)]$.

Satisfying α-Mon ensures that collecting a larger sample is not a disastrously bad idea. Surprisingly, some standard hypothesis tests fail to satisfy even this weak requirement. Chernick and Liu [2002] noticed non-monotonic behavior in the power function of textbook tests of the binomial proportion, and proposed heuristic software solutions. The test exhibited in the proof of Lemma 3.2.1 also displays dramatic non-monotonicity (Figure 3.1). Others have raised worries of non-monotonicity in consumer safety regulation, vaccine studies, and agronomy [Musonda, 2006, Schaarschmidt, 2007, Schuette et al., 2012].

We now articulate a notion of statistical verifiability that requires α-monotonicity. Write $a_n \downarrow 0$ if the sequence (a_n) converges monotonically to zero. Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible tests of $H^{\mathsf{c}} \subseteq W$ is an *α-monotonic verifier* of H iff

MVanErr. For all $\mu \in H^{\mathsf{c}}$, there exists a sequence (α_n) such that each $\alpha_n \leq \alpha$, $\alpha_n \downarrow 0$, and $\mu^n[\lambda_n^{-1}(H)] \leq \alpha_n$;

LimCon. For all $\mu \in H$, $\mu^n[\lambda_n^{-1}(H)] \xrightarrow{n} 1$;

α-Mon. For all $\mu \in W$, $\mu^{n_1}[\lambda_{n_1}^{-1}(H)] - \mu^{n_2}[\lambda_{n_2}^{-1}(H)] < \alpha$, if $n_1 < n_2$.

Say that H is *α-monotonically verifiable* iff there is an α-monotonic verifer of H. Say that H is *monotonically verifiable* iff H is α-monotonically verifiable, for every $\alpha > 0$.

It is clear that every α-monotonic verifier of H is also an α-verifier in chance. However, not every α-monotonic verifier of H is an almost sure α-verifier. As is clear from Figure 3.1, the converse also does not hold.

Defining monotonic refutability requires no new ideas. Say that H is *α-monotonically refutable in chance* iff there is an α-monotonic verifier of H^{c}. Say that H is *montonically refutable* iff H^{c} is α-monotonically verifiable for every $\alpha > 0$.

3.3.2 Characterization Theorems

The central theorem of this section states that every statistically verifiable hypothesis is also monotonically verifiable.

Theorem 3.3.1. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-verifiable in chance for some $\alpha > 0$;*
2. *H is monotonically verifiable;*
3. *H is almost surely verifiable;*
4. *H is open in the weak topology on W.*

The characterization of monotonic refutability follows immediately.

Theorem 3.3.2. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-refutable in chance for some $\alpha > 0$;*
2. *H is monotonically refutable;*
3. *H is almost surely refutable;*

4. *H is closed in the weak topology on W.*

The proof of Theorem 3.3.1 is somewhat involved. Theorem 3.2.1 states that 1, 3 and 4 are equivalent. The fact that 2 implies 1 is immediate from the definitions. To prove that 4 implies 2, we proceed largely as before, with greater attention to details. First, we prove the somewhat technical Lemma 3.3.2, which says that if you have a countable collection of α_i-monotonic verifiers $(\lambda_n^1), (\lambda_n^2), \ldots$, of hypotheses $A_1, A_2, \ldots$, then it is possible to construct a countable collection of *mutually independent* monotonic verifiers for the A_i. The idea behind the construction is very rudimentary: you can always render methods independent by splitting up the sample in the appropriate way. Then, we show in Lemma 3.3.3, that for almost surely decidable A, the hypothesis $\{\mu : \mu(A) > r\}$ is monotonically verifiable. That entails that every element of the subbasis for the weak topology exhibited in Theorem 3.1.3 is almost surely verifiable. Finally, we show in Lemmas 3.3.4, and 3.3.5, that the monotonically verifiable propositions are closed under finite intersections, and countable unions, which completes the proof. It remains to prove the four crucial Lemmas.

Lemma 3.3.2. *Suppose that I is a countable index set and that for $i \in I$, $(\psi_n^i)_{n \in \mathbb{N}}$ is an α_i-monotonic verifier of A_i. Then, there exist $(\lambda_n^1), (\lambda_n^2), \ldots$ where each $(\lambda_n^i)_{n \in \mathbb{N}}$ is an α_i-monotonic verifier of A_i, and for each n, $(\sigma(\lambda_n^i), i \in I)$ are mutually independent.*

Proof of 3.3.2. The basic idea is to render the (ψ_n^i) mutually independent by splitting the sample and feeding it to the individual verifiers according to a triangular dovetailing scheme. Represented in tabular form:

$$\begin{array}{ccccc}
(\psi_n^1) & (\psi_n^2) & (\psi_n^3) & (\psi_n^4) & \cdots \\
 & & & & \\
\omega_1 & & & & \\
\omega_2 & \omega_3 & & & \\
\omega_4 & \omega_5 & \omega_6 & & \\
\omega_7 & \omega_8 & \omega_9 & \omega_{10} & \\
\omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ldots
\end{array}$$

The samples in the i^{th} column of the table are the samples that are fed to the i^{th} verifier. Since the verifiers are essentially functions of disjoint samples, they are mutually independent. Since the samples are i.i.d. all of the desirable statistical properties of the verifiers are preserved. Before we formalize this idea, we develop some fundamentals.

If J is a subset of $\{1, \ldots, n\}$, let p_J^n be the projection from Ω^n to $\Omega^{|J|}$ with

$$p_J^n(\omega_1, \ldots, \omega_n) = (\omega_i, i \in J).$$

To get our bearings, we show that the projection function is $(\mathcal{B}^{\otimes n}, \mathcal{B}^{\otimes |J|})$ measurable. It is clear that the collection

$$\{A \in \mathcal{B}^{\otimes |J|} : (p^n_J)^{-1}(A) \in \mathcal{B}^{\otimes n}\}$$

is a λ-system. Furthermore, the collection of rectangular sets

$$\mathcal{R} = \{\times_{i \in J} A_i : A_i \in \mathcal{B}\}$$

is a π-system that generates $\mathcal{B}^{\otimes |J|}$.[16] Therefore, by the $\pi - \lambda$ theorem it is sufficient to show that $(p^n_J)^{-1}(A) \in \otimes^n_{i=1}\mathcal{B}$ for every $A \in \mathcal{R}$. But that is straightforwardly true, since

$$(p^n_J)^{-1}(\times_{i \in J} A_i) = \times^n_{i=1} B_i,$$

where $B_i = A_i$ if $i \in J$, and $B_i = \Omega$, otherwise.

Next we show that for $A \in \otimes^{|J|}_{i=1}\mathcal{B}$,

$$\mu^n[(p^n_J)^{-1}(A)] = \mu^{|J|}(A). \tag{3.1}$$

This is another straightforward application of the $\pi - \lambda$ theorem. Let

$$\mathcal{S} = \{A \in \mathcal{B}^{\otimes |J|} : \mu^n[(p^n_J)^{-1}(A)] = \mu^{|J|}(A)\}.$$

First we show that $\mathcal{S}$ is a λ-system. It is clear that $\Omega^{|J|} \in \mathcal{S}$. Next, we show that $\mathcal{S}$ is closed under countable disjoint union. Suppose that $A_1, A_2, \ldots$ are disjoint and in $\mathcal{S}$. Since the preimage of a disjoint union is the disjoint union of the preimages: $\mu^n[(p^n_J)^{-1}(\sqcup A_i)] = \mu^n[\sqcup (p^n_J)^{-1}(A_i)] = \sum_{i=1}^{\infty} \mu^n[(p^n_J)^{-1}(A_i)] = \sum_{i=1}^{\infty} \mu^{|J|}(A_i) = \mu^{|J|}[\sqcup A_i]$. Finally, we show that $\mathcal{S}$ is closed under complements. Suppose that $A \in \mathcal{S}$. Since the preimage of the complement is the complement of the preimage: $\mu^n[(p^n_J)^{-1}(\Omega^{|J|} \setminus A)] = \mu^n[\Omega^n \setminus (p^n_J)^{-1}(A)] = 1 - \mu^n[(p^n_J)^{-1}(A)] = 1 - \mu^{|J|}[A] = \mu^{|J|}[\Omega^{|J|} \setminus A]$. Therefore $\mathcal{S}$ is a λ-system. It remains to show that the π-system $\mathcal{R} \subseteq \mathcal{S}$. Recall that

$$(p^n_J)^{-1}(\times_{i \in J} A_i) = \times^n_{i=1} B_i,$$

where $B_i = A_i$ if $i \in J$, and $B_i = \Omega$, otherwise. Therefore, by the definition of the product measure, $\mu^n[(p^n_J)^{-1}(\times_{i \in J} A_i)] = \prod_{i \in J} \mu(A_i) = \mu^{|J|}(\times_{i \in J} A_i)$.

Let $J_1, J_2, \ldots, J_k$ be disjoint subsets of $\{1, \ldots, n\}$. We show that

$$\sigma(p^n_{J_1}), \ldots, \sigma(p^n_{J_k}) \text{ are mutually } \mu^n\text{-independent.} \tag{3.2}$$

By Lemma 3.2.3, it suffices to show that there are mutually μ^n-independent generating π-systems $\pi(p^n_{J_1}), \ldots, \pi(p^n_{J_k})$. Let

$$\mathcal{R}_{J_i} = \{\times_{i \in J_i} A_i : A_i \in \mathcal{B}\},$$

[16] To see that it is a π-system recall that the intersection of a cartesian product is the cartesian product of the component-wise intersections.

and $\pi(p^n_{J_i}) = (p^n_{J_i})^{-1}(\mathcal{R}_{J_i})$.[17] For $i \in \{1, \ldots k\}$, let $E_i \in \pi(p^n_{J_i})$. Then $E_i = (p^n_{J_i})^{-1}(\times_{k \in J_i} A_{i,k}) = \times^n_{k=1} B_k$, where $B_k = A_{i,k}$ if $k \in J_i$ and $B_k = \Omega$, otherwise. Therefore, $\cap^k_{i=1} E_i = \times^n_{k=1} B_k$, where $B_k = A_{i,k}$ if there is J_i such that $k \in J_i$, and $B_k = \Omega$, otherwise. By the definition of the product measure, $\mu^n(\cap^k_{i=1} E_i) = \prod^k_{i=1} \mu^{|J_i|}(\times_{k \in J_i} A_{i,k})$. By (1), above, $\prod^k_{i=1} \mu^{|J_i|}(\times_{k \in J_i} A_{i,k}) = \prod^k_{i=1} \mu^n(E_i)$, and we are done.

Recall the table presented at the beginning of this proof. Notice that the indices along the diagonal are given by the triangular numbers $T_i = \binom{i+1}{2}$. Define:

$$\begin{aligned} n_{i,1} &= T_i; \\ n_{i,j} &= n_{i,j-1} + i + j - 2. \end{aligned}$$

So, for example, the indices along the second column are given by $n_{2,1} = 3, n_{2,2} = 5, n_{2,3} = 8, \ldots$ Define:

$$\begin{aligned} k(n) &= \max_i \ T_i \leq n; \\ n_i(n) &= \max_j \ n_{i,j} \leq n. \end{aligned}$$

The function $k(n)$ returns the number of verifiers which are fed samples at sample size n. For example, $k(8) = 3$, since the third column is the rightmost column "visited" by sample size 8. The function $n_i(n)$ returns the effective sample size for the i^{th} method at sample size n. For example, $n_1(8) = 4, n_2(8) = 3, n_3(8) = 1$, since these are the lowermost rows "visited" in columns $1, 2$ and 3, by sample size 8. Define, I^i_n, the set of indices of samples fed to the i^{th} method at sample size n, as follows:

$$I^i_n = (n_{i,1}, \ldots, n_{i,n_i(n)}).$$

Let

$$\lambda^i_n(\omega_1, \ldots, \omega_n) = \begin{cases} \psi^i_{n_i(n)} \circ p^n_{I^i_n}(\omega_1, \ldots, \omega_n), & \text{if } k(n) \geq i, \\ W, & \text{otherwise.} \end{cases}$$

Since $\sigma(\lambda^i_n) \subseteq \sigma(p^n_{I^i_n})$, the σ-algebras $\sigma(\lambda^1_n), \sigma(\lambda^2_n), \ldots$ are mutually μ^n-independent by (2), above.

It remains to show that each (λ^i_n) is an α_i-monotonic verifier of A_i. Since (ψ^i_n) is an α_i-monotonic verifier of A_i, for each $\mu \in A^\complement_i$ there is (α_n) such that

[17] To check that $\pi(p^n_{J_i})$ is a π-system note that if $\mathcal{C}$ is a π-system and f is a function, then $f^{-1}(\mathcal{C})$ is a π-system. To check that $\pi(p^n_{J_i})$ generates $\sigma(p^n_{J_i})$ note that if $\mathcal{C}$ is a collection of sets and f is a function, then $f^{-1}(\sigma(\mathcal{C})) = \sigma(f^{-1}(\mathcal{C}))$. So we have that $\sigma(p^n_{J_i}) = (p^n_{J_i})^{-1}(\mathcal{B}^{\otimes |J_i|}) = (p^n_{J_i})^{-1}(\sigma(\mathcal{R}_{J_i})) = \sigma((p^n_{J_i})^{-1}(\mathcal{R}_{J_i}))$.

each $\alpha_n < \alpha_i$, $\alpha_n \downarrow 0$ and $\mu^n[\psi_n^{-1}(A_i)] \leq \alpha_n$. For $n < n_{i,1}$, $\mu^n[(\lambda_n^i)^{-1}(A_i)] = 0$. And for $n \geq n_{i,1}$,

$$\begin{aligned}\mu^n[(\lambda_n^i)^{-1}(A_i)] &= \mu^n[(p^n_{I_n^i})^{-1}((\psi^i_{n_i(n)})^{-1}(A_i))] \\ &= \mu^{n_i(n)}[(\psi^i_{n_i(n)})^{-1}(A_i)] \\ &\leq \alpha_{n_i(n)} \leq \alpha_i,\end{aligned}$$

where we have invoked (1), above, to get to the second line. Therefore, letting

$$\alpha'_n = \begin{cases}\alpha_i, & \text{if } n < n_{i,1} \\ \alpha_{n_i(n)}, & \text{otherwise,}\end{cases}$$

we have that each $\alpha'_n \leq \alpha_i$, $\alpha'_n \downarrow 0$ and $\mu^n[(\lambda_n^i){-}1(A_i)] \leq \alpha'_n$. So the (λ_n^i) satisfy MVanErr. Furthermore, for $\mu \in A_i$, $\lim_n \mu^n[(\lambda_n^i)^{-1}(A_i)] = \lim_n \mu^{n_i(n)}[(\psi^i_{n_i(n)})^{-1}(A_i)] = 1$. Therefore, the (λ_n) satisfy LimCon. Finally, to see that the (λ_n^i) satisfy α_i-Mon, notice that, for $n_1 < n_{i,1}$,

$$\mu^{n_1}[(\lambda^i_{n_1})^{-1}(A_i)] - \mu^{n_2}[(\lambda^i_{n_2})^{-1}(A_i)] \leq 0 < \alpha_i.$$

And for $n_1 \geq n_{i,1}$,

$$\begin{aligned}&\mu^{n_1}[(\lambda^i_{n_1})^{-1}(A_i)] - \mu^{n_2}[(\lambda^i_{n_2})^{-1}(A_i)] = \\ &\qquad = \mu^{n_i(n_1)}[(\psi^i_{n_i(n_1)})^{-1}(A_i)] - \mu^{n_i(n_2)}[(\psi^i_{n_i(n_2)})^{-1}(A_i)] \\ &\qquad \leq \alpha_i,\end{aligned}$$

as required. □

Lemma 3.3.3. *Suppose that B is almost surely decidable for every $\mu \in W$. Then, for all real b, the hypothesis $H = \{\mu : \mu(B) > b\}$ is monotonically verifiable.*

Proof of Lemma 3.3.3. We restrict attention to the non-trivial cases where $b \in (0, 1)$. The idea is to take an almost sure α-verifier of H and modify it slightly to ensure α-monotonicity. Let $(\lambda_n)_{n\in\mathbb{N}}$ be the a.s. α-verifier exhibited in the proof of Lemma 3.2.1, i.e letting $t_n = \sqrt{\frac{1}{2n}\ln(\pi^2 n^2/6\alpha)}$,

$$\lambda_n(\vec{\omega}) = \begin{cases}H, & \text{if } \sum_{i=1}^n \mathbb{1}_B(\omega_i) \geq \lceil n(b + t_n)\rceil, \\ W, & \text{otherwise.}\end{cases}$$

Let

$$\beta_n(\theta) = \sum_{\lceil n(b+t_n)\rceil}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Then, $\beta_n(\mu B) = \mu^n[\lambda_n^{-1}(H)]$. It is something of a nuisance that $\beta_n(\theta)$ is exactly zero for n such that $\lceil n(b + t_n)\rceil > n$. Since the t_n converge monotonically to

0, this is the case for only finitely many initial sample sizes n. For example, for $b = .5, \alpha = .05$, $\beta_n(\theta)$ is non-trivial for samples sizes larger than 20. Let $n_0 = \min\{n : \lceil n(b + t_n)\rceil \leq n\}$.

