# Is the Most Likely Model likely to be the *Correct* Model?



by
Beracah Yankama
B.S. Electrical Engineering
University of Michigan, 2000

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2009





Author .....
Department of Electrical Engineering and Computer Science
September 4, 2009

Certified by .....
Robert C. Berwick
Professor of Computer Science and Engineering/Comp. Linguistics
Thesis Supervisor

Certified by .....
Whitman A. Richards
Professor of Brain and Cognitive Science/Media Arts and Science
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Professor of Electrical Engineering
Chairman, Committee for Graduate Students

# Is the Most Likely Model likely to be the *Correct* Model?

by

Beracah Yankama

Submitted to the Department of Electrical Engineering and Computer Science
on September 4, 2009, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

## Abstract

In this work, I test the hypothesis that the 2-dimensional dependencies of a deterministic model can be correctly recovered via hypothesis-enumeration and Bayesian selection for a linear sequence, and what the degree of 'ignorance' or 'uncertainty' is that Bayesian selection can tolerate concerning the properties of the model and data. The experiment tests the data created by a number of rules of size 3 and compares the implied dependency map to the (correct) dependencies of the various generating rules, then extends it to a composition of 2 rules of total size 5. I found that 'causal' belief networks do not map directly to the dependencies of actual causal structures. For deterministic rules satisfying the condition of multiple involvement (two tails), the correct model is not likely to be retrieved without augmenting the model selection with a prior high enough to suggest that the desired dependency model is already known – simply restricting the class of models to trees, and placing other restrictions (such as ordering) is not sufficient. Second, the identified-model to correct-model map is not 1 to 1 – in the rule cases where the correct model is identified, the identified model could just as easily have been produced by a different rule. Third, I discovered that uncertainty concerning identification of observations directly resulted in the loss of existing information and made model selection the product of pure chance (such as the last observation). How to read and identify observations had to be agreed upon a-priori by both the rule and the learner to have any consistency in model identification. Finally, I discovered that it is not the rule-observations that discriminate between models, but rather the noise, or uncaptured observations that govern the identified model.

In analysis, I found that in enumeration of hypotheses (as dependency graphs) the differentiating space is very small. With representations of conditional independence, the equivalent factorizations of the graphs make the differentiating space even smaller. As Bayesian model identification relies on convergence to the differentiating space, if those spaces are diminishing in size (if the model size is allowed to grow) relative to the observation sequence, then maximizing the likelihood of a particular hypothesis may fail to converge on the correct one. Overall I found that if a learning mechanism

either does not know how to read observations or know the dependencies he is looking for a-priori, then it is not likely to identify them probabilistically.

Finally, I also confirmed existing results – that model selection always prefers increasingly connected models over independent models was confirmed, as was the knowledge that several conditional-independence graphs have equivalent factorizations. Finally Shannon's Asymptotic Equipartition Property was confirmed to apply both for novel observations and for an increasing model/parameter space size.

These results are applicable to a number of domains: natural language processing and language induction by statistical means, bioinformatics and statistical identification and merging of ontologies, and induction of real-world causal dependencies.

Thesis Supervisor: Robert C. Berwick
Title: Professor

Thesis Supervisor: Whitman A. Richards
Title: Professor

# Acknowledgments

First and foremost, I would like to thank my family: my brother Timothy, and my parents Dr(s). Andrew and Rachel Yankama. Without their loving support and unwillingness to accept anything less than a reach for the moon, nothing I do would be possible. Also, I will probably never know how they did it, but somehow they instilled a policy not to admit conceptual or logical inconsistency into my own knowledge base, which has granted a somewhat unique way of looking at the world from which I will always reap rewards and I will always be grateful.

I would like to substantially thank my advisor, Professor Bob Berwick – without his belief in me, careful advice, and opportunities be involved in the MIT community, I would not even be here right now. The substantial intellectual freedom that he allowed me also provided me time to seek conceptual consistency and reach my own conclusions free from interference or pressure.

I would also like to thank Professor Whitman Richards for his substantial corrective input into this work – while no mistakes that I make can be attributed to him, the nearly infinite amount of mistakes that I *did not* make certainly can be! His guidance, clarity, motivation and confidence in me has helped me to achieve a confidence in myself (not to mention completion of this work) that I did not posess, and surely would not have otherwise.

I would like to thank Professor C. Forbes Dewey for his continuous prompting to 'get done', produce something, and not procrastinate. His efforts to involve me with ontologies and Bioinformatics really helped me to understand how the problems that I am interested in penetrate other fields. And finally, Professor Sanjoy Mitter, from whom references and pointers *always* turned out to be strikingly relevant, and whos friendly-yet-unyielding criticism forced large increases in the complexity of my knowledge base. Overall, the sharpness and unwillingness to accept or perpetuate incorrect knowledge of all of my committee members has impressed and taught me enormously.

I would also like to thank the now Professor, Dr. Navid Sabbaghi for his constant

reasurance and always telling me, "you can do it man!' at a time that I really wanted to give up, and reminding me to 'just widen the bandwidth!', and all of my friends in the D-D-D-D712, in addition to a number of special individuals whom I know would prefer not to be named.

If I've forgotten anyone, I'd like to apologize and blame directly the long, sleepless, overcaffeinated nights that have affected my memory. The effect of the large number of people believing in and supporting me has been so necessary and helpful that I cannot begin to take personal credit without being thankful.

I would also like to thank the Dean Lerman and the Office of the Dean for Graduate Education for accepting my petition.

WITH GOD ALL THINGS ARE POSSIBLE.

# Contents

# Chapter 1

# Introduction

This thesis experimentally tests the idea that the causal dependencies of a deterministic model, rule, or ontology (in the form of a 2-d graph) can be recovered via probabilistic means and assumptions from a 1-d linearized stream. It differs from existing work in that the deterministic rule is known beforehand – there is no uncertainty as to what entails 'correct', and rules are tested progressively to identify and understand where and why the probabilistic model selected from an enumeration of models deviates from the correct model. It attempts to demonstrate what happens to overall correct model induction as complexity rises if probabilities are used to represent uncertainty and justify relaxed correctness.

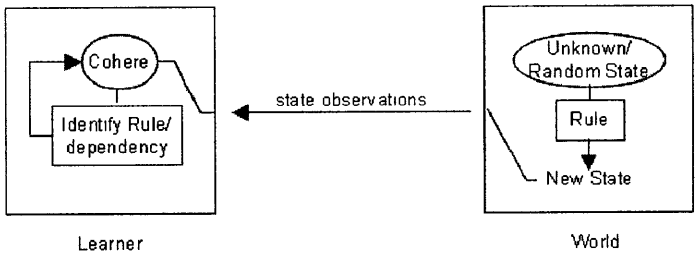



Figure 1-1: A world in unknown, potentially random states, which are transformed by deterministic rules into different states. State observations, consistent with the rule are transmitted from the World to a Learner (or resolver), who attempts to identify dependencies which cohere in some sense to the observations received and to the world's transformation rules. The observation can also be made that the World can be replaced by another agent.

It has become increasingly the case that the idea of inducting a real-world model

(for instance, a language model, an ontology, or a causal model) is equatable with inferring a "probabilistic" model. For many, the probabilistic handling of uncertainty is well-defined and can be relied upon to deliver stable results. But it is of interest whether adopting such a method leads us to identify the *correct* model or if the probabilistic model is increasingly likely to be incorrect.

From the start of learning to work with probabilities, most children find them to be immediately distasteful. It quickly becomes clear to them that even when distributions are properly selected, they are not likely to correctly predict a new observation. It takes many a while to master the notion that what is actually being "predicted" is the *distribution* of possible observations. And even after that is learned, it takes a substantial amount of study and subtlety to understand that the predicted distribution is only from those observations that have been *seen before*. In other words, the distribution does not a-priori account for any kind of exception, or "novel" observation – methods of adding a novel observation can require recomputation of the entire history of prior observations (if the novel observation forces an increase parameter size). Clearly that is impossible for humans, computationally intractable for computers, not productively useful for modeling the real world, and of philosophical question whether enough events even exist in the real-world to enumerate all the possible observation-combinations needed to prevent spareseness in the distributions.

In many fields, probabilities find justification under the wide umbrella of 'representing uncertainty'. But how far does probabilistic uncertainty or 'belief' go, and to what extent is there uncertainty in nature? A little introspection suggests that uncertainty is a human invention that describes the discrepancies between our internal predictive model and the actual of transformational rules of the real-world. From that frame, 'uncertainty' may be more about incompleteness of predictive model and a lack of observability, than it is about a collapsed, monotonic 'belief' or 'frequency' function. In other words, probabilities: frequentist, belief, or otherwise, are used to represent the lack of knowledge, or rather, a lack of consistency of evidence around these two pieces of information.

Over time, this problem has seemingly been overcome by a number of intellectual

contributions that can be combined in the following (trivially illustrated) argument. Laplace demonstrated that Bayes Rule can be used to select between two competing hypotheses – the hypothesis that 'best captures' the evidence (or is 'most likely') will eventually be converged to [27, 45]. Reichenbach contributed a mathematical ground that several forms of induction resolve to "inductive enumeration" [32, 43]. Given these observations, directed graphs are a convenient representation for transitional probabilities and conditional probabilities. Perhaps it is the perception of transition, along with the logical dependency implication of the arrow that completes the overall picture that dependency models can be selected probabilistically.

Intuitively, the above argument is limited in the following respects: Bayesian selection only works when the hypothesis subspaces do not overlap substantially for the given evidence (there is some differentiating hypothesis range [52]), hypothesis enumeration can be always found to be incomplete when presented with novel evidence, intersection between true dependency and a transition probability or independence map is not clear, as will be shown in Chapter 2, and finally that some aspect of the rule and world is random.

But if the world (and our internal representation of it) is deterministic, then science is faced with a problem. By approaching the model identification probabilistically, the observational (novel observations or inhibition from surfacing) and variational (of generating model) uncertainties are flattened into a single measure, and inadequacies of a predictive model are no longer considered on the basis of the type, location, or reason for the shortcoming, but on the basis of how well the model 'performs' overall. By analogy, as everyone who has taken exams knows, sometimes it is a lot easier to achieve better performance by writing down a lot of random equations for partial credit than it is to demonstrate mastery on one exam question. But random equations are *not* mastery, and the more complex the problem, the less likely the student is to have productive model adequate to the task.

## 1.1 Why is this problem important?

The assumption that the world and knowledge is in some way random has some significant effects on the problem of dependency identification which surface in a number of domains. Even the idea that the dependencies or determinitic rules may be surrounded in some form of random noise has ramifications on dependency identification that are not well studied.

Natural language, Ontologies in Life Sciences, Mathematics, and physical (Causal) models of the world all are affected. In Natural Language, relaxing the constraint of correctness in favor of uncertainty has led to increased reliance on statistical models of language formation, which though mathematically elegant, are difficult to explain, impoverished predictively, and make mistakes that children simply do not make [41]. In Life sciences, failure to identify dependencies (or identification of incorrect ones) has a measurable and costly effect – drugs may end up further down the development pipeline than they would have been had unwanted dependencies or effects been correctly identified. And finally, in causal models, incorrect dependency identification may either trivially select the wrong knowledge structure for the path a leaf takes falling (suggesting randomness), or may create the wrong dependency argument for a country's involvement in terrorist events! In other words, the cost of wrong answers can be extremely high!

The problem of dependency and rule identification is at the heart of all productive knowledge discovery. When dependencies have been identified correctly, connecting up new knowledge structures grants even greater predictive power into the future, and explanatory power into the past.

Nonetheless, researchers have continued to advance the art of randomness, suggesting that unobserved distributions can be inferred from observed ones, and then that the models that transform our world are merely distributions with a conditional 'dependency' structure that can be inferred just as readily as the unobserved distributions can be inferred. They've suggested that probabilistic or 'degree of belief' knowledge is just as valid – if not more so – than 'certain' knowledge, and have ad-

vocated data driven, rather than knowledge driven analyses. For example, in fMRI analysis, the 'perceived success' of probabilistic methods has been observed to lead to a relaxing of experimental constraints which establish causal relations [30].

I contend however, that what knowledge affords to operators in their domain is *certainty*, and the counter-proof is simple:

> *If knowledge does not necessarily afford certainty, even in probabilistic domains, then one should not place certainty in knowledge such as Bayes Rule, which transforms distributions, but instead believe that there is some probability (or have some degree of belief) that Bayes Rule does not work for the current class of distributions. Certainty in Bayes Rule itself requires an infinite amount of existing evidence for the current distributions to approach a likelihood of* $p = 1.0$*, against which there would be no test of correctness, rendering its use unproductive.*

This counter-example does not assert that Bayes Rule is unproductive, just that the idea that knowledge itself is probabilistically uncertain is self-contradictory. Even the operators on which probabilities depend require certainty to be applied.

Without certainty, there is no knowledge, no prediction, and possibly most importantly, no way to extend scientific models given additional evidence. This is not to say that certainty of knowledge always yields *correctness*, but what it does yield is immediate certainty of some *incorrectness* (i.e. Popper's falsifiability [40]).

Finally, two processes are at work: there are an increasing crossover between disciplines, a greater number of researchers (or students) not formally trained in their domain seek to make contributions, and a large increase in the production of data by different parties. Probabilistic model selection provides a tool by which the researchers can remain ignorant about the underlying physical processes, and turn to 'parameter tweaking' of the size, utility functions, or priors of their model to maximize what it captures (without having a background-prior on the models), without knowing or understanding if there are particular conditions or properties of the domain where the right model is unlikely to be retrieved. The vast production of data greatly increases

the chance of inconsistency and a probabilistic analysis is increasingly expensive.

## 1.2 What makes this problem hard?

Mapping statistical models to real-world models (trees or otherwise) is difficult to lay out and understand [6]. In the real world, no two events are truly independent, much less conditionally independent – even the traditional coin flip example is non-probabilistically dependent on millions of external variables: wind speed, starting side, initial force and directions, etc. Even if one assumes that two separate events are truly independent, it is unclear whether the independence is one of *state*, disruptive *cause*, or transformational *model*. Consider the example that two separate rocks are at the very peak of two separate mountains on opposite sides of the Earth. Simultaneously, one rock is disrupted by the wind, the other by a butterfly. It is arguable that these two events and their states are sufficiently independent to imply that their transformational models & outcomes should be as well. And yet, the two completely independent, partitioned graphs of the rocks follow the same transformational process, and the rocks roll down the mountains according to the rules of *gravity*. The models and graph structures are not independent.

Beyond that, statistical models have an innate "stateless" or locally exchangeable property, and require "complete" observation information (that we have completely enumerated all variables), which are fixed at the start of analysis. While there is something disquieting about fixing the model at the outset of an analysis that implies substantial a-priori knowledge (despite some Bayesians' assertions to the contrary), the assumptions of exchangeability and "complete observations" have special ramification to the problem of tree induction and any process consisting of transformational rules.

And it is easy to confuse the idea of selecting observations from a bucket of those seen before with a "predictive" model. In some sense, human knowledge comes from what has been seen before, and the human faculty for model discovery and alignment is unparalleled. But probabilistic methods produce models and structures

(in language for instance) that humans do not produce. In fact, it is almost trivial to come up with "real world" situations where a statistical model for learning and induction would yield certain death for a human.

This idea may fly somewhat in the face of the notion of convergence, but according to Salmon, "*Any* value of the relative frequency in an observed initial section [sequence of observations] of *any* length is compatible with *any* value for the limit. [...] and we cannot be sure that such [non-converging] sequences do not occur in nature"[46], which is effectively a statement about the probabilistic admissability of *any* generating rules.

Finally, even supposing that the map from statistical models to real-world models was clear, the problem of model identification still contains the graph-matching and resolution problem, both on the nodes and the relations which define the edges, which for bipartite matching is a combinatorial optimization problem – polynomial in size (the Hungarian method [20]).

## 1.3 Background

Though probabilities are used sometimes to "account" for more observations than we have models for, or to "justify" an existing set of relations, I do not want to reduce the importance of probabilities or of treating the observations as a sample space. These assumptions allow us to analyze observations in terms of the *capacity* necessary to "carry" them [50]. When no relational model is known and capacity (or "focus") is limited, knowing the characteristics of the observations can insure that the majority are transmitted by excluding those not likely to be important. In terms of tree identification, this could be useful for excluding superfluous or noisy information. But a question of equal importance, is what (if any) properties or features are lost by treating observations in that manner.

The suggestion that an underlying model may not be recoverable probabilistically exists in a number of forms. Most notably, in the famous debates between Einstein and Heisenberg and in the innate existence of some kind of physical "uncertainty".

Einstein et al. attempted to quantify "certainty" [13] in terms of a 1-1 relation between observables and hidden nodes, but the question of what uncertainty itself is, within the frame of complete determinism, has not been adequately addressed.

Karl Popper often attacked probabilistic interpretations of uncertainty directly, both at a quantum-mechanical level and at a logical level. He did so quite compellingly, but most arguments against him take the form that predictive uncertainty and incomplete observation from the past are identical.

According to him, probabilistic uncertainty, as a class of random sample spaces in probability theory, and then the subclass "subjective uncertainty" [40] or the "gradient middle" [59] between categorical points, has taken over, despite the fundamental work in that area which suggests otherwise. It is not explicitly stated, but careful reading of a proof of the Strong Law of large numbers [16] suggests that there are cases where *divergence* not convergence is guaranteed: for *finite* data sets in a single sample space, or as I believe, when the number of sought "representations" (i.e. relations between multiple sample spaces) grows faster than the convergence rate from additional observations.

Kolmogorov complexity, while not computable for real data [26], also suggests it – given that the minimal program needed to compress a random number is also infinite, the existence of a model or transformational rule implies the existence of a minimal program [37, 8]. Tree representations imply multiple levels of compression and abstraction, and allow the simplest representations of ideas or laws [19].

Even Chomsky addresses the idea that an n-th order statistical approximation of English "will exclude (as more and more improbable) an ever-increasing number of grammatical sentences" [9], which is an application of Shannon's Asymptotic Equipartition Property [50]. Chomsky also captures, without explicitly linking it to a divergence of probabilistic description, that every n-order approximation of sentences is a finite state Markov grammar, which cannot generate or "predict" additional correct observation sequences and will contain (or carry) a large number of incorrect ones [9].

In the same vein, Wolfram provides an elegant demonstration that the *appearance*

of randomness combined with limited observability, can be derived from deterministic laws [57].

If probabilities are thus abandoned, the problem of identification of data consistent structures is extremely hard – represented by an automaton Gold found to be NP-complete given a finite source of observations (& polynomial in the limit) [23]. This alone might be seen as a reason to employ probabilities – to prefer searching one set of transitions over another.

There is substantial evidence that "good" results can be achieved with probabilities. One of the most notable came from the introduction of the Lyapounov, which seems to leverage commutativity of operations across a phase space to show that stochastic processes can result in a deterministic long-term outcome [34].

In fact, the field of machine learning is full of methods designed to capture hidden variables and structure (HMMs, SVMs, Bayesian, MDL, etc.) [29, 45, 28, 26, 37] Probabilistic methods are applied in new ways frequently to recover the underlying form (or at least provide a "plausible account" [14] of data and variation [59]), for instance Optimality Theory in generative grammars, Bayesian statistics for syntactic structure [15], or Bayesian networks for ontology alignment [54]. MDL is generally accepted to provide good, or at least 'well founded' results, where a composition chosen models to describe a sequence is of the shortest length – while still providing a good fit – conceptually similar to Occam's Razor [26].

The gold model may be carried within the most probable model, and I would like to find out to what extent. It also may be that recent unique approaches, such as Eisner's [15] may have a chance to posit the correct heirarchy by restricting model formation to first-order "triangles", but may possibly still be susceptible to incorrect dependency direction which would propagate throughout the hierarchy. In this experiment, I am primarily concerned with correct identification of dependencies, which as Pearl notes, if one has access to the dependency structure, then one can perform structured equation modeling (sem) to recover individual paramaters of 'causal' change [39] (it is of note that the 'parameters' are of functions of change in expectations).

