# The Minimalist Mind:
## On Minimality in Learning, Reasoning, Action, and Imagination

Ardavan Salehi Nobandegani

Department of Electrical & Computer Engineering
McGill University

December 2017

---

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

© 2017 Ardavan Salehi Nobandegani

To Aida, Zhila, Ardeshir
*for their love*

&

To Judea Pearl, Allen Newell, Noam Chomsky
*for greatly inspiring me through their work*

## Abstract

For a notion to be perceived as pivotal in the history of human thought, it ought to appear and reappear in a multitude of human endeavors — from physics to metaphysics, arts to natural sciences, etc. Any truly fundamental notion leaves its trace on many domains of knowledge; it only evolves through time yet never disappears. The notion of *minimality* (which can be replaced, via relaxation, by *sufficiency*) is of such nature. It is only natural then to ask where minimality meets cognition.

This dissertation explores, mainly at Marr's computational and algorithmic levels of analysis, the fundamental role the notion of minimality plays in human cognition in the contexts of *probabilistic reasoning*, *causal reasoning*, *action*, *control*, *learning*, and *imagination*. The work comprises seven chapters, the content of each is briefly discussed next.

Chapter 1 provides an introduction to the dissertation. Chapter 2 addresses, for the first time in the literature, how the notion of minimality can be applied to probabilistic reasoning under partial knowledge. To this end, drawing on the notion of bounded rationality manifested in a reasoner's limited attention span and scope, Chapter 2 presents a novel graphical model, termed the Multi-Context Model (MCM), to represent the reasoner's state of partial knowledge of a domain. MCM occupies a middle ground between Probabilistic Logic, Bayesian Logic, and Probabilistic Graphical Models. Also, drawing on the quintessence of Bayesian networks (BNs), i.e., the concept of conditioning, MCM generalizes BN to the realm of partial knowledge. Importantly, MCM also serves as the first *normative*, *probabilistic*, representational-level account of an important developmental shift in infant information processing, between four and ten months of age.

Inspired by Simon's bounded rationality and drawing on the notion of minimality, Chapter 3 provides a novel algorithmic perspective to the causal variant of the frame problem (CFP), a deep puzzle in philosophy of mind and epistemology. Chapter 3 begins by introducing a notion called Potential Level (PL). PL generalizes the graph-theoretic concept of topological sorting, and extends the fundamental notion of Lamport's logical clock to causal Bayesian networks (CBNs). Drawing on the psychological literature on causal judgment, Chapter 3 substantiates the claim that PL may bear on how *time* is encoded in the mind. Using PL, Chapter 3 then proposes an inference framework, called the PL-based Inference Framework (PLIF), permitting a boundedly-rational approach to the CFP, formally articulated at Marr's algorithmic level of analysis. PLIF is also shown to be consistent with a

wide range of findings in the causal judgment literature. Interestingly, the ideas explored in Chapters 2 and 3 demonstrate how the old concept of imprecise probabilities naturally emerges out of Simon's bounded rationality.

Chapter 4 pursues the notion of minimality in the key context of action and control. Chapter 4 studies, for the first time in the literature, the problem of probabilistic controllability in CBNs. Probabilistic controllability extends the fundamental concept of controllability in control theory to probabilistic CBNs. More specifically, the aim of Chapter 4 is two-fold: (i) to introduce and formalize the problem of probabilistic structural controllability in CBNs, and (ii) to identify a sufficient set of driver variables for the purpose of probabilistic structural controllability of a generic CBN. Furthermore, Chapter 4 elaborates on the nature of minimality that the identified set of driver variables satisfies. The results of Chapter 4 have important implications for a line of work in developmental psychology concerning causal learning by young children in pedagogical settings. Also, the formalism developed in Chapter 4 establishes, for the first time in the literature, a rational, algorithmic-level account of a curious behavior demonstrated by young children called *overimitation*, generally taken as evidence for children's irrationality. Chapter 4 concludes by exploring the computational complexity of the problem under study and presenting $\mathcal{NP}$-hardness results for it.

Chapter 5 revisits the fundamental notion of conditional probabilistic independence as the core concept which gives rise to minimality in probabilistic settings. Chapter 5, for the first time in the literature, proposes an asynchronous, distributed, message-passing algorithm—akin, in spirit, to Pearl's Belief Propagation scheme—so as to implement Pearl's key notion of $d$-separation. Also, through the introduction of a key graph-theoretic notion, termed minimal refutation-module, Chapter 5 shows how the notion of minimality manifests itself in a distributed, message-passing implementation of $d$-separation. The proposed algorithm exhibits intriguing properties which position it as a plausible candidate for the implementation of $d$-separation at Marr's algorithmic level of analysis. Furthermore, the proposed algorithm outperforms all the previously proposed algorithms in the literature in terms of worst-case running time, and serves as the first rational, *distributed*, process-level account of how humans handle probabilistic independence.

Chapter 6 explores the notion of minimality in the context of learning and imagination. Chapter 6, for the first time in the literature, proposes a *neurally-plausible* and *computationally-efficient* framework which allows to transform any deterministic, discriminative neural network (e.g., deep convolutional neural networks and multilayer perceptron)

into a probabilistic, generative model. Using this framework, cascade-correlation neural networks (CCNNs)—a class of self-organized, deterministic, discriminative models which have been successful in accounting for a variety of psychological phenomena—are converted into probabilistic generative models, thereby enabling CCNNs to probabilistically generate samples from a category of interest. Importantly, the proposed framework: (1) suggests a modular account of human imagination which is supported by studies on learning and imaginative abilities of hippocampal amnesic patients as well as a growing body of brain imaging studies showing that perception and imagery share neural representation, (2) gives rise to *self-organized* generative models, (3) strongly suggests that, contrary to a widely-held view, the boundary between discriminative and generative models is blurry, (4) bridges computational, algorithmic, and implementational levels of analysis, and finally, (5) connects two dominant schools of thought in cognitive sciences, namely, connectionism and Bayesian cognition.