It is worth pointing out some additional features of the $\beta_n(\theta)$ that we will be appealing to in the following. It is evident that, since $\beta_n(\theta)$ is a polynomial, it is a continuous function of the parameter θ. Although it is "obviously" true that, for $n \geq n_0$, $\beta_n(\theta)$ is stricly increasing in θ, it is surprisingly non-trivial to demonstrate. For an elegant proof of this fact, see Gilat [1977]. It is a standard fact of analysis that that if $[a, b]$, $[c, d]$ are closed real intervals and $f : [a, b] \to [c, d]$ is a continuous real function, then f is bijective iff it is strictly increasing. Therefore, for $n \geq n_0$, $\beta_n : [0, 1] \to [0, 1]$ is bijective.

There are two important properties of the collection $(\beta_n(\theta))_{n\in\mathbb{N}}$ that follow from the fact that (λ_n) is an almost sure verifier: (1) for $\theta > b$, $\lim_n \beta_n(\theta) = 1$; (2) $\sum_n \sup_{\theta\leq b} \beta_n(\theta) = \sum_n \beta_n(b) < \alpha$. The first property follows from LIM-CON. The second property follows from σ-BNDERR and the fact that $\beta_n(\theta)$ is increasing. Define $\alpha_n = \beta_n(b)$. Note that for $n \geq n_0$, $\alpha_n > 0$. Let $\alpha_n^* = \min\{\alpha_m : n_0 \leq m \leq n\}$. For $n \geq n_0$, let

$$K_n = \max\{k \in \mathbb{N} : \alpha + k\alpha_n < 1 - \alpha_n^*\}.$$

Now, define the increasing step function:

$$\phi_n(\theta) = \begin{cases} 0, & \beta_n(\theta) < \alpha, \\ \alpha + k\alpha_n, & \alpha + (k-1)\alpha_n \leq \beta_n(\theta) < \alpha + k\alpha_n, \quad k \in \{1, \ldots, K_n\}, \\ 1 - \alpha_n^*, & \beta_n(\theta) \geq \alpha + K_n\alpha_n. \end{cases}$$

We first show that $\beta_n(\theta) - \phi_n(\theta) < \alpha$. If $\beta_n(\theta) < \alpha$, then $\beta_n(\theta) - \phi_n(\theta) \leq \beta_n(\theta) < \alpha$. Furthermore, $\phi_n(\theta) > \beta_n(\theta)$ for θ such that $\alpha \leq \beta_n(\theta) < \alpha + K_n\alpha_n$. Finally, for θ such that $\beta_n(\theta) \geq \alpha + K_n\alpha_n$, $\beta_n(\theta) - \phi_n(\theta) \leq 1 - \phi_n(\theta) = \alpha_n^* < \alpha$.

For $n \geq n_0$, let $\theta_{n,i} = \beta_n^{-1}(\alpha + i\alpha_n)$, for $i \in \{0, \ldots, K_n\}$. (The $\theta_{n,i}$ are well defined because $\beta_n(\theta)$ is surjective whenever $n \geq n_0$.) Note that since for each i, $\beta_n^{-1}(\theta_{n,i}) \geq \alpha$, each $\theta_{n,i} > b$. Since $\lim_n \beta_n(\theta) \to 1$ for all $\theta > b$, there is a least $N > n$, such that for all $i \in \{0, \ldots, K_n\}$, $\beta_N(\theta_{n,i}) > \phi_n(\theta_{n,i})$. For all n, define $\sigma(n)$ to be the least such N. Define $\sigma^0(n) := n, \sigma^1(n) := \sigma(n)$, and $\sigma^m(n) := \sigma(\sigma^{m-1}(n))$.

We show that $\beta_{\sigma(n)}(\theta) \geq \phi_n(\theta)$. Suppose that $\theta < \theta_{n,0}$. Then $\beta_n(\theta) < \alpha$, and $\phi_n(\theta) = 0 \leq \beta_{\sigma(n)}(\theta)$. Suppose that $\theta_{n,i} \leq \theta < \theta_{n,i+1}$ for $i \in \{0, \ldots, K_n - 1\}$. Then, $\beta_{\sigma(n)}(\theta) \geq \beta_{\sigma(n)}(\theta_{n,i}) > \phi_n(\theta_{n,i}) = \phi_n(\theta)$. Finally, suppose that $\theta \geq \theta_{n,K_n}$. Then, similarly, $\beta_{\sigma(n)}(\theta) \geq \beta_{\sigma(n)}(\theta_{n,K_n}) > \phi_n(\theta_{n,K_n}) = \phi_n(\theta)$.

Now we show that $\phi_{\sigma(n)}(\theta) \geq \phi_n(\theta)$. Suppose that $\beta_{\sigma(n)}(\theta) < \alpha$. Then, since, $\phi_n(\theta) < \beta_{\sigma(n)}(\theta)$, it follows that $\phi_n(\theta) < \alpha$, and therefore, $\phi_n(\theta) = 0 \leq \phi_{\sigma(n)}(\theta)$.

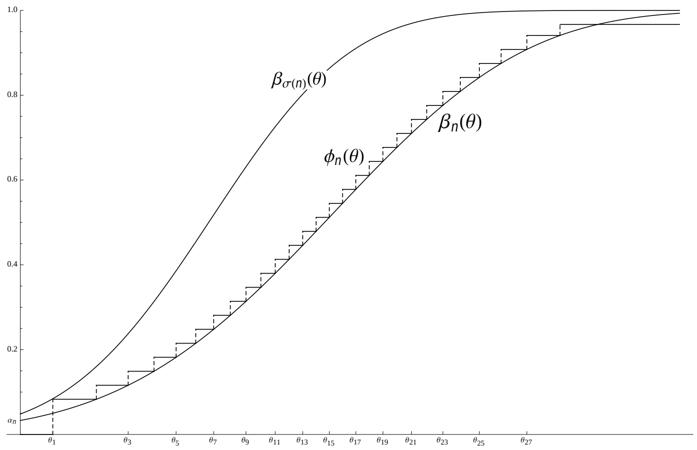


Figure 3.2: The basic idea of the proof is encapsulated in the figure. For any sample size n, we construct a step function ϕ_n that "almost dominates" the power function β_n. The step function dominates β_n for all θ except those for which $\beta_n(\theta)$ is less than α, or close to 1. Then, since the set of steps is finite, and the original test satisfies LimCon, there must be a sample size $\sigma(n)$ such that the power function $\beta_{\sigma(n)}$ strictly dominates the step function. Since $\beta_{\sigma(n)}$ dominates the step function, $\beta_{\sigma(n)}(\theta)$ can only be less than $\beta_n(\theta)$ if $\beta_n(\theta)$ is less than α, or they are both close to 1. The loss of power from n to $\sigma(n)$ is thereby bounded by α. Iterating this process we get a sequence of "good" sample sizes $n, \sigma(n), \sigma^2(n), \ldots$ such that the power is "almost increasing". To ensure α-monotonicity, it remains only to interpolate the intermediate sample sizes with test methods that throw out data points until they arrive at the nearest "good" sample size.

Suppose that $\alpha \leq \beta_{\sigma(n)} < \alpha + K_{\sigma(n)}\alpha_{\sigma(n)}$. Then, $\phi_n(\theta) \leq \beta_{\sigma(n)}(\theta) < \phi_{\sigma(n)}(\theta)$. Finally, if $\beta_{\sigma(n)} \geq \alpha + K_{\sigma(n)}\alpha_{\sigma(n)}$, then $\phi_\sigma(n) = 1 - \alpha^*_{\sigma(n)} \geq 1 - \alpha^*_n \geq \phi_n(\theta)$.

It is now easy to show that $\beta_n(\theta) - \beta_{\sigma^m(n)}(\theta) < \alpha$, since $\beta_n(\theta) - \beta_{\sigma^m(n)}(\theta) \leq \beta_n(\theta) - \phi_{\sigma^{m-1}(n)}(\theta) \leq \beta_n(\theta) - \phi_n(\theta) < \alpha$. Therefore, for an increasing sequence of "good" sample sizes, $n, \sigma(n), \sigma^2(n), \ldots$, the verifier $\{\lambda_n\}$ is α-monotonic. Furthermore, since the sequence $(\beta_{\sigma^m(n)}(b))_{m\in\mathbb{N}}$ converges to zero, there exists a subsequence $(\beta_{n^*_i}(b))_{i\in\mathbb{N}}$ that converges monotonically to zero. We use these facts to construct a verifier that is α-monotonic, by patching over the "bad" sample sizes.

Let $\pi(n) = \max\{n^*_i : n^*_i \leq n\}$. Let $\lambda^*_n(\omega_1, \ldots, \omega_n) := \lambda_{\pi(n)}(\omega_1, \ldots, \omega_{\pi(n)})$. We have taken pains to ensure that (λ^*_n) satisfies α-Mon. Since

$$\sup_{\mu\in H^c} \mu^n[\lambda_n^{-1}(H)] \leq \beta_{\pi(n)}(b) \downarrow_n 0,$$

(λ_n) satisfies MVanErr. Since (λ_n) is an almost sure verifier, the set $C =$

$\{\omega \in \Omega^\infty : \lambda_n(\omega|_n) \to H\}$ has $\mu^\infty(C) = 1$ for every $\mu \in H$. But if the sequence $(\lambda_n(\omega|_n))$ converges to H then so does the subsequence $(\lambda_{\pi(n)}(\omega|_{\pi(n)}))$. Therefore $C \subseteq C^* = \{\omega \in \Omega^\infty : \lambda_{\pi(n)}(\omega|\pi(n)) \to H\}$. Therefore, $\mu^\infty[\liminf(\lambda^*_n)^{-1}(H)] = 1$, if $\mu \in H$, and (λ^*_n) satisfies SLimCon. A fortiori, it also satisfies LimCon. Since $\alpha > 0$ was arbitrary, we are done. □

Lemma 3.3.4. *The monotonically verifiable propositions are closed under finite conjunctions.*

Proof of Lemma 3.3.4. By Lemma 3.3.2, it suffices to show that if $(\lambda^1_n), \ldots, (\lambda^k_n)$ are mutually independent α_i-monotonic verifiers of $A_1, \ldots, A_k$, with $\alpha = \sum_i \alpha_i$, then

$$\lambda_n(\vec{\omega}) = \begin{cases} \cap_{i=1}^k A_i, & \text{if } \lambda^i_n(\vec{\omega}) = A_i \text{ for all } i \in \{1, \ldots, k\}, \\ W, & \text{otherwise,} \end{cases}$$

is an α-monotonic verifier of $\cap_{i=1}^k A_i$.

Consider the case where $k = 2$. We first demonstrate that (λ_n) satisfies MVanErr. Suppose $\mu \notin A_1 \cap A_2$. Without loss of generality, suppose that $\mu \notin A_1$. By assumption there exists a sequence (α^1_n) such that $\alpha^1_n < \alpha_1$, $\alpha_n \downarrow 0$ and $\mu^n[(\lambda^1_n)^{-1}(A_1)] < \alpha^1_n$. Noticing that

$$\begin{aligned} \mu^n[\lambda_n^{-1}(A_1 \cap A_2)] &= \mu^n[(\lambda^1_n)^{-1}(A_1) \cap (\lambda^2_n)^{-1}(A_2)] \\ &\leq \mu^n[(\lambda^1_n)^{-1}(A_1)] \\ &\leq \alpha^1_n, \end{aligned}$$

we see that (λ_n) also satisfies MVanErr.

Suppose that $\mu \in A_1 \cap A_2$. We demonstrate that (λ_n) satisfies LimCon.

$$\begin{aligned} \mu^n[\lambda_n^{-1}(A_1 \cap A_2)] &= \mu^n[(\lambda^1_n)^{-1}(A_1) \cap (\lambda^2_n)^{-1}(A_2)] \\ &= 1 - \mu^n[(\lambda^1_n)^{-1}(W) \cup (\lambda^2_n)^{-1}(W)] \\ &\geq 1 - \mu^n[(\lambda^1_n)^{-1}(W)] + \mu^n[(\lambda^2_n)^{-1}(W)]. \end{aligned}$$

But since $\mu^n[(\lambda^1_n)^{-1}(W) + \mu^n[(\lambda^2_n)^{-1}(W)] \to 0$, it follows that

$$\mu^n[\lambda_n^{-1}(A_1 \cap A_2)] \to 1.$$

It remains to show that $(\lambda_n)_{n\in\mathbb{N}}$ satisfies α-Mon. To make the expressions more manageable, we make the following substitutions:

$$a_i^{n_j} = \mu^{n_j}[(\lambda^i_{n_j})^{-1}(A_i)].$$

Under the assumption of independence:

$$\mu^{n_1}[\psi_{n_1}^{-1}(A_1 \cap A_2)] - \mu^{n_2}[\psi_{n_2}^{-1}(A_1 \cap A_2)] = a_1^{n_1}a_2^{n_1} - a_1^{n_2}a_2^{n_2}.$$

By assumption, $a_1^{n_1} - a_1^{n_2} < \alpha/2$ and $a_2^{n_1} - a_2^{n_2} < \alpha/2$. Therefore:

$$\begin{aligned} a_1^{n_1}a_2^{n_1} - a_1^{n_2}a_2^{n_2} &< a_1^{n_1}a_2^{n_1} - a_1^{n_2}(a_2^{n_1} - \alpha/2) \\ &= a_1^{n_1}a_2^{n_1} - a_1^{n_2}a_2^{n_1} + a_1^{n_2} \cdot \alpha/2 \\ &= a_2^{n_1}(a_1^{n_1} - a_1^{n_2}) + a_1^{n_2} \cdot \alpha/2 \\ &< a_2^{n_1} \cdot \alpha/2 + a_1^{n_2} \cdot \alpha/2 \\ &\leq \alpha. \end{aligned}$$

The case when $k > 2$ follows immediately by induction. □

Lemma 3.3.5. *The monotonically verifiable propostions are closed under countable disjunction.*

Proof of Lemma 3.3.5. By Lemma 3.3.2, it suffices to show that if $(\lambda_n^1), \ldots, (\lambda_n^k), \ldots,$ are mutually independent α_i-monotonic verifiers of $A_1, \ldots, A_k, \ldots,$ such that $\sum_{i=1}^\infty \alpha_i$ converges to α, then,

$$\lambda_n(\vec{\omega}) = \begin{cases} \cup_{i=1}^\infty A_i, & \text{if } \lambda_n^i(\vec{\omega}) = A_i \text{ for some } i \in \{1, \ldots, n\}, \\ W, & \text{otherwise,} \end{cases}$$

is an α-monotonic verifier of $\cup_{i=1}^\infty A_i$.