## 1.4 Thesis Overview

This thesis proceeds by discussing why the assumptions are chosen the way they are, and explains what exact features lead me to believe that I will discover something important (Chapter 2), then explains the experimental setup (Chapter 3), delves into the results and what they mean in Chapter 4, and finally concludes (Chapter 5), providing a 'big picture' of the findings, how they were limited, and future work.

# Chapter 2

# Models, Representations, and Uncertainty

In Chapter 1, I touched upon the idea that the goal of much of the use of probabilities and probabilistic model selection is to capture 'uncertainty'. In this chapter, I explain how the structure of different models handle uncertainty innately (if they do), and provide some examples demonstrating that what (in terms of model identification) probabilities are attempting to capture is not uncertainty, but 'inconsistency'. Finally, I provide some examples that show that the 'intermediate' or 'intersecting' form between representations of probability and representations of dependency is inadequate to represent either, possibly explaining why a great deal of ambiguity is introduced into model induction, and why I expect to find something interesting.

## 2.1 Models

A 'model' is a 'representation of reality' that contains the important features necessary to correctly predict an outcome, or transmit some features and knowledge with certainty. What can be transmitted therefore is highly dependent upon the model chosen. Perhaps it may be better understood in terms of 'perception': if a learner is receiving some symbols transmitted by a model, what can actually be integrated from the model are only those symbols that are consistent with the model itself (self-

consistent), with the learner's model, or with the real-world. In the latter case, the errors that (features of) the model makes are objective and are measurable in terms of probability of error – some of the transmitted symbols are better than others [44]. In the case where the transmitting model and the testing model are subjective (the objective truth is elsewhere), as it is with learning a grammar, or combining two ontologies, the method of representing inconsistency and uncertainty is important, as this representation may later provide the tool and motivation for increasing complexity, and thus predictive cohesion against an objective truth.

### 2.1.1 Describing Models & Structural Uncertainty

Different models have different methods of managing uncertainty. In graphs, trees, and ontologies, uncertainty that cannot be resolved produces inconsistency, which cannot be represented in the structure.

Consider a simple scenario where one tries to use probabilities to handle the uncertainty that trees naturally handle well. For trees, "uncertainty" is *structural.* Predictive uncertainty is handled as branch insertions, extensions, or replacements (Fig. 2-1). Explanatory uncertainty (in the present) surfaces as the number of paths through the tree that converge on the present observation/node. Incompleteness of those paths is handled again as branch additions & extensions. Incomplete observability of a branch node does not inhibit prediction or usage of child nodes. But possibly the biggest case for tree-uncertainty is the partitioning: negative evidence can be used to destroy an entire branch linkage (or move/transform it intact elsewhere) without affecting any existing relations at all – the same is true of prediction – prediction down one branch is fully partitioned from another. In many ways, uncertainty for trees is not "uncertain" at all. This reduces the kind of uncertainty that trees have to handle to the lexicon/symbol matching problem, and conflicting dependency direction.

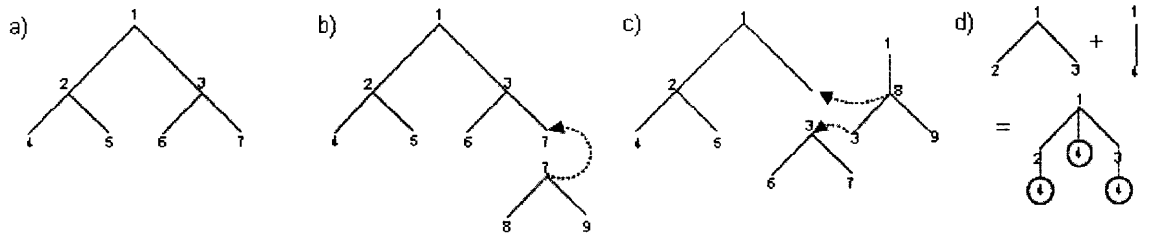



Figure 2-1: a) An initial tree structure – {6,7}, and {4,5} may inherit the lack of ordering from regular graphs, but both branches are completely partitioned. b) Uncertainty in augmenting knowledge is handled simply as branch extension. c) Uncertainty around adding a 'increased complexity' dependencies of the $1 \rightarrow 8 \rightarrow 3$ is just a branch insertion. d) Uncertainty in adding knowledge that can result in consistency problems. 4 is a subclass of 1, but it could be a direct subclass or a derived subclass – a member of 2 or 3. Making 4 a direct child of 1 is the least inconsistent representation that maintains dependency.

## 2.2 Problems with Graph Representations

Though we use graph representations, trees or otherwise, all the time to convey connectivity, dependencies, and separations (cliques), neither graph representations of probabilistic models, or of deterministic models are complete. In this section, I will describe some of the reasons that graph representations are limited and inconsistent for each domain/model type, and explain why connecting probabilistic representations to deterministic representations through the graph representation is unlikely to produce predictable or consistent results.

### 2.2.1 Probabilistic Models

The problems that probabilities introduce are not limited to trees, but surface whenever one tries to augment the edges of a graph with probabilities. Probabilities are deeply associative and connectionist representations, relating the frequency of every event to every other event. But when used as a measure of uncertainty, as on the edges of a graph, the unordered probabilistic connections strain against the desired graph connectivity. It can become almost trivial to compose examples where the 1st-order logical connections of a graph become inconsistent with probabilities. For example, consider a graph $G = \{A, B, C\}$, and that A–C are deemed 'independent', below a

level of significance $S = .1$. If A–B, and B–C are non-independent, using probabilities P1, P2, respectively, then $P1 \times P2$ must be less than $S = .1$ to maintain consistency. In other words, $P2 \leq \frac{S}{P1}$ – the probability of graph edge B–C is dependent upon both A–B and the chosen A–C threshold $S$. By extension, the first order edges assertions are constrained by second order, third order, and so on, probabilities. Even if nodes are independent, the probabilities and math certainly aren't.

Secondly, even if the conditional independence graph is consistent, the conversion of a causal process to that graph is not entirely straightforward. Despite the large amount of literature tying probabilistic models and causal models together (Pearl/Wright's path coefficients[39, 58], Bayesian networks[45, 28], etc.) and the mountain of 'rain, sprinkler, and wet-grass' examples to that effect, causal examples can be composed quite easily who's map to a conditional network yields a conflicting structure. Neglecting causal loops, which have no conditional network analog, but can generally be easily understood in terms of consumption of state (like a rock rolling down a hill), there are spreading tree examples as well.

Consider the case of a forest of trees (T) growing on some hill in one location, and then sufficiently far away as to be totally (causally) independent, a hill covered with bunches of rocks (R). Directly in the middle is a volcano (V) which erupts, sending out shockwaves, pyroclastic flows, sulfur dioxide, and ejecting large chunks of obsidian. It is clear that the tree formations and the rock contents are partially causally dependent upon volcanic eruption (Fig. 2-2 b). (i.e. the shockwaves disrupts growth patterns, kill different species, sulfur content changes supportable plants, etc.). If however, one takes this same graph and interprets it in terms of conditional independence, one would say that Trees and Rocks are independent given knowledge of the existence (possibly influence) of the Volcano (and its values). Are they? Certainly not. While one could argue that once you know the volcano is erupting or not, the trees provide no more information about the rock states or expected values and vice verse, but that is untrue. Because the Volcano is causally involved in the process that results in the distribution of trees and rocks, by inspection, the silica content of the rocks tells us something about the volcano, which tells us about the explosive

power, the shockwave size, the sulfur proportion, and by extension, the trees on the other side. Simultaneously, if we have no knowledge of the volcano's existence, the graph tells us that our state of information/knowledge without the volcano would lead us to believe that the trees and rocks are non-independent – something that is also untrue. We would believe because of separation, and all other variables aside, that the tree growth is nearly completely independent from the rocks on the other. The conditional independence graph that the "information" is most logically consistent is not the one aligned to the causal model, but the inverse (Fig. 2-2 c). This is a demonstration of the confusion between the idea of causal 'effect', and changes in the state of information.

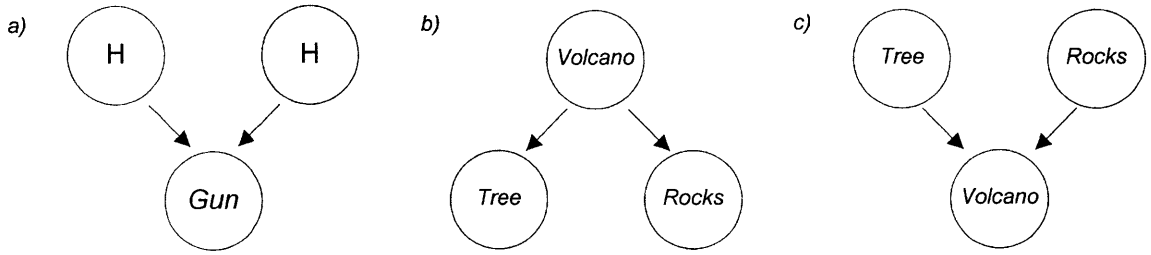



Figure 2-2: a) An example of two coins being flipped; A gun (G) is fired when two separate coins come up heads (H). Evidence of the Gun being fired 'provides evidence' (makes the values of the two coin flips non-independent). It should be noted that neither flips of the coins are actually *causally involved* in the production of the gun fire. b) A Volcano is causally involved in the Trees and Rock distributions. c) Knowledge of the Volcano makes the Rocks and Trees non-independent. The preferred conditional independence model is inverted.

Ultimately this has become a problem largely because of the words that are chosen to describe the graphs. While the dependency graphs in Figure 2-2 are 'conditionally independent' given the existence or non-existence of some piece of information[28, 45], over time (perhaps due to the arrows) they have come to be referred to as 'conditionally *dependent*', and thus suggest some sort of causal or structural dependency between the observations. This is incorrect. To account for the inconsistencies between truly dependent networks and conditional networks (CN), the CN's are best referred to as 'describing' how the 'information of observations' changes knowing a particular observational value or having *no information about an observation at all.* This means that each node in a conditional network is actually making 2 sep-

arate graph statements: about the non/independence of information from adjacent nodes given the node exists and has values, or when it does not exist at all. In the graph above (Fig. 2-2 b), the growth pattern of trees on one side of a volcano is deemed to be conditionally independent of the rock patterns on the other, given observational information of a volcanic eruption (this is sometimes said to *block* the flow of evidence[45, 39]). Besides the obvious intuition that the rocks and trees are not independent given that information, this is an assertion about the invariance of the expectations of the two different distributions (rocks and trees) with respect to observations of the volcano – that the current expectations are the 'natural' or 'independent' ones. The counter assertion made by the graph is that failing to observe that the volcano even exists, the rocks' and trees' distributions will vary from their 'natural' (with volcanic information) value, and that variation is non-independence. In many ways, it does not describe the underlying processes at all, but our 'state of information' relative to some new parameter.

Finally, Bouckaert demonstrated that not all independence statements can be represented by the first order edges of a conditional network, and that some in/dependency statements cannot be deduced from the structure [7].

### 2.2.2 Deterministic Models

This is not to say that a pure graph is a complete representation of a deterministic "model", either. Transformation functions have input states, output states, "causes" that select them, and orders of operation. A featureless node is not sufficient to capture all three – at best it could only represent a "state". But in this particular case, a tree supports structural dependencies, and order of operations that make it of interest.

When converting to a pure graph representation – even a tree, the edge augmented feature of state/cause and the node feature of 'transform' is lost. In other words, the 'why' and the 'how' a process occurs is not represented in the graph – only the participants (nodes), and a loose approximation of outcome dependency remains. In the causal example of a shoemaker constructing shoes (Figure 2-3 b), the shoemaker

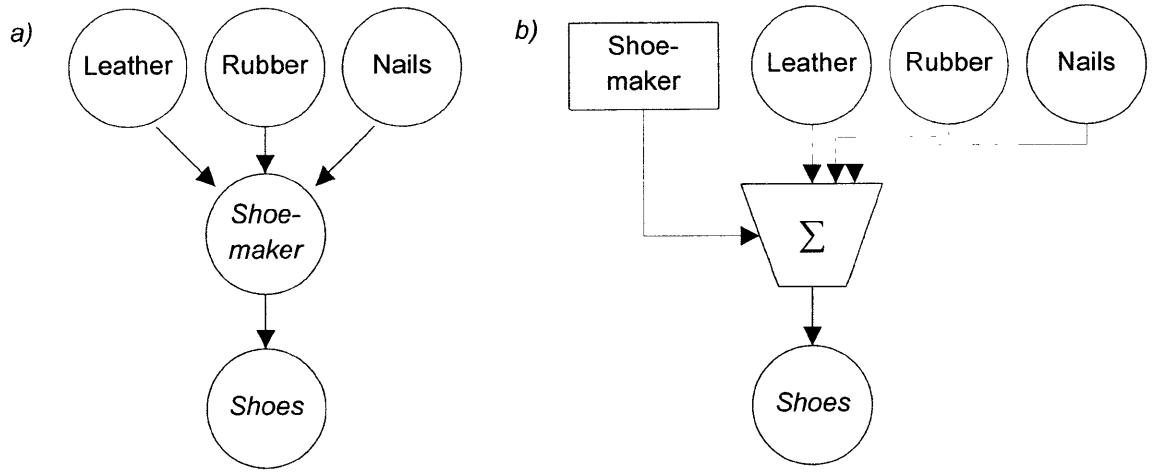



Figure 2-3: a) A possible graph representation for the dependencies between the shoe, shoemaker, and materials. b) A partial causal representation of the conversion of materials into a shoe. Before the shoe-maker 'causes' the sum/composition function to be applied, the input materials are in a static (unmodified) state. The top is the input state, the left is 'cause' or disrupting input.

'causes' the composition ($\sum$ function) of raw-elemental materials into the shoe. But in the simple graph (Fig 2-3 a), the function that the shoemaker actually performs is unknown – it is only a human interpretation of the semantic meaning of 'shoe-maker' that provides any insight. If it were not for the knowledge that a 'shoemaker' 'makes shoes', the graph would not provide information as to whether the shoemaker is replacing, trading, wearing, teleporting, or ... *inhibiting* the production of shoes. Would the shoes naturally form themselves if it weren't for the shoemaker? A 'causal' graph should provide information about the 'final states' or 'final observables' were it not for a transformation function.

An even more exacerbated example is one of causal loops (Figure 2-4), which we can represent in a graph, but cannot handle in a belief network [45, 28] (though it can be handled in a Markov Chain). In a purely directed graph, the cycle, though it may have multiple entries, implies an infinite loop without a stopping/exit point, whereas a causal representation (Figure 2-4 b) may consume an input and stop (similar to an FST).

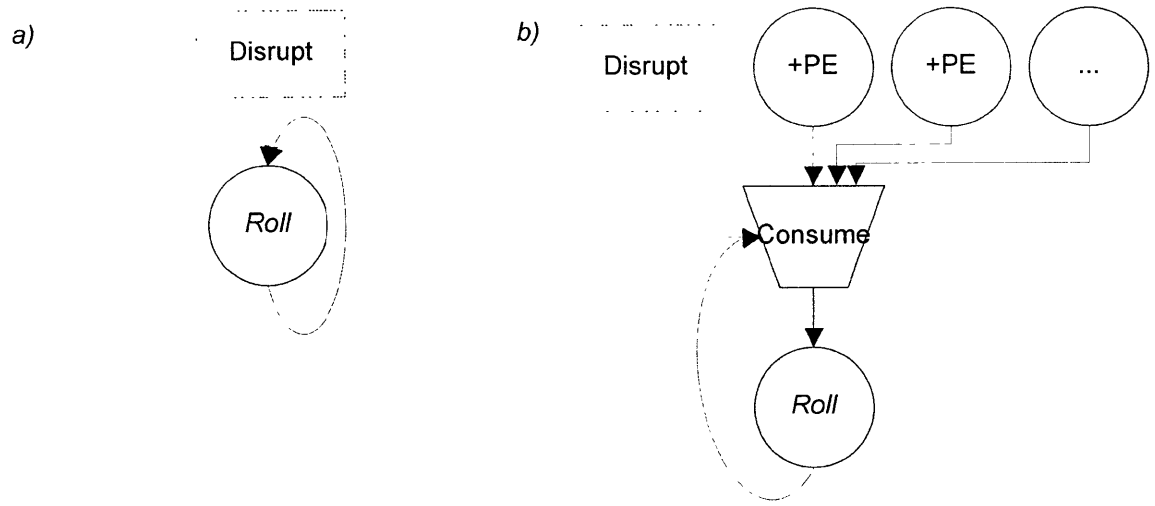



Figure 2-4: a) A graph representation of the 'rolling' process, perhaps of a rock down a hill. This cycle has no completion or exit and does not describe anything about the process besides that a 'roll' now is dependent upon roll in the past. Augmenting this graph with potential energy, where would it go? b) A state-augmented graph of the same causal loop. Roll continues as long as it is in a state where potential energy (PE) is available for consumption. Interrupting the roll leaves the rock in a state with the remaining potential energy unconsumed.

## 2.3 Uncertainty

Rather than delve into the esoteric philosphies about what "uncertainty" entails, and if "certainty" even really exists, I would like to enumerate a number of concrete ways in which uncertainty surfaces in *discrete* observations and analysis.

In model formation and identification, uncertainty can be present in both the observations themselves, and in the dimensions that the model space can take on. And while I might feel inclined to separate explicitly observed dependencies into a different space, dependence between observations can be seen as a kind of observation itself, which restricts ordering, but inherits the same kinds of uncertainties.

### 2.3.1 Observations

Uncertainties that affect observations, and by extension, observations of dependence are listed below:

1. Observation Order

2. Novel Observations (totally unique)
3. Grouping of observations – either state dependence, or number of joined observations in a branch.
4. Similar to a different observation – i.e. sharing of features, *has-part*, weighted-Observations/Co-informing.
5. Variation in measurement around Observation true value – could also end up as similar to a different observation (4).
6. Inconsistently Observed

   O is not always observed; it is either obscured or not present in the source. This could also be the transmitter's rule or knowledge-base.
7. Number of Observations – the number of O to obtain a 'representative sample', if clustered by similarity.

### 2.3.2 Model Correctness

Assuming no uncertainty in the observations themselves, in the production of a predictive (or even just a descriptive) model, there are still a number of innate unknowns that surface as uncertainties and affect model identification. Consider a sequence of observations:

$$\texttt{A B C B B A B C B B A A B A B C B A A A B C} \tag{2.1}$$

Then a partial enumeration of the dimensions of uncertainties:

- Parameter Size (Window Size on the sequence of observations – the separation on which observations are irrelevant or how far back in time one need not look) $[A], [AB], [ABC], [ABCB], [ABCBB] \ldots$
- Temporal Dependency Space

  There is some delay $t$ before a dependent observation surfaces – for instance, B

at time 5 may surface $t = 2$ later than its triggering state of C at time 3.

$$A_1\ B_2\ C_3\ B_4\ B_5\ A_6$$

- Delineation of Observation-Experimental Groupings – which observation symbols represent the *control* variables, and which are the outcome variables. This is a very strong assumption, especially in experiments with state.
- Parameter-Parameter Dependencies – these are dependencies between groups of observations deemed to be 'similar'.
- Observation→State Aggregation – the combinations of observation-symbols constitute the relevant state of the experiment. For instance, in the above example, $B_5$ could be dependent upon $\{A_1, B_2, B_4\}$, or any other combination of prior observations and position one might concoct.
- Weighted Contribution to an Overall Theory (Coherence).
- Transforms on discrete or continuous observational state. For instance, consider the transform $F(X, X+1) => X+2$. Then $(A, B) => C$ , and $(B, C) => D$, and so on. In a sequence of observations, this is a generator function [57].

### 2.3.3 Completeness of Model Space

If I follow the reasoning that induction is merely the process of enumerating all possible outcomes and choosing the most probable, then I would want to take all of the 'uncertainties' listed above and create a model space {M} that contains them all.

It can be shown that full enumeration of all possible dependency directions between all possible observations contains all variations of temporal delay t before an observation surfaces, but after removal of cycles for Bayesian networks[45], does not contain all possible mixtures of models (Markov cycles are not contained). The final primary restriction is on state aggregation and generator functions, which are not contained in the graphs as described earlier.