Taken together, the ideas explored in this dissertation suggest that pursuing the notion of minimality is a fruitful endeavor for understanding cognition at the computational and algorithmic levels, and equally importantly, demonstrate how pursuing the notion of minimality allows for developing new computational problems as well as introducing novel algorithms, data structures, and several algorithmic concepts. Inspired by this, Chapter 7 concludes the dissertation by proposing a new mode of enquiry, termed the Rational Minimalist Program, outlining a principled, rational methodology for studying cognition at the algorithmic level of analysis.

## Résumé

Pour qu'une notion soit perçue comme essentielle dans l'histoire de la pensée humaine, elle devrait apparaître et réapparaître dans une multitude d'efforts humains - de la physique à la métaphysique, des arts aux sciences naturelles, etc. Toute notion qui est vraiment fondamentale laisse sa trace sur de nombreux domaines de connaissances; elle ne disparaît jamais mais seulement évolue dans le temps. La notion de minimalité (qui peut être remplacée, par relaxation, par suffisance) jouit d'une telle nature. Il est naturel alors de demander où la minimalité rencontre la cognition.

Cette thèse explore, principalement aux niveaux d'analyse computationnelle et algorithmique de Marr, le rôle fondamental que joue la notion de minimalité dans la cognition humaine dans les contextes du raisonnement probabiliste, du raisonnement causal, de l'action, du contrôle, de l'apprentissage et de l'imagination. Cet ouvrage comprend sept chapitres dont le contenu est brièvement présenté dans la suite.

Le 1$^{\text{er}}$ chapitre introduit la thèse. Le 2$^{\text{ème}}$ chapitre traite, pour la première fois dans la littérature, comment la notion de minimalité peut être appliquée au raisonnement probabiliste à base de connaissance partielle. À cette fin, en s'appuyant sur la notion de rationalité limitée qui se présente dans la portée et l'étendue de l'attention limitée d'un raisonneur, le 2$^{\text{ème}}$ chapitre présente un nouveau modèle graphique, appelé le modèle multi-contexte (MCM pour "Multi-Context Model" en anglais), pour représenter l'état de connaissance partielle du domaine d'un raisonneur. Le MCM occupe une position intermédiaire entre la logique probabiliste, la logique bayésienne et les modèles graphiques probabilistes. De plus, en s'appuyant sur la quintessence des réseaux bayésiens (BN pour "Bayesian Networks" en anglais), c'est-à-dire le concept de conditionnement, MCM généralise BN dans le domaine de connaissance partielle. Il est important de noter que MCM est également le premier compte-rendu normatif, probabiliste et de niveau représentatif d'un important changement de développement dans le traitement d'information chez les nourrissons entre quatre et dix mois.

Inspiré par la rationalité limitée de Simon et en s'appuyant sur la notion de minimalité, le chapitre 3 offre une nouvelle perspective algorithmique à la variante causale du problème de frame (CFP) comme étant un puzzle profond dans la philosophie de l'esprit et de l'épistémologie. Le 3$^{\text{ème}}$ chapitre commence par introduire une notion appelée niveau potentiel (PL pour "Potential Logic" en anglais). PL généralise le concept de théorie graphique du tri topologique et étend la notion fondamentale de l'horloge logique de Lamport aux CBNs.

En s'appuyant sur la littérature psychologique sur le jugement causal, le 3ème chapitre justifie l'affirmation selon laquelle PL pourrait expliquer comment le temps est représenté dans l'esprit. À l'aide de PL, le 3ème chapitre propose une structure d'inférence basée sur PL appelée PLIF (PLIF pour "PL-based Inference Framework" en anglais), permettant une approche à base de rationalité limitée pour traiter le CFP, auparavant articulé au niveau algorithmique d'analyse de Marr. La PLIF est également compatible avec une large gamme de résultats dans la littérature sur le jugement causal. Fait intéressant, les idées explorées dans le 2ème et 3ème chapitre démontrent comment l'ancien concept de probabilités imprécises surgit naturellement de la rationalité limitée de Simon.

Le 4ème chapitre poursuit la notion de minimalité dans le contexte clé d'action et de contrôle. Le chapitre 4 étudie, pour la première fois dans la littérature, le problème de la contrôlabilité probabiliste dans les réseaux bayésiens causaux (CBN). La contrôlabilité probabiliste étend le concept fondamental de contrôlabilité rencontré dans la théorie du contrôle aux CBN probabilistes. Plus précisément, l'objectif du 4ème chapitre est double: (i) introduire et formaliser le problème de la contrôlabilité structurelle probabiliste dans les CBN, et (ii) identifier un ensemble suffisant de variables pilote dans le but de la contrôlabilité structurelle probabiliste d'un CBN générique. En outre, le 4ème chapitre élabore sur la nature de la minimalité que l'ensemble identifié de variables pilote satisfait. Les résultats du 4ème chapitre ont des implications importantes pour une ligne de travail en psychologie du développement concernant l'apprentissage causal chez les jeunes enfants dans les milieux pédagogiques. En outre, le formalisme développé dans ce chapitre établit, pour la première fois dans la littérature, un compte-rendu rationnel et algorithmique d'un comportement curieux démontré par les jeunes enfants, appelé la sur-imitation, généralement considéré comme preuve pour l'irrationalité des enfants. Le 4ème chapitre conclut en explorant la complexité computationnelle du problème considéré et en présentant les résultats correspondants de NP-difficulté.