We first demonstrate that (λ_n) satisfies MVanErr. Suppose that $\mu \notin \cup_{i=1}^\infty A_i$. Then, for each i there is (α_n^i) such that $\alpha_n^i < \alpha_i$, $\alpha_n^i \downarrow 0$, and $\mu^n[(\lambda_n^i)^{-1}(A_i)] \leq \alpha_n^i$. Therefore,

$$\begin{aligned} \mu^n[\lambda_n^{-1}(\cup_{i=1}^\infty A_i)] &= \mu^n[\cup_{i=1}^n (\lambda_n^i)^{-1}(A_i)] \\ &\leq \sum_{i=1}^n \mu^n[(\lambda_n^i)^{-1}(A_i)] \\ &\leq \sum_{i=1}^\infty \alpha_n^i \leq \alpha. \end{aligned}$$

It remains to show that $S_n = \sum_{i=1}^\infty \alpha_n^i$ converges monotonically to zero as $n \to \infty$. Since $\alpha_n^i \geq \alpha_{n+1}^i$ for each i, we have that $S_n \geq S_{n+1}$. Therefore, the sequence (S_n) is decreasing and bounded below by zero. By the monotone convergence theorem, the sequence (S_n) converges to its infimum. We show that the infimum is zero. Let $\epsilon > 0$. Since the tail of a convergent series tends to zero, there is K such that $\sum_{i=K}^\infty \alpha_i < \epsilon/2$. Therefore, $T_n = \sum_{i=K}^\infty \alpha_n^i < \epsilon/2$. Since $S_n = T_n + \sum_{i=1}^{K-1} \alpha_n^i$, and $\sum_{i=1}^{K-1} \alpha_n^i \to 0$ as $n \to \infty$, there is N such that $S_n < \epsilon$ for $n \geq N$. Since ϵ was arbitrary, we are done.

Suppose that $\mu \in \cup_{i=1}^\infty A_i$. We show that (λ_n) satisfies LimCon. Since $\mu \in \cup_{i=1}^\infty A_i$, there is k such that $\mu \in A_k$. For $n \geq k$,

$$\begin{aligned}\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] &= \mu^n[\cup_{i=1}^{n} (\lambda_n^i)^{-1}(A_i)]\\ &\geq \mu^n[(\lambda_n^k)^{-1}(A_k)].\end{aligned}$$

But since $\mu^n[(\lambda_n^k)^{-1}(A_k)] \to 1$, we have that $\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] \to 1$.

It remains to show that (λ_n) satisfies α-MON. By the inclusion-exclusion formula:

$$\begin{aligned}&\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] = \mu^n[\cup_{i=1}^{n} (\lambda_n^i)^{-1}(A_i)] =\\ &= \sum_{i=1}^{n} \mu^n[(\lambda_n^i)^{-1}(A_i)] - \sum_{i<j<n} \mu^n[(\lambda_n^i)^{-1}(A_i) \cap (\lambda_n^j)^{-1}(A_j)]\\ &+ \sum_{i<j<k<n} \mu^n[(\lambda_n^i)^{-1}(A_i) \cap (\lambda_n^j)^{-1}(A_j) \cap (\lambda_n^k)^{-1}(A_k)] + \cdots\\ &+ (-1)^{n-1}\mu^n[\cap_{i=1}^{n} (\lambda_n^i)^{-1}(A_i)].\end{aligned}$$

To make the expressions more manageable, we make the following substitutions:

$$a_i^n = \mu^n[(\lambda_n^i)^{-1}(A_i)].$$

Since the verifiers are mutually independent:

$$\mu^n[\lambda_n^{-1}(\cup_i A_i)] = \sum_{i=1}^{n} a_i^n - \sum_{i<j<n} a_i^n a_j^n + \cdots + (-1)^{n-1} a_1^n a_2^n \cdots a_n^n.$$

Or, in closed form:

$$\mu^n[\lambda_n^{-1}(\cup_i A_i)] = \sum_{j=1}^{n} \left((-1)^{j-1} \sum_{\substack{I \subset \{1,\ldots,n\}\\ |I|=j}} \prod_{i \in I} a_i^n \right).$$

Furthermore, for $n_1 < n_2$,

$$\begin{aligned}\mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] &= \mu^{n_2}[\cup_{i=1}^{n_2} (\lambda_{n_2}^i)^{-1}(A_i)]\\ &\geq \mu^{n_2}[\cup_{i=1}^{n_1} (\lambda_{n_2}^i)^{-1}(A_i)]\\ &= \sum_{j=1}^{n_1} \left((-1)^{j-1} \sum_{\substack{I \subset \{1,\ldots,n_1\}\\ |I|=j}} \prod_{i \in I} a_i^{n_2} \right).\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mu^{n_1}[\lambda_{n_1}^{-1}(\cup_i A_i)] - \mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] = \\
&= \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\prod_{i\in I}a_i^{n_1}\right) - \sum_{j=1}^{n_2}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_2\}\\|I|=j}}\prod_{i\in I}a_i^{n_2}\right) \\
&\le \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\prod_{i\in I}a_i^{n_1}\right) - \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\prod_{i\in I}a_i^{n_2}\right) \\
&= \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\left(\prod_{i\in I}a_i^{n_1} - \prod_{i\in I}a_i^{n_2}\right)\right).
\end{aligned}$$

Next, we demonstrate that

$$\prod_{i\in I}a_i^{n_1} - \prod_{i\in I}a_i^{n_2} = \sum_{i\in I}\left((a_i^{n_1}-a_i^{n_2})\prod_{j\in I, j<i}a_j^{n_2}\prod_{j\in I, j>i}a_j^{n_1}\right).$$

By induction on $|I|$. Let $k = \max i \in I$.

$$\begin{aligned}
&\prod_{i\in I}a_i^{n_1} - \prod_{i\in I}a_i^{n_2} = a_k^{n_1}\prod_{i\in I\setminus\{k\}}a_i^{n_1} - a_k^{n_1}\prod_{i\in I\setminus\{k\}}a_i^{n_2} + (a_k^{n_1}-a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= a_k^{n_1}\left(\prod_{i\in I\setminus\{k\}}a_i^{n_1} - \prod_{i\in I\setminus\{k\}}a_i^{n_2}\right) + (a_k^{n_1}-a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= a_k^{n_1}\left(\sum_{i\in I\setminus\{k\}}\left((a_i^{n_1}-a_i^{n_2})\prod_{j\in I\setminus\{k\}, j<i}a_j^{n_2}\prod_{j\in I\setminus\{k\}, j>i}a_j^{n_1}\right)\right) + (a_k^{n_1}-a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= \sum_{i\in I\setminus\{k\}}\left((a_i^{n_1}-a_i^{n_2})\prod_{j\in I, j<i}a_j^{n_2}\prod_{j\in I, j>i}a_j^{n_1}\right) + (a_k^{n_1}-a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= \sum_{i\in I}\left((a_i^{n_1}-a_i^{n_2})\prod_{j\in I, j<i}a_j^{n_2}\prod_{j\in I, j>i}a_j^{n_1}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mu^{n_1}[\lambda_{n_1}^{-1}(\cup_i A_i)] - \mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] \leq \\
&\leq \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\ |I|=j}} \left(\prod_{i\in I} a_i^{n_1} - \prod_{i\in I} a_i^{n_2}\right) \\
&= \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\ |I|=j}} \sum_{i\in I} \left((a_i^{n_1} - a_i^{n_2}) \prod_{k\in I, k<i} a_k^{n_2} \prod_{k\in I, k>i} a_k^{n_1}\right) \\
&= \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{i=1}^{n_1}(a_i^{n_1} - a_i^{n_2}) \sum_{\substack{I\subset\{1,\ldots,n_1\}\\ |I|=j, i\in I}} \prod_{k\in I, k<i} a_k^{n_2} \prod_{k\in I, k>i} a_k^{n_1} \\
&= \sum_{i=1}^{n_1}(a_i^{n_1} - a_i^{n_2}) \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\ |I|=j, i\in I}} \prod_{k\in I, k<i} a_k^{n_2} \prod_{k\in I, k>i} a_k^{n_1}
\end{aligned}$$

Noticing that

$$\begin{aligned}
&\sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\ |I|=j, i\in I}} \prod_{k\in I, k<i} a_k^{n_2} \prod_{k\in I, k>i} a_k^{n_1} = \\
&= \sum_{j=1}^{n_1-1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\setminus\{i\}\\ |I|=j}} \prod_{k\in I, k<i} a_k^{n_2} \prod_{k\in I, k>i} a_k^{n_1} \\
&= \sum_{j=1}^{n_1-1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\setminus\{i\}\\ |I|=j}} \mu^{n_2}\left[\bigcap_{k\in I, k<i} (\lambda_{n_2}^k)^{-1}(A_k)\right] \mu^{n_1}\left[\bigcap_{k\in I, k>i} (\lambda_{n_1}^k)^{-1}(A_k)\right] \\
&= \sum_{j=1}^{n_1-1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\setminus\{i\}\\ |I|=j}} \mu^{n_1}\times\mu^{n_2}\left[\bigcap_{k\in I, k>i} (\lambda_{n_1}^k)^{-1}(A_k) \times \bigcap_{k\in I, k<i} (\lambda_{n_2}^k)^{-1}(A_k)\right] \\
&= \mu^{n_1}\times\mu^{n_2}\left[\bigcup_{k>i} (\lambda_{n_1}^k)^{-1}(A_k) \times \Omega^{n_2} \cup \bigcup_{k<i} \Omega^{n_1} \times (\lambda_{n_2}^k)^{-1}(A_k)\right] \leq 1,
\end{aligned}$$

where the last equality follows from the inclusion-exclusion principle. It follows that

$$\begin{aligned}\mu^{n_1}[\lambda_{n_1}^{-1}(\cup_i A_i)] - \mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] &\leq \sum_{i=1}^{n_1}(a_i^{n_1} - a_i^{n_2}) \\ &\leq \sum_{i=1}^{n_1} \alpha_i < \alpha,\end{aligned}$$

as required.

□

3.3.3 Monotonic Verification of Statistical Dependence

In light of the results of the previous section, it is straightforward to show that hypotheses of conditional dependence are monotonically verifiable.

Theorem 3.3.3. *For almost surely clopen events A, B, C and $r \in [0, 1]$, the following hypotheses are monotonically verifiable:*

1. $\{\mu : \mu(A) > r\}$;

2. $\{\mu : \mu(A) < r\}$;

3. $\{\mu : \mu(A)\mu(B) > r\}$;

4. $\{\mu : 0 < \mu(A)\mu(B) < r\}$;

5. $\{\mu : \mu(A \cap B) \neq \mu(A)\mu(B)\}$;

6. $\{\mu : \mu(C) > 0, \frac{\mu(A)}{\mu(C)} > r\}$;

7. $\{\mu : \mu(C) > 0, \frac{\mu(A)}{\mu(C)} < r\}$;

8. $\{\mu : \mu(C) > 0, \frac{\mu(A)\mu(B)}{\mu(C)} > r\}$;

9. $\{\mu : \mu(C) > 0, 0 < \frac{\mu A \mu B}{\mu C} < r\}$;

10. $\{\mu : \mu(C) > 0, \frac{\mu(A\cap B\cap C)}{\mu(C)} \neq \frac{\mu(A\cap C)\mu(B\cap C)}{\mu(C)}\}$.

Proof of Theorem 3.3.3. The arguments are identical, except we invoke Lemma 3.3.3 in the place of Lemma 3.2.1, and Lemmas 3.3.5 and 3.3.4 in the place of Lemma 3.2.2. □

Theorem 3.3.4. *Suppose that X, Y, Z are random variables taking values in an aribtrary measurable space, and that the σ-algebras $\sigma(X), \sigma(Y), \sigma(Z)$ are generated by countable, almost surely clopen bases $\mathcal{I}(X), \mathcal{I}(Y), \mathcal{I}(Z)$. Then the hypothesis of conditional dependence, $\{\mu : X \not\perp\!\!\!\perp_\mu Y|Z\}$, is monotonically verifiable.*

Proof. We have demonstrated, in the proof of Theorem 3.2.7 that

$$\{\mu : X \not\perp\!\!\!\perp_\mu Y|Z\} =$$
$$= \bigcup_{A\in\mathcal{A}(X), B\in\mathcal{A}(Y), C\in\mathcal{A}(Z)} \left\{\mu : \mu(C) > 0, \frac{\mu(A \cap B \cap C)}{\mu(C)} \neq \frac{\mu(A \cap C)\mu(B \cap C)}{\mu(C)}\right\}.$$

By part (10) of Theorem 3.3.3, we have expressed the hypothesis of conditional dependence as a countable union of monotonically verifiable hypotheses. Therefore, by Lemma 3.3.5, it is monotonically verifiable. □

It poses no additional difficulty to prove analogous results for the setting of causal graphical models. Suppose that M is a set of measures on a measurable space $(\Omega, \mathcal{B})$. Let $\mathcal{V}$ be a fixed, finite set of random variables $X_1, X_2, \ldots, X_n$ taking values in measurable spaces $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n)$. Assume that each X_i is $(\mathcal{B}, \mathcal{S}_i)$ measurable. Let DAG be the set of all directed acyclic graphs on variables in $\mathcal{V}$. Let $W \subset M \times \mathsf{DAG}$.

Theorem 3.3.5. *Suppose that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the following hypotheses are monotonically refutable:*

1. *The true measure is Markov to $\mathcal{G} \in \mathsf{DAG}$;*
2. *The true measure is Markov to $[\mathcal{G}]$, for $\mathcal{G} \in \mathsf{DAG}$;*
3. *The true measure is Markov to some $\mathcal{G} \in \mathsf{DAG}$.*

Proof of Theorem 3.3.5. The proof is identical to the proof of Theorem 3.2.10, except we invoke Theorem 3.3.4 in the place of Theorem 3.2.7, and Lemma 3.3.5 in the place of Lemma 3.2.2. □

Corollary 3.3.1. *Suppose that the causal Markov and faithfulness assumptions hold, and that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the causal hypothesis*

$$\{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] \preceq [\mathcal{G}]\}$$

is monotonically refutable.