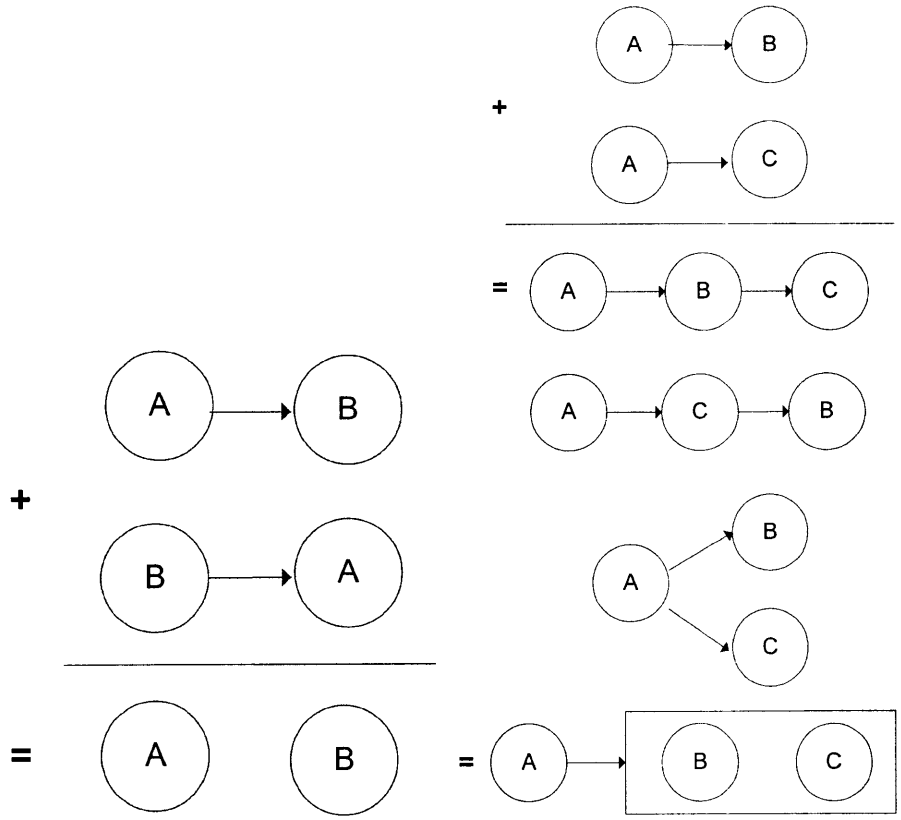
### 2.3.4 Case for Sequences of Observations

In merging dependencies, ordering, and hierarchies, we are faced with the question: *what do we do with conflicting information, ordering, and knowledge?* In fact, this is the only question of interest. If our two knowledge representations are consistent, then there is no problem, and nothing interesting to show (in the same way that a completely consistent graph can be recovered from two consistent partial orderings [4]). If they are inconsistent, however, we are faced with the *merge* problem that crops up in ontology alignment, and indeed, all forms of learning and science.

The basic problem with probabilities is that they treat observations as ultimately stateless (i.e. exchangeable). In the stateless combination of dependencies, merging conflicting dependencies breaks down to a single case: $(A \rightarrow B) + (B \rightarrow A)$ (Fig. 2-5a). In the case where two knowledge structures make *entirely different* dependency assertions, the merge order is still not clear – the only consistent information – which observation is the 'head' remains, but the dependencies become pure, unordered observations as well (Fig. 2-5b). In other words, pure observations are the lowest common denominator of knowledge.

Secondly, many times in scientific inquiry, observations are all one has – whether they are DNA sequences (read: '`G A T T A C A`'), or classifying and tagging words with other known observations. For example, 'The dog ran.' becomes '`D N V .`' (Determiner, Noun, Verb, Period). Similarly, parsing introduces extra symbols (bracketing) into the observation stream to represent dependencies '`S ( D N V )`'.

Finally, in an ontology, as the graph's edges are augmented by features that can describe the 'type' of relation or dependency (i.e. 'is-a', 'has-part', 'regulates' (Gene Ontology [5])), themselves are augmented by specifications that describe the transitivity and reasoning of the nodes those features connect. But in the case where the logical features conflict (i.e. A 'is-a' B, and B 'is-a' A), but the dependency direction and the features are inconsistent and cannot be represented except by the observations themselves (A 'is equivalent' to B, but no such relation/feature exists).




(a) Merge of $(A \rightarrow B) + (B \rightarrow A)$. In the 'most consistent' form, the conflicting dependency is removed and $\{A, B\}$ become just separate, exchangeable observations.

(b) Merge of $(A \rightarrow B) + (A \rightarrow C)$. In the 'most consistent' form, $A$ remains as the head, but since no information exists about the relation between nodes $\{C, D\}$, three possible dependency orders exist: $A \rightarrow ((B \rightarrow C) + (C \rightarrow B) + \{B, C\})$. $B, C$ reduce to a case of Fig 2-5a, pure observation.

Figure 2-5: Cases of conflicting dependency reduce to pure (unordered) observation.

## 2.4 Going Forward

In short, part of the reason that this problem is hard is because the graph, which is used as a simple diagrammatic representation through which probabilistic descriptions and predictive models are often connected, is an underrepresentation of both, in orthogonal directions. Unfortunately, this has not stopped others from referring directly to the "additional features" of causal structures probabilistically [29, 45], suggesting that the graph contains them.

Going forward, I've made the case that the greatest common information in a

sans-state dependency conflict reduces to pure observations. Additionally, I've argued that most model combinations and dependencies are completely contained within an enumeration of all graph models of a given size, though as we remove cycles to posit valid conditional networks, this is reduced. Finally, I expect to discover something interesting due to the poor overlap between the deterministic and the probabilistic graph representation. And in trying to recover the deterministic rule, I will focus on the only aspect that the graphs seem to have in common – dependency.

# Chapter 3

# Experiment

In Chapter 2, I made the case that resolution of partial or conflicting knowledge structures with unknown dependencies can, under best conditions, be seen to be an ordered sequence of observations. In this experiment, I will take a 1-D sequence of semi-random-observations which have a known deterministic model which relate some of the observations, and after applying a 'veil of forgetfulness' (as a colleague called it) about all generating knowledge, I will try to recover the 2-D dependencies of the deterministic model, if not the full model, probabilistically.

The question that I want to answer is if the process of selecting a dependency model probabilistically from some hypothesis space (treating it as a typical machine-learning problem) is likely to identify necessary aspects of the correct deterministic model (and discriminate incorrect aspects as Valiant does [56]), and if not, why?

I shall do this by establishing a number of 'gold' generating models, which are fully deterministic and operate on a sequence of observations, transforming subsets of the sequence into different sequences. The deterministic transform will repeat all the way through the sequence, satisfying redundancy, and each transform's appearance will be independent of observations not transformed and other transforms' appearance.

The transforms' inputs, outputs, and position within the string will not be known a-priori, meaning that all the probabilistic recovery mechanism will have access to is the final sequence of observations. Explicit dependency relations will also be invisible. While this is a substantial limitation, in many domains it is the only information that

we have access to: DNA/protein sequences, Natural Language Processing, or just a linear sequence of static observations made from the real world "broken glass on floor, footprints, overturned chair", and thus this is a common approach towards making dependency assertions between observations.

This section proceeds by listing out the domain definitions, how the observations are generated from the source model, how the probabilitic models are selected and tested, goes through a toy example, and finally, some of the limitations of this approach are discussed and what is not likely to be discovered.

## 3.1 Experimental Overview

With the goal to obtain insight towards the above questions, the steps of the experiment are:

1. Pick a domain that mimicks observational & uncertainty properties of the "real-world".
2. Transform portions of the sequence of observations from that domain using a deterministic rule into a different sequence of observations.
3. Setting aside all prior knowledge and using only the new observation sequence as the source of evidence, try to recover the dependencies of the rule probabilistically, and test the dependencies of the recovered rule against the original rule.
4. Accumulate observation frequency counts and compute joint probability distribution for observation-states within a window of fixed size Q – the ordering and dependency is unknown. (i.e. Q=2 is adjacent observations) : Vary Q up to 5.
5. Create a hypothesis space consisting of all dependency models of size Q. Test each one's conditional independence map to find which one is most likely to have generated the observation sequence from the joint probability distribution (Most Likely Model $M_{ml}$).
6. Map the most likely model $M_{ml}$ to the most likely observation $O_{ml}$ values to determine if the identified model's distributions map to the correct observation

& dependency structure.

7. Continuously increase the length of O (add more observations) and test the direction of the edges in the most-likely-observation-dependency model $M_{ml} \cdot O_{ml}$ to determine if the identified model is the correct (gold) one, how correct it is, how far from correct it is, how sensitive it is to new evidence, how quickly it is converged to, and how likely the correct model is to be selected overall.
8. Do 2-7 for a number of models & different Q sizes.

## 3.2 Domain Definition/Assumptions

Because in this experiment, observation generation (real world domain) and model selection (limited problem domain) are treated separately, it is important that the domain assumptions for probabilistic induction (model selection) be defined as closely to the real-world "generating" domain as possible, and to understand where specifically it deviates.

### 3.2.1 Intermixed Observations & Outcomes

Before breaking down the domain, it is important to describe some of the limitations of a necessary assumption – that of the sequence of observations. I may have argued the case for usage of the sequence, but the sequence has some properties that are both helpful and detrimental to the purpose of my experiment.

Separating input observations and outcome observations allows us to more easily establish and test a *control*. Specifically, it will allow us to answer under what circumstances the transformational rules are visible, partially visible, or completely inaccessible. If provably inaccessible, then no probabilistic method will ever be able to recover it. If the rules are accessible to a separated input-output analysis, it will help provide insight as to what (if any) information is being lost with probabilistic model selection.

However, mixing observations in a sequence has some advantages as well. I believe it to more accurately represent the problem of 'incorporating' new observations.

Specifically, as observations are added from different sources, with different characteristics, the learner may not know whether an observation is an input or an output, or connected via a causal relation – only additional evidence allows them to posit some theory. Effectively it makes it easier to add new observations as part of the experiment, and describe 'convergence' as a function of the number of observations. Secondly, provided that observations are 'chunked' up into a sequence and ordered consistently, then the sequence contains the case of separated inputs and outcomes, while providing a consistent set of observations to reuse throughout the experiment. When the sequence is initially generated by a random set of symbols, it also emulates (in the limit) the full space/complexity of state possibilities that the world could be in for a rule to apply. Finally it makes for a clear representation of the complexity of the problem of 'position dependence' or 'temporal dependence' within the stream – that an outcome may surface some time later than it's dependent state.

### 3.2.2 Real-World Domain

In the Real-World domain, the scope of the full model selection problem is characterized.

GIVEN: A linear stream of observations.

THE PROBLEM: To determine Causal and State dependencies among observations.

THE DOMAIN: States may be aggregated to create new states

CONSTRAINTS: Different observations are produced by different states (the same state must be observed identically, unless altered in some way).

UNDER-CONSTRAINED: Dependent observations may not surface immediately – there may be some unknown delay $t$).
Different states need *not* lead to different observations.
States are not necessarily *consistently* or *completely* observed.
State Observations may be made in any order.
State Dependencies may be made in *any consistent* order.

Novel observations may be of an 'unknown type' and properties.

If the 'real world' is random, the distributions are unknown a-priori.

VARIABLES: The total number of observations $O$.

Number of state dependencies $D$.

Number of states $S$.

### 3.2.3 Limited Problem Domain

In the limited domain, I make a number of assumptions that make the problem more amenable to probabilities, that are intended to serve as a "best case", while still displaying the aspects of the problem I am seeking to investigate.

GIVEN: A linear stream of observation symbols.

THE PROBLEM: To determine the dependencies between observations.

THE DOMAIN: All states are atomic and elemental states – states may *not* be aggregated to create new states.

CONSTRAINTS: Different observation symbols may be produced by the same state (distribution of observations).

A state may account for only 1 observation at a time.

Dependent observations surface immediately (zero delay).

State inputs are position-specific relative to dependent observation

UNDER-CONSTRAINED: Different states need not lead to different observations.

States are not necessarily *consistently* or *completely* observed.

State Observations may be made in any order.

State Dependencies may be made in *any* order.

Novel observations may come from any distribution.

Distributions are not necessarily uniform.

VARIABLES: The total number of observations

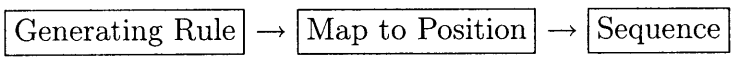
Number of observation/state dependencies.

Number of states.

In the separation between the real world domain and the domain used for induction, a logical problem emerges. If we require all individual observations to be elemental states, and do not combine them into more complex states, then the required functions (that I will use, such as $AB \rightarrow C$) cannot be produced. Because it is sometimes assumed that forms of fuzzy logic (including probabilities and belief representations) can represent both AND and OR of boolean logic [60], and because a graph of $[A\ or B] \rightarrow C$ is the closest that a single elemental model could be to correct, the selection of an independent A or B with the correctly dependent $C$ will be acceptable. However, connection of A and B will be considered to be *unacceptable*, because even though it is obvious that A and B are correlated, a dependency assertion between them could be turned into a logical dependency – that C is dependent (derives logical evidence) upon both and A and A (via B).

## 3.3 Generating Models

The generating models are the list of rules that create and transform the observation sequence into the one for analysis. It consists of two parts: production + alignment of the initial observations from the random distribution(s), and the transformation of subsequences of observations with deterministic rules.




### 3.3.1 Initial Observation Sources

The input alphabet $\Sigma$ is the source from which initial observations to be transformed are selected. Before tranformation by the rules $\Gamma$, each observation o is selected from the discrete uniform distribution with $p(o) = \frac{1}{|\Sigma|}$.

### Random

In a randomly produced sequence, the alphabet of input observations $\Sigma$ is selected from a uniform distribution of alphabet, and if the prior observations in the sequence satisfy the state requirements of the transform rules $\Gamma$, then $\Gamma$'s output is inserted into the sequence.

For instance, consider $\Gamma = \{A, B \rightarrow C\}$, and $\Sigma = \{A, B\}$. Then A and B are selected from a uniform distribution of .5:.5, and whenever the sequence AB is seen, C is inserted. The position of the rule throughout the sequence is not pre-assigned, and it could surface in any (non recursive) position.

```
A B C B A B C B B A A A B C A B C B B B B A          (3.1)
```

### Partitioned & Aligned

In an aligned and partitioned sample source, the distributions that each observation position maps into are asserted a-priori. This is conceptually identical to the formal separation of input and output observations, and is effectively equivalent to 'naming' observations, or rather the distributions that they come from in advance.

```
A B C, B B A, A B C, B A A, A B C, A B C, B B B      (3.2)
```

## 3.3.2 Transform Models

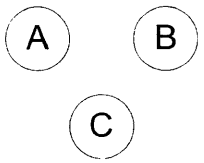
It is of note that Models 1-3 are generated using a consistent size of $Q = 3$, so that an understanding of the errors made in model selection may be built progressively on those earlier models, without having to control for changes in size.

### Model 1: No Dependencies

This test is effectively a control on the model selection – can fully independent model with no transform rules, and thus no dependencies, be identified?

$\Sigma$ : $A\ B\ C$

$\Gamma$ : $\emptyset$



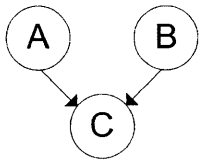

*Example* : B A B A B C C A C B C C A A B

### Model 2: Simple Transform

The simple transform is the composition of two observations into the insertion of a new observation. The task is the correct selection of the known dependencies, and the child node.

$\Sigma$ : $A\ B$

$\Gamma$ : $A\ B \rightarrow C$




Ordered (ABC) Example B A B C B A A B C A B C A A A A A A

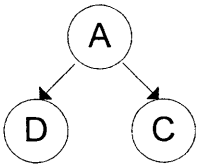
Aligned Example B B A, B B A, A A B, A B C, A B C, B B B

### Model 3: Shared Process (Dual Dependency)

This is our Volcano example; two observations are dependent upon the existence of one. Observations from one distribution are involved in the production of observations in other distributions. Apart from the production of D and C, A and B are generated uniformly.

$\Sigma$ : $A\ B$

$\Gamma$ : $A \rightarrow D\ C$



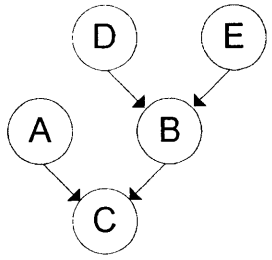

Aligned Example B B A, B B B, A D C, A D C, B A B, A D C

### Model 4: Hierarchical Model

The hierarchical model tests the composition of 2 tree models of the 'simple transform' (Model 2) type to identify if the multiple dependencies can be recovered, and it also

satisfies the 'randomness of branch position' (but within valid branch bracketing positions).

$$\Sigma : A\ B\ D\ E$$
$$\Gamma : D\ E \rightarrow B$$
$$A\ B \rightarrow C$$




Aligned Example `E D B A C,...,D E B A C,...,A E D B C,...,A D E B C`

In this hierarchical model, positions of {D,E} may alternate, as may positions of {A,B}, while the root outcome C stays fixed. B must also stay fixed relative to D and E, given the 0 delay assumptions in domain constraints. This means that positions of A may be displaced relatively, as may be positions of D and E relative to the root C. Note the sequences which are *not produced*: `[ED]ABC, B[DE]AC,AB[DE]C...`. Though the observation positions within the branch are equivalent, the dependency structure of the tree prevents truly random observations from surfacing when linearized.

## 3.4 Model Identification

Identification of the most likely model $M_{ml}$ from a sequence of observations formally consists of 3 parts:

1. Reading observations from the sequence into source distributions. Source distributions are position dependent.

2. Identifying position dependencies finds the model $M_{ml}$ between the distributions that "best fits" the observation sequence (most likely to have produced the observations). This also represents the most likely position dependency model within the sequence.

3. Identify observation dependencies by mapping the best fitting position dependency model to the most likely observations $O_{ml}$.

DEFINITIONS:

$O$ - Let O be the entire observation sequence.

$S$ - Let S be the state size, in number of observations. In this experiment, S=1

$N$ - Let $N$ be the number (length) of the entire observation sequence, and subscript$_n$ an index into it.

$E$ - Let $E$ be the observation at the $n$'th position (sequence index $O_n$).

$Q$ - Let Q be the "complexity" of model – both the window size on the sequence, and the number of parameters/distributions.

$i$ - Let $i$ be an index into a distribution or parameter.

$D_i$ - be the positional distribution from which a single $O_n$ (or $E$) is drawn.

$\{M\}_Q, M, g$ - Let $\{M\}_Q$ be the set of all independency models of sizes Q, $M$ be a single set of models of a fixed size $Q$, and $g$ an index into an $M$ from which a single model $M_g$ is selected.

$m, n, o$ - Are the read distributions $(D_2, D_1, D_0)$ for convenience and readability.

### 3.4.1 Reading Observations

Because everying about working with probabilities is about working with distributions, the process of reading observations from the source sequence is about reading observations into target distributions – in order to calculate joint probability distributions for different models. It is common to make a number of assumptions regarding independence within the observation sequence, but one must be careful. The set of all graphs already contains the set of all possible independence assertions (independence maps), which is what I will be testing, so I must avoid duplicating independence assertions to prevent the "best independence match' from being the one that best matches my *assumptions.*

In reading observations, the only assumption that I will make is the Markovian one – that all observations within the window size Q are independent of those that came

before. While this is a partially fair assumption (that will be broken by unpartitioned and free reading orders), in a general case it may not be a valid assumption. It is entirely possible for long distance (and temporal) dependencies to exist within the sequence, that are outside of capture by the window size. Secondly, 'state dependence' can be seen as an unknown width of composited observations which may not fit within the window size Q, regardless of whether we allow grouped observations (i.e. "AB" ) to be read into a single distribution. So this assumption basically excludes long distance dependencies or larger compositions of state from being captured.