Le 5ème chapitre revisite la notion fondamentale de l'indépendance probabiliste conditionnelle en tant que concept central qui donne lieu à une minimalité dans les contextes probabilistes. Le 5ème chapitre, pour la première fois dans la littérature, propose un algorithme asynchrone, distribué, de transmission de message - semblable, en esprit, au schéma de propagation de croyance de Pearl - afin d'implémenter la notion clé de d-séparation de Pearl. En outre, grâce à l'introduction d'une notion clé de théorie graphique, appelée module de réfutation minimale, le 5ème chapitre montre comment la notion de minimalité se

manifeste dans une implémentation distribuée et transmise par message de la d-séparation. L'algorithme proposé présente des propriétés intrigantes qui le positionnent comme un candidat plausible pour la mise en œuvre de la notion de d-séparation au niveau algorithmique d'analyse de Marr. En outre, l'algorithme proposé surpasse tous les algorithmes précédemment proposés dans la littérature en temps d'exécution dans le pire des cas et sert de premier compte-rendu rationnel, distribué, de niveau du processus, de la façon dont les humains doivent gérer l'indépendance probabiliste.

Le 6ème chapitre explore la notion de minimalité dans le contexte de l'apprentissage et de l'imagination. Le chapitre, pour la première fois dans la littérature, propose un cadre plausible au niveau neural et efficace en terme de calculs qui permet de transformer tout réseau neuronal déterministe et discriminatif (par ex., réseaux neuronaux convolutionnels profonds et perceptron multicouches) en un modèle probabiliste et génératif. À l'aide de ce cadre, les réseaux neuronaux en cascade-corrélation (CCNN pour "cascade-correlation neural networks" en anglais) - une classe de modèles auto-organisés, déterministes et discriminatifs qui ont réussi à expliquer une variété de phénomènes psychologiques - sont transformés en modèles génératifs probabilistes, permettant ainsi aux CCNN de générer (de manière probabiliste) des échantillons d'une catégorie d'intérêt. Il est important de souligner que le cadre proposé: (1) offre un compte-rendu modulaire de l'imagination humaine, soutenu par des études sur l'apprentissage et les capacités imaginatives des patients amnésiques de l'hippocampe ainsi que par un nombre croissant d'études d'imagerie cérébrale suggérant que la perception et l'imagerie partagent une représentation neurale ( 2) donne lieu à des modèles génératifs auto-organisés, (3) suggère fortement que, contrairement à une vue largement répandue, la limite entre les modèles discriminatifs et génératifs est floue, (4) rapproche les niveaux d'analyse informatique, algorithmique et d'implémentation, et enfin, (5) relie deux écoles de pensée dominantes dans les sciences cognitives, c'est-à-dire le connexisme et la cognition bayésienne.

Ensemble, les idées explorées dans cette thèse suggèrent que la poursuite de la notion de minimalité est un effort fructueux pour comprendre la cognition aux niveaux informatique et algorithmique et, tout aussi important, démontrent comment la notion de minimalité permet de développer de nouveaux problèmes de calcul ainsi qu'introduire de nouveaux algorithmes, structures de données et plusieurs concepts algorithmiques. Inspiré par cela, le chapitre 7 conclut la thèse en proposant un nouveau mode d'enquête, appelé le programme rationnel minimaliste, décrivant une méthodologie rationnelle et fondée sur des principes pour étudier

la cognition au niveau des processus.

## Acknowledgments

My graduate studies have been anything but predictable, an extraordinary journey from utter confusion to pleasing clarity. All I knew when it all started, roughly six years ago, was that it had better touch on three things — math, mind, and of course, philosophy — or I would certainly be unhappy once I get to the end of it.[1] Having absolutely no idea of what I would have to be doing throughout my PhD, I simply followed two ideas: (1) *"[t]he scientific problem chooses you, you don't choose it"*[2], (2) *"[s]tudy hard what interests you the most in the most undisciplined, irreverent and original manner possible"*[3]. It turned out that these two pieces of advice were virtually all I needed. But carrying out (2) would have been almost impossible without having an extraordinarily understanding adviser. I was extremely fortunate to have Yannis Psaromiligkos as my advisor, and I am forever indebted to him for giving me the freedom to explore my interests wherever they would take me, and of course, for his unwavering friendship; I simply couldn't have wished for a better adviser. Thank you Yannis for everything! You are undoubtedly one of the best things that have ever happened to me in life!

I was unbelievably lucky to have Tom Shultz and Luc Devroye in my committee. Purely by chance, one day I dropped by Tom's computational psychology class, and the day after, his cognitive science course, and I immediately came out of that experience knowing that Tom is the one! I soon joined Tom's lab, and it didn't take me long to realize that it was simply one of best things that could've happened to me. I will certainly miss Friday morning meetings, with Tom always enthusiastically listening to my crazy ideas even when they were completely out of scope! Thank you Tom for all your patience, support, and contagious positivity! My first encounter with Luc Devroye was also purely by chance, when one day I dropped by his graduate algorithm class. I wasn't making much progress on the problem that I was working on at the time, and I thought to myself perhaps dropping by a class would be a good way of restarting my mind's Markov chain Monte Carlo (MCMC) search process.[4] Seeing Luc was all I needed for restarting my MCMC on that day—and perhaps for all my future MCMCs. The level of passion with which he was presenting his material,

[1] Truth be told, had these elements been missing, I wouldn't have had the motivation to complete my thesis.

[2] Allen Newell (1991).

[3] Richard Feynmann (1965).

[4] The phrase "restarting mind's MCMC" is due to a very close friend of mine, Falk Lieder, which he jokingly used at the end of one of our discussions in CogSci'17.

and the breadth of topics he was drawing on were, in simple terms, greatly inspiring. Very soon I asked him to join my committee, which he very kindly agreed, and ever since, every single meeting of mine with Luc has been a true joy. Without Luc, my PhD life would've been, in the true sense of the word, "incomplete!"