Proof of Corollary 3.3.1. The proof is identical to the Proof of Corollary 3.2.2, except we invoke Theorem 3.3.5 in the place of Theorem 3.2.9. □

Theorem 3.3.6. *Suppose that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the following hypotheses are statistically verifiable:*

1. *The true measure is faithful to $\mathcal{G} \in \mathsf{DAG}$;*
2. *The true measure is faithful to $[\mathcal{G}]$, for $\mathcal{G} \in \mathsf{DAG}$;*
3. *The true measure is faithful to some $\mathcal{G} \in \mathsf{DAG}$.*

Proof of Theorem 3.3.6. The proof is identical to the proof of Theorem 3.2.10, except we invoke Theorem 3.3.4 in the place of Theorem 3.2.7, and Lemma 3.3.5 in the place of Lemma 3.2.2. □

Corollary 3.3.2. *Suppose that the causal Markov and faithfulness assumptions hold, and that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the causal hypothesis*

$$\{(\mu, \mathcal{G}') \in W : [\mathcal{G}] \preceq [\mathcal{G}']\}.$$

is monotonically verifiable.

Proof of Corollary 3.3.1. The proof is identical to the Proof of Corollary 3.2.3, except we invoke Theorem 3.3.6 in the place of Theorem 3.2.10. □

3.4 Limiting Statistical Verification, Refutation, and Decision

3.4.1 Defining the Success Concepts

We now weaken the preceding two criteria of statistical verifiability to arrive at statistical notions of limiting verifiability. These notions are properly *inductive*, because they drop any requirement of a short-run bound on the chance of error.

Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible methods is a *limiting verifier in chance* of $H \subseteq W$ iff

1. $\mu \in H$ iff there is $H' \subseteq H$, s.t. $\lim_{n\to\infty} \mu^n[\lambda_n^{-1}(H')] = 1$;

2. $\mu \notin H$ iff for all $H' \subseteq H$, $\lim_{n\to\infty} \mu^n[\lambda_n^{-1}(H')] = 0$.

Say that $H \subseteq W$ is *limiting verifiable in chance* iff there is a limiting verifier in chance of H.

Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible methods is a *limiting almost sure verifier* of $H \subseteq W$ iff

1. $\mu \in H$ iff there is $H' \subseteq H$, s.t. $\mu^\infty[\liminf_{n\to\infty} \lambda_n^{-1}(H')] = 1$;

2. $\mu \notin H$ iff for all $H' \subseteq H$, $\mu^\infty[\limsup_{n\to\infty} \lambda_n^{-1}(H')] = 0$.

Say that $H \subseteq W$ is *limiting a.s. verifiable* iff there is a limiting a.s. verifier of H.

We say that an hypothesis H is *limiting refutable in chance*, or *limiting a.s. refutable*, iff its complement is limiting verifiable in the appropriate sense.

A method is a limiting verifier of H if it converges to some "reason" H' entailing H, if H is true, and eventually eliminates every such reason if H is false. In possibilities where H is false, a limiting verification method may have a high chance of outputting *some* reason entailing H at *every* sample size. If that is a shortcoming, it is remedied by the two-sided notion of statistical decision in the limit.

Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible methods is a *limiting decision procedure in chance* for $H \subseteq W$ iff it is a limiting verifier in chance of H and H^{c}. Say that $H \subseteq W$ is *limiting decidable in chance* iff there exists a limiting decision procedure in chance for H. Say that $(\lambda_n)_{n\in\mathbb{N}}$ is a *limiting almost sure decision procedure* for H iff it is a limiting almost sure verifier of H and H^{c}. Say that $H \subseteq W$ is *limiting almost sure decidable* iff there exists a limiting almost sure decision procedure for H.

3.4.2 Characterization Theorems

The central theorem of this section states that, for samples spaces with countable, almost surely clopen bases, limiting verifiability in chance and limiting almost sure verifiability are equivalent to being a countable union of locally closed sets in the weak topology.

Theorem 3.4.1. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is limiting verifiable in chance;*
2. *H is limiting almost surely verifiable;*
3. *H is a countable union of locally closed sets in the weak topology.*

The characterization of statistical refutability follows immediately.

Theorem 3.4.2. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is limiting refutable in chance;*
2. *H is limiting almost surely refutable;*
3. *H^{c} is a countable union of locally closed sets in the weak topology.*

We also characterize they hypotheses that are statistically decidable in the limit.

Theorem 3.4.3. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is limiting decidable in chance;*
2. *H is limiting almost surely decidable;*
3. *Both H and H^{c} are countable unions of locally closed sets in the weak topology.*

In light of Theorems 3.4.1, 3.4.2, 3.4.3 we will say that a hypothesis is simply *statistically verifiable/refutable/decidable in the limit* when the precise sense of limiting verifiability/refutability is not relevant.

The Proof of Theorem 3.4.1 proceeds as follows. It is immediate from the definitions that 2 entails 1. It is also, easy to show that 1 entails 3.

Proof of Theorem 3.4.1. 1 entails 3. Suppose that $(\lambda_n)_{n\in\mathbb{N}}$ is a limiting verifier in chance of H. Let $\alpha \in (0,1)$. Let $\mathsf{rng}(\lambda_n)$ be the range of λ_n. For every $H' \in \mathsf{rng}(\lambda_n)$, let:

$$\begin{aligned} \mathsf{trig}_n(H') &= \{\mu : \mu^n[\lambda_n^{-1}(H')] > \alpha\}; \\ \mathsf{def}_n(H') &= \cup_{m>n}\{\mu : \mu^m[\lambda_m^{-1}(H')^{\mathsf{c}}] > 1-\alpha\}. \end{aligned}$$

Lemma 3.1.4 and the feasibility of the λ_n entail that $\mathsf{trig}_n(H')$ and $\mathsf{def}_n(H')$ are both open in the weak topology. We claim that:

$$H = \bigcup_{n=1}^{\infty} \bigcup_{\substack{H'\in\mathsf{rng}(\lambda_n)\\ H'\subseteq H}} \mathsf{trig}_n(H') \setminus \mathsf{def}_n(H').$$

Observe that $\nu \in H$ iff there is $H' \subseteq H$ and $m \in \mathbb{N}$, such that for all $n \geq m$, $\nu^n[\lambda_n^{-1}(H')] > \alpha$ iff $\nu \in \mathsf{trig}_n(H') \setminus \mathsf{def}_n(H')$. Therefore, proposition H is a countable union of locally closed sets. □

To show that 3 entails 1, we prove Lemma 3.4.1. Say that a hypothesis is *verifutable* iff it is the conjunction of an a.s. verifiable and an a.s. refutable hypothesis. By Theorem 3.2.1, every locally closed set in the weak topology is verifutable. The first part of Lemma 3.4.1 shows that every verifutable hypothesis is limiting a.s. decidable, and therefore, limiting verifiable in chance. The third part of Lemma 3.4.1 shows that countable unions of verifutable propositions are limiting a.s. verifiable, completing the proof.

Lemma 3.4.1.

1. *Verifutable propositions are limiting almost surely decidable.*
2. *The limiting almost surely decidable propositions form an algebra.*
3. *If $(A_i)_{i\in\mathbb{N}}$ is a collection of limiting almost surely decidable hypotheses, then $\cup_{i=1}^{\infty} A_i$ is limiting almost surely verifiable.*

Proof of Lemma 3.4.1. Proof of 1. Suppose that $A = V \cap R$, where V is almost surely verifiable and R is almost surely refutable. We show that A is almost surely decidable. Let (τ_n) be an almost sure α-verifier of V. Let (δ_n) be an almost sure α-verifier of R^{c}. Define:

$$\lambda_n(\vec{\omega}) = \begin{cases} A, & \text{if } \tau_n(\vec{\omega}) = V \text{ and } \delta_n(\vec{\omega}) = W, \\ A^{\mathsf{c}}, & \text{otherwise.} \end{cases}$$

We show that:

1. for $\mu \in A$, $\mu^\infty[\liminf_{n\to\infty}(\lambda_n)^{-1}(A)] = 1$,
2. for $\mu \in A^{\mathsf{c}}$, $\mu^\infty[\liminf_{n\to\infty}(\lambda_n)^{-1}(A^{\mathsf{c}})] = 1$.

Suppose that $\mu \in A$. Then:

$$\begin{aligned}
&\mu^\infty \left[\liminf_{n\to\infty}(\lambda_n)^{-1}(A)\right] = \\
&= \mu^\infty \left[\liminf_{n\to\infty}(\tau_n)^{-1}(V) \cap (\delta_n)^{-1}(W)\right] \\
&= 1 - \mu^\infty \left[\limsup_{n\to\infty}(\tau_n)^{-1}(W) \cup (\delta_n)^{-1}(R^{\mathsf{c}})\right] \\
&= 1 - \mu^\infty \left[\limsup_{n\to\infty}(\tau_n)^{-1}(W) \cup \limsup_{n\to\infty}(\delta_n)^{-1}(R^{\mathsf{c}})\right] \\
&\geq 1 - \mu^\infty \left[\limsup_{n\to\infty}(\tau_n)^{-1}(W)\right] - \mu^\infty \left[\limsup_{n\to\infty}(\delta_n)^{-1}(R^{\mathsf{c}})\right],
\end{aligned}$$

where the final inequality follows from the Union Bound. But since $\mu \in V$, $\mu^\infty \left[\limsup_{n\to\infty}(\tau_n)^{-1}(W)\right] = 0$. Furthermore, since $\mu \in R$,

$$\sum_{n=1}^{\infty} \mu^\infty \left[(\delta_n)^{-1}(R^{\mathsf{c}})\right] \leq \alpha.$$

Therefore, by the Borel-Cantelli Lemma, $\mu^\infty \left[\limsup_{n\to\infty}(\delta_n)^{-1}(R^{\mathsf{c}})\right] = 0$. So $\mu^\infty \left[\liminf_{n\to\infty}(\lambda_n)^{-1}(A)\right] = 1$.

Now, suppose that $\mu \in A^{\mathsf{c}}$. Then either $\mu \notin V$, or $\mu \in R^{\mathsf{c}}$. In the first case:

$$\begin{aligned}
\mu^\infty \left[\liminf_{n\to\infty}(\lambda_n)^{-1}(A^{\mathsf{c}})\right] &= \mu^\infty \left[\liminf_{n\to\infty}(\tau_n)^{-1}(W) \cup (\delta_n)^{-1}(R^{\mathsf{c}})\right] \\
&\geq \mu^\infty \left[\liminf_{n\to\infty}(\tau_n)^{-1}(W)\right].
\end{aligned}$$

But $\mu^\infty \left[\liminf_{n\to\infty}(\tau_n)^{-1}(W)\right] = 1$ by the Borel-Cantelli Lemma. In the second case, $\mu \in R^{\mathsf{c}}$. Therefore:

$$\begin{aligned}
\mu^\infty \left[\liminf_{n\to\infty}(\lambda_n)^{-1}(A^{\mathsf{c}})\right] &= \mu^\infty \left[\liminf_{n\to\infty}(\tau_n)^{-1}(W) \cup (\delta_n)^{-1}(R^{\mathsf{c}})\right] \\
&\geq \mu^\infty \left[\liminf_{n\to\infty}(\delta_n)^{-1}(R^{\mathsf{c}})\right] \\
&= 1.
\end{aligned}$$

Proof of 2. Clearly, the limiting almost surely decidable propostions are closed under complements. We show that they are closed under finite unions and intersections. Suppose that A', A'' are limiting almost surely decidable. Let $A = A' \cup A''$. Let (λ'_n) be an almost sure decision procedure for A' and let (λ''_n) be an almost sure decision procedure for A''. Define:

$$\lambda_n(\vec{\omega}) = \begin{cases} A, & \text{if } \lambda'_n(\vec{\omega}) = A' \text{ or } \lambda''_n(\vec{\omega}) = A'', \\ A^{\mathsf{c}}, & \text{otherwise.} \end{cases}$$

Suppose that $\mu \in A' \cup A''$. Then, either $\mu \in A'$ or $\mu \in A''$. In the first case:

$$\begin{aligned}\mu^\infty \left[\liminf_{n\to\infty} (\lambda_n)^{-1}(A)\right] &= \mu^\infty \left[\liminf_{n\to\infty} (\lambda'_n)^{-1}(A') \cup (\lambda''_n)^{-1}(A'')\right] \\ &\geq \mu^\infty \left[\liminf_{n\to\infty} (\lambda'_n)^{-1}(A')\right] = 1.\end{aligned}$$

In the second case:

$$\begin{aligned}\mu^\infty \left[\liminf_{n\to\infty} (\lambda_n)^{-1}(A)\right] &= \mu^\infty \left[\liminf_{n\to\infty} (\lambda'_n)^{-1}(A') \cup (\lambda''_n)^{-1}(A'')\right] \\ &\geq \mu^\infty \left[\liminf_{n\to\infty} (\lambda''_n)^{-1}(A'')\right] = 1.\end{aligned}$$

Now, suppose that $\mu \in A^{\mathsf{c}}$. Then:

$$\begin{aligned}&\mu^\infty \left[\liminf_{n\to\infty} (\lambda_n)^{-1}(A^{\mathsf{c}})\right] = \\ &= \mu^\infty \left[\liminf_{n\to\infty} (\lambda'_n)^{-1}(A'^{\mathsf{c}}) \cap (\lambda''_n)^{-1}(A''^{\mathsf{c}})\right] \\ &= 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda'_n)^{-1}(A') \cup (\lambda''_n)^{-1}(A'')\right] \\ &= 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda'_n)^{-1}(A') \cup \limsup_{n\to\infty} (\lambda''_n)^{-1}(A'')\right] \\ &\geq 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda'_n)^{-1}(A')\right] - \mu^\infty \left[\limsup_{n\to\infty} (\lambda''_n)^{-1}(A'')\right] \\ &= 1.\end{aligned}$$

We now show that limiting almost surely decidable propositions are closed under finite conjunctions. Let $B = A' \cap A''$. Define:

$$\lambda_n(\vec{\omega}) = \begin{cases} B, & \text{if } \lambda'_n(\vec{\omega}) = A' \text{ and } \lambda''_n(\vec{\omega}) = A'', \\ B^{\mathsf{c}}, & \text{otherwise.} \end{cases}$$

Suppose that $\mu \in B$. Then:

$$\begin{aligned}&\mu^\infty \left[\liminf_{n\to\infty} (\lambda_n)^{-1}(B)\right] = \\ &= \mu^\infty \left[\liminf_{n\to\infty} (\lambda'_n)^{-1}(A') \cap (\lambda''_n)^{-1}(A'')\right] \\ &= 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda'_n)^{-1}(A'^{\mathsf{c}}) \cup (\lambda''_n)^{-1}(A''^{\mathsf{c}})\right] \\ &= 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda'_n)^{-1}(A'^{\mathsf{c}}) \cup \limsup_{n\to\infty} (\lambda''_n)^{-1}(A''^{\mathsf{c}})\right] \\ &\geq 1 - \mu^\infty \left[\limsup_{n\to\infty} (\lambda'_n)^{-1}(A'^{\mathsf{c}})\right] - \mu^\infty \left[\limsup_{n\to\infty} (\lambda''_n)^{-1}(A''^{\mathsf{c}})\right] \\ &= 1.\end{aligned}$$

Now suppose that $\mu \notin B$. Suppose, without loss of generality, that $\mu \notin A'$. Then:

$$\begin{aligned}\mu^\infty \left[\liminf_{n\to\infty}(\lambda_n)^{-1}(B^\mathsf{c})\right] &= \mu^\infty \left[\liminf_{n\to\infty}(\lambda'_n)^{-1}(A'^\mathsf{c}) \cup (\lambda''_n)^{-1}(A''^\mathsf{c})\right]\\ &\geq \mu^\infty \left[\liminf_{n\to\infty}(\lambda'_n)^{-1}(A'^\mathsf{c})\right] = 1.\end{aligned}$$

Proof of 3. Suppose that $(A_i)_{i\in\mathbb{N}}$ are limiting almost surely decidable. Let (λ^i_n) be a limiting decision procedure for A_i. Now, define: $\sigma_n(\vec{\omega}) = \min\{i : i \leq n \text{ and } \lambda^i_n(\vec{\omega}) = A_i\}$. Let

$$\lambda_n(\vec{\omega}) = \begin{cases} A_{\sigma(\vec{\omega})}, & \text{if } \sigma(\vec{\omega}) < \infty \\ W, & \text{otherwise.}\end{cases}$$

Since the verdict of λ_n depends only on the output of finitely many λ^i_n, it is feasible. Suppose that $\mu \in H$. Let $i = \min\{i : \mu \in A_i\}$.

$$\begin{aligned}\mu^\infty \left[\liminf_{n\to\infty}(\lambda^i_n)^{-1}(A_i)\right] &= \mu^\infty \left[\liminf_{n\to\infty} \cap_{j<i}(\lambda^j_n)^{-1}(A^\mathsf{c}_j) \cap (\lambda^i_n)^{-1}(A_i)\right]\\ &= 1 - \mu^\infty \left[\limsup_{n\to\infty} \cup_{j<i}(\lambda^j_n)^{-1}(A_j) \cup (\lambda^i_n)^{-1}(A^\mathsf{c}_i)\right]\\ &\geq 1 - \sum_{j<i}\mu^\infty[\limsup_{n\to\infty}(\lambda^j_n)^{-1}(A_j)] - \mu^\infty[\limsup_{n\to\infty}(\lambda^i_n)^{-1}(A^\mathsf{c}_i)].\end{aligned}$$

But for $j < i$, $\mu^\infty[\limsup_{n\to\infty}(\lambda^j_n)^{-1}(A_j)] = 0$. And furthermore,

$$\mu^\infty[\limsup_{n\to\infty}(\lambda^i_n)^{-1}(A^\mathsf{c}_i)] = 0.$$

Therefore $\mu^\infty \left[\liminf_{n\to\infty}(\lambda^i_n)^{-1}(A_i)\right] = 1$.