The experiment that will be run here is a calculation of the maximum Kullback-Leibler (KL) divergence (relative entropy) between the read distributions. While Kullback-Leibler is not symmetric between two distributions $P_1$ and $P_2$, when KL $= 0$, $P_1 = P_2$. In this case, if a KL divergence converges on 0 this will tell us that the reading stategy being tested (partitioned, ordered, free) has a diminishing 'distinguishing' information content between all of the reading distributions. (In general, $D_{KL}(P_1|P_2)$ tells us the 'information', or additional bits needed to encode an observation from $P_1$ when $P_2$ is used) [25].

$$P_a, P_b \in \{D_{0...Q-1}\} : \max D_{KL}(P_a||P_b) = \max \sum_E P_a(E) log \frac{P_a(E)}{P_b(E)} \tag{3.3}$$

$D_{KL}(P_a||P_b) = 0$ does not mean that there is no information *within* each distribution or that the entropy within a distribution $P$ has been maximized, only that two separate distributions are effectively equivalent.

### Partitioned (Aligned)

In testing an aligned method of reading observations into distributions, the observation stream is chunked into aligned sequences of size Q, and the observation at each position is read into a distribution at that position.

Consider sequence 3.2, repeated here and sequences of size Q=3 mapped into distributions (D) $m$, $n$, and $o$ (labeled for convenience):

| D: | m n o | m n o | m n o | m n o | m n o | m n o | m n o | m n o |
|---|---|---|---|---|---|---|---|---|
| O: | A B C | B B A | A B C | B A A | A B C | A B C | B B B | B B A |

Then distribution $m$ contains 4 counts of A, 4 of B, and 0 of C. Distribution $o$ contains 4 C's, 3 A's, and 1 B.

This is similar to what most statistical natural language processing, or simple Bayesian usage examples are. The distributions have been named and pre-assigned, separating the observations right at the start.

### Ordered

In 'partitioned', I arbitrarily enforced a strict mapping of source distributions to the observation sequence. But testing unpartitioned reading of the observation sequence is the very first form of "uncertainty" to handle – that we do not know what 'state', 'input' or 'distribution' an observation belongs to, or where to start! In many experiments, this information (the target distribution) is not available; outcome observations may be intermixed within the observation sequence directly, and novel observations may not be named and could come from any distribution. In this test, the mapping of the distributions to the sequence is assumed to be ordered, but not partitioned. In many ways, I expect it to be like listening to music – if I start the song in the wrong place, it should take longer to 'get'.

Consider the same observation sequence, size, distributions ($m$, $n$, $o$), and different positional mappings $p_1$, $p_2$, $p_3$:

| O: | A B C | B B A | A B C | B A A | A B C | A B C | B B B | B B A |
|---|---|---|---|---|---|---|---|---|
| $p_1$: | m n o | m n o | m n o | m n o | m n o | m n o | m n o | m n o |
| $p_2$: | m n | o m n | o m n | o m n | o m n | o m n | o m n | o ... |
| $p_3$: | m | n o m | n o m | n o m | n o m | n o m | n o m | n o ... |

In this case, when the exact delineation of rule observations ABC is not known, each of the distributions $m$, $n$, $o$ derive some evidence from each position offset. The problem here is that the joint probabilities of $m \cap n$, $n \cap o$, and $o \cap m$ will begin

to converge. The contribution of the evidence at each position into distributions $m$, $n$, $o$ will be made uniformly (though any weight choice is possible). Another way of looking at it is in the size of overlap of the read stream. Each chunk of 3 is read after incrementing the read index by 1 – meaning that there is an overlap of 2.

Reading strategies may thus be seen in terms of 'overlap' of the number of distributions that each observation goes into. In the completely aligned case, there is an overlap of 0 and each observation goes into a single distribution. In the Ordered, but unknown case, there is an overlap of 2, and each observation goes into 3 distributions (with the exception of the endpoints). Finally, it is possible to choose an overlap of 1, where some observations go into 2 distributions.

#### Free

In a free assignment of observations, any observation may go into any distribution (optionally with a weight). Because this can yield inconsistent reading orders (and thus any dependency model will have no meaning), and because it seems pretty obvious that as $O$ gets longer, the distributions will converge, I will not use this method for reading. I believe that the trend suggested between aligned and ordered (unpartitioned) will be sufficient.

### 3.4.2 Dependency Models

In the selection of the most likely independency model $M_{ml}$, the parameters being optimized for are the independency between the positional distributions obtained from reading the observation sequence. The discrete distributional shape is unknown a-priori.

Each independency model $M_g$ will represent a 'hypothesis', and all possible independency models of size $Q$ will be enumerated. For hypothesis selection, there are a number of other types of models/hypotheses to enumerate – for instance, one could enumerate all possible deterministic rules. But one of the major goals of this experiment is to 'capture uncertainty', not deterministically succeed or fail (resulting in a

log-Likelihood of $-\infty$).

### Bayesian Hypothesis Selection

In Bayesian Hypothesis Selection, I enumerate all possible independency models $M$ as the 'hypotheses' within the available parameter space of size $Q$ ($\{M\}_Q$)and select the model $M_{ml}$ that is most likely to have produced the observation sequence $O$. Formally, we leverage Bayes rule and choose an individual model $M_g$ that maximizes the probability of the observation sequence.

$$P(M_{ml}) = \max_g \left( P(M_g|O) = \frac{P(O|M_g) \cdot P(M_g)}{\sum_g P(O|M_g)P(M_g)} \right) \tag{3.4}$$

The marginal probability ($\sum_g P(O|M_g)P(M_g)$) is the total probability that a model (over all $M$) of size $Q$ generates the data $O$. For a fixed $Q$, the marginal probability acts as a constant normalizing factor $\alpha$ (this cannot be done if $Q$ is varied & compared), reducing the selection of the most probable model to maximizion the posterior and prior probabilities:

$$\max P(M_g|O) = \max P(O|M_g) \cdot P(M_g) \cdot \alpha(Q) \tag{3.5}$$

### Probability of Observation Sequence Given Model

Generally, the probability of generating a particular sequence given a Model is commonly referred to the probability of the Data $D$ given a model $M_g$, which breaking it down into non-overlapping independent observation window subsequences of size Q, is the product of the probabilities of N/Q subsequences in sequence O given the model.

$$P(D|M_g) = \prod_{a=0}^{N/Q} P(O[Q \cdot a : Q(a+1)]|M_g) \tag{3.6}$$

Given that all dependency orderings are allowed within the window (none outside), the probability of any subsequence from $O_a$ to $O_{a+Q}$ is the product of the probability of each individual observation given the values of all other observations related by

the conditional independency model $M_g$. Some observations will be independent according to the model and others will not.

$$P(O[Q \cdot a : Q(a+1)]|M_g) = \prod_{i=0}^{Q} P(D_i = O_{Q \cdot a+i}|M_g, D_{\{0 \ldots Q \cap \bar{i}\}}) \tag{3.7}$$

**Priors**

Priors are generally considered the "a-priori knowledge" or "belief" that selects one model over another more quickly, or if all other evidence is equal. Naively, I may place an increased prior on a class of models drawn from $M$ (for instance, 'trees'), counting on the prior to return correct results early in the observation sequence. But because priors are in direct contention with evidence, they only change the rate at which additional evidence is needed to select (or *overcome* a bad prior) individual models, but they do not actually change the knowledge being gained. Experimentally, they may actually lessen it – by choosing a prior of unprincipled value, information is lost about the relative rates of convergence for any particular model over another. Secondly, this experiment is about the "best case" "lack of knowledge" performance of probabilistic dependency identification. In a real-world scenario, it is unlikely that we would know the structure of the model – that it is a tree or otherwise, so placing an increased prior on one model class is not justifiable. Finally, it is may be the case that a particular graph structure has an over-representation in the hypothesis space (class of models) [6], but this does not actually change the selection of the most likely model – each individual $M_g$ is in competition with every other one.

PRIOR: Uniform.

With a uniform prior over M, $P(M_{g_1}) = P(M_{g_2})$ and the $P(M_g)$ becomes another normalization factor $\beta(Q)$ – a function of Q size, reducing the calculation of the distribution probabilities for models $M_g$ given $O$ (within a fixed Q) to:

$$P(M_g|O) = P(O|M_g) \cdot \alpha(Q)\beta(Q) \tag{3.8}$$

or:

$$P(M_g|O) \propto P(O|M_g) \tag{3.9}$$

The loss due to these simplifications is the ability to compare models of differing sizes – where the total probability that a model of size Q generated the data, or the prior, despite being uniform, is diminishing with model size Q.

**Enumerating Hypotheses: Models $\{M\}_Q$**

Up to this point we've expounded on the selection of the most likely model from $M$, but not what those models are, or where they come from.

When the state size S is fixed at 1 (elemental observations), Q is both the window size on the observation sequence *and* the number of distributions. The individual distributions can act as nodes in a graph, between which we can draw arrows that represent conditional independence assertions. For Bayesion model selection and two distributions, $D_1$ and $D_2$, there are 3 kinds of arrows: $D_1\{\rightarrow, \leftarrow, \nleftrightarrow\}D_2$ (disregarding the cyclic $\leftrightarrows$). In otherwords, $D_1$ is conditionally independent given $D_2$, $D_2$ is conditionally independent given $D_1$, or $D_1$ and $D_2$ are independent (or conditionally independent given some third node). In chapter 2, some problems with this representation of independence were exposed, but at this point, for consistency with other works I will retain this notation.

To enumerate models in M, we consider a set of nodes size $Q$, consisting of $N = \{n_1, n_2, \ldots, n_Q\}$:

1. Choose all possible 2-combinations from $N$, call this new set: $\{N_2\}$ – it will have size $\binom{N}{2}$. For 4 nodes (Q=4), this is 6.

2. Create trinary table of size $3^{\binom{N}{2}}$, where each entry represents a graph, and each value a direction/type of edge. For Q=4, this is 729. This is the model space $\{M\}_Q$ (Table A.1), and does not include first order cycles $\rightleftarrows$ (the table size would be $4^{\binom{N}{2}}$).

For an example graph/model enumeration, see Appendix A.1.

Removing Cycles

The enumeration of all possible graphs creates a number of cycles, which present problems for Bayesian model selection [28]. I use the following strategy to detect cycles and prune those graphs from $\{M\}_Q$.

```
foreach model Mg in M
   while there is a node with no children
      delete node and edges pointing to it
   if |nodes| greater than 0
      cycle = true
      remove Mg
```

### Best Fit Observations

Finally the set of 'most probable' models ($\{M_{mp}\}$) (which contains $M_{ml}$), equivalently captures the most frequently surfacing independency-relationships within the sequence $O$. This basically tells 'how to read' the dependencies within $O$ in the 'most consistent' way. I would like to know what the 'best fit' observation sequence is to the selected dependencies – what are the relations between actual observations that are 'best captured'? This can be seen to be a prediction of the model back onto the world. Does the best fitting independency relationship predict a different set of outcomes than the deterministic rule?

$$\max P(O_{x_i}|M_g) \tag{3.10}$$

```
foreach obs-comb-map (Oxi) in (unique-obs choose Q):
   compute P(Oxi|Mg)
return list of obs-comb-map satisfying MAX P(Oxi|Mg)

% (A number of top equiprobable mappings are possible)
```

This will become important to see what kinds of sequences the observation-dependencies map prefers to generate. For instance, a distribution-dependency map

of $(D_1 = A) \rightarrow (D_2 = A)$, if selected as the top deterministic rule $(A \rightarrow A)$ will generate the infinite sequence in 3.11.

$$\mathbf{AAAAAAAA}...\infty \tag{3.11}$$

## 3.5 Computing Correctness

To determine correctness of the selected models, a number of tools will be used depending upon the experiment run. I would like to choose an evaluation scheme which is consistent with the use of probabilities. It is easy to say "yes" or "no" that a singly returned model is exactly correct, but I'd like to know what the "likelihood" is of selecting the correct model over is, and "how correct" it is. Finally, I'd like to answer if the removal of probabilities in the top model selection results in a deterministic model with degenerate properties, incapable of reproducing the data (infinite loops, etc.)

### 3.5.1 Edge Overlap (O)

Computing edge overlap is in many ways, the least informative (and very nearly asinine) of all possible metrics for computing correctness. It does however, mimic what I believe to be the "lazy-human" evaluation of correctness – that readers gloss their eyes over and casually glance at a selected model and say "that more or less looks correct", without actually carefully evaluating dependency structure. It only tests that an edge *exists* where it is supposed to, not direction, and only inside the gold model. Meaning, it only tests the percentage of overlap on the gold model against the selected model.

```
foreach selected model SM:
   overlap[SM] = 0
   for pair nodes in comb(gold_nodes):
      nodes connected in both Gold and SM:
         overlap[SM]++
      nodes not connected in both Gold and SM:
         overlap[SM]++
   overlap[SM] /= total possible Gold edges
```

Because a number $n$ of equiprobable selected models can be identified, in keeping with the approach of uninformed identification, the true overlap likelihood is the *expectation* of overlap $E_o$:

$$E_o = \frac{1}{n} \sum_{1...n} O_n$$

```
expected_overlap = 0.0
foreach selected model SM:
   expected_overlap += overlap[SM]
expected_overlap /= total number selected models
```

This metric provides the likelihood of uninformedly selecting the correct model from the identified models.

### 3.5.2 Edge Direction

The second test of correctness – test all of the edges within the gold model to identify what proportion have the correct direction. This will be tested over the entire observation sequence O to determine what proportion of the returned models are completely directionally correct (as a function of $n$).

```
foreach selected model SM:
   direction[SM] = 0
   for edge in comb(gold_nodes):
      edge-direction, Gold == SM:
         direction[SM]++
      edge not connected in both Gold and SM:
         direction[SM]++
   direction[SM] /= total possible Gold edges
```

### 3.5.3 Convergence

I would like to graph the convergence to selection of single model as additional observations are added. Convergence will provide insight into sensitivity of model selection, as a function of overlap with the gold model. For instance, does the model selection converge on the right or the wrong model as observation length goes up to 100,000, and is it that selection stable?

## 3.6 Simple Example

Of the listed tests for model building, this test uses Bayesian Networks, and the goal is to end up with graph that nominally contains (A $\rightarrow$ C $\leftarrow$ B) I'd like to determine what the "most probable" model of observation states is, knowing no parameters ($\theta$) or prior structure. This requires me to test parameter counts of N=$\{1, 2, 3, 4\}$, where N=3 is matched to the number of unique observations, 4 assumes 1 hidden node/param, and N=1 param offloads all unique observations to a distribution. I'll use 3 params and unaligned reading for this test case.

Considering 2 models, 3 parameters each $\theta_{n-2} = m, \theta_{n-1} = n, \theta_n = o$:

$M_1$. (m n o) (all independent)

$M_2$. (m$\rightarrow$n o) (one edge)

Mapping to the input stream:

```
m  n  o
-------------------------
A  B  C  A  B  C
-------------------------
   m  n  o          (θ moving down the stream)
```

Each model generates different conditional probability tables (CPT).

For Model 1 ($M_1$): All independent, the CPTs for all parameters are:

`P({m,n,o}={A,B,C}):` `= 1/3` each value

For Model 2 ($M_2$):

$o$ is independent, as is $m$, but according to the graph, $n$ is conditionally dependent on $m$, so the produced CPT is:

```
P(o) = { A,B,C } = 1/3
P(n|m) = { P[n|m=A] , P[n|m=B] , P[n|m=C] }
     B 1.0        B 0.0     B 0.0
     A 0.0        A 0.0     A 1.0
     C 0.0        C 1.0     C 0.0
P(m) = { A,B,C } = 1/3
```

Calculating $P(M_g|O)$ for Models 1 & 2:

Model 1 ($M_1$):

$P(O|M_1)$ = probability of this particular sequence (there are 3 of them), when selected from the random variables.

$= (P(m = \texttt{A}) \cdot P(n = \texttt{B}) \cdot P(o = \texttt{C}))^3 = (1/3 * 1/3 * 1/3)^3$

$P(M_1|O) \propto P(O|M_1) = \mathbf{(1/27)^3}$

Model 2 ($M_2$):

$P(O|M_2)$ = The probability of the sequence for two independent distributions (m & o) and one dependent (n).

$= (P(m = \texttt{A}) * P(n = \texttt{B}|m = \texttt{A}) * P(o = \texttt{C}))^3 = (1/3 * 1.0 * 1/3)^3$

$P(M_2|O) \propto \mathbf{(1/9)^3}$

This indicates that Model 2 (m→n o) is a more likely fit than Model 1. Mapping the variables to the model yields M2 = (A → B C). It can be seen right away that as O increases in length, and the reading method is *unaligned*, (A→B C) becomes equally probable a prediction for Model 2 as (B→C A).

## 3.7 Alternate Methods & Experiment Limitations

There are different ways of approaching this problem, and this experiment is incomplete in a number of dimensions. Most notably, testing the model selection within a complete space of an expanding window size, all the way up to Q=length of observation sequence will not be done.

In terms of the performance measure being maximized (over potential dependencies), the overall Likelihood is not necessarily the best measure – the hit rate – the ratio of correct predictions to false positives provides a better method to rank models in terms of their mappings to actual events/observations within the sequence. In Signal Detection [24], maximizing the posterior ratio against the false positives is proposed, as a method of discriminating rule against noise (d-prime), which would provide a measure for whether I am capturing rule occurrences or noise (non-rule). This is also somewhat similar to Valiant's method [56], where the identification of the discriminative program or vector (in disjunctive normal form) can be achieved via the use of an oracle.

While for this experiment, I will not be using Hidden Markov Models (due to time and space constraints), it is important to make a few statements about them. In this experiment, there are two obvious places where they could be used. The first is in reading observations from O – there could be 2 sequences: Noise sequences (N) and Rule (R). Accumulating the transition probabilities is done to determine which observations are relevant, and which observations are not. This could achieve 2 purposes – alignment of the observations into distributions, and boosting the positive observations. Identifying the transition probabilities between observations also seems to be a more 'natural' way to model implied state transitions observation to observation.

# Chapter 4

# Results & Analysis

In this section I describe the results obtained from the multiple experiments, and then I seek to explain what they mean in terms of knowledge representations, dependency recovery, and overall model building.

## 4.1 Reading Observations

The results from the attempts to read the observations into distributions are confusing, but are necessary to comprehend the results of Model Selection. This section presents the results from different reading strategies using Model 2 ($AB \rightarrow C$). Model 2 is the simplest rule being evaluated, so the reading strategy we wish to select is one that maximizes the information available for the model selection step.

As touched upon in section 3.4.1, the maximum KL Divergence $\max_{P,Q} D_{KL}(P||Q)$ between the read distributions provides a measure of 'information content' or 'distinguishability' [25] contained between the different distributions that observations are read into. For all strategies of reading that either overlap, or in which the rule's appearances are not specifically aligned to a known distribution (Fig. 4-2 b, c), the $\max_{P,Q} D_{KL}(P||Q)$ drives to zero – meaning that all reading distributions are converging and becoming equal – the differentiating information content is disappearing. Introduction of an additional reading "uncertain" parameter (size = 4) also drives convergence, as that parameter disrupts alignment and knowledge of the distribution into which a new observation should go.