A great many wonderful people helped me in my PhD studies, sometimes by only a few words of wisdom and/or comfort: Thomas Icard, Falk Lieder, Emma Tecwyn, Benoit Champagne, Daphna Buchsbaum, Jad Kabbara, Aida Nematzadeh, Aditya Mahajan, Francois Cote, Morteza Dehghani, Ayoub Saab, Ali Shahrad, Peter Helfer, Marcel Montrey, Ahmad Rashid, Milad Kharratzadeh, Kevin da Silva Castanheira, Billy Campoli, Artem Kaznatcheev, and many more; I am truly thankful to all of you.

Special thanks go to Jad Kabbara with whom I spent a great deal of time chatting and contemplating wonderful ideas; I enjoyed every single moment of it. Thanks for everything Jad; you are one of the nicest people I have ever seen in my life!

I also would very much like to thank Tim O'Donnell and Alan Jern, who graciously agreed to review my dissertation. Thank you Tim for your wonderful comments and for your illuminating questions in my defence.

And, finally, to three wonderful people, Aida, Zhila, Ardeshir, whose love is everything to me in life: Thanks for being you!

# Contents

# List of Figures

# List of Acronyms

**ACE** automatic causal encoding
**BFS** breadth first search
**BN** Bayesian network
**BP** belief propagation
**CBN** causal Bayesian network
**CCNN** cascade-correlation neural network
**CFP** causal frame problem
**CPD** conditional probability distribution
**CUG** color update grammar
**DAG** directed acyclic graphs
**D-RMP** dual rational minimalist program
**FIFO** first-in first-out
**fMRI** functional magnetic resonance imaging
**IP** intervention policy
**IT** inference threshold
**IT-RS** inference threshold root set
**JPD** joint probability distribution
**KBCC** knowledge-based cascade-correlation
**LDPC** low-density parity-check
**LP** linear programming
**LSM** locally structurally minimal

## Acronyms

**LTI** linear time invariant
**LTM** long term memory
**LTWM** long-term working memory
**MAL** Markov-Adjusted Langevin
**MCM** multi-context model
**MCMC** Markov chain Monte Carlo
**M-WM** multi-component model of working memory
**OIP** optimal intervention policy
**PDP** parallel-distributed-processing
**PGM** probabilistic graphical models
**PL** potential level
**PLIF** potential-level-based inference framework
**RMP** rational minimalist program
**RV** random variable
**SCT** structure control theory
**SDCC** sibling-descent cascade-correlation
**TCS** theoretical computer science
**TPS-controllability** targeted probabilistic structural controllability
**USM** uniformly structurally minimal

# Chapter 1

# Introduction

*"Truth is ever to be found in the simplicity,*
*and not in the multiplicity and confusion of things."*
— Isaac Newton, Treatise on Revelation

- What are the minimal set of conditions a planet must possess for life to emerge? (The central question in Biophysics.)
- What are the minimal set of physical laws that ought to be invoked to explain our universe? (The elusive unified theory yet to be discovered.)
- What is the minimal set of properties a Boolean formula must possess so that its satisfiability can be decided in polynomial time by a deterministic Turing machine? (Answering this question settles the famous $\mathcal{P}$ versus $\mathcal{NP}$ debate in computational complexity.)
- What is the minimal amount of innate knowledge a child must be equipped with to find their way to adult-level intelligence? (The famous Empiricism vs. Innatism debate in psychology and philosophy of mind.)
- What is the minimal set of syntactic rules/operations generative grammar should posit to account for the syntax of all human language? (In their recent book, Berwick and Chomsky (2015) argue that only one suffices, namely, Merge.)

The above questions, *prima facie*, seem to be as detached as one could possibly imagine—at the very least, each is the business of a whole *field* of its own. However, upon a closer look, an intricate commonality, echoing through the word *minimal*, starts to coalesce. And yet, the story goes the same for all profound notions in the history of human thought. In a heroic act, they come to rescue us from a world of disparity to a world of harmony, from a world of individuality to a world of unity. The notion of minimality is of such nature. A glance at the list outlined above, which is far from being exhaustive, attests to this claim.

Occam's razor, Kolmogorov's complexity, the minimum description length principle, the principle of least effort, Chomsky's Minimalist Program in linguistics, the notion of necessary-and-sufficient condition in logic, the principles of least time and least action in physics, the notion of minimal sufficiency in statistics, and finally the statement "everything should be made as simple as possible, but not simpler" (often attributed to Albert Einstein) all essentially bear on the notion of minimality. In arts, a school of thought named *Minimalism* adheres to the maxim of "reducing everything to its bare essentials" and "stripping away any redundancies."

There indeed lies a deep sense of perfectionism at the heart of minimality. In this vein, the notion of *sufficiency* is invoked as a mere relaxation of this elusive notion, allowing one to aim at a potentially more attainable goal. Let us characterize the notions of minimality and sufficiency as follows.

**Def 1.1.** Set $S$ is *sufficient* for task $\mathfrak{T}$ iff using the members of $S$ (i.e., the information contained in $S$), $\mathfrak{T}$ can be accomplished.

**Def 1.2.** Set $S$ is *minimal* for task $\mathfrak{T}$ iff (i) $S$ is sufficient for $\mathfrak{T}$, and (ii) any proper subset of $S$ is not sufficient for $\mathfrak{T}$.

The above characterization is broad and, admittedly, open to a variety of interpretations. However, it provides the reader with a sense—albeit imprecise—of the notions under discussion.[1] In what follows, we safely focus our discussion on minimality rather than sufficiency with this key understanding in mind that the latter can be invoked as a mere relaxation of the former at will.

[1]The line of work pursued in this dissertation lends itself to the provided characterization. However, by the time the reader will finish his/her journey through the dissertation, we hope that he/she will have reached the verdict that the notion of minimality is even deeper and more elusive than the provided characterization, begging for a yet broader and perhaps more elegant characterization.