Now, suppose that $\mu \notin H$. Then:

$$\begin{aligned}\mu^\infty \left[\liminf_{n\to\infty}(\lambda^i_n)^{-1}(A_i)\right] &= \mu^\infty \left[\liminf_{n\to\infty} \cap_{j<i}(\lambda^j_n)^{-1}(A^\mathsf{c}_j) \cap (\lambda^i_n)^{-1}(A_i)\right]\\ &\leq \mu^\infty \left[\liminf_{n\to\infty}(\lambda^i_n)^{-1}(A_i)\right]\\ &= 0.\end{aligned}$$

□

Proof of Theorem 3.4.3. It is immediate from the definitions that 2 implies 1. If H is limiting decidable in chance, it is both limiting verifiable and refutable in chance. Therefore, by Theorem 3.4.1, 1 implies 3. It remains to show that 3 implies 2. The proof is nearly identical to the proof of Theorem 3.4.1. By assumption, there is a countable collection of locally closed A_i such that $W =$

$\cup_i A_i$ and, for each i, $A_i \subseteq H$ or $A_i \subseteq H^{\mathsf{c}}$. Each A_i can be expressed as a difference of open sets:

$$A_i = \mathsf{ext}(\mathsf{frnt}A_i) \setminus \mathsf{ext}A_i.$$

For each A_i, let (τ_n^i) be an a.s. α-verifier of $\mathsf{ext}(\mathsf{frnt}A_i)$ and let (δ_n^i) be an a.s. α-verifier of $\mathsf{ext}A_i$. Define

$$\lambda_n^i(\vec{\omega}) = \begin{cases} A_i, & \text{if } \tau_n^i(\vec{\omega}) = \mathsf{ext}(\mathsf{frnt}A_i) \text{ and } \delta_n^i(\vec{\omega}) = W, \\ A_i^{\mathsf{c}}, & \text{otherwise.} \end{cases}$$

Then, as was shown in the proof of Theorem 3.4.1, we have that:

1. for $\mu \in A_i$, $\mu^\infty[\liminf_{n\to\infty}(\lambda_n^i)^{-1}(A_i)] = 1$,
2. for $\mu \in A_i^{\mathsf{c}}$, $\mu^\infty[\liminf_{n\to\infty}(\lambda_n^i)^{-1}(A_i^{\mathsf{c}})] = 1$.

Now, define: $\sigma_n(\vec{\omega}) = \min\{i : i \leq n \text{ and } \lambda_n^i(\vec{\omega}) = A_i\}$. Let

$$\lambda_n(\vec{\omega}) = \begin{cases} A_{\sigma(\vec{\omega})}, & \text{if } \sigma(\vec{\omega}) < \infty \\ W, & \text{otherwise.} \end{cases}$$

Since the verdict of λ_n depends only on the output of finitely many λ_n^i, it is feasible. Suppose that $\mu \in W$. Let $i = \min\{i : \mu \in A_i\}$.

$$\begin{aligned} \mu^\infty\left[\liminf_{n\to\infty}(\lambda_n^i)^{-1}(A_i)\right] &= \mu^\infty\left[\liminf_{n\to\infty} \cap_{j<i}(\lambda_n^j)^{-1}(A_j^{\mathsf{c}}) \cap (\lambda_n^i)^{-1}(A_i)\right] \\ &= 1 - \mu^\infty\left[\limsup_{n\to\infty} \cup_{j<i}(\lambda_n^j)^{-1}(A_j) \cup (\lambda_n^i)^{-1}(A_i^{\mathsf{c}})\right] \\ &\geq 1 - \sum_{j<i} \mu^\infty[\limsup_{n\to\infty}(\lambda_n^j)^{-1}(A_j)] - \mu^\infty[\limsup_{n\to\infty}(\lambda_n^i)^{-1}(A_i^{\mathsf{c}})]. \end{aligned}$$

But for $j < i$, $\mu^\infty[\limsup_{n\to\infty}(\lambda_n^j)^{-1}(A_j)] = 0$. And furthermore,

$$\mu^\infty[\limsup_{n\to\infty}(\lambda_n^i)^{-1}(A_i^{\mathsf{c}})] = 0.$$

Therefore $\mu^\infty\left[\liminf_{n\to\infty}(\lambda_n^i)^{-1}(A_i)\right] = 1$.

□

3.4.3 Application: Causal Graphical Models

Returning to the setting of causal graphical models, suppose that M is a set of measures on a measurable space $(\Omega, \mathcal{B})$. Let $\mathcal{V}$ be a fixed, finite set of random variables $X_1, X_2, \ldots, X_n, \ldots$, taking values in measurable spaces $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n), \ldots$. Assume that each X_i is $(\mathcal{B}, \mathcal{S}_i)$ measurable. Let DAG be the set of all direted acyclic graphs on the fixed variable set $\mathcal{V}$. Let $W \subset M \times \cup_i\mathsf{DAG}$. Then, we have the following:

Theorem 3.4.4. *Suppose that for each X_i, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the following hypotheses are verifutable, and therefore, statistically decidable in the limit:*

1. *The true meaure is Markov and faithful to $\mathcal{G} \in \mathsf{DAG}$;*

2. *The true measure is Markov and faithful to $[\mathcal{G}]$ for $\mathcal{G} \in \mathsf{DAG}$.*

Proof of Theorem 3.4.4. By Theorem 3.2.9, the hypothesis

$$\{\mu \in W : \mu \text{ is Markov to } \mathcal{G}\}$$

is a.s. refutable. By Theorem 3.2.10, the hypothesis

$$\{\mu \in W : \mu \text{ is faithful to } \mathcal{G}\}$$

is a.s. verifiable. Therefore, the hypothesis

$$\{\mu \in W : \mu \text{ is Markov to } \mathcal{G}\} \cap \{\mu \in W : \mu \text{ is faithful to } \mathcal{G}\}$$

is verifutable. By Theorem 3.4.1, it is limiting a.s. decidable in the limit. □

Corollary 3.4.1. *Suppose that the causal Markov and faithfulness assumptions hold, and that for each X_i, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the causal hypothesis:*

$$\{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] = [\mathcal{G}]\}$$

is verifutable, and therefore statistically decidable in the limit.

Proof of Corollary 3.4.1. The hypothesis $\{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] = [\mathcal{G}]\} =$

$$\{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] \preceq [\mathcal{G}]\} \cap \{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] \succeq [\mathcal{G}]\}.$$

By Corollaries 3.2.2 and 3.2.3, it is verifutable, and therefore statistically decidable in the limit. □

3.5 Problems and Solutions

Say that a *statistical problem* is a countable partition $\mathcal{Q}$ of the worlds in W into a set of *answers*. For $\mu \in W$, let $\mathcal{Q}_\mu$ denote the answer true in μ, i.e. $\mathcal{Q}_\mu$ is the unique element of $\mathcal{Q}$ containing μ. Say that a family $(\lambda_n)_{n \in \mathbb{N}}$ of feasible methods is a *solution in chance* to the *question* $\mathcal{Q}$ iff for every $\mu \in W$, $\lim_{n\to\infty} \mu^n[\lambda_n^{-1}(\mathcal{Q}_\mu)] = 1$. Say that $\mathcal{Q}$ is *solvable in chance* iff there exists a solution in chance to $\mathcal{Q}$. A family $(\lambda_n)_{n \in \mathbb{N}}$ of feasible methods is an *almost sure solution* to $\mathcal{Q}$ iff for every $\mu \in W$, $\mu^\infty[\liminf_{n\to\infty} \lambda_n^{-1}(\mathcal{Q}_\mu)] = 1$. Furthermore, say that $\mathcal{Q}$ is *almost surely solvable* iff there exists an almost sure solution to $\mathcal{Q}$.

3.5.1 A Characterization of Solvable Problems

Theorem 3.5.1. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, the following are equivalent:*

1. *$\mathcal{Q}$ is solvable in chance;*

2. *$\mathcal{Q}$ is almost surely solvable;*

3. *The elements of $\mathcal{Q}$ are countable unions of locally closed sets in the weak topology.*[18]

The fact that 2 entails 1 follows directly from the definitions. It is also straightforward to show that 1 entails 3. Suppose that $(\lambda_n)_{n \in \mathbb{N}}$ is a solution in chance to $\mathcal{Q}$, and that A is an answer to $\mathcal{Q}$. Then, $(\lambda_n)_{n \in \mathbb{N}}$ is a limiting verifier in chance of A. Therefore, by Theorem 3.4.1, A is a countable union of locally closed sets in the weak topology. To show that 3 entails 1, we first prove Lemma 3.5.1, which states that every questions consisting of limiting a.s. decidable answers is a.s. solvable. If every answer A to $\mathcal{Q}$ is a countable union of locally closed sets, then, since $\mathcal{Q}$ is countable, A^{c} must also be a countable union of locally closed sets. Therefore, by Theorem 3.4.3, every answer to $\mathcal{Q}$ is limiting a.s. decidable. Invoking Lemma 3.5.1 completes the proof.

Lemma 3.5.1. *If every answer to question $\mathcal{Q}$ is limiting almost surely decidable, then $(W, \mathcal{Q})$ is almost surely solvable.*

Proof of Lemma 3.5.1. Let $A_1, A_2, \ldots$ enumerate the elements of $\mathcal{Q}$. For each $i \in \mathbb{N}$, let (λ_n^i) be an almost sure decision procedure for A_i. Define

$$\sigma_n(\vec{\omega}) = \min\{i : i \leq n \text{ and } \lambda_n^i(\vec{\omega}) = A_i\}.$$

Let

$$\lambda_n(\vec{\omega}) = \begin{cases} A_{\sigma_n(\vec{\omega})}, & \text{if } \sigma(\vec{\omega}) < \infty, \\ W, & \text{otherwise.} \end{cases}$$

Since the verdict of λ_n depends on the verdict of only finitely many $\lambda_n^{i,j}$, it is feasible. Suppose that $\mu \in A_k$. Then:

[18] A similar result is proven in [Dembo and Peres, 1994, Theorem 2] under different conditions. Dembo and Peres do not require their methods to be feasible, so Theorem 3.4.1 does not straightforwardly generalize their result. It is not difficult to reprove Theorem 3.4.1 without that requirement to obtain a generalization of the result in Dembo and Peres [1994]. But since requirement of feasibility has a strong independent motivation, I do not take this to be a shortcoming vis-a-vis the older result.

$$\begin{aligned}
&\mu^\infty\left[\liminf_{n\to\infty}\lambda_n^{-1}(A_k)\right] = \\
&= \mu^\infty\left[\liminf_{n\to\infty}\bigcap_{i<k}\left(\lambda_n^i\right)^{-1}(A_i^{\mathsf{c}})\cap\left(\lambda_n^k\right)^{-1}(A_k)\right] \\
&= 1-\mu^\infty\left[\limsup_{n\to\infty}\bigcup_{i<k}\left(\lambda_n^i\right)^{-1}(A_i)\cup\left(\lambda_n^k\right)^{-1}(A_k^{\mathsf{c}})\right] \\
&= 1-\mu^\infty\left[\bigcup_{i<k}\limsup_{n\to\infty}\left(\lambda_n^i\right)^{-1}(A_i)\cup\limsup_{n\to\infty}\left(\lambda_n^k\right)^{-1}(A_k^{\mathsf{c}})\right] \\
&\geq 1-\sum_{i<k}\mu^\infty\left[\limsup_{n\to\infty}\left(\lambda_n^i\right)^{-1}(A_i)\right]-\mu^\infty\left[\limsup_{n\to\infty}\left(\lambda_n^k\right)^{-1}(A_k^{\mathsf{c}})\right].
\end{aligned}$$

Since each $(\lambda_n^i)_{n\in\mathbb{N}}$ is a limiting a.s. decision procedure for A_i, and $\mu\notin\cup_{i<k}A_i$, it follows that $\mu^\infty\left[\limsup_{n\to\infty}(\lambda_n^i)^{-1}(A_i)\right]=0$ for $i<k$. Therefore,

$$\mu^\infty\left[\liminf_{n\to\infty}\lambda_n^{-1}(A_i)\right]\geq 1-\mu^\infty\left[\limsup_{n\to\infty}\left(\lambda_n^k\right)^{-1}(A_k^{\mathsf{c}})\right].$$

But since $\mu\in A_k$:

$$\mu^\infty\left[\limsup_{n\to\infty}\left(\lambda_n^k\right)^{-1}(A_k^{\mathsf{c}})\right]=0,$$

so $\mu^\infty\left[\liminf_{n\to\infty}\lambda_n^{-1}(A_k)\right]=1$, as required.

□

3.5.2 Solving the Markov Class Problem for Causal Graphs

Spirtes et al. [2000] describe several algorithms for discovering the Markov equivalence class of the true causal graph. These algorithms are provably correct given reliable procedures for making the requisite statistical decisions about conditional indendence. Spirtes et al. [2000] cite appropriate tests for the linear Gaussian case, and the discrete case. One may still wonder, however, whether appropriate procedures exist in general. In this section, we leverage the results of the previous section to show that under a weak condition, there exists a pointwise consistent method for discovering the true Markov equivalence class. These results hold for discrete variables, variables with density functions, and any mixture of the two.