It is of note that even though the 'uncertain' distributions converge in the first order,

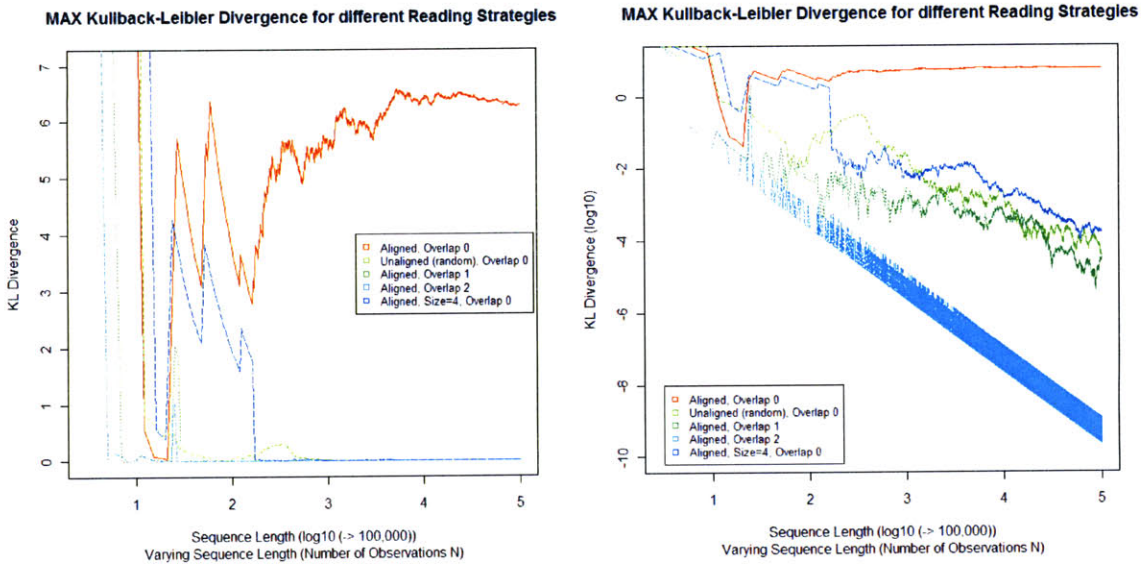
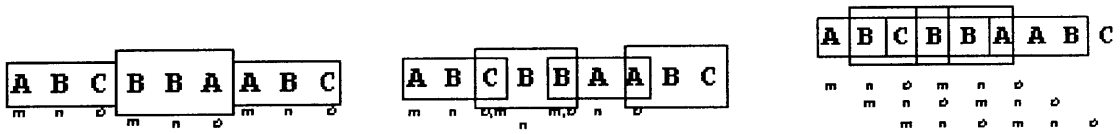



Figure 4-1: a) The maximum Kullback-Leibler Divergence between the distributions that the observations are read into, up to 100,000 observations ($log_{10}$). All distribution Divergences drive to 0 (converge) for all reading strategies except the completely aligned case. b) The same graph, scaled $log_{10} - log_{10}$, zoomed to -10 ($10^{-10}$) – The relative rates that the maxmum Kullback-Leibler Divergence drives to zero for the "uncertain" methods of reading.

having the same proportions of A, B, and C, that does not mean that they are independent. In fact, it is entirely possible that 3 distributions have exactly the same contents, but be non-independent – i.e. they could have joint probabilities of 1.0. $P(m = x \cap n = x) = 1.0$ etc. However, we know from the high variance in the KL divergence for small length observation sequences (small number of observations) that the distributions do not have a joint probability of 1.0 a-priori.

Unaligned reading may be the general case for reading and interpreting new observations, but aligned provides the greatest information content between distributions, and if there is *any* uncertainty in the way observations are read, then it drives distribution convergence. The more uncertainty in the process of reading – as in "uncertainty into which distribution an observation belongs" – the faster the distributions converge and the faster information content drives to zero (Fig. 4-1 b).

This test does not cover or enumerate *all* reading strategies, which can be made arbitrarily complex, but I believe it to cover the general case where an observation either overlaps and could be deposited into any distribution or does not. If it is the case that one deposits




a) b) c)

Figure 4-2: Reading Strategies for observations into distributions $m, n$, and $o$. a) Aligned, Overlap 0 – Every observation is uniquely assigned to one distribution. b) Aligned, Overlap 1 – Every other observation contributes to 2 distributions, except end-points. c) Aligned, Overlap 2 – All observations contribute to all distributions, except the end-points.

an observation into multiple distributions (lets say $D_1, D_2$), then as $D_1$ and $D_2$ converge, any desired dependency between a different distribution $D_d$ and $D_1$ would be duplicated on $D_2$. If *all* distributions converge (as they do in all the above reading strategies), then all dependencies would be duplicated between all distributions.

This suggests that if a person does not 'know how to read', interpret, or separate (classify) observations a-priori, then the use of reading distributions and probabilities will only guarantee a lack of information content. Moreover, in this experiment if the reader chooses to try all reading strategies simultaneously, choosing the one which maximize $D_{KL}$, then it is not until around the 200'th observation that one strategy is appreciably superior (Fig. 4-1 b). For less than around 50 observations, at least 3 strategies intermingle and the aligned strategy is not the superior one.

Overall, this shows us that if one ignores the 'phrase structure' of each observation-subsequence triple, then the information is being degraded. In terms of Natural Language, it shows us that the periods and commas in a sentence are important, not just as another observation-token to be read into a distribution, but to *align* tokens into the correct distributions. From an ontological perspective, it means that the branches whos nodes serve as the observations must also have an a-priori known span size to assign them into correct distributions as well. Finally, from a 'causal'-'real-world' mapping (not to be confused with probabilistic 'causal' reasoning), it shows us that if one uses probabilities for 'learning' in the 'real-world', the 'codes' or the 'features' that are communicated from the real world to the observer have to be agreed upon a-priori. Right away, the total size of the dependencies, width and depth, have to be known at the start. In other words, if he is using probabilities, the learner (or the observer) has to already correctly *know and recognize* those items that he is meant to learn.

## 4.2 Model Selection

Given that all distribution contents converge in this experiment for all "uncertain" methods of reading, Model selection will be conducted with the *aligned* data source and method of reading observations, which 'maximize' the 'information' between the distributions. This is logically equivalent to if all inputs and outcome observations are separated a-priori. This already adds one dimension in which our 'veil of forgetfulness' has to be violated.

### 4.2.1 Model 1: Independent (No Rule)

All of the models that follow Model 1 embed a deterministic model within the random observation sequence. Model 1 however contains no rule whatsoever, and the pseudo-random observations are generated independently. The test is if model selection can identify the fully independent reading model (dependencies between the positional-distributions that the observations are read into), rather than test if it can predict correct output dependencies between the observations themselves.

#### Identified Models

An example of the identified read model after 100,000 points is contained in Figure 4-3.

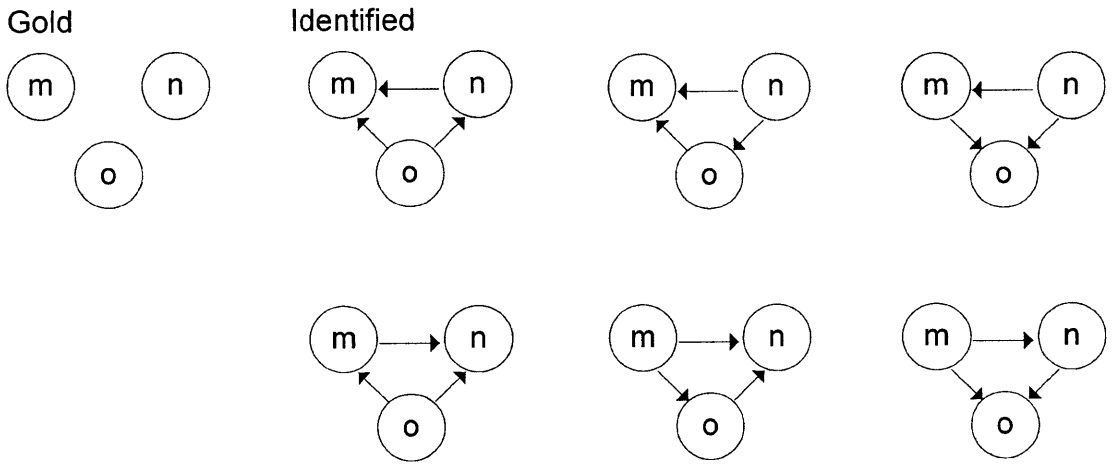



Figure 4-3: After 100,000 points, all 6 fully connected graphs were identified as equally likely representations of the dependencies between the read-distributions.

The 'reading' model is first selected, before mapping to physical observations, and given the aligned observation order, the distributions correspond exactly to positions within the

sequence, and thus the selected model to a positional corellation.

The identified model which has the highest relative likelihood is *not* the fully *independent* model. While most analysis of these results is being held off until the next section, I feel that it is important to partially explain the correctness of this particular result.

Two random variables $X$, and $Y$ are considered to be independent if the conditional probability of one variable given the other is equivalent to the joint probability.

$$P(X|Y) = P(X)P(Y) \tag{4.1}$$

In 'real' data, it is nearly impossible to achieve a the perfect sample where the independent probabilities are exactly equal, and because of that, conditional probabilities 'carry' higher probability density than independent-joint probabilities – the likelihood produced by a conditional probability will always be higher (or more preferred) than that produced by a joint probability, except in the case where the random variables are truly independent, when the joint and conditional probabilities become equal (thus equivalently likely in the identified models). Though we can see them converging in figure 4-4; even after 100,000 observations, there is still enough of a difference to prefer a model of conditional independence. This result is supported by Jaynes, where model selection that is not augmented by prior knowledge (a non-uniform prior) will always prefer the surer thing [27].

For future tests, the positional dependency model between the distributions will be mapped back to observations to produce and test the 'observation dependency' model.

### 4.2.2 Model 2: Simple Rule (Converging Dependency)

Model 2 is an interesting model, and there are a couple of ways of looking at it: that it contains multiple strict dependencies, that it has a 'distance' dependency, or that it exhibits the case of multiple correlations.

The relative rates of Model 2's rule appearance within the data are shown in Table 4.1. The sequences of ABA and ABB are completely replaced by ABC. By 100 points, ABC is not the most frequent observation – AAA is, but by 100,000 points ABC's rate of appearance is greater that (approximately double) the rates of the other sequences.

Because what we are working with is relative likelihood, the relative rate of approximately double should be sufficient to 'prefer' a model capturing A, B, and C. However,

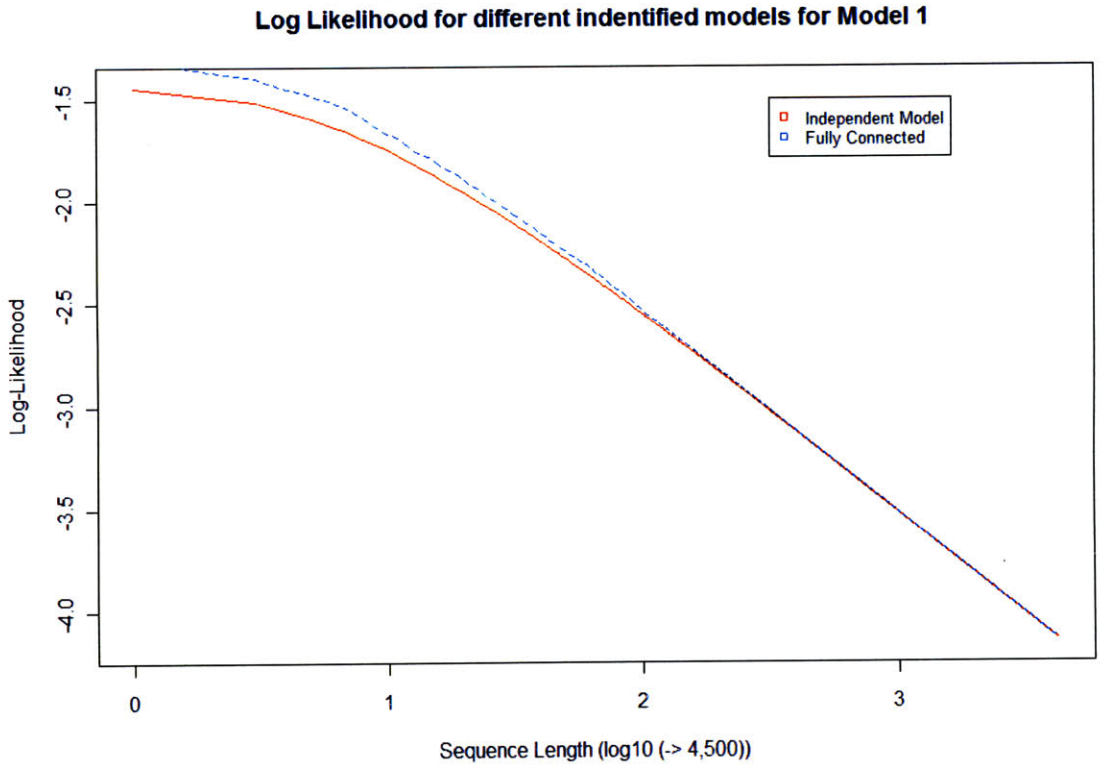



Figure 4-4: Graph comparing and demonstrating the convergence of the log-likelihood of an identified model (fully connected) to the correct 'independent' model in Model 1. This graph is on a $log_{10}, -log_{10}(-L)$ scale. Higher is 'more likely' – the fully connected model is preferred completely to the independent model.

if one considers all of the non-ABC sequences to be noise, the total probability of those sequences is 3/4, which is greater than the ABC probability.

## Identified Models

After mapping back to observations, and choosing from the full class of model hypotheses, the most likely identified models can be seen in Figure 4-5. Six models are identified as equiprobable, and all the observations are connected, implying that there is no independence of state. Models that represent C as the outcome state or the head state are represented twice each, as are models where C is an intermediary in the dependency path. However, the correct representation of the outcome of C is still accompanied by the false positive edge between A and B.

| m | n | o | $f_{100}$ | $rate_{100}$ | $f_{100,000}$ | $rate_{100,000}$ | exact $P$ |
|---|---|---|---|---|---|---|---|
| A | A | A | 6 | 0.182 | 4199 | 0.126 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| A | A | B | 5 | 0.152 | 4306 | 0.129 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| **A** | **B** | **C** | 5 | **0.152** | 8247 | **0.247** | $\frac{\mathbf{1}}{\mathbf{2}}\frac{\mathbf{1}}{\mathbf{2}}\mathbf{1} = \frac{\mathbf{1}}{\mathbf{4}}$ |
| B | A | A | 5 | 0.152 | 4093 | 0.123 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| B | A | B | 4 | 0.121 | 4066 | 0.122 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| B | B | A | 5 | 0.151 | 4246 | 0.127 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| B | B | B | 3 | 0.091 | 4176 | 0.125 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |

Table 4.1: Relative frequences and rates of rule appearance within the data for Model 2 vs. all other unique sequences, after read into distributions (m, n, and o) using an aligned strategy. $f_{100}$ is after N=100, and $f_{100,000}$ is after N=100,000. The exact probability of appearance of the sequences A,B,C in the distributions. C appears in distribution $o$ IFF both A and B appeared in $m$ and $n$, otherwise A or B are selected randomly with probability 1/2.

Given the results from the fully independent model, it is of little surprise that the fully connected model is preferred over a heirarchical tree model that separates A and B. What is a surprise is that *all* six fully connected models from the size=3 model space capture the observations with equivalent likelihood. This means that when the 'fit' between the independence model and the data is maximized, all knowledge concerning directional dependency, independence, and overall structure is lost.

To retrieve model information, lets see what adding a little more certainty can do. If we bend the veil of forgetfulness a little more and restrict the tested models to the class of trees (adding knowledge about the class of models that the desired model belongs to), then the identified model is substantially better (Fig. 4-6).

LOG-LIKELIHOOD: -63596.19 (correctly identified model, tree restriction)

The correct dependency structure is identified right away and does not fluctuate much relatively (Fig. 4-7). This happens so quickly that it is almost uninteresting in its efficiency. Personally, I really did not expect this to happen, largely because graph 'dependency' implied by arrow directions for probabilistic models is not about dependency at all, but about conditional independency. But in this case (with this constraint), the arrows of the conditional independence graph do overlap with my declared dependencies.

Partly this is because the restriction that the model is drawn from the class of trees is a fairly strong restriction – it is an assertion that in the size of 3, at least two of the position

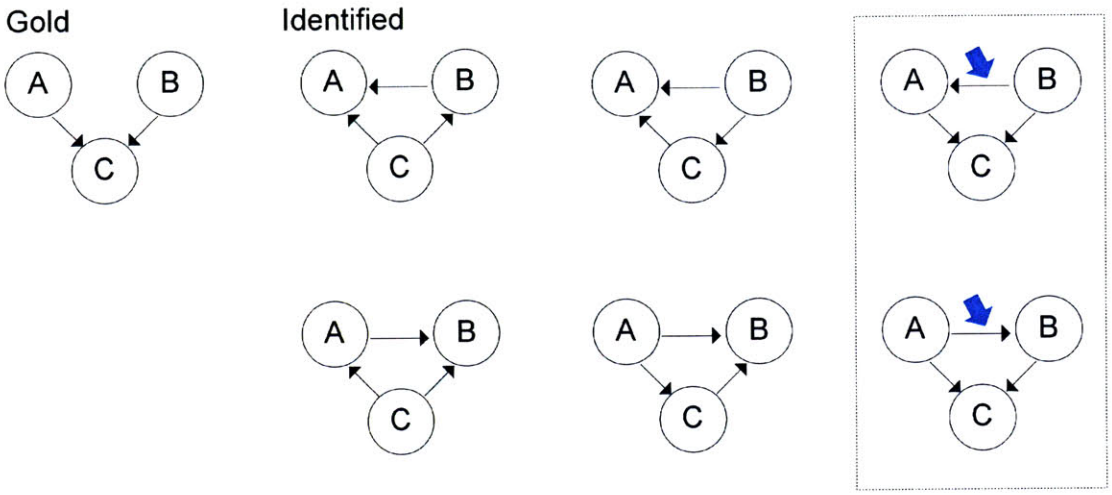



Figure 4-5: Model 2 - Identified most likely models from the full class models after an observation length 100,000. All 6 fully connected models are identified as equiprobable. In the two models containing the correct dependency direction, an extra dependency exists between A–B.

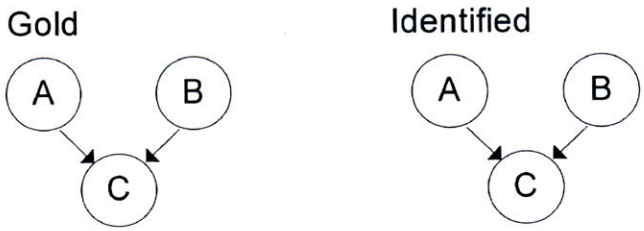



Figure 4-6: Model 2 - Identified model restricted to the class of trees, after observation length 100,000. The correct model is identified.

distributions are independent in *some way* (whether with or without information). Beyond that, in the acyclic models, there are 6 fully connected graphs which are removed directly. Secondly, model selection universally prefers more conditional edges to less, which means that only partially connected graphs (fully disconnected, and single disconnection) of a total of 7 will be removed as well. This means that the 'tree' restriction reduces model selection to a class of 12. By the same token though, the model selection without a restriction (from the full set) is limited by the preference of greater connectivity, to the 6 fully connected models.