## 1.1 Minimality in Cognition, Philosophy, and Philosophy of Mind

In philosophy of mind, the notion of minimality is well captured, arguably, in the quest for understanding the *essence* of an entity—the truly essential features (aka *attributes*) of an entity without which it would simply lose its *identity*. The above ideas can be restated in the context of the classical view of *concept* in epistemology and cognitive psychology as follows: Let an entity $e$, possessing a set of features $F$, belong to the category (or concept) $C$. The notion of essence concerns the following questions: What subset of $F$ is responsible for $e$ being categorized as a member of $C$, the removal of which would render $e$ to be no longer a member of $C$?[2] To clarify, let us consider the following example. Let $x$ be a member of the concept 'tree.' Now could $x$ have no leaves and still be a tree or is having leaves essential for an entity to be categorized as a tree? Could knowing the shape, color, or texture affect $x$ being categorized as a tree? All these questions bear on the notion of essence.

But, perhaps, nothing better than the notion of *simplicity* captures what is really at the core of the notion of minimality. The history of philosophy (and that of science, alike) is filled with remarks on the privileged status of simplicity, elevating it to a theoretical virtue in and of itself. It is present in Aristotle's writings when he writes in his *Posterior Analytics* "We may assume the superiority *centeris paribus* [other things being equal] of the demonstration which derives from fewer postulates or hypothesis;" and its connection to minimality explicitly manifests itself in Thomas Aquinas's writings when he writes:

> "If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments when one suffices" (Aquinas, 1945).

In the same vein, Immanuel Kant (1781), in *Critiques of Pure Reason*, advocates the dictum that "rudiments or principles must not be unnecessarily multiplied." Issac Newton (1687), in *Principia Mathematica*, remarks that "Nature is pleased with simplicity, and affects not the pomp of superfluous causes;" and Galilei (1632) writes "Nature does not multiply things unnecessarily; that she makes use of the easiest and simplest means for producing her effects; that she does nothing in vain and the like." In sciences, simplicity is often evoked by appealing to Occam's razor principle, which goes as follows: "Entities are not to be multiplied without necessity." A large body of work in philosophy, and, particularly, philosophy of science, has

[2] In computational complexity theory, one is only concerned with the hardness of deciding whether an instance $x$ is a member of set (aka language) $C$, and not with the notion of essence.

investigated possible rational grounds for simplicity as a guiding principle in theorizing about nature (see, e.g., Baker, 2016, for a review of such arguments).

Now, a fundamental question presents itself: "Are simplicity and minimality essentially the same?" Although we admit that this question certainly deserves much contemplation,[3] for the purpose of this dissertation, we view minimality as a formal way to operationalize simplicity, in our attempt to develop a formal, principled, rational methodology for studying cognition at the algorithmic level of analysis, which we outline in the epilogue chapter to this dissertation (Chapter 7); also, it is worth noting that the line of work pursued in Chapters 2 to 6 can all be viewed as instantiations of that methodology.

### 1.1.1 On the Connection to Bounded Rationality

There is a deep connection between the notion of minimality and Simon's (1957) bounded rationality. According to bounded rationality, a reasoner is inevitably bounded in time and computational resources and has to accomplish a task of interest $\mathfrak{T}$ subject to this constraint. The boundedly-rational reasoner then has to strive to only attend to the information (either already at her disposal or receiving it in real time through sensory organs) which is deemed relevant to the task $\mathfrak{T}$, and ignore possibly all[4] that is irrelevant to $\mathfrak{T}$. In this vein, the notion of *attention* and that of *executive functions*—as a set of mechanisms responsible for guiding attention—bear on the interplay between minimality and bounded rationality. For example, in the case of visual attention, a boundedly-rational reasoner strives to attend only to the parts of visual stimuli deemed relevant to the task of interest (say, a face recognition task).

The realization of the fact that the reasoner is inevitably boundedly-rational signifies the role that minimality plays in cognition. Indeed, had the reasoner possessed unbounded time and computational resources, the notion of minimality would be of no significance for cognition.

### 1.1.2 On the Connection to the Frame Problem

The frame problem is a puzzle in philosophy of mind and epistemology which, in its most generality, is characterized by the Stanford Encyclopedia of Philosophy as follows: *"How do*

[3] Nonetheless, the observation that simplicity is very often expressed by appealing to minimality strongly suggests that the latter is more fundamental than the former.

[4] Depending on the task of interest, this may sound too demanding for the reasoner. The notion of sufficiency then comes into play, as we discussed earlier, to relax it.

*we account for our apparent ability to make decisions on the basis only of what is relevant to an ongoing situation without having explicitly to consider all that is not relevant?"* The frame problem, at its core, is essentially concerned with minimality. This understanding becomes more evident if we characterize the notion of minimality (and, by extension, sufficiency) in terms of the notions of *relevance* and *irrelevance* as follows.

**Def. 1.3.** Set $S$ is *sufficient* for task $\mathfrak{T}$ iff $S$ contains all that is *relevant* to $\mathfrak{T}$.

**Def. 1.4.** Set $S$ is *minimal* for task $\mathfrak{T}$ iff (i) $S$ is *sufficient* for $\mathfrak{T}$, and (ii) $S$ contains nothing that is *irrelevant* to $\mathfrak{T}$ (hence, no redundancy in $S$).

It is worth emphasizing that, akin to the notion of minimality, the frame problem is of any significance for cognition insofar as the reasoner is presumed to be boundedly-rational. That is, it is the very self-awareness of a cognitive system of its bounded rationality that leads the system to act in the manner portrayed by the characterization of the frame problem given above.