Suppose that M is a set of measures on a measurable space $(\Omega,\mathcal{B})$. Let $\mathcal{V}$ be a fixed, finite set of random variables $X_1,X_2,\ldots,X_n$ taking values in measurable spaces $(S_1,\mathcal{S}_1),\ldots,(S_n,\mathcal{S}_n)$. Assume that each X_i is $(\mathcal{B},\mathcal{S}_i)$ measurable. Let DAG be the set of all direted acyclic graphs on the fixed variable set $\mathcal{V}$. Let $W\subset M\times\mathsf{DAG}$.

By Theorem 3.4.4, the hypotheses

$$\mathsf{MarkovClass} = \{\{\mu \in M : \mu \text{ is Markov and faithful to } [\mathcal{G}]\} : \mathcal{G} \in \mathsf{DAG}\},$$

are disjoint and verifutable. However, these hypotheses do not necessarily exhaust the measures in M, since some measures may fail to be Markov and faithful to any $\mathcal{G} \in \mathsf{DAG}$. However, since $\cup\mathsf{MarkovClass}$, is a finite union of verifutable hypotheses, both $\cup\mathsf{MarkovClass}$ and the catchall hypothesis $\mathsf{Catchall} = W \setminus \cup\mathsf{MarkovClass}$ are limiting a.s. decidable by Lemma 3.4.1. Therefore, letting $\mathcal{Q} = \mathsf{MarkovClass} \cup \mathsf{Catchall}$, we have that each answer to $\mathcal{Q}$ is limiting a.s. decidable and therefore, by Lemma 3.5.1, $(M, \mathcal{Q})$ is a solvable statistical problem. We have proven the following:

Theorem 3.5.2. *Suppose that for each X_i, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, every answer to the statistical problem $(M, \mathcal{Q})$, where $\mathcal{Q} = \mathsf{MarkovClass} \cup \mathsf{Catchall}$, is decidable in the limit, and $(M, \mathcal{Q})$ is solvable in the limit.*

The causal Markov class question is given by:

$$\mathsf{CausalMarkovClass} = \{\{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] = [\mathcal{G}]\} : \mathcal{G} \in \mathsf{DAG}\}.$$

Under the casual Markov and faithfulness assumption, each element of $\mathsf{CausalMarkovClass}$ is decidable in the limit by Corollary 3.4.1. Therefore, by Lemma 3.5.1, it is solvable in the limit. We have proven the following:

Theorem 3.5.3. *Suppose that the causal Markov and faithfulness assumptions hold, and that for each X_i, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, every answer to the causal inference problem, $(W, \mathcal{Q})$, where $\mathcal{Q} = \mathsf{CausalMarkovClass}$, is decidable in the limit, and therefore $(W, \mathcal{Q})$ is solvable in the limit.*

3.6 Progress, Simplicity and Ockham's Razor

3.6.1 Simplicity

Popper [1959] proposed that A is as simple as B, which we abbreviate with $A \preceq B$, iff A is at least as falsifiable as B, i.e. if every verifiable proposition inconsistent with B is also inconsistent A. In the equivalent, contrapositive formulation: $A \preceq B$ iff any verifiable proposition consistent with A is consistent with B. Popper's proposal was widely criticized (see e.g. Fitzpatrick [2013b]) for being inapplicable in statistical settings. The typical objection was that, on his definition, all statistical hypotheses are equally simple since, for example, every real-valued random sample is logically consistent with any generating normal distribution. That difficulty is resolved when one replaces propositional falsification with statistical falsification.

Theorem 3.6.1. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then, for $A, B \subseteq W$, the following are equivalent:*

1. $A \subseteq \mathsf{cl}B$;[19]

2. *Every statistically verifiable proposition inconsistent with B is inconsistent with A;*

3. *Every statistically verifiable proposition consistent with A is consistent with B.*

Proof of Theorem 3.6.1. 2 and 3 are equivalent by contraposition. By Theorem 3.2.1, the statistically verifiable propositions are exactly the open sets in the weak topology. To see that 1 and 3 are equivalent, note that $A \subseteq \mathsf{cl}(B)$ iff every open set compatible with A is compatible with B iff every statistically verifiable proposition compatible with A is compatible with B. □

As we discussed in Section 2.5, Popper's definition has the unfortunate feature of confounding relations of logical entailment with relations of empirical underdetermination. Recall that, since $A \subseteq \mathsf{cl}B$ whenever A entails B, any hypothesis is simpler than its logical consequences, and W, the trivial hypothesis, is maximally complex. That problem is eliminated if we restrict the simplicity relation to competing hypothesis and set $A \preceq B$ iff $A \subseteq \mathsf{frnt}B = \mathsf{cl}B \setminus B$. However, on that definition, simplicity relations can be obscured by disjoining irrelevant possibilities, e.g. if $A \preceq B$, and $\mu \notin \mathsf{frnt}B$, then $A \cup \{\mu\} \not\preceq B$. Those sorts of considerations (see Section 2.5 for more discussion) suggest the following defintion: say that A is as simple as B, written $A \triangleleft B$ iff $A \cap \mathsf{frntr}(B) \neq \varnothing$, which says that there is $\mu \in A$, where B is false, but all statistically verifiable propositions true in μ are consistent with B. Say that A is *simplest* iff there is no B such that $B \triangleleft A$. Then, maximal simplicity has the following correspondence with statistical refutability.

Theorem 3.6.2. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Then for $A \subseteq W$, A is simplest iff A is statistically refutable.*

Proof. By Lemma 2.1.1, A is closed iff $\mathsf{frnt}A = \varnothing$, iff there is no $B \triangleleft A$. By Theorem 3.2.2, A is closed iff A is statistically refutable. □

When A and B are competing hypotheses, the simplicity relation has intuitive consequences for hypothesis testing. Suppose that $A \triangleleft B$. Then there is $\mu \in A \cap \mathsf{frnt}B$. Let $\psi : \Omega \to \{W, A^{\mathsf{c}}\}$ be a feasible test of A. By assumption, there is a sequence (μ_n) lying in B such that $\mu_n[\psi^{-1}(A^{\mathsf{c}})] \to \mu[\psi^{-1}(A^{\mathsf{c}})]$. Therefore, $\inf_{\mu \in B} \mu[\psi^{-1}(A^{\mathsf{c}})] \leq \sup_{\mu \in A} \mu[\psi^{-1}(A^{\mathsf{c}})]$, which is to say that for any test of A, the worst-case power in B does not exceed the significance level.

[19] All topological operators are with respect to the weak topology.

Simplicity in Linear Models

In discussions of linear regression modeling, one often hears that models with "fewer" predictors are simpler than models with more. In this section we show that the simplicity relation defined in the previous section agrees with intuition in a precise sense.

Let $X_1, X_2, \ldots X_k$ be a set of observable random variables. Let ϵ be an unobserved noise term. Let $X_0 = 1$. The set of worlds W is the set of random vectors $(Y, X_1, X_2, \ldots, X_k, \epsilon)$, where

$$Y = \sum_{i=0}^{k} \beta_i X_i + \epsilon.$$

Each world may be expressed as a product of matrices BX where $X = (1, X_1, X_2, \ldots, X_k, \epsilon)^T$, and B is the equal to the identity matrix, except in the first row. For example, if $k = 2$, and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, then:

$$(Y, X_1, X_2, \epsilon) = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \epsilon \end{bmatrix}$$

Of course, some of the β_i may be equal to zero. The problem of finding out exactly which β_i are zero is called the problem of *subset selection.* If $I \subseteq \{0, 1, \ldots, k\}$, let W_I be the set of worlds $(Y, X_1, \ldots, X_k, \epsilon)$ where $Y = \sum_{i \in I} \beta_i X_i + \epsilon$, and each β_i is non-zero.

We show that $W_I \lhd W_J$, if $I \subset J$, i.e. that models with a strict superset of predictors are more complex. Let $BX \in W_I$. The sequence of matrices $\{B_N\}$ is defined by:

$$B_{N\,i,j} = \begin{cases} \frac{1}{N}, & \text{if } i = 1, \text{ and } j \in J \setminus I, \\ B_{i,j}, & \text{otherwise.} \end{cases}$$

By construction, $BX \in W_I$ and each $B_N X \in W_J$. Since the B_N are converging in Euclidean norm to B, it follows, a fortiori, that the B_N are converging to B in probability. By Slutsky's theorem (Theorem 3.1.5), $B_N X \Rightarrow BX$. Therefore $BX \in \mathsf{frnt} W_J$. Since BX was arbitrary, $W_I \subseteq \mathsf{frnt} W_J$, and $W_I \lhd W_J$.

Simplicity in Linear Causal Models

A *linear causal model* is a random vector $Z = (X_1, X_2, \ldots, X_k, e_1, e_2, \ldots, e_k)$ such that the following hold:

1. There is an ordering $k(i)$ of the observable variables X_i such that each variable is a linear function of variables earlier in the order, plus an un-

observed noise term e_i. That is:

$$X_i(\omega) = \sum_{k(j)<k(i)} b_{ij} X_j(\omega) + e_i(\omega).$$

2. The noise terms $e_1, \ldots, e_k$ are mutually independent.

Let LIN be the set of all linear causal models $(X_1, \ldots, X_k, e_1, \ldots, e_k)$ defined on a common space $(S, \mathcal{S})$. Let DAG be the set of all directed acyclic graphs on the set of indices $\{1, \ldots, k\}$. Each linear causal model corresponds to a DAG in a natural way: the directed graph $\mathcal{G}$ of a linear causal model $Z = (X_1, \ldots, X_k, e_1, \ldots, e_k)$ has a directed edge from i to j iff $b_{ij} \neq 0$. If $\mathcal{G}$ is the DAG for a linear model X, then $\mathcal{L}(Z)$ is Markov for $\mathcal{G}$ [Spirtes, 1995, Theorem 1]. Let W be the set of all pairs $(Z, \mathcal{G})$ where $Z \in \mathsf{LIN}, \mathcal{G} \in \mathsf{DAG}$ and $\mathcal{G}$ is the canonical DAG for Z. If $\mathcal{G} \in \mathsf{DAG}$, let $W_{\mathcal{G}}$ be the set $\{(Z, \mathcal{G}') \in W : \mathcal{G}' = \mathcal{G}\}$. We may think of a DAG $\mathcal{G}$ as a set of pairs $\{(i, j) : i \text{ is a direct cause of } j \text{ in } \mathcal{G}\}$. If $\mathcal{G}'$ has a strict subset of the edges of $\mathcal{G}''$, write $\mathcal{G}' \subset \mathcal{G}''$.

It is natural to think that if $\mathcal{G}' \subset \mathcal{G}''$, then $W_{\mathcal{G}'}$ is simpler than $W_{\mathcal{G}''}$. We show that the simplicity relation defined in Section 3.6.1 agrees with this fundamental intuition. Let $(Z, \mathcal{G}') \in W_{\mathcal{G}'}$. For each linear causal model, the set of observable variables $X = (X_1, \ldots, X_k)$ can be expressed as the system of equations:

$$X = BX + e,$$

where e is the vector of noise terms $(e_1, \ldots, e_k)$, and B is the matrix of coefficients b_{ij}. The matrix B can be permuted to a strictly lower triangular matrix (a matrix with zeros on and above the diagonal) if the true causal order is known. Shimizu et al. [2006] observe that, solving for X, one obtains

$$X = (I - B)^{-1} e.$$

Let

$$B_{N\,i,j} = \begin{cases} \frac{1}{N}, & \text{if } (j,i) \in \mathcal{G}'' \setminus \mathcal{G}', \\ B_{i,j}, & \text{otherwise.} \end{cases}$$

Let $X_n = (I - B_N)^{-1} e$ and $Z_n = (X_n, e)$. By construction, $B_{N\,ij} \neq 0$ iff $(j, i) \in \mathcal{G}''$. By a standard result in graph theory, since $\mathcal{G}''$ is a DAG, there is an ordering of its vertices $k(i)$ such that if $(j, i) \in \mathcal{G}''$, $k(j) < k(i)$. Therefore, if $b_{N\,ij} \neq 0$, then $k(j) < k(i)$, as required to satisfy condition 1. Furthermore, since the distribution of the errors in e is unchanged, they are mutually independent, satisfying the second condition. Therefore, the Z_n are in LIN, and $(Z_n, \mathcal{G}'') \in W$.

By construction, the $\{B_N\}$ are converging to B in the Euclidean norm. By the continuous mapping theorem, it follows that $(I - B_N)^{-1} \to (I - B)^{-1}$. By Slutsky's Theorem (Theorem 3.1.5), it follows that $X_n \Rightarrow X$. If we endow the set W with the topology inherited from the weak topology on the observable variables, it follows that $W_{\mathcal{G}'} \subseteq \mathsf{frnt} W_{\mathcal{G}''}$, and therefore $W_{\mathcal{G}'} \lhd W_{\mathcal{G}''}$.

3.6.2 Progressive Solutions

A perennial hope of the philosophy of science is to give reason to believe that science makes progress. In the context of a particular empirical question, that hope could be vindicated by exhibiting methods whose objective chance of producing the true answer increases monotonically with sample size. In light of Lemma 3.3.1, we know that is not typically possible. But if that is infeasible, one could at least show that there exist solutions where the chance of producing the true answer never decreases by more than α, where α is small. Call a method satisfying that latter property *α-progressive*. Theorem 3.6.3 shows that for a wide class of problems, there exist *α-progressive* methods for every $\alpha > 0$.

Say that a solution in chance $\{\lambda_n\}_{n\in\mathbb{N}}$ is *α-progressive* iff for all $n_1 < n_2$,

α-Prog. $\mu^{n_2}[\lambda_{n_2}^{-1}(\mathcal{Q}_\mu)] + \alpha > \mu^{n_1}[\lambda_{n_1}^{-1}(\mathcal{Q}_\mu)]$.