Secondly, the interpretation of the identified conditional probability graph is somewhat counter-intuitive – it effectively states that "A and B's appearance are independent in the

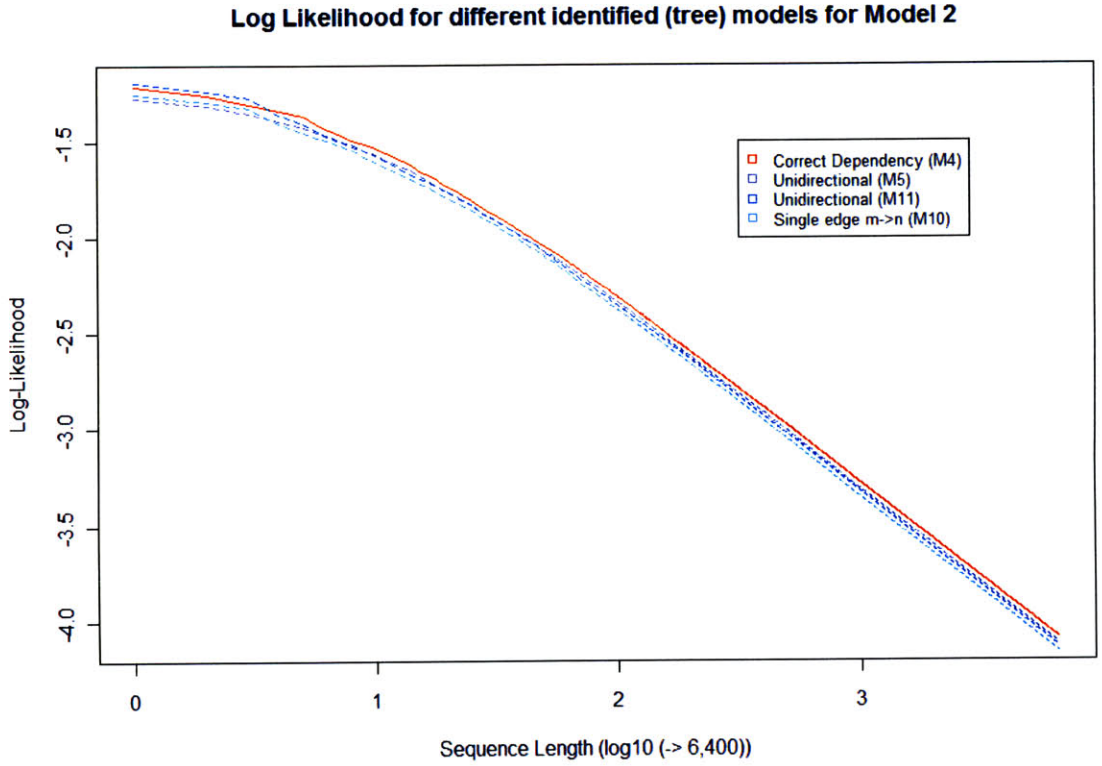



Figure 4-7: Model 2 - Log likelihood of different tree dependency structures. Model 4 (solid line) is correctly identified very quickly, where it surpasses the unidirection Model 11. Plotted on a $log_{10}, -log_{10}(-L)$ scale.

data, but given knowledge of C, it may not be independent". This is actually correct for the Model 2 experiment, but it is not a statement of dependence of C on A and B – it is a statement about the effect that the information that C has occurred upon the information about A & B's relative appearance. Because mapping to A, B, and C takes place after model selection, what it really says is that 'position-distribution $m$ is independent from $n$, but not necessarily independent from each other if a value of $o$ is known'. If $o$ is unknown, unobserved, or unread from the sequence, then $m$ and $n$ are independent, which given a perfect sample, would be true. With our real data, it would prefer to connect $m$ and $n$ due to small variations from the expected values of $m$ and $n$. It should be noted however, that the observation of $\{m, o\}$ or $\{n, o\}$ are *not* independent, so when *forcing the choice* between different independence maps each containing at least 1 independence, the selection of $\{m, o\}$ and $\{n, o\}$ as being connected is a far better fit than $m$ and $n$.

If we analyze the relative rates of appearance, considering first $o = C$, we'll see that

$P(m = A, n = B|o = C) = 1.0$. And looking at A and B's conditional independence given $o = C$, we'll see that the $P(m = A|n = B) = 1.0$, which is the same as $P(m = A) \cdot P(n = B) = 1.0$ (complete enumeration of all observations for m and n given $o = C$). In other words, 2 points are evident: in model selection, the independent and the conditionally independent model should both be identified as equivalant; second, their equivalence means that $m$ and $n$ are independent given $o = C$. And, in exact probability, $m$ and $n$ are also independent. (i.e. $P(m = A) = \frac{1}{2} = P(m = A|n = B) = \frac{(1/4)}{(1/4+2/8)}$) In actual rates however, at $f_{100,000}$ : $P(m = A) = .502$ and $P(m = A|n = B) = .495$ are not equal and not independent. It is also of note that the model selection is benefitting from a *uniform* distribution of A and B. With a non-uniform distributions, it may be more difficult to demonstrate the independence of A and B's appearance within the read-distributions $m$,$n$, and $o$.

## Discriminating

In trying to understand why the dependency model identified for Model 2 appears correct, we can calculate the probability of rule P(R) (Table 4.2) as captured by the different conditional assertions of the different graphs. One can see that the because the rule is deterministic, the probability of the sequence ABC ($P(R)$) is 1/4 for several markedly different models. This means that it is not the probability of rule, or how well the model replicates the rule in its overall Likelihood of the data ($P(D|M)$) that governs model selection, but it is the proportion of alternate (non-rule) observations that do. In Table 4.3, the displacement ratio, or rather, the new expected values for the non-rule observations given the outcome $o$ are different and unequal, which means that $M_8$ ($P(mn|o)$) is eliminated becauses for non-rule observations surfacing in $o$, $m$ and $n$ are not independent.

This is an important observation, because even though the relative rate of rule appearance is higher than any other subsequence, it basically means that in this experiment, noise drives model selection – in an experiment that is 'open' either to new observations, new knowledge structures, new phrase structures, the 'noise' may effectively spread to uniformity, leading to loss of the identified dependency structure.

By extension, if the rule is the only observation sequence available,

$$ABCABCABC...$$

| | $m$ | $n$ | $o$ | $p(m)$ | $p(n)$ | $p(o\|m \cap n)$ | $P(R)$ |
|---|---|---|---|---|---|---|---|
| $M_4$ | A | B | C | 1/2 | 1/2 | 1.0 | 1/4 |
| | $m$ | $n$ | $o$ | $p(m\|o)$ | $p(n\|o)$ | $p(o)$ | |
| $M_8$ | A | B | C | 1 | 1 | 1/4 | 1/4 |
| | $m$ | $n$ | $o$ | $p(m)$ | $p(n\|m)$ | $p(o\|n)$ | |
| $M_{10}$ | A | B | C | 1/2 | 1/2 | 1 | 1/4 |

Table 4.2: When comparing the various tree dependence models for their probability of rule it can be seen that the probability P(R) ($\frac{1}{4}$) is equivalent across several models.

| | $E(m\|o)$ | $E(n\|o)$ | $o = \{A, B\}$ |
|---|---|---|---|
| A | 1/3 | 2/3 | |
| B | 2/3 | 1/3 | |

Table 4.3: The displacement ratio of expected values for $M_8$ and $o$ = A or B (not C, which is the rule). The original ratios for $m$ and $n$ are both 1/2 A, B each.

and there is no variation, then *all 25 models* are as identified equiprobable with $P(R) = 1.0$.

This observation helps to suggest that there may be a 'range' over which the the proportion of rule observations, from 0.0 – 1.0 affects the models that are identified. Also, given that restriction on the space of models seems to achieve a correct dependency model for Model 2 suggests that the use of a non-uniform prior would yield desired results while maximizing prior knowledge. However, the combination of the preference towards fully connected models along with the non-discriminating probability of rule shows that for this experiment, no non-uniform prior choice can be justified. It will always be overwhelmed by the distribution of noise and converge towards a fully connected (non-independent) model.

### Sparsity (Widening observation space)

Restricting the class of models being evaluated to trees, and knowing that it isn't the probability of rule that discriminates one rule from another, lets see what happens when the space of non-rule is widened and sparse, as integrating novel observations or a (nearly) unlimited supply of new rules/structures (that we have not yet accounted for) might accomplish in our model. To accomplish this, the exact same observation sequence will be used, with the exception that *all* non-$AB \rightarrow C$ sequences will be randomly replaced with one of 89 characters from the ascii character set (not including

{A,B, or C}). This reduces the availability and variability of the {A,B} sequences possible, so their use to indicate whether the read models $m$, and $n$ are independent will be reduced.

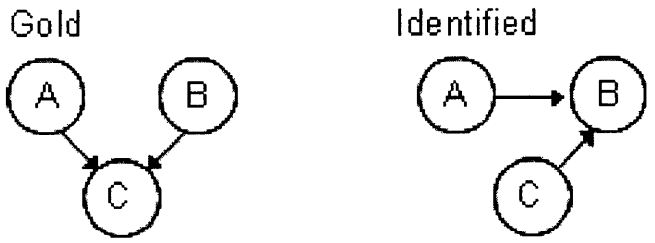



Figure 4-8: The dependency model identified for a spread out noise space (non-rule sequences), restricted to trees. A converging dependency model is selected with B as the dependent node.

Overall, this is the same as surrounding the rule with a large volume of novel, but irrelevant evidence. Because the frequency of rule sequence is the same as for the other model 2 tests, this helps to visualize what happens when more information is added that is unaccounted for.

### 4.2.3 Model 2: Random Position - ..ABC...BAC..

In the initial Model 2, the sequence ABC is fixed, which leads one to wonder why a model $A \rightarrow B \rightarrow C$ (which could either be a surface model, or two rules $A \rightarrow B, B \rightarrow C$) is not preferred to the hierarchical model $A \rightarrow C, B \rightarrow C$. To test this, I will use the sequence with the two parents {A, B} in randomly alternating observation positions. This might be seen to be the 'most correct' data representation of the tree, as it does not necessarily enforce any ordering between A and B in the single branch case, but it is also of note that it means that the 'observation-position' is random, which it would not be in a real-world model, and in natural language, would be immediately violated with compositions of trees/branches. Even if individual branch members can be in random positions, the adjacency (or rather bracketing) requirements would eliminate various 'fully random' representations of surfacing data – so this positioning-test would be violated.

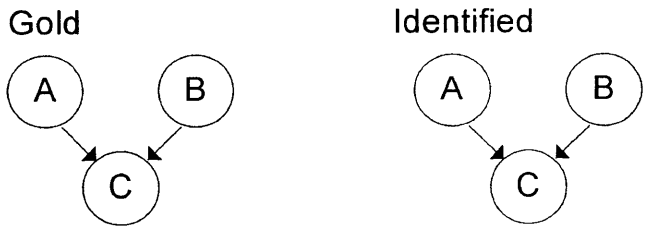



Figure 4-9: Model 2, with a random position data sequence that includes both 'BAC' and 'ABC'. The Identified model restricted to the class of trees, after observation length 100,000. The correct model is identified.

LOG-LIKELIHOOD: -57757.47

| O length | $D_{KL}(m\|\|n)$ |
|---|---|
| 99 | 0.0074 |
| 999 | 0.00065 |
| 9999 | $5.8\ 10^{-5}$ |
| 99999 | $4.4\ 10^{-5}$ |

Table 4.4: The KL divergence between positions $m$, and $n$, which contain both A and B. The 'information content' between the read-positions is driving to 0. (Sequences read in chunks of 3, hence '99','999',... instead of in factors of 10).

Though the correct model is identified (Fig. 4-9, and identified with greater log-likelihood (has greater probability of producing the data) than with the 'more consistent' strictly ABC data sequence, the actual information content that would 'separate' observations A and B is going to 0. At this point, we might start believing that we are looking at a Q=2 (2 parameter) rule instead, where the state that produces C is can have values A or B.

In this method of reading and model, the veil has been pierced just a little more – knowledge that A and B could appear in either position, but C *could not.* This is effectively a position and bracketing restriction.

### 4.2.4 Model 3: Dual Dependency

Model 3 is the most interesting case of the size Q=3 models, as it can be seen as simultaneous multiple distance dependencies, multiple (probabilistic) influences, or

as the conjunction of two separate dependencies that happen to take the same input. From Chapter 2, we know the conditional representation of diverging evidence to be the least 'consistent' with a causal representation.

As before, rates of the rule's appearance within the data are:

| m | n | o | $f_{100}$ | $rate_{100}$ | $f_{100,000}$ | $rate_{100,000}$ | exact $P$ |
|---|---|---|---|---|---|---|---|
| **A** | **D** | **C** | 18 | **0.545** | 16738 | **0.502** | $\mathbf{\frac{1}{2}1 \cdot 1 = \frac{1}{2}}$ |
| B | A | A | 2 | 0.061 | 4205 | 0.126 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| B | A | B | 2 | 0.061 | 4082 | 0.122 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| B | B | A | 6 | 0.182 | 4145 | 0.124 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |
| B | B | B | 5 | 0.152 | 4163 | 0.125 | $\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$ |

Table 4.5: Relative frequences and rates of rule appearance within the data for Model 3 vs. all other unique sequences, after read into distributions (m, n, and o) using an aligned strategy. $f_{100}$ is after N=100, and $f_{100,000}$ is after N=100,000. The exact probability of appearance of the sequences involving A,B,C,D in the distributions. C and D appear in distribution $n$ and $o$ IFF A appeared in $m$, otherwise A or B are selected randomly with probability 1/2.

In this sample, the frequency of the rule appearance vastly dominates the appearance of other sequences, and has greater total rate (0.502) ($P = 1/2$) as well. All observations vary independently within the same space, with the exception of the rule, which deposits different observations into the position distribution. For this experiment, I will deal explicitly with the dependencies between read-distributions $m$, $n$, and $o$, because only $n$ contains the outcome D of the rule, and only $o$ contains C. That means that if the dependencies between $\{m, n, o\}$ are incorrect, so also will be the dependencies between the events and the rule.

### Identified Models

There are 12 graphs models identified from the full class as equally likely after a sequence length of 100,000. It includes all fully connected graphs and 6 of the 12 double-edge (tree) graphs. When the class of models is restricted to tree, those 6 are shown in Fig. 4-10.

After 100,000 points, the selection identified 6 models as equally likely (none matching the gold model), which suggests that the observation sequence does not

possess the necessary evidence to establish the independency of $n$ and $o$. This is not a surprise, as in the data, $n$ and $o$ are forced into a high 'correlation' by the rule. Correlation is the correct description here, and not some form of dependency, as the model selection identified an equivalent number of models where $o$ is conditioned upon $n$ $(n|o)$, and vice verse $(o|n)$.

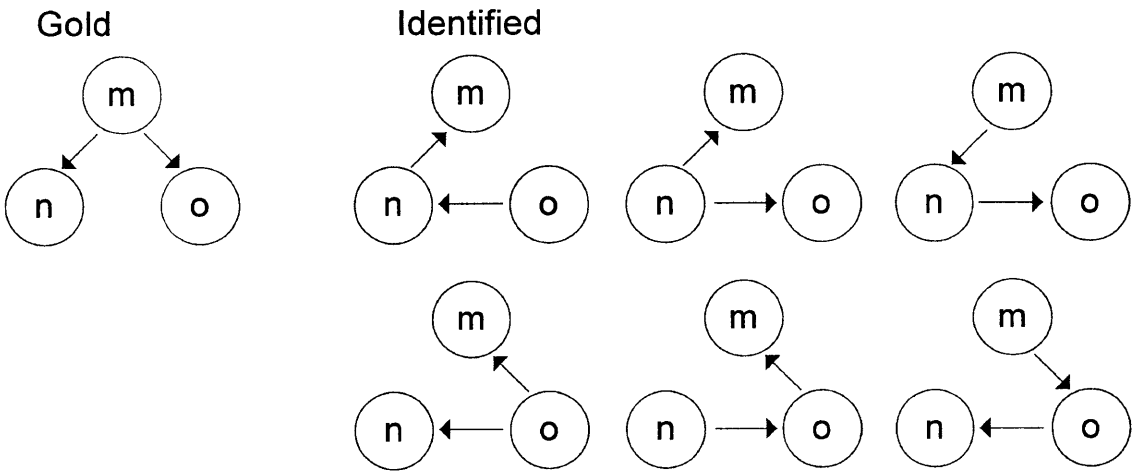



Figure 4-10: After 100,000 points, there are 6 read-position dependency graphs, which means that when mapped back to observations, which contain unique D and C observations, there will also be just as many graphs.

Over 38,700 observations, the total number of correctly identified dependency models is only 6 (Fig 4-11). This test demonstrates that even given variations in evidence over the entire sequence, the correct model could not be identified. The full 100,000 observations was not reached due to a slower compute time, evaluating the change in models for every observation sequence read.

The identification of multiple models as equivalent in probability has some ramifications on model identification for other models. It means that if one of the six models is identified in a general model selection problem, the identified model could either be correct or one of the 6 multiple models for headed conditional probability (Model 3). We might seek a policy whereby if 6 multiple models are identified, then the generating, original model must be similar to Model 3, but this would not work in the general case where multiple generating models are conjoined, distributing probability and changing the identification of substructures.

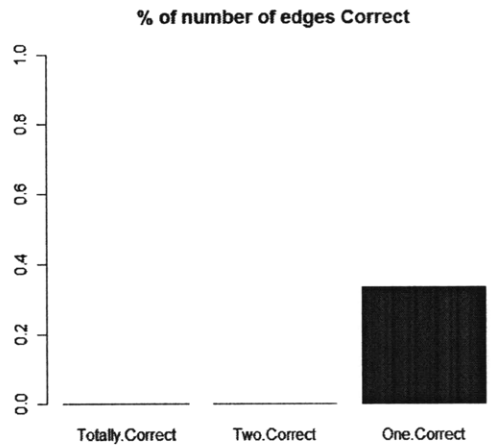
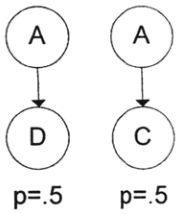



Figure 4-11: After 38,700 points, the number of models identified was 227,971, and returned that were totally correct is 6. The number that had 2 edges correct was 19, and the number with only 1 edge correct was 76,009 (33%).

### Rule Separation/Partial Inhibition

I'd like to test the idea of whether evenly distributing within the outcomes (D, C) of the $A \rightarrow DC$ rule helps with the identification of the read-model with edges in the appropriately from $m \rightarrow n$ and $m \rightarrow o$. That is, does making D and C conditionally independent within A still allow identification of the correct model? This experiment partially replicates the idea of 'inhibition', that part of the rule's dependencies do not always surface (and the reason is inaccessible to us) – the shared 'cause' is probabilistically involved in its children.




Before:

$BBA\mathbf{ADC}BBB\mathbf{ADC}BAA$

After (ex):

$BBA\mathbf{A\underline{A}C}BBB\mathbf{AD\underline{B}}BAA$

Though it is closer, identification of the correct model ($M_{12}$) is still not forthcoming (Fig. 4-12), so what happens if we add an extra constraint such as linear ordering to the model class? This is a justifiable constraint as 'temporal' ordering within the sequence, 'headedness' in NLP, or directional rooting in an ontology. This additional constraint restricts the total class of acyclic models of size 3 from 25 down to 7: 1

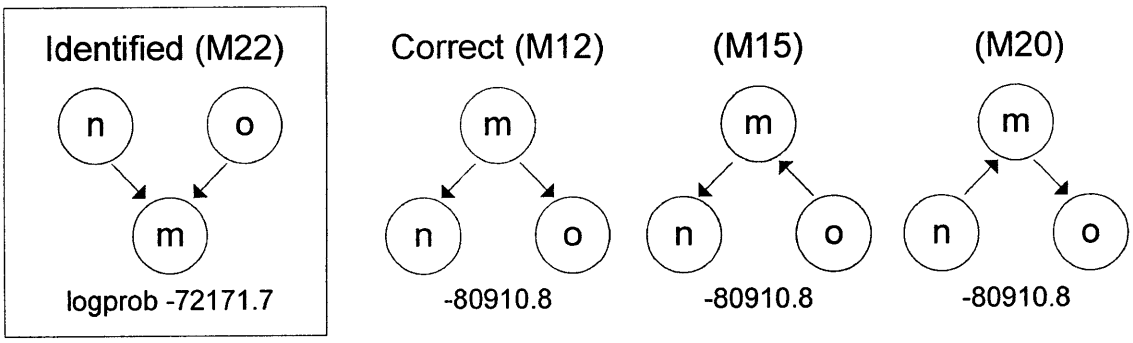



Figure 4-12: After 100,000 points, the most probable model is not the correct model $M_{12}$, but is a similar structure for our experiment with Model 2; $n$ and $o$ converging on $m$. The log probability of the correct model and its equiprobable structures are also shown. This is because the $M_{15}$ and $M_{20}$ are equivalent factorizations of conditional probability as $M_{12}$. Note that between $M_{22}$ and $M_{12}$, there are several other models.

fully independent, 3 with 1 directional edge, and 3 with 2 edges.

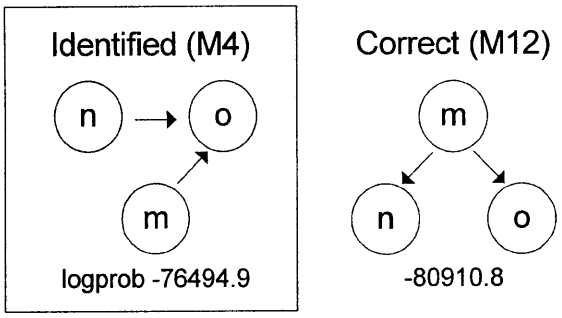



Figure 4-13: After 100,000 points, restricting the class of models to those that maintain a temporal ordering of dependency from $m$ to $n$ to $o$, the correct model is the second-most probable, behind M4, converging on $o$.