## 1.2 When Minimality Meets other Key Notions

As we go through this dissertation and investigate the notion of minimality (and its relaxation, sufficiency) in different settings, we will encounter other key notions, e.g., *symmetry*, *invariance*, *scale-invariance* (aka *self-similarity*), *locality*, *maximal-informativeness*, *anytime algorithms*, and *nestedness*. We will also encounter *asynchronous distributed algorithms* as mechanisms requiring no coordination between computing agents.[5] We will revisit these notions and highlight their significance in Chapter 7 where we conclude the dissertation. In the mean time, we would like to ask the reader to attend to these notions and their interplay as he/she walks through the dissertation.

## 1.3 Dissertation Outline

This dissertation is comprised of four main parts. Part I explores the notion of minimality in the contexts of *probabilistic reasoning* and *causal reasoning*. Part II explores how the notion of minimality emerges in the context of *action* and *control*. Part III revisits the key notion of (probabilistic) conditional independence—as the core concept which gives rise to minimality

[5] Asynchronous, distributed algorithms grant the least level of *control* to be exercised throughout an execution, as computing agents *autonomously* engage in message-passing. We will see more on this in Chapter 5.

in probabilistic settings—and particularly, Pearl's graph-theoretic notion of $d$-separation, as a fundamental concept for verifying independence statements. Finally, Part VI explores the notion of minimality in the contexts of learning and imagination. The content of each part is explicated in more detail next.

### Part I: On Minimality in Reasoning

Part I is comprised of Chapters 2 and 3, the contents of which are discussed below.

Chapter 2 addresses, for the first time in the literature, how the notion of minimality can be applied to probabilistic reasoning under partial knowledge. To this end, drawing on the notion of bounded rationality manifested in a reasoner's limited attention span and scope, Chapter 2 puts forth a novel graphical model, termed the Multi-Context Model (MCM), to represent the reasoner's state of partial knowledge of a domain. MCM occupies a middle ground between Probabilistic Logic, Bayesian Logic, and Probabilistic Graphical Models. Also, drawing on the quintessence of Bayesian networks (BNs), i.e., the concept of *conditioning*, MCM generalizes BN to the realm of partial knowledge. Importantly, MCM serves as the first *rational*, *probabilistic*, representational-level account of an important developmental shift from features in isolation to correlations between those features, in infants between four and ten months of age.

Inspired by Simon's bounded rationality and drawing on the notion of minimality, Chapter 3 provides a novel algorithmic perspective to the causal variant of the frame problem (CFP), a deep puzzle in philosophy of mind and epistemology. Chapter 3 begins by introducing a notion called potential level (PL). PL generalizes the graph-theoretic concept of topological sorting, and extends the fundamental notion of Lamport's logical clock to causal Bayesian networks (CBNs). Drawing on the psychological literature on causal judgment, Chapter 3 substantiates the claim that PL may bear on how *time* is encoded in the mind. Using PL, Chapter 3 then proposes an inference framework, called the PL-based inference framework (PLIF), permitting a boundedly-rational approach to the CFP, formally articulated at Marr's algorithmic level of analysis. PLIF is also shown to be consistent with a wide range of findings in the causal judgment literature. To our knowledge, PLIF is also the first inference framework that capitalized on time to constrain the scope of causal reasoning over CBNs, and importantly, can handle any inference mechanism. Interestingly, the ideas explored in Chapter 2 and 3 demonstrate how the old concept of *imprecise probabilities* naturally emerges out of Simon's bounded rationality.

### Part II: On Minimality in Action

Part II is comprised of Chapter 4, the content of which is discussed below.

Chapter 4 pursues the notion of minimality in the key context of action and control. Chapter 4 studies, for the first time in the literature, the problem of probabilistic controllability in CBNs. Probabilistic controllability extends the fundamental concept of controllability in control theory to probabilistic CBNs. More specifically, the aim of Chapter 4 is two-fold: (i) to introduce and formalize the problem of probabilistic structural controllability in CBNs, and (ii) to identify a sufficient set of driver variables for the purpose of probabilistic structural controllability of a generic CBN. Furthermore, Chapter 4 elaborates on the nature of minimality that the identified set of driver variables satisfies. The results of Chapter 4 have important implications for a line of work in developmental psychology concerning causal learning by young children in pedagogical settings. Also, the formalism developed in Chapter 4 establishes, for the first time in the literature, a rational, algorithmic-level account of a curious behavior demonstrated by young children called *overimitation*, generally taken as evidence for children's irrationality. Chapter 4 concludes by exploring the computational complexity of the problem under study and presenting $\mathcal{NP}$-hardness results for it.

### Part III: Conditional Independence, $d$-separation, and Minimality

Part III is comprised of Chapter 5, the content of which is discussed below.

Chapter 5 revisits the fundamental notion of conditional probabilistic independence as the core concept which gives rise to minimality in probabilistic settings. Chapter 5, for the first time in the literature, proposes an asynchronous, distributed, message-passing algorithm—akin, in spirit, to Pearl's Belief Propagation scheme—so as to implement Pearl's key notion of $d$-separation. Also, through the introduction of a key graph-theoretic notion, termed minimal refutation-module, Chapter 5 shows how the notion of minimality manifest itself in a distributed, message-passing implementation of $d$-separation. The proposed algorithm exhibits intriguing properties which position it as a plausible candidate for the implementation of $d$-separation at Marr's algorithmic level of analysis. Furthermore, the proposed algorithm outperforms all the previously proposed algorithms in the literature in terms of worst-case running time, and serves as the first rational, *distributed*, process-level account of how humans handle probabilistic independence.

### Part IV: On Minimality in Learning and Imagination

Part IV is comprised of Chapter 6, the content of which is discussed below.