First, we prove a lemma:

Lemma 3.6.1. *Suppose that α and $a_1, \ldots, a_n$ are in $[0,1]$, and that $a_i > \frac{\alpha}{2^i}$. Then*

$$\prod_{i=1}^{n}(a_i - \frac{\alpha}{2^i}) \geq \prod_{i=1}^{n} a_i - \sum_{i=1}^{n} \frac{\alpha}{2^i}.$$

Proof of Lemma 3.6.1. By induction on n. The base case is trivial. For the inductive step, note that:

$$\begin{aligned}
\prod_{i=1}^{n+1}(a_i - \frac{\alpha}{2^i}) &= (a_{n+1} - \frac{\alpha}{2^{n+1}}) \cdot \prod_{i=1}^{n}(a_i - \frac{\alpha}{2^i}) \\
&= a_{n+1} \cdot \prod_{i=1}^{n}(a_i - \frac{\alpha}{2^i}) - \frac{\alpha}{2^{n+1}} \cdot \prod_{i=1}^{n}(a_i - \frac{\alpha}{2^i}) \\
&\geq a_{n+1} \cdot \prod_{i=1}^{n}(a_i - \frac{\alpha}{2^i}) - \frac{\alpha}{2^{n+1}} \\
&\geq a_{n+1} \cdot \left(\prod_{i=1}^{n} a_i - \sum_{i=1}^{n} \frac{\alpha}{2^i}\right) - \frac{\alpha}{2^{n+1}} \\
&= \prod_{i=1}^{n+1} a_i - a_{n+1} \cdot \sum_{i=1}^{n} \frac{\alpha}{2^i} - \frac{\alpha}{2^{n+1}} \\
&\geq \prod_{i=1}^{n+1} a_i - \sum_{i=1}^{n+1} \frac{\alpha}{2^i},
\end{aligned}$$

where we have used the inductive hypothesis to get to the fourth line. □

Theorem 3.6.3. *Suppose (1) that $\mathcal{I}$ is a countable base (2) that W is a set of Borel measures on $(\Omega, \mathcal{I})$, and (3) that $\mathcal{I}$ is almost surely clopen in every $\mu \in W$. Suppose that $\mathcal{Q}$ is a partition of W, and that there exists $A_1, A_2, \ldots,$ an* *enumeration of $\mathcal{Q}$ agreeing with the simplicity relation $\triangleleft$, i.e. an enumeration such that if $j < i$ then $A_i \ntriangleleft A_j$.*[20] *Then there exists an α-progressive solution to $\mathcal{Q}$, for every $\alpha > 0$.*

Proof of Theorem 3.6.3. We show that under the conditions in the antecedent, and for all i, $\cup_{j \leq i} A_j$ is monotonically refutable. The result follows by Lemma 3.6.2. By Theorem 3.3.2, it is sufficient to show that $\cup_{j \leq i} A_j$ is closed in the weak topology on W. Suppose for a contradiction that $\cup_{j \leq i} A_j$ is not closed. Then there is some $\mu \in A_k$ such that $k > i$ and $\mu \in \mathsf{cl}(\cup_{j \leq i} A_j)$. But since the closure of a finite union is the union of the closures, there must be $j < k$ such that $\mu \in \mathsf{cl} A_j$. But then $A_k \cap \mathsf{frnt} A_j \neq \varnothing$. Contradiction. □

Lemma 3.6.2. *Suppose that there exists $A_1, A_2, \ldots,$ an enumeration of $\mathcal{Q}$ such that for all i, $\cup_{j \leq i} A_j$ is monotonically refutable. Then there exists an α-progressive solution to $\mathcal{Q}$, for every $\alpha > 0$.*

Proof of Lemma 3.6.2. Suppose that there exists $A_1, A_2, \ldots,$ an enumeration of $\mathcal{Q}$ such that for all i, $\cup_{j \leq i} A_j$ is monotonically refutable. By Theorem 3.3.2, there exist $(\lambda_n^1), (\lambda_n^2), \ldots,$ mutually independent, $\frac{\alpha}{2^i}$-monotonic refutation methods for $\cup_{j \leq 1} A_j, \cup_{j \leq 2} A_j, \cup_{j \leq 3} A_j \ldots$. Let $\sigma_n(\vec{\omega}) = \min\{i \leq n : \lambda_n^i(\vec{\omega}) = W\}$. Let

$$\lambda_n(\vec{\omega}) = \begin{cases} A_{\sigma_n(\vec{\omega})}, & \text{if } \sigma(\vec{\omega}) < \infty \\ W, & \text{otherwise.} \end{cases}$$

First we show that (λ_n) is a solution in chance. Suppose that $\mu \in A_i$. Then, by assumption, $\mu \notin \cup_{k \leq j} A_k$ for $j < i$. Therefore, since the (λ_n^i) are mutually independent,

$$\begin{aligned} \lim_{n \to \infty} \mu^n[\lambda_n^{-1}(A_i)] &= \lim_{n \to \infty} \mu^n[\cap_{j<i} (\lambda_n^j)^{-1}(\cap_{k \leq j} A_k^{\mathsf{c}}) \cap (\lambda_n^i)^{-1}(W)] \\ &= \lim_{n \to \infty} \mu^n[(\lambda_n^i)^{-1}(W)] \cdot \prod_{j<i} \mu^n[(\lambda_n^j)^{-1}(\cap_{k \leq j} A_k^{\mathsf{c}})] \\ &= 1. \end{aligned}$$

It remains to show that for $n_1 < n_2$,

$$\mu^{n_2}[\lambda_{n_2}^{-1}(A_i)] > \mu^{n_1}[\lambda_{n_1}^{-1}(A_i)] - \alpha.$$

Note that if $\mu^{n_1}[(\lambda_{n_1}^j)^{-1}(\cap_{k \leq j} A_k^{\mathsf{c}})] \leq \frac{\alpha}{2^j}$ for any $j < i$, then $\mu^{n_1}[\lambda_{n_1}^{-1}(A_i)] < \alpha$, and we are done. So suppose that $\mu^{n_1}[(\lambda_{n_1}^j)^{-1}(\cap_{k \leq j} A_k^{\mathsf{c}})] > \frac{\alpha}{2^j}$ for all $j < i$.

[20] A sufficient condition for the existence of such an enumaration is that the question is a *stratification.*

Since the (λ^i_n) are mutually independent, and $\alpha/2^i$-monotonic:

$$\begin{aligned}
\mu^{n_2}[\lambda_{n_2}^{-1}(A_i)] &= \mu^{n_2}\left[\cap_{j<i}(\lambda^j_{n_2})^{-1}(\cap_{k\leq j}A_k^{\mathsf{c}}) \cap (\lambda^i_{n_2})^{-1}(W)\right] \\
&= \mu^{n_2}[(\lambda^i_{n_2})^{-1}(W)] \cdot \prod_{j<i} \mu^{n_2}[(\lambda^j_{n_2})^{-1}(\cap_{k\leq j}A_k^{\mathsf{c}})] \\
&\geq \mu^{n_2}[(\lambda^i_{n_2})^{-1}(W)] \cdot \prod_{j<i} \left(\mu^{n_1}[(\lambda^j_{n_1})^{-1}(\cap_{k\leq j}A_k^{\mathsf{c}})] - \frac{\alpha}{2^j}\right)
\end{aligned}$$

Since $\mu \in A_i$,

$$\mu^{n_2}[(\lambda^i_{n_2})^{-1}(W) \geq 1 - \frac{\alpha}{2^i} \geq \mu^{n_1}[(\lambda^i_{n_1})^{-1}(W)] - \frac{\alpha}{2^i}.$$

Therefore,

$$\mu^{n_2}[\lambda_{n_2}^{-1}(A_i)] \geq \left(\mu^{n_1}[(\lambda^i_{n_1})^{-1}(W)] - \frac{\alpha}{2^i}\right) \cdot \prod_{j<i} \left(\mu^{n_1}[(\lambda^j_{n_1})^{-1}(\cap_{k\leq j}A_k^{\mathsf{c}})] - \frac{\alpha}{2^j}\right)$$

By Lemma 3.6.1:

$$\begin{aligned}
\mu^{n_2}[\lambda_{n_2}^{-1}(A_i)] &\geq \mu^{n_1}[(\lambda^i_{n_1})^{-1}(W)] \cdot \prod_{j<i} \mu^{n_1}[(\lambda^j_{n_1})^{-1}(\cap_{k\leq j}A_k^{\mathsf{c}})] - \sum_{j=1}^{i} \frac{\alpha}{2^j} \\
&= \mu^{n_1}[\lambda_{n_1}^{-1}(A_i)] - \sum_{j=1}^{i} \frac{\alpha}{2^j} \\
&\geq \mu^{n_1}[\lambda_{n_1}^{-1}(A_i)] - \alpha.
\end{aligned}$$

□

3.6.3 Progress and Ockham's Razor

In this section we demonstrate that every α-progressive method satisfies a probabilistic version of Ockham's razor. If progressiveness sounds like a weak property, demonstrating that Ockham's razor is a necessary condition of that weak condition is a very *strong* justification — if it is necessary for that very weak success notion, it is also necessary for every stronger one.

Say that a statistical method $\{\lambda_n\}_{n\in\mathbb{N}}$ is *α-Ockham* iff the chance that it conjectures an answer more complex than the truth is bounded by α, i.e.:

α-Ockham. If $A \in \mathcal{Q}$ and $\mu \in \mathsf{frnt}A$, then $\mu^n[\lambda_n^{-1}(A)] \leq \alpha$.

Theorem 3.6.4. *Every α-progressive solution is α-Ockham.*

Proof. Suppose that $\mu \in \mathsf{frnt}A$, but $\mu^{n_1}[\lambda_{n_1}^{-1}(A)] \geq \alpha + \epsilon > \alpha$. Since $\{\lambda_n\}_{n\in\mathbb{N}}$ is a solution, there is n_2 such that $\mu^{n_2}[\lambda_{n_2}^{-1}(\mathcal{Q}_\mu)] > 1 - \epsilon$. Since $\{\lambda_n\}_{n\in\mathbb{N}}$ is feasible,

$$E := \{\nu : \nu^{n_1}[\lambda_{n_1}^{-1}(A)] > \alpha + \epsilon\} \cap \{\nu : \nu^{n_2}[\lambda_{n_2}^{-1}(\mathcal{Q}_\mu)] > 1 - \epsilon\}$$

is open in the weak topology. Furthermore, $\mu \in E$. Since $\mu \in \mathsf{frnt}A$, there is $\nu \in E \cap A$. Therefore $\nu_{n_1}[\lambda_{n_1}^{-1}(\mathcal{Q}_\nu)] > \alpha + \epsilon$ and $\nu_{n_2}[\lambda_{n_2}^{-1}(\mathcal{Q}_\nu)] < \epsilon$. It follows that $\nu_{n_1}[\lambda_{n_1}^{-1}(\mathcal{Q}_\nu)] - \nu_{n_2}[\lambda_{n_2}^{-1}(\mathcal{Q}_\nu)] > \alpha$. □

3.6.4 A Progressive Solution to the Markov Class Problem

Constraint-based methods for solving the causal Markov class problem typically start out by conjecturing sparse graphs with few causal relations, and are driven to introduce new causal relationships only when the relevant conditional independencies (and thereby d-separations) are statistically refuted. Algorithms such as SGS and PC often proceed by nesting sequences of conditional independence tests. Although several such methods are known to converge to the true Markov equivalence class in the limit of infinite data, there are infinitely many others strategies that would have the same limiting performance, but make drastically different decisions on finite samples. Some of these alternative methods may even reverse the usual preference for sparse graphs for arbitrarily many sample sizes. What could justify these seemingly reasonable procedures? Spirtes et al. [2000] note that none of the usual comforts of hypothesis testing are present:

> Most of the algorithms we have described require statistical decisions which, as we have just noted, can be implemented in the form of hypothesis tests. But the parameters of the tests cannot be given their ordinary significance. The usual comforts of a statistical test are the significance level ... and the power against an alternative .... Except in very large samples, neither the significance level nor the power of tests used within the search algorithms to decide statistical dependence measures the long run frequency of anything interesting about the search. What does?

Spirtes et al. [2000] proceed to list several error probabilities that one might want to know, but despair of obtaining any analytical answers to these questions. In this section, I suggest that it *is* possible to give an interpretation to the error probabilities of the nested hypothesis tests, although it is not the usual one. The results of the previous section show that if one carefully manages the error probabilities of nested and, crucially, *monotonic*, hypothesis tests, one can ensure that the search procedure is α-progressive, for any $\alpha > 0$. Moreover, any progressive solution must, by Theorem 3.6.4, obey a probabilistic Ockham's razor. In this section we demonstrate how to apply these results to the setting of causal search.

Suppose that M is a set of measures on a measurable space $(\Omega, \mathcal{B})$. Let $\mathcal{V}$ be a fixed, finite set of random variables $X_1, X_2, \ldots, X_n$ taking values in measurable spaces $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n)$. Assume that each X_i is $(\mathcal{B}, \mathcal{S}_i)$ measurable. Let DAG be the set of all direted acyclic graphs on the fixed variable set $\mathcal{V}$. Let $W \subset M \times \mathsf{DAG}$. Recall that the causal Markov class problem is $(W, \mathsf{CausalMarkovClass})$, where $\mathsf{CausalMarkovClass} = \{\{(\mu, \mathcal{G}') \in W : [\mathcal{G}'] = [\mathcal{G}]\} : \mathcal{G} \in \mathsf{DAG}\}$.

Theorem 3.6.5. *Suppose that the causal Markov and faithfulness assumptions hold, and that for $X_i \in \mathcal{V}$, $\sigma(X_i)$ is generated by a countable, almost surely clopen basis $\mathcal{I}(X_i)$. Then, the causal inference problem, $(W, \mathsf{CausalMarkovClass})$ has an α-progressive solution, for every $\alpha > 0$.*

Proof of Theorem 3.6.5. By the order-extension principle, there is a total order $\preceq^*$ compatible with the I-map partial order on causal graphs. Let $[\mathcal{G}_1], [\mathcal{G}_2], \ldots$ be an enumeration of the hypotheses in $\mathsf{CausalMarkovClass}$ such that $i \leq j$ iff $[\mathcal{G}_i] \preceq^* [\mathcal{G}_j]$. To show that $(W, \mathsf{MarkovClass})$ has an α-progressive solution for every $\alpha > 0$, it suffices, by Lemma 3.6.2, to show that $\bigcup_{j \leq i}\{(\mu, \mathcal{G}) : [\mathcal{G}] = [\mathcal{G}_j]\}$ is monotonically refutable. We argue that

$$\bigcup_{j \leq i}\{(\mu, \mathcal{G}) : [\mathcal{G}] = [\mathcal{G}_j]\} = \bigcup_{j \leq i}\{(\mu, \mathcal{G}) : [\mathcal{G}] \preceq [\mathcal{G}_j]\}.$$

By Corollary 3.3.1, the rhs is a finite union of monotonically refutable hypotheses. By Lemma 3.3.5, the rhs is monotonically refutable. Suppose that $(\mu, \mathcal{G})$ is an element of the lhs. Then there is $j \leq i$ such that $[\mathcal{G}] = [\mathcal{G}_j]$. A fortiori, $[\mathcal{G}] \preceq [\mathcal{G}_j]$. Therefore, $(\mu, \mathcal{G})$ is an element of the rhs. To show that the rhs is contained in the lhs, suppose that $(\mu, \mathcal{G})$ is not an element of the lhs. Then, $[\mathcal{G}] \not\preceq^* [\mathcal{G}_i]$. Suppose for a contradiction that $(\mu, \mathcal{G})$ is an element of the rhs. Then $[\mathcal{G}] \preceq [\mathcal{G}_j]$ for some $j \leq i$. But then $[\mathcal{G}] \preceq^* [\mathcal{G}_j] \preceq^* [\mathcal{G}_i]$. By transitivity, $[\mathcal{G}] \preceq^* [\mathcal{G}_i]$. Contradiction. □

3.7 Related Work

Chapter 2 recapitulates foundational results in topological learning theory. Results stated in that section appear previously in de Brecht and Yamamoto [2009], Genin and Kelly [2015], Kelly et al. [2016], Genin and Kelly [2018], and Baltag et al. [2016]. All novel results are contained in Chapter 3, although some have already been published in Genin and Kelly [2017]. In statistical terminology, Theorem 3.2.1 provides necessary and sufficient conditions for the existence of a Chernoff consistent test. Although there is extensive statistical work on pointwise consistent hypothesis testing, I am unaware of any topological result analogous to Theorem 3.2.1. The closest work I am aware of is Ryabko [2011], where a topological characterization is given for consistent hypothesis testing of ergodic processes with samples from a discrete, finite alphabet. That result is incomparable with my own, because, although my work is done in the i.i.d setting, I allow samples to take values in an arbitrary, second-countable space. Furthermore, the topology employed in Ryabko [2011] is not the weak topology, but the topology of distributional distance. The existence of uniformly consistent tests is investigated topologically in Ermakov [2013], where some sufficient conditions are given. Limiting statistical solvability, or *discernability*, as it is known in the literature, has been investigated topologically in Dembo and Peres [1994] and Kulkarni and Zeitouni [1995]. The results of Dembo and Peres [1994] are generalized to ergodic processes in Nobel [2006]. Although the setting is slightly different, Theorem 3.5.1 gives a simpler back-and-forth condition than the one given in Dembo and Peres [1994] and is arrived at more systematically, by building on the fundamental Theorem 3.2.1. The weak topology is used in Dembo and Peres [1994], but Theorem 3.2.1 shows that the weak topology is the *unique* topology for which the open sets are exactly the statistically verifiable propositions. My result shows, therefore, that the weak topology is more than just a convenient technical device. The results on conditional independence testing are, so far as I know, also new.[21] Theorem 3.2.7 is a theoretical improvement on previous non-parametric results given by Gretton and Györfi [2010], Györfi and Walk [2012] which, while guaranteeing the existence of tests that are consistent in the limit of infinite data, do not guarantee finite-sample bounds on the chance of error. Theorem 3.5.3 is anticipated in Spirtes et al. [2000], although it is something of an improvement because it does not rely on the existence of "plug in" tests for conditional independence. So far as I know, all results pertaining to monotonic tests and progressive methods, although inspired by Chernick and Liu [2002], are without precedent in the existing literature. I have in mind especially Theorems 3.3.1, 3.6.3, 3.6.4 and 3.6.5, which I take to be substantive.