Once again, though the correct model is not identified, it is extremely close (Fig. 4-13). Either additional restrictions are necessary or aspects of the rule and experiment have to be changed. While at this point I do not know of a *justifiable* restriction that can be made on the class; any such choice would have to assert the separation and conditional independence of $n$ and $o$. This restriction would eliminate 3 additional models, leaving the 2-edge class containing only the $M_4$ model hypothesis.

### 4.2.5 Model 4: Hierarchical Model

The hierarchical model is an interesting model because it tests the validity of dependency identification for compositions of Model 2, which was the 'best case' match between the graph's conditional independency arrows, and the dependency arrows of the rule. In other words, though model identification worked for the simplest converging case (Model 2 – with 'limited' constraint), does it still work for more complex compositions. as generating rule size grows?

The short answer is, *not necessarily.*

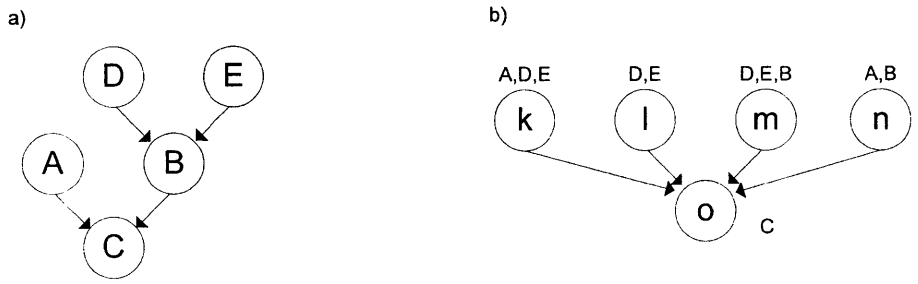



Figure 4-14: Most likely model restricted to 4 edges – class of trees, plus some 4-edge forests, 20000 pts. a) The original model b) The identified read-position model (using aligned read-distributions [*klmno*]) is a tree of maximal width, and no edges exist between l and m, or m and n to even contain dependencies between the original observations A,B,C,D, and E. (the observations are mapped onto the distribution-positions that they can appear in)

Using 'similar constraints' as before, that the dependency structure be restricted to the class of trees – which I emulated by admitting dependency models with 4 edges between the 5 nodes – this includes both trees and some compositions of trees (4 edge forests). With this restriction in place, the most likely model is shown in Figure 4-14 b. The identified tree is one of maximal width, for a number of reasons. Between Models 2 and 3, I saw that the 'preferred' model is one of converging dependence (conditional non-independence given outcome) – when restricted to trees, it tries to 'account' for as much non-independence as possible with a single observation value. Secondly, if the probabilities are allowed to represent fuzzy logic, accounting for both *and* and *or* [60, 51] of boolean logic, distribution of terms both allows the tree structure to be

flattened: $\{Z = (V \cup Y), V = (W \cup X)\} \Rightarrow Z = ((W \cup X) \cup Y) = Y \cup W \cup X$, which with idempotence is the same as $V \cup Y \cup W \cup X$. The same applies for $\wedge/\cap$.

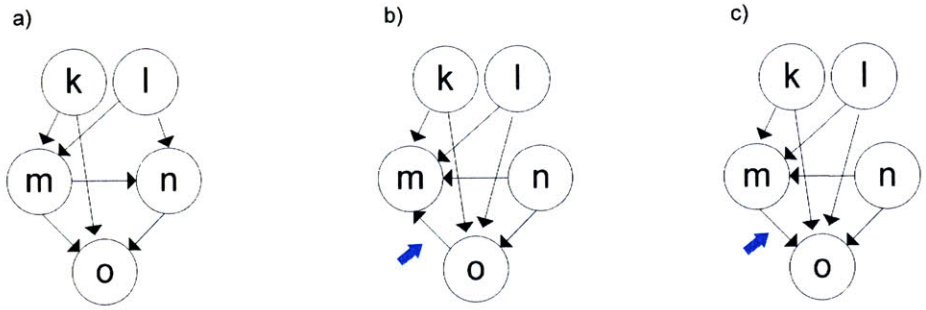



Figure 4-15: a) The composite of the rule dependencies onto positions, 20000 pts, restricting the class of model to 7 edges a) The rule $DE \rightarrow B$ maps onto $k, l$, and $m$, and from branch rotation of $AB \rightarrow C$ has the positions $l, m$, and $n$ also available. Similarly, $AB \rightarrow C$ has positions $k, n, o$ and $m, n, o$ available (Fig 4-14 b & c) The identified models as equiprobable do a lot better, but $m, o$ is changing direction (with equal prob), but $m, n$ and $l, n$ are strictly incorrect. Both returned models seek a higher number of convergent nodes (3 and 4) than the gold model, which has 2,2, and 3.

But in inspection, one may realize that equivalence of rotation of dependencies within the branch actually allows observations of the rule to surface in a number of different read positions, creating an overall non-tree dependency structure (Fig 4-15 a). What happens when the models are constrained to 7 edges? In figure 4-15 b/c, the most likely models (equiprobable) for combined surfacing position and dependency structure show a great deal in common with the gold model. The mistakes are the returned models indicate $o$ (containing C) is equally likely to be the final dependency as is $m$, containing D,E, and B. This may seem like a trivial mistake, but if it persists as observations grow, it means that a knowledge/tree structure built upon the final dependencies cannot be more generally correct, even it is acceptable that a model be locally incorrect within some smaller subset of observation-distribution space. Second, the lack of an edge between $l$ and $n$ is somewhat interesting. From Model 1 and 2 constraints, I know that conditional dependencies are preferred to independencies, but in this Model (with the 7 edge constraint), it has preferred the addition of the conditional dependency of $l \rightarrow o$ 'minus' (not strictly) the loss from $n$'s independence

to the conditional dependency of $l \rightarrow n$. In this experiment, it suggests that positions $l$ and $m$ are more strongly correlated given $o$'s (as is $l, o|m$ ) values than is $l$ and $n$.

This also might be seen as support for having the property of seeking more inbound edges, as in Figure 4-14 and Models 2 & 3, where the number of inbound edges to $\{m, o\}$ is {3,4}, balanced between the two equiprobable models.

Together, the constraints of 4-edges and 7-edges show me something interesting – that if one is 'overinformed' about the actual dependencies between the events (the observation values themselves), then the wrong constraints constraints can be selected. I also had to take into account knowledge of how the dependencies surface in the observation-positions to achieve the best match between the most likely model and the true model. The 'veil of forgetfulness' had to be completely pierced to be able to enumerate all of the ways in which the dependencies can (and cannot) surface in the sequence and thus come up with the right edge-constraints for the reading-model. To push it one step further, in Models 2 and 3, there were only 3 possible edges, so constraining the number of edges alone provided a 'reasonable' 1/3 chance I would get it right by just guessing on edge counts. But with the size 5 model, there are 10 possible edges and to just 'guess' 7 as the correct number of edges has only a 1/10 chance. As the models become larger and more complex, the likelihood of guessing the total number of dependency counts correctly continues to decrease, and as is shown for this experiment, even when the counts are guessed correctly, the correct dependency model is not always achieved.

Note: Model 4 was only computed for 20,000 observations rather than the full 100,000 due to time and computational tractability (took over 30 hours to process).

### 4.2.6 Overall

All of these experiments explicitly contain a lack of knowledge, which is evident by what knowledge and restrictions have to be encoded in. By the end of Model 3, a number of additional (but no incompatible with prior Models) restrictions had to be added, in addition to a change in how the rule surfaced (observability & separation) to come close to the selection of the correct dependency. By the end, any final restriction

that would guarantee selection also eliminates *all* competing models except for the fully independent model and 2 single edge (single independent) models. Now in this experiment, I used a uniform prior and the hard-boundary of selection restrictions acted as prior. Any prior strong enough to guarantee the correct selection of the dependencies of the Rule through the evidence and posterior which prefers other models such as $M_{22}$ and $M_4$, and disproportionately favors $M_{12}$ over its equivalant factorizations will run into the same problem and be encoding the same information. In the end, to recover Model 3, I had to know *a-priori* what the correct model was.

And combining the multiple-equivalent identifications of Model 3 with Model (Rule) 2, which is quickly identified, any of the identified models could belong to Rule 3 or Rule 2, and there is no way to tell which it is without already knowing first.

Therefore, for some classes of rules, to select the correct dependency structure between groups of observations probabilistically, one has to already know what the structure is. And knowing that, the test of model selection by enumerating hypotheses (models in this experiment), reduces simply to the pairwise test of independence between several models. Beyond that, the conditional independence structure graphs do not entirely map directly onto deterministic, causal, or other ontological graphs, meaning something a bit different.

Finally, it is of importance to note that from the perspective of the model identification from data using conditional probabilities, Model 2 (ABC) and Model 3 (ADC) are identical. They both contain a set of symbols that appear in their positions only together. Though ADC may be produced through rule $A \rightarrow \{D, C\}$, because they appear simultaneously within a fixed, stateless window size, it would be equivalent if they had been produced by $\{D, C\} \rightarrow A$.

For completeness of size=3 models, it should be observed that the data would be the same if the rules were fully linear: $A \rightarrow D \rightarrow C$. Identifying the correct rule dependencies follows identically as in Model 3 – meaning that there are 3 equivalent factorization-graphs and the same number of graphs with a higher likelihood as in Model 3 (Fig. 4-12).

This means that not only are the actual rules' dependencies inaccessible to a fixed

size probabilistic analysis, but in hypothesis enumeration and model selection, the conditional probability models do not distinguish between positions and symbols (i.e. it does not matter if they are labeled as 'grass' and 'sprinkler' vs. 'clouds') – they universally prefer (for 'real' data) fully connected models $\succ P(X|Y,Z)$ (converging 'dependence' (non-independent given evidence)) $\succ P(X,Z|Y)$ (independence given evidence) models.

## 4.3 Analysis

This analysis is not intended to construe a proof, but rather provide some insight as to why I think that I observed these results, to help predict what one might expect as models become larger, and finally to motivate future work.

### 4.3.1 Observational Uncertainty

In this experiment, when the reading strategy was fully uncertain, it effectively represented the effect of a uniform spread on the distribution that symbol could go into. By allowing a symbol to enter stochastically into other distributions (i.e. it need not be deposited into, or contributed evidence to all at once), the dynamical evolution of the system of read-distributions is *mixing*, and the strategy by which one deposits evidence into a distribution can be composed in terms of a Markov blanket (defining a group of Markov processes), where addition of evidence either is made in terms of the evidence alone, or some joint probability with the current state of $> 1$ read distribution. (Without the addition of evidence, or certainty of only 1 distribution, there is no process by which the system would evolve). According to Misra et al.[34], the mixing of n-distributions is shown to converge deterministically to a joint expectation of the distributions involved in the process. The Lyapounov then is monotonically decreasing with additional evidence $e$ (or as the observation length increases), instead of T (time).

In many ways, this explains why a transmitter and a receiver (in this experiment, the world and the learner) have to agree on codes/symbols being transmitted. In the

presence of uncertainty around the symbols (that is being stochastically updated), the distributions containing the symbols converge. This does not mean that the transmitter cannot transmit 'information' about the existence of a new symbol, which is added to the receiver's alphabet, but uncertainty around the symbols themselves, if tolerated at all, can only be so to a limited extent.

### 4.3.2 Novel Observations

One of the majors tests of a model is how well it handles 'novel' or new observations. In the general case, novel observations (or novel dependencies) are either entirely new events, or sequences of existing events that we have never seen before. Consider the case of a single novel observation $z$. If $z$ is an entirely new event (coming from some source of novel observations Z), then by definition we do not know what read-distribution it should belong to. Even if we start with a sharply modal graph (Fig. 4-16 a), as new $z$'s are added, the read distributions' probability mass spreads out, increasingly overlaps, and the expectations converge (Fig. 4-16 b). A relevant side note: many times the probabilities attached to the read-distributions are associated with *fuzzy logic*, and the differences between the intersections and overlaps are used to make inferences about boolean logic {*and*, *or*} (for instance to identify a 'circuit diagram' probabilistically in genomics [51, 38]) – but in the case of the distributions, the total probability of the set members must sum to 1, so fuzzy set logic constructed of *min* and *max* will only yield identity [60]. In this experiment, novel observations of this type were simulated by Model 2 and widening noise.

Secondly, the introduction of novel evidence in a position suggests that we also may not know how to read the existing observation sequence – either $z$ replaces an expected observation (as above), or it displaces the sequence by 1. If it displaces the sequence, then the model size must be increased (or the evidence ignored). According to Shannon Entropy, the maximum information per model will be obtained when they all occupy an equal area in hypothesis space, which diminishes with the number of models, so the probability of selecting a model from the space, in terms of a model size Q:

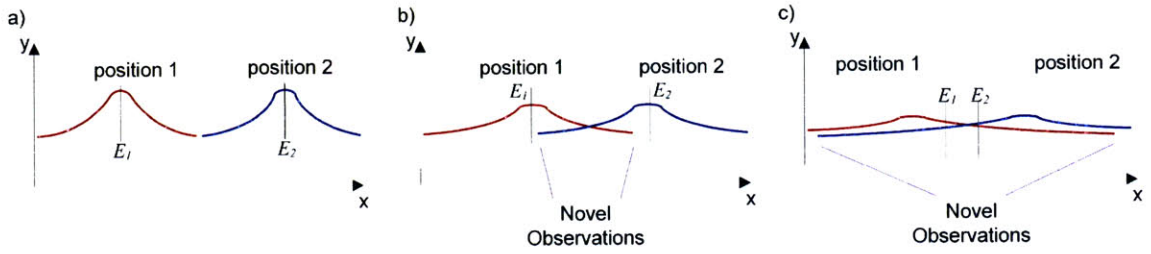



Figure 4-16: A diagram demonstrating the effect of novel observations upon positional-read distributions. According to AEP[50] and as I observed in this experiment, novel observations cause the distributions to spread out. a) the distributions are highly modal, containing completely different observations $\Rightarrow$ high $D_{KL}$. b) Novel observations contribute to both distributions due to uncertainty, the two positional-distribution sets begin to overlap. c) Novel observations consist of the majority of evidence – differentiating, 'known' observation area is approaching 0. Distributions almost completely overlap, and expectations converge.

$$\{Q > 1\} : P(Q) = \frac{1}{3^{\binom{Q}{2}} - C(Q)} \tag{4.2}$$

Where $C(Q)$ is the number of cycles removed from the model space of size Q (For $Q = 4, C(Q) = 250$).

| Size (Q) | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| Total $3^{\binom{Q}{2}}$ | 27 | 729 | 59,049 | 14,348,907 |
| Cycles $C(Q)$ | 2 | 250 | 38,648 | 12,346,772 |
| Acyclic $M(Q)$ | 25 | 479 | 20,401 | 2,002,135 |

Table 4.6: P(Q) is $1/M(Q)$, the number of acyclic dependency hypotheses, which grows > exponential rate with Q.

The area occupied by any individual model in the model space will diminish more quickly than an exponential as model size is increased with novel observations. Convergence to the correct model with additional observations (containing a novel source) may be difficult to achieve, as the observations are added linearly, either spreading the evidence-distribution space, or spreading the model space.

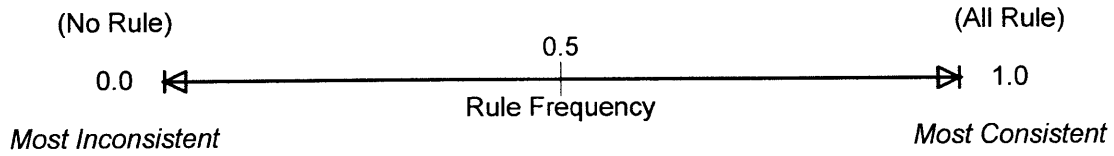



Figure 4-17: The model selection can be seen in terms of the frequency of rule in the O. When the observation sequence consists only of the rule ($f = 1.0$), it represents the 'most consistent' set of observations. All hypothesized models $M_G$ are equiprobable. As the frequency of rule approaches 0.0, or as the number of possible rules continues to rise, the observation sequence enters it's most 'inconsistent' state. If all subsequences selected from O are equiprobable, then the entropy of O is at a maximum.

### 4.3.3 Model Identification & Shannon Entropy

In trying to understand why the most likely dependency model when mapped to actual observations is sometimes wrong, it is helpful to look at the types observations in terms of Shannon (Information) Entropy.

Basically, in any experiment, there are 3 types of observations: observations that completely represent our rule (R), observations that partially represent the rule (L), and those observations that are "noisy" (N) or do not capture the any aspect of current rules (in the future I hope to deal with this as a case of 'evidence for a new model'). Using Model 3 ($AB \rightarrow C$), examples of R, L, and N are as follows:

| | Observations | | | |
|---|---|---|---|---|
| **R:** | ABC | | | (Rule) |
| **L:** | A$\overline{B}C$ | $\overline{A}BC$ | | (Partial) |
| **N:** | AAA | BBA | ... | ('Noise') |

The Bayesian selection for any particular Model is driven by the Model's 'ability' to replicate the data (or rather, it's probability of replicating the data - $P(D|M)$ ).

Taking a sequence of observations which consists of r occurrences of R type observations, l occurrences of L type observations, and n occurrences of N type observations. Then the probability of a single Model generating some observation sequence (or data D; we used 'O' in the experiment):

$$P(D) = P(R)^r P(L)^l P(N)^n$$

Knowing that $r + n + m = t$, the total number of observations,

$$P(D)^{\frac{1}{t}} = P(R)^{\frac{r}{t}} P(L)^{\frac{l}{t}} P(N)^{\frac{n}{t}}$$

Well, for the single model case, the ratio of r/t is just P(R), as n/t is P(N):

$$P(D)^{\frac{1}{t}} = P(R)^{P(R)} P(L)^{P(L)} P(N)^{P(N)}$$

If we take the $-log_2$ of both sides, we are left with the equation for Entropy, and if we select some model M with the minimum entropy *from a closed class* of {M}, then we are effectively selecting the maximum, or 'most probable' model.

To understand how minimum entropy affects model selection, lets forget about the partial observation type L for now and assume that there are no partial observations represented within the data. Then,

$$P(D) = P(R)^r P(N)^n$$

and entropy E given some model M:

$$E = -P(R) \cdot log_2 P(R|M) - P(N) \cdot log_2 P(N|M)$$

If we use a concrete example, with a couple of cases say, with a model where the rule is captured perfectly, another model where the noise is slightly more captured, and finally a model where the ratio of the rule-noise R:N is matched perfectly. Lets also assume a ratio of occurrences r:n of 6:4, total occurrences t=10.

| Model | $P(R\|M)$ | $P(N\|M)$ | $P(D)^{1/t}$ | $E$ |
|---|---|---|---|---|
| Rule | 1.0 | 0.0 | 0.0 | $\infty$ |
| Noise | 0.4 | 0.6 | .47 | 1.08 |
| **Ratio** | **0.6** | **0.4** | **.51** | **.97** |

In this example, we can see that the probabilistic model that 'perfectly' represents the Rule that *we believe exists* has zero ability (likelihood) to capture the data, and entropy of infinity ($log(0) = -\infty$); it will *never* be selected, except when there are no noise observations N (where $(P(N) = 0)^0 = 1$). Otherwise, the model favoring Noise fares far better, but the Model selected (E=.97) is always the one that best matches the full ratio of observations.

Because this minimization of entropy penetrates down to the individual events in the distributions and observations that make up my O sequence, and minimizing the difference between the actual probability of each observation and our represented one, any choice of factorization of distributions and evidence will obtain the same result. This means that no joint probability distribution can be chosen that separates 'certain' dependencies from a random variable – a probabilistic model cannot be considered to contain (or factorized into) some known rule and a random variable – all selected joint distributions must contain the random variable at the observed ratios for entropy to be minimized.

As humans, we often think of the 'most probable model' as being the model that 'best captures' some deterministic rule within our data, if our assumptions of redundancy and independence are met, but the most probable model is not the one with the highest likelihood of reproducing the data, it is the one with the highest probability of replicating the *distribution* or relative ratios of the data.

We might like to think that if we have some deterministic model + some representation of noise, and that they are optimized and selected separately, but the distribution must be represented from top to bottom, and the rule no longer explicitly exists.