Humans are not only adept in recognizing what class an input instance belongs to (i.e., classification task), but perhaps more remarkably, they can imagine (i.e., *generate*) plausible instances of a desired class with ease, when prompted. Inspired by this, Chapter 6 explores the notion of minimality in the contexts of learning and imagination. Chapter 6, for the first time in the literature, proposes a *neurally-plausible* and *computationally-efficient* framework, allowing to transform any deterministic, discriminative neural network (e.g., deep convolutional neural networks and multilayer perceptron) into a probabilistic, generative model. The proposed framework is based on a Markov chain Monte Carlo (MCMC) method, called the Metropolis-adjusted Langevin (MAL) algorithm, which capitalizes on the gradient information of the target distribution to direct its explorations towards regions of high probability, thereby achieving good mixing properties. (It is crucial to note that our proposed framework can accommodate any gradient-based MCMC method in order to achieve good mixing and convergence properties.) Using this framework, cascade-correlation neural networks (CCNNs)—a class of deterministic, discriminative neural networks which construct their topology in a minimal fashion and have been successful in accounting for a variety of psychological phenomena—are converted into probabilistic generative models, thereby enabling CCNNs to probabilistically generate samples from a category of interest. Importantly, the proposed framework: (1) suggests a modular account of human imagination which is supported by studies on learning and imaginative abilities of hippocampal amnesic patients as well as a growing body of brain imaging studies showing that perception and imagery share neural representation, (2) gives rise to *self-organized* generative models, (3) strongly suggests that, contrary to a widely-held view, the boundary between discriminative and generative models is blurry, (4) bridges computational, algorithmic, and implementational levels of analysis, and finally, (5) connects two dominant schools of thought in cognitive sciences, namely, connectionism and Bayesian cognition.

Finally, inspired by the results of Chapters 2 to 6, Chapter 7 concludes the work by proposing a new mode of enquiry, termed the Rational Minimalist Program, which integrates Anderson's rational analysis methodology and the key notion of minimality. Concretely, Rational Minimalist Program outlines a principled, rational methodology for studying cognition at Marr's algorithmic level of analysis.

## 1.4 Contributions & Publications

The work presented in this dissertation is an original contribution of the author, and parts of it have appeared in a number of publications, the list of which is given below.

- **A. S. Nobandegani** & I. N. Psaromiligkos; Multi-Context Models for Reasoning under Partial Knowledge: Generative Process and Inference Grammar, In *Proc. of the* $31^{st}$ *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.
- **A. S. Nobandegani** & I. N. Psaromiligkos; The Causal Frame Problem: An Algorithmic Perspective, In *Proc. of the* $39^{th}$ *Annual Conference of the Cognitive Science Society (CogSci)*, 2017.
- **A. S. Nobandegani** & T. R. Shultz; Converting Cascade-Correlation Neural Nets into Probabilistic Generative Models, In *Proc. of the* $39^{th}$ *Annual Conference of the Cognitive Science Society (CogSci)*, 2017.

Throughout, the line of work presented in this dissertation simultaneously follows two persistent themes: (1) How ideas and observed effects in cognitive psychology can be used to develop cognitively-inspired algorithms, machine leaning concepts, and human-like artificial intelligence, and equally importantly, (2) How theoretical computer science (TCS) mindset (often advocated by echoing the term *algorithmic lens*, and mainly concerned with rigorous definitions, formalization, axiomatization, design and analysis of algorithms, data structures, and computational complexity theory) allows for developing formal, mathematically-rigorous foundations for ideas and observed effects in cognitive psychology, thereby enriching our understanding of their computational underpinnings. Hence, this dissertation is an instantiation of a research program which aims to bridge TCS and cognitive psychology, allowing these two fields to communicate, to benefit from each other's history as well as advances, and importantly, to engender new advances in each other through synergistic interactions. (It is worth noting that contemporary computational cognitive science borrows its mathematical tools predominantly from statistical machine learning and computational statistics, and relatively rarely makes contact with the ideas in TCS alluded to above.) I will return to the idea of bridging TCS and cognitive psychology in the epilogue chapter to this dissertation (Chapter 7), where I formally articulate a new mode of enquiry, called Rational Minimalist Program, as a principled, rational methodology for studying cognition at Marr's algorithmic level of analysis.

Finally, in the following two subsections, I outline the main contributions of the dissertation to theoretical computer science, artificial intelligence, and machine learning, on the one hand, and to cognitive psychology, neuroscience, and computational cognitive science, on the other.

### 1.4.1 Main Contributions to Theoretical Computer Science, Artificial Intelligence, and Machine Learning

In the following, the main contributions of each chapter pertaining to theoretical computer science (TCS), artificial intelligence, and machine learning are outlined.

- ▷ **Chapter 2**: Chapter 2, for the first time in the literature, explores how the notion of minimality can be applied to probabilistic reasoning under partial knowledge. Concretely, the main contributions of Chapter 2 are as follows:
  - Formally presenting the first graphical model specifically tailored toward capturing the state of partial knowledge in probabilistic settings, called multi-contex model (MCM). MCM occupies a middle ground between Probabilistic Logic, Bayesian Logic, and Probabilistic Graphical Models, and generalizes Bayesian networks (BNs) to the realm of partial knowledge.
  - Formally presenting a generative process allowing to form partial beliefs over a domain, in a gradual, contradiction-free manner.
  - Introducing and formalizing tow key concept of *nestedness* and *transformation* in the context of MCM, allowing for computationally efficient inference in MCM.
  - Presenting a computationally-efficient algorithm for handling evidential inference in MCM, outputting optimal bounds to any given query of interest.
  - Introducing the key notions of *scale-invariance* in the context of MCM, allowing for efficient, lifted inference in MCM.
- ▷ **Chapter 3:** Chapter 3 formally presents a novel algorithmic perspective to the causal variant of the frame problem, a deep puzzle in epistemology and philosophy of mind. Concretely, the main contributions of Chapter 3 are as follows:

- Formally introducing potential level (PL), a generalization of the two important concepts of topological sorting and Lamport's logic clock to causal Bayesian networks (CBNs).
- Endowing CBNs with PL, thereby introducing a new data structure which allows for efficient submodel selection over CBNs (which, in turn, permits efficient, targeted, inference over CBNs).
- Formally introducing the key notion of *maximally-informativeness*, a broadly applicable information-theoretic performance guarantee for anytime algorithms.
- Formally presenting a novel, anytime, algorithmic approach to the causal frame problem, which satisfies the maximally-informativeness property.

▷ **Chapter 4:** Chapter 4, for the first time in the literature, studies the problem of probabilistic structural controllability in CBNs. Concretely, the main contributions of Chapter 4 are as follows:

- Introducing and formalizing the problem of probabilistic structural controllability in CBNs.
- Identifying a set of driver variables for the purpose of probabilistic structural controllability of a generic CBN, which is shown to be *both* minimal and optimal.
- Presented a linear-time algorithm $\mathcal{C}^*$ for identifying the aforesaid set of driver nodes, which easily lends itself to an asynchronous message-passing implementation. Surprisingly, $\mathcal{C}^*$ is among the rare cases of *correct*, greedy algorithms (i.e., involving no approximations).
- Formally introducing the notions of *i-subsumability* and *i-domination*, broadly applicable for problems involving strategic planning and policy making using CBNs.
- Formally introducing two important structural notions of minimality, namely, *local structural minimality* and *uniform structural minimality*, broadly applicable for problems involving strategic planning and policy making using CBNs.
- Characterizing the computational complexity of the task under study, and presenting $\mathcal{NP}$-harness results for it. Interestingly, the $\mathcal{NP}$-hardness results are established using a special class of (degenerate) CBNs for which any Exact Inference or Maximum A-Posterior (MAP) query can be answered in poly-time (hence, tractable).

▷ **Chapter 5:** Chapter 5 revisits the fundamental notion of conditional probabilistic independence as the core concept which gives rise to minimality in probabilistic settings. Concretely, the main contributions of Chapter 5 are as follows:

- Proposing, for the first time in the literature, an asynchronous, distributed, message-passing algorithm for implementing Pearl's key notion of $d$-separation.
- Importantly, the proposed algorithm outperforms all past algorithms in the literature in terms of worst-case running time.
- Formally showing how the notion of minimality manifests itself in a distributed, message-passing implementation of $d$-separation, through the introduction of a key graph-theoretic notion, called *minimal refutation-module*.
- Showing the fruitfulness of separately studying the runtime of an algorithm on NO-instances and YES-instances, even for polynomially-solvable problems (i.e., problems in the complexity class $\mathcal{P}$).
- Showing how the graph-theoretic notion of minimal refutation-module can be used as a natural parameter for studying $d$-separation in the context of parameterized complexity.
- Formally demonstrating how pursuing the notion of minimality makes contact with the key notion of shortest disproof (shortest proof) for the complexity class co$\mathcal{NP}$ ($\mathcal{NP}$) in automated theorem-proving.

▷ **Chapter 6:** Chapter 6 explores the notion of minimality in the context of learning and imagination. Concretely, the main contributions of Chapter 6 are as follows:

- Presenting, for the first time in the literature, a neurally-plausible and computationally efficient framework which allows to transform any deterministic, discriminative neural network (e.g., deep convolutional neural networks and multilayer perceptron) into a probabilistic, generative model.
- Given that the hierarchical structure of deterministic, discriminative neural networks permits efficient computation of gradient and higher-order derivatives, we showed how gradient-based MCMCs and deterministic, discriminative neural networks can be naturally paired up for computationally-efficient handling of example generation tasks.

- Given that learning probabilistic models (e.g., Restricted Boltzman Machines and Deep Boltzman Machines) is computationally intractable in general, our framework offers a much more computationally efficient way of obtaining probabilistic models.

▷ **Chapter 7:** Chapter 7 articulates a new mode of enquiry, called Rational Minimalist Program (RMP), as a principled, rational methodology for studying cognition at Marr's algorithmic level of analysis. Concretely, the main contributions of Chapter 7 are as follows:

- RMP permits bridging between TCS and cognitive psychology, allowing for developing cognitively-inspired algorithms and human-like artificial intelligence.
- RMP makes contact with a broad range of topics in TCS, most notably: design and analysis of algorithms, exact and approximation algorithms, parameterized complexity, fixed-parameter tractability, inapproximability, shortest proof and shortest disproof for complexity classes $\mathcal{N}P$ and co$\mathcal{N}P$.

### 1.4.2 Main Contributions to Cognitive Psychology, Neuroscience, and Computational Cognitive Science

In the following, the main contributions that each chapter makes with regard to cognitive psychology, neuroscience, and computational cognitive science are outlined.

▷ **Chapter 2:** Drawing on the notion of bounded rationality manifested in a reasoner's limited attention span and scope, Chapter 2 presents a novel graphical model, termed MCM, to represent the reasoner's state of partial knowledge of a domain, where the term partial signifies a reasoner's complete lack of knowledge as to parts of the underlying dependency structure of the domain. Concretely, the main contributions of Chapter 2 are as follows:

- Given the prominent role of BN in Bayesian models of cognition (Gopnik et al., 2004, *inter alia*), our proposed model generalizes BN to the realm of partial knowledge.
- To our knowledge, MCM is the first *normative*, *parsimonious*, representational-level model for capturing the state of partial knowledge.

- The proposed algorithm, $\mathcal{I}^*$, can be viewed as a rational process model (Griffiths et al., 2012) for inference under partial knowledge, in a domain modeled by MCM.
- MCM serves as the first *normative*, *probabilistic*, representational-level account of an important developmental shift in infant information processing, between four and ten months of age.

▷ **Chapter 3:** Chapter 3 proposes an inference framework, called the PL-based inference framework (PLIF), permitting a boundedly-