[21] The simplicity of the methods employed suggests to me that I must be missing something.

Bibliography

Samson Abramsky. *Domain Theory and the Logic of Observable Properties.* PhD thesis, University of London, 1987.

Francis Bacon. *Novum Organum Scientarium.* Leiden: Wyngaerden and Moiardum, 1645. Found in *A Critical Edition of the Major Works,* edited by Brian Vickers. Oxford University Press, 2000.

Alexandru Baltag, Nina Gierasimczuk, and Sonja Smets. On the solvability of inductive problems: A study in epistemic topology. *Electronic Proceedings in Theoretical Computer Science*, 215:81–98, 2016. doi: 10.4204/EPTCS.215.7.

Yehoshua Bar-Hillel and Rudolf Carnap. Semantic information. *The British Journal for the Philosophy of Science*, 4(14):147–157, 1953.

Isaiah Berlin. Hume and the sources of German anti-rationalism. In Henry Hardy, editor, *Against the Current: Essays in the History of Ideas*, pages 162–187. The Viking Press, 1980.

Patrick Billingsley. *Probability and Measure.* John Wiley & Sons, second edition, 1986.

Patrick Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, second edition, 1999. doi: 10.1002/9780470316962.

Lorenzo Carlucci and John Case. On the necessity of U-Shaped learning. *Topics in Cognitive Science*, 5(1):56–88, 2013. doi: 10.1111/tops.12002.

Lorenzo Carlucci, John Case, Sanjay Jain, and Frank Stephan. Non U-shaped vacillatory and team learning. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 241–255. Springer Berlin Heidelberg, 2005. doi: 10.1007/11564089_20.

Rudolf Carnap. On inductive logic. *Philosophy of Science*, 12(2):72–97, 1945.

Michael R. Chernick and Christine Y. Liu. The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. *The American Statistician*, 56(2):149–155, 2002. doi: 10.1198/000313002317572835.

A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41:1–31, 1979. doi: 10.2307/2984718.

Matthew de Brecht and Akihiro Yamamoto. Interpreting learners as realizers for Σ^0_2-measurable functions. *Special Interests Group on Fundamental Problems in Artificial Intelligence (SIG-FPAI)*, 74:39–44, 2009. URL `http://www.iip.ist.i.kyoto-u.ac.jp/member/matthew/Sigma2_full_version.pdf`.

Amir Dembo and Yuval Peres. A topological criterion for hypothesis testing. *The Annals of Statistics*, 22(1):106–117, 1994. doi: 10.1214/aos/1176325360.

René Descartes. *Discours de la methode pour bien conduire sa raison, et chercher la verité dans les sciences: plus la dioptrique, les meteores, et la geometrie, qui sont des essais de cete method*. Leiden: Jan Maire, 1637. English translation Clarke, Desmond M. *Discourse on Method and Related Writings*, London: Penguin, 1999.

Bernd Droge. Minimax regret analysis of orthogonal series regression estimation: selection versus shrinkage. *Biometrika*, 85(3):631–643, 1998. doi: 10.1093/biomet/85.3.631.

Pierre Duhem. *La théorie physique son objet et sa structure*. Paris: Chevalier et Rivière, second edition, 1914. English translation Phillip Wiener, *The Aim and Structure of Physical Theory*, Princeton University Press, 1954.

Mikhail Ermakov. On distinguishability of hypotheses. *arXiv preprint arXiv:1308.4295*, 2013.

Simon Fitzpatrick. Kelly on Ockham's razor and truth-finding efficiency. *Philosophy of Science*, 80(2):298–309, 2013a. doi: 10.1086/670298.

Simon Fitzpatrick. Simplicity in the philosophy of science. In J. Fiser, editor, *The Internet Encyclopedia of Philosophy*. 2013b. URL `http://www.iep.utm.edu/simplici/`.

Luciano Floridi. Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2):351–370, 2005. doi: 10.1111/j.1933-1592.2005.tb00531.x.

Luciano Floridi. *The Philosophy of Information*. Oxford University Press, 2011.

Malcolm R. Forster. The new science of simplicity. In Arnold Zellner, Hugo A. Keuzenkamp, and Michael McAleer, editors, *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*, pages 83–117. Cambridge, 2002.

Malcolm R. Forster and Elliott Sober. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994. doi: 10.1093/bjps/45.1.1.

Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994. doi: 10.1214/aos/1176325766.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013. doi: 10.1111/j.2044-8317.2011.02037.x.

Konstantin Genin and Kevin T. Kelly. Theory choice, theory change, and inductive truth-conduciveness. In Ramaswamy Ramanujam, editor, *Proceedings Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 111–121, 2015. URL `https://www.imsc.res.in/tark/TARK2015-proceedings.pdf`.

Konstantin Genin and Kevin T. Kelly. The topology of statistical verifiability. In Jérôme Lang, editor, *Proceedings Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 236–250, 2017. doi: 10.4204/EPTCS.251.17.

Konstantin Genin and Kevin T. Kelly. Learning, theory choice, and belief revision. *Studia Logica*, 2018. doi: 10.1007/s11225-018-9809-5.

David Gilat. Monotonicity of a power function: An elementary probabilistic proof. *The American Statistician*, 31(2):91–93, 1977. doi: 10.2307/2683050.

Clark Glymour. *Theory and Evidence.* Princeton University Press, 1980.

Arthur Gretton and László Györfi. Consistent nonparametric tests of independence. *The Journal of Machine Learning Research*, 11:1391–1423, 2010. URL `http://www.jmlr.org/papers/v11/gretton10a.html`.

László Györfi and Harro Walk. Strongly consistent nonparametric tests of conditional independence. *Statistics & Probability Letters*, 82(6):1145–1150, 2012. doi: 10.1016/j.spl.2012.02.023.

Ian Hacking. *The Logic of Statistical Inference.* Cambridge University Press, 1965.

Carl G. Hempel. Studies in the logic of confirmation (ii.). *Mind*, 54(214):97–121, 1945. URL `https://www.jstor.org/stable/2250948`.

David Hume. *A Treatise of Human Nature*. London: John Noon, 1739. Found in *A Treatise of Human Nature,* edited by David Fate Norton and Mary J. Norton, Oxford University Press, 2000.

Edward L. Ionides, Alexander Giessing, Yaacov Ritov, and Scott E. Page. Response to the ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 71(1):88–89, 2017. doi: 10.1080/00031305.2016.1234977.

K. T. Kelly. Simplicity, truth and probability. In P. S. Bandyopadhyay and M. Forster, editors, *Handbook of the Philosophy of Science Volume 7: Philosophy of Statistics*, pages 983–1027. North Holland, 2011.

Kevin T. Kelly. *The Logic of Reliable Inquiry.* Oxford University Press, 1996.

Kevin T. Kelly. Justification as truth-finding efficiency: how Ockham's razor works. *Minds and Machines*, 14(4):485–505, 2004. doi: 10.1023/B:MIND.0000045993.31233.63.

Kevin T. Kelly. A new solution to the puzzle of simplicity. *Philosophy of Science*, 74(5):561–573, 2007. doi: 10.1086/525604.

Kevin T. Kelly and Clark Glymour. Convergence to the truth and nothing but the truth. *Philosophy of Science*, 56(2):185–220, 1989. doi: 10.1086/289483.

Kevin T. Kelly and Clark Glymour. Why probability does not capture the logic of scientific justification. In Chris Hitchcock, editor, *Contemporary Debates in the Philosophy of Science*, pages 94–114. Blackwell, 2004.

Kevin T. Kelly and Conor Mayo-Wilson. Causal conclusions that flip repeatedly and their justification. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 277–285, 2010. URL `https://event.cwi.nl/uai2010/papers/UAI2010_0164.pdf`.

Kevin T. Kelly, Konstantin Genin, and Hanti Lin. Realism, rhetoric, and reliability. *Synthese*, 193(4):1191–1223, 2016. doi: 10.1007/s11229-015-0993-9.

Thomas S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, second edition, 1962.

Sanjeev R. Kulkarni and Ofer Zeitouni. A general classification rule for probability measures. *The Annals of Statistics*, 23(4):1393–1407, 1995. doi: 10.1214/aos/1176324714.

Watson Ladd. Some applications of martingales to probability theory. Technical report, University of Chicago, 2011. URL `http://www.math.uchicago.edu/ may/VIGRE/VIGRE2011/REUPapers/Ladd.pdf`.

Imre Lakatos. Falsification and the methodology of scientific research programmes. In Imre Lakatos and Alan Musgrave, editors, *Proceedings of the International Colloquium in the Philosophy of Science: Criticism and the Growth of Knowledge*, pages 91–196. Cambridge, 1970.

Imre Lakatos. The role of crucial experiments in science. *Studies in History and Philosophy of Science Part A*, 4(4):309–325, 1974. doi: 10.1016/0039-3681(74)90007-7.

Larry Laudan. *Progress and its Problems: Towards a Theory of Scientific Growth.* University of California Press, 1978.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005. doi: doi.org/10.1017/S0266466605050036.

Isaac Levi. *Gambling with Truth.* New York: Afred A. Knopf, 1967.

Hanti Lin and Jiji Zhang. How to tackle an extremely hard learning problem: Learning causal structures from non-experimental data without the faithfulness assumption or the like. *arXiv preprint arXiv:1802.07051*, 2018.

Deborah G. Mayo and David R. Cox. Frequentist statistics as a theory of inductive inference. In Javier Rojo, editor, *Optimality: The Second Erich L. Lehmann Symposium*, pages 77–97. Institute of Mathematical Statistics, 2006. doi: 10.1214/074921706000000400.

David Miller. Popper's qualitative theory of verisimilitude. *The British Journal for the Philosophy of Science*, 25(2):166–177, 1974. doi: 10.1093/bjps/25.2.166.

Walter M. Miller Jr. *A Canticle for Leibowitz.* New York: Bantam, 1959.

Patrick Musonda. *The self-controlled case series method: performance and design in studies of vaccine safety.* PhD thesis, Open University, 2006. URL `http://statistics.open.ac.uk/sccs/thesis_patrick_musonda.pdf`.

Ilkka Niiniluoto. *Critical Scientific Realism.* Oxford University Press, 1999.

Ilkka Niiniluoto. Scientific progress. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, summer 2015 edition, 2015. URL `https://plato.stanford.edu/archives/sum2015/entries/scientific-progress/`.

Illka Niiniluoto. *Truthlikeness*, volume 185 of *The Synthese library.* D. Reidel, 1987.

Shailesh Niranjan and James F. Frenzel. A comparison of fault-tolerant state machine architectures for space-borne electronics. *IEEE Transactions on Reliability*, 45(1):109–113, 1996. doi: 10.1109/24.488925.

Andrew B. Nobel. Hypothesis testing for families of ergodic processes. *Bernoulli*, 12(2):251–269, 2006. doi: 10.3150/bj/1145993974.

Graham Oddie. *Likeness to Truth*, volume 30 of *The University of Western Ontario series in philosoph of science.* D. Reidel, 1986.

Kalyanapuram Rangachari Parthasarathy. *Probability measures on metric spaces*, volume 3 of *Probability and Mathematical Statistics.* New York: The Academic Press, 1967.

Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.

Judea Pearl and Thomas S Verma. A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics*, 134:789–811, 1995. doi: 10.1016/S0049-237X(06)80074-1.

Karl R. Popper. *The Logic of Scientific Discovery.* London: Hutchinson, first english edition, 1959.

Karl R Popper. *Objective Knowledge: An Evolutionary Approach.* Oxford University Press, revised edition, 1979.

Hilary Putnam. Trial and error predicates and the solution to a problem of Mostowski. *Journal of Symbolic logic*, 30(1):49–57, 1965. doi: 10.2307/2270581.

Hans Reichenbach. *Wahrscheinlichkeitslehre: eine Untersuchung über die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung.* AW Sijthoff, 1935.

Hans Reichenbach. *The theory of probability, an inquiry into the logical and mathematical foundations of the calculus of probability.* University of California Press, second edition, 1949. English translation of Reichenbach [1935].

Daniil Ryabko. *Learnability in Problems of Sequential Inference.* PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2011. URL `https://hal.inria.fr/EC-LILLE/tel-00675680v2`.

Sadahiro Saeki. A proof of the existence of infinite product probability measures. *The American Mathematical Monthly*, 103(8):682–683, 1996. doi: 10.1080/00029890.1996.12004804.

Leonard J Savage. *The Foundations of Statistics.* New York: Dover, second revised edition, 1972.

Frank Schaarschmidt. Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. *Communications in Biometry and Crop Science*, 2(1):32–40, 2007. URL `http://agrobiol.sggw.waw.pl/ cbcs/articles/CBCS_2_1_5.pdf`.

Paul Schuette, C. George Rochester, and Matthew Jackson. Power and sample size for safety registries: new methods using confidence intervals and saw-tooth power curves. In *8th International R User Conference*, 2012. URL `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/UseR-2012/54-Schuette.pdf`.

Oliver Schulte. Means-ends epistemology. *The British Journal for the Philosophy of Science*, 50(1):1–31, 1999. doi: 10.1093/bjps/50.1.1.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006. URL `http://www.jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf`.

Elliott Sober. *Ockham's Razors.* Cambridge University Press, 2015.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In Philippe Besnard and Steve Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 491–498. San Mateo: Morgan Kaufmann Publishers, 1995. URL `http://arxiv.org/abs/1302.4982`.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT press, second edition, 2000.

Daniel Steel. What if the principle of induction is normative? Formal learning theory and Hume's problem. *International Studies in the Philosophy of Science*, 24(2):171–185, 2010. doi: 10.1080/02698595.2010.484544.

Pavel Tichý. On Popper's definitions of verisimilitude. *The British Journal for the Philosophy of Science*, 25(2):155–160, 1974. doi: 10.1093/bjps/25.2.155].

Thomas Verma and Judea Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In D. Dubois, M. Wellman, B. D'Ambrosio, and P. Smets, editors, *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, pages 323–330. San Mateo: Kaufmann Publishers, 1992. URL `http://arxiv.org/abs/1303.5435`.

Steven Vickers. *Topology Via Logic.* Cambridge University Press, 1996.

Alfred North Whitehead. *Science and the Modern World: Lowell Lectures, 1925.* The New American Library of World Literature, first Pelican Mentor Books edition, 1948.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Fabio Cozman and Avi Pfeffer, editors, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artifical Intelligence*, pages 804–814, 2011. URL `http://arxiv.org/abs/1202.3775`.