# Chapter 5

# Conclusion

In this thesis, I tested the use of probabilities to approach a problem that humans are clearly capable of surmounting. The problem is generally mimicked by statistical induction of natural language, merging of ontological dependency structures, and induction of causal dependency from observation – the identification of a correct 2-d dependency model from a linear sequence. The problem was primarily phrased in terms of insertion of symbol(s) into an otherwise random observation sequence due to a deterministic rule and the rule's dependency recovery. This allowed me to test different levels of 'uncertainty' and different combinations of the rule to determine what constraints (if any) were necessary for recovery of the dependency structure of the rule. As a result, I gained some insight into the minimum information required to probabilistically identify a hierarchical dependency model.

In the larger problem, those of ontologies and causality, the experiment represented a world consisting of an unknown internal state, but operating with a known ontology. Transmissions of symbols from the world's state either have access to observations from the ontology, or do not, and from the symbols, the learner must construct his own ontology or representation that mimicks the correct (the world's) ontology.

While this problem was mostly limited to identification of ontology/dependency structures of size 3, it also showed the effect and meaning of a uniform prior on model selection. The test of what would happen if one did not know a-priori the correct structure size (i.e. if it was either larger or smaller) was done, and finally a brief

analysis demonstrated that the area maximally available to distinguish the models is decreasing with window size at greater than an exponential rate, a point exacerbated by the equivalent factorizations of conditional independence.

As generally as possible, what I found was:

1. If one does not know how to 'read' (identify) their observations a-priori then the adoption of probabilities guarantees that the observation distributions will become indistinguishable over time. In otherwords, the information is being degraded. This result is supported by the idea that a transmitter and receiver must agree on codes/symbols, and the central limit of the Lyapounov [34].

2. Similarly, I learned that hypothesis formulation and dependency identification cannot be separated from observation reading. If one does not know how to read, then the space of all possible single-symbol hypotheses (including 1st order cycles) grows at a rate of $4^{\binom{N}{2}}$ of the number of observations N – at a rate far higher than the rate at which observations are added – which means that the model/hypothesis space spreads out far more quickly than convergence might be achieved in terms due to additional observations.

3. The minimum information necessary to recover the correct dependency model from independence models probabilistically is... knowledge of the correct dependency model. The conversion of identified conditional-independency structure to a graph structure is not necessarily 1:1 and because the majority of the rules ($A \rightarrow C, D$) have equivalent factorizations and are not the 'most likely', the constraints (or the prior) that one has to put on the model space eliminate most or all competing dependency models (Model 3).

4. Fully enumerating a hypothesis space of greater than 2 does not make sense probabilistically. The hypothesis space of conditional independency between random variables does not map to explicit dependencies, the graph/conditional independencies have multiple equivalent factorizations which overlap with competing hypotheses, truly independent variables are not likely to be identified

automatically in real data, and the discriminating evidence between hypotheses is diminishing as the size and space of hypotheses grows.

5. Even when the correct graph is identified probabilistically, it is the shape/displacement of the *noise* rather than the existence of the rule that identifies it. In the case where there are novel observations, either through productions of uncaptured rules, or from new rules, the noise displacement drives down, and the distribution approaches uniformity, which is supported by asymptotic equipartition [50].

6. Finally, as suggested by Model 4, it is not just the randomness of observations and position, or even the inconsistency that helps to identify the tree structure – it is also the distribution of what we *do not see*. For instance, what has the power to give away the data from Model 4 as being generated from a tree is the fact that though we see observations in a few random positions, we do not see those same 'random' observations in other positions.

Practically, what these results tell me is that in the concrete case of merging two biological ontologies, and the case of identifying multiple common factors or modifiers in two biological pathways cannot be done probabilistically. Where identification of common elements in rate equations comes down to identifying dependencies in two different distributions, similar to the rule $A \rightarrow C, D$, many other structures are preferred. Similarly, a cause-effect, real-world dependency model faces identification problems. This suggests (though not proved exhaustively) to me that if we are faced with the problem of merging/seeking coherence between conflicting models without knowing the dependencies, we must seek alternate methods than a probabilistic one.

In conclusion, where 'knowledge' is defined as certainty of dependencies or observation, probabilities are an explicit representation of a lack of knowledge. Not only is it a representation of the lack of knowledge of dependencies between observations (and the resulting inability to recover them), it also an explicit abandonment of the larger proportion of observations, whos only purpose is to be the noise displaced in favor of conditional independencies between observations of the rule. When we consider the

non-rule observations to actually be random noise, as they are in this experiment, that may be acceptable. But in the case where the variation of observations actually is produced by compositions of other rules, and not randomly generated, discarding the large proportions of evidence is not acceptable.

Conditional probabilities are about information about how a particular observations changes what is not known about other observations. Though the graphs for conditional probabilities can 'look' the same as causal or ontological dependency graphs, they have nothing to with them a-priori. This problem is exacerbated by the observation that for some generating models and strong restrictions, the graphs can overlap and appear identical.

Similarly, probabilities, despite being a measure of lack of knowledge, seem to provide a mathematical tool by which we can 'minimize our ignorance'. But looked at as a measure of 'inconsistency', it can be seen that when we minimize inconsistency, all possible models become equally likely (Sect. Model 2, discrimination). In this manner, probabilities are perhaps uninforming – when the observations are fully inconsistent, any number of models may be justified, when they are consistent, all models are admissable (in that they admit the data equally well).

In the end, even probabilities require 'certainty' to be mean anything; too many uncertainties and distributions overlap and run afoul of the central limit theorem in unintended dimensions. While the law of large numbers guarantees convergence, it does not guarantee convergence to the generating rule – it guarantees convergence to the mean of the distributions.

This conforms to logic and intuition, that one cannot recover structure from unstructured information, and though it may conflict with the results from other works, it confirms the results by Sewall Wright in Correlation and Causation [58].

## 5.1 Future Work

Originally, I had intended to fully (exhaustively) capture the space of all possible identified models, for all parameter choice size Q. This idea turned out to be quite

impractical considering the computational complexity and how quickly the number possible dependency models grows with Q, combined with testing the changes in all of the read-distributions with each additional observation. This is still an important experiment to perform, considering that it allows us to exhaustively examine the surface of all size models chosen for overall correctness, saddle points, min/max behavior, and the overall recoverability as the model size grows. As a result, I am not completely convinced that this experiment constituted the most general or complete case.

In part 1 (reading observations) of the experiment, I found that the uncertainty around reading the observations caused convergence; it would be a good extension to exactly model the rate of convergence in terms of the 'shape' of the uncertainty – i.e. non-uniform/etc. and the size of the model space. For instance, can one characterize 'how much' they are not likely to be able to retrieve probabilistically based upon information that they know about their experiment.

Second, an exhaustive comparison of the divergence for different reading strategies, combined with an exploration of the total variation of evidence points introduced by different reading strategies. It is clear that as the read size approaches the length of the full observation sequence, the overall variation goes down, and though memorization of the model has been maximized, all possible models have become equally possible (Model 3 experiment).

Third, this experiment was limited to just a few dependency models, which while they demonstrated exceptions to convergence and identification, did not test all possible rules/dependency structures for recoverability. The entire dependency space within a single rule size Q should be tested, along with more complex effects of composition with other rules (beyond just the one size 5 tree).

Fourth, in analysis, I became aware of the possibility that enumerating strict independence maps of size $> 2$ (including conditional independence graphs) may produce inconsistent independence statements. I spent a great deal of time attempting to show and prove this (finding that the only consistent cases of independence graphs are combined fully connected and fully disjoint nodes), but in the end I decided that

the proofs were not rigorous enough at this stage to be included in this thesis. A clear extension is to finish that analysis and determine if a majority of enumerated independence hypotheses do completely overlap in area, which would tell us if the correct hypothesis can be converged on in principle.

Finally, a more complex analysis of the 'state' that the world could be in before emitting a symbol needs to be done. Much of the original intent of this work was to demonstrate how state is a necessary participant in the induction of models, and how reversing (or disregarding) it leads to conflicting model selection. Unfortunately, I became mired in the the problem of reading observations into distributions, which is part of the very same problem. For this experiment I had to assume a limited state-symbol size of 1, though in practice, the state that the world may be in before transmitting a symbol could be any combination of prior symbols into complex states. Given an observation sequence of length L, the number of possible (relevant) prior states is $\sum_{n=1}^{L} \binom{L}{n}$. This also entails a deeper exploration of 'descriminating evidence' – evidence that either identifies one model over another (differentiating likelihood), or forces an increase in model complexity and size.

To achieve that comparison, there also needs to be a more exhaustive test of what different deterministic models produce, and a more rigorous analysis of what entails a 'dependency', though quite possibly this will require a full expansion of the definitions of causality.

## 5.2 Closing Remarks

This work is not intended to suggest that the use of probabilities is 'not productive' or to cast doubt on many of the effective uses of probabilities. After all, it is reported that Laplace used probabilities (through Bayes' Theorem) quite effectively to select astronomical problems to work on [27]. While there was no *control* for Laplace, so we will never know if it was just his diligent nature that led him to a productive lifetime and not the existing discrepency between observation and prediction (an example of conditional probabilities contra-indicating their own use), this experiment shows that

the *choice* or decision of dependencies, certainty of observation, etc. is the researcher's to make, not the probabilities', conditional or otherwise.

From that standpoint, it is my personal assertion that 'machine learning' should not be maintain the word 'learning' as part of its moniker, as the identification of dependencies may not be accomplished probabilistically, as 'machine learning' would seem to imply. This is not to overstep the results of this work, as probabilities can be used to help identify features that may be of interest, optimize capacities and usage, in additional to many other uses both current and yet to be discovered. Beyond that, we may have to be careful about using probabilities by *default* to do things that non-probabilistic methods (ranking, for example) can do also [59].

Finally, I am quite certain that readers will find problems with this work; not just because of its limited scope, but because the process of science is one of strong, certain statements, which are then refuted, driving increases in complexity of knowledge.

## 5.3 What was Implemented

A number of algorithms and procedures were (newly) implemented for this work. First, a Bayesian class (Existing Bayes' modules in Python are somewhat limited, and development on the leading version (OpenBayes) has been stalled for around 3 years.) and iterator was constructed in Python that allows independency maps (Bayesian Networks) to be built and tested on the fly given only the parameter size, along with inspection of the likelihood of data for each independence graph at each new evidence point. Some attempts were made to distribute the likelihood computation for each model across a compute cluster, but it turned out that copying of classes and data cost more in computation time than local solving. This is not the general case for this problem, as there are many different ways to factor and distribute the problem. Second, an efficient graph enumerator that produces all possible graphs for a given size, along with an identification of acyclic graphs (an example described in [33]). Finally, applications to compare the produced models against the declared correct models (and plot them in R), along with a graphic visualizer to ease inspection of

the selected dependency model was built using my own PyGraph libraries, which are available under MIT License.

To replicate this experiment, the necessary components may be found at:

Graphs - http://alpha-leonis.lids.mit.edu/~beracah/masters/graphs

Data - http://alpha-leonis.lids.mit.edu/~beracah/masters/data

Models - http://alpha-leonis.lids.mit.edu/~beracah/masters/models

Algorithms - http://alpha-leonis.lids.mit.edu/~beracah/masters/code

# Bibliography


[1] Hirotogu Akaike, *A new look at the statistical model identification*, IEEE TRANSACTIONS ON AUTOMATIC CONTROL **AC-19, NO. 6, DECEMBER** (1974), 716–723.

[2] Dana Angluin, *Learning regular sets from queries and counterexamples*, Information and Computation **75** (1987), no. 2, 87–106.

[3] Michael A. Arbib, *Brains, machines, and mathematics*, Springer-Verlag, 197.

[4] Kenneth J. Arrow, *A difficulty in the concept of social welfare*, Social Choice and Individual Values/RAND Corporation/Cowles Commission for Research, 1948.

[5] Jennifer Williams Barry Smith and Steffen Schulze-Kremer, *The ontology of the gene ontology*, Proceedings of AMIA Symposium (2003).

[6] Bezáková, Kalai, and Santhanam, *Graph model selection using maximum likelihood*, (2006).

[7] Remco R. Bouckaert, *Conditional dependencies in probabilistic networks*, In Proceedings 4th International Workshop on AI and Statistics, 1993.

[8] Carlos A. P. Campani and Paulo Blauth Menezes, *On the application of kolmogorov complexity to the characterization and evaluation of computational models and complex systems.*

[9] Noam Chomsky, *Three models for the description of language*, IRE Transactions on Information Theory **2** (1956), 113–124.

[10] Joy Christian, *Disproof of bell's theorem by clifford algebra valued local variables*, (2007), 4.

[11] Thomas G. Dietterich, *Machine learning for sequential data: A review.*

[12] David Draper, *Assessment and propagation of model uncertainty.*

[13] A. Einstein, B. Podolsky, and N.Rosen, *Can quantum mechanical description of physical reality be considered complete?*, Physical Review **47** (1935), 4.

[14] Jason Eisner, *Review of optimality theory (rené kager)*, Computational Linguistics **26** (2000).

[15] ———, *Discovering syntactic deep structure via bayesian statistics*, Cognitive Science **26** (2002), 255–268.

[16] N. Etemadi, *An elementary proof of the strong law of large numbers*, Z. Wahrscheinlichkeitstheorie verw. Gebiete/Springer-Verlag **55** (1981), 119–122.

[17] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio, *Statistical learning theory: A primer*, International Journal of Computer Vision **38** (2000), 9–13.

[18] Jacob Feldman, *Formal constraints on cognitive intrerpretations of causal structure*, IEEE Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems (1995).

[19] Richard Feynman, *The character of physical law: Seeking new laws*, Modern Library, 1965 (1994).

[20] Andras Frank, *On kuhns hungarian method a tribute from hungary*, Tech. report, Egervary Research Group on Combinatorial Optimization, 2004.

[21] Kurt Gödel, *On formally undecidable propositions in principia mathematica and related systems i*, (1931).

[22] E. Mark Gold, *Language identification in the limit*, Information and Control **10** (1967), 447–474.

[23] ———, *Complexity of automaton identification from given data*, Information and Control **37** (1978), no. 3, 302–320.

[24] David M. Green and John A. Swets, *Signal detection theory and psychophysics*, John Wiley and Sons, Inc., 1966.

[25] Peter D. Grünwald, *the minimum description length principle*, The MIT Press, 2007.

[26] Mark H. Hansen and Bin Yu, *Model selection and the principle of minimum description length*, Journal of the American Statistical Association **96** (2001), 746–774.

[27] E. T. Jaynes, *Bayesian methods: General background - an introductory tutorial*, vol. Maximum Entropy and Bayesian Methods in Applied Statistics, Cambridge University Press, August 1984, pp. 1–25.

[28] Finn V. Jensen and Thomas D. Nielsen, *Bayesian networks and decision graphs*, Springer Verlag, 2007.

[29] Daniel Jurafsky and James H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.*, non-citable (not yet released), 2007.

[30] Yongwook Bryce Kim, *Comparison of data-driven analysis methods for identification of functional connectivity in fmri*, Master's thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2008.

[31] L. Laera, V. Tamma, J. Euzenat, T. Bench-Capon, and T. Payne, *Arguing over ontology alignments*, ISWC Workshop - OM 2006, 2006.

[32] Eugene M. Lashchyk, *Scientific revolutions*, Ph.D. thesis, University of Pennsylvania, 1969.

[33] K. Mehlhorn and P. Sanders, *Data structures and algorithms; the basic toolbox*, Springer-Verlag, May 2008.

[34] B. Misra, I. Prigogine, and M. Courbage, *From deterministic dynamics to probabilistic descriptions*, Proc. Natl. Acad. Sci. USA **76, No. 8** (1979), 3607–3611.

[35] ———, *Lyapounov variable: Entropy and measurement in quantum mechanics*, Proc. Natl. Acad. Sci. USA **76, no. 10** (1979), 4768–4772.

[36] Sayan Mukherjee, *Statistical learning: Algorithms and theory.*

[37] Volker Nannen, *A short introduction to model selection, kolmogorov complexity and minimum description length (MDL)*, April 2003.

[38] Shoudan Liang Patrik Dhaeseleer and Roland Somogyi, *Genetic network inference: from co-expression clustering to reverse engineering*, BIOINFORMATICS **16** (2000), 707–726.

[39] Judea Pearl, *Causality - models, reasoning, and inference*, Cambridge University Press, 2000.

[40] Karl R. Popper, *Objective knowledge*, Oxford University Press, 1973.

[41] Andrew Radford, Martin Atkinson, David Britain, Harald Clahsen, and Andrew Spencer, *Linguistics, an introduction*, Cambridge University Press, 1999.

[42] Adrian Raftery, David Madigan, and Jennifer Hoeting, *Model selection and accounting for model uncertainty in linear regression models*, Tech. report, Journal of the American Statistical Association, 1993.

[43] Hans Reichenbach, *The rise of scientific philosophy*, University of California Press, 1951.

[44] W. Richards, J. Feldman, and A. Jepson, *From features to perceptual categories*, British Machine Vision Conference (1992).

[45] Stuart Russell and Peter Norvig, *Artificial intelligence: A modern approach - chapter 20: Statistical learning methods*, Prentice Hall, 2003.

[46] Wesley C. Salmon, *The foundations of scientific inference*, University of Pittsburg Press, 1966,1967.

[47] Stefan Schaal, Christopher G. Atkeson, and Sethu Vijayakumar, *Real-time robot learning with locally weighted statistical learning*, International Conference on Robotics and Automation (2000).

[48] Glenn Shafer, *Constructive decision theory*, December 1982.

[49] ———, *Savage revisited*, Statistical Science **1, No 4** (1986), 463–501.

[50] Claude E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal **27, July, October** (1948), 379423, 623656.

[51] Stefanie Fuhrman Shoudan Liang and Roland Somogyi, *Reveal, a general reverse engineering algorithm for inference of genetic network architectures*, Pacific Symposium on Biocomputing (1998).

[52] Silvia, *Data analysis - a bayesian tutorial*, Clarendon Press, Oxford University Press, 1996.

[53] David M. Sobel and Natasha Z. Kirkham, *Interactions between causal and statistical learning*, 2007.

[54] Ondrej Sváb and Vojtech Svátek, *Combining ontology mapping methods using bayesian networks*, ISWC Workshop (2006), 206–210.

[55] Ching-Huei Tsou, *A statistical learning framework for data mining pf large-scale systems: Algorithms, implementation, and applications*, Ph.D. thesis, Massachusetts Institute of Technology, 2007.

[56] L.G. Valiant, *A theory of the learnable*, Communications of the ACM **27** (1984), no. 11, 1134–1142.

[57] Stephen Wolfram, *A new kind of science*, Wolfram Media, 2002 (English).

[58] Sewall Wright, *Correlation and causation*, Journal of Agricultural Research **7** (1921), 557–585.

[59] Charles Yang, *The great number crunch*, Journal of Linguistics 00 (0000) 124. (2008).

[60] L.A. Zadeh, *Fuzzy sets*, INFORMATION AND CONTROL **8** (1965), 338–353.

# Appendix A

# Appendix

## A.1 Graph Enumeration Table

$\{\mathrm{M}\}_Q$:

| $g$ | $n_{1,2}$ | $n_{1,3}$ | $n_{1,4}$ | $n_{2,3}$ | $n_{2,4}$ | $n_{3,4}$ |
|---|---|---|---|---|---|---|
| 0: | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ |
| 1: | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\rightarrow$ |
| 2: | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\leftarrow$ |
| 3: | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\not\leftrightarrow$ | $\rightarrow$ | $\not\leftrightarrow$ |
| | | | | ... | | |
| 728: | $\leftarrow$ | $\leftarrow$ | $\leftarrow$ | $\leftarrow$ | $\leftarrow$ | $\leftarrow$ |

Table A.1: An enumeration of all possible graphs of size $Q = 4$, from g=0 to g=728: $n_{i,j}$ is the connection between nodes $n_i$ and $n_j$. g=0 represents a fully independent graph. Key: $\not\leftrightarrow$: $n_{i,j}$ are fully disconnected. ; $n_{i,j} \rightarrow$: means that $n_i \rightarrow n_j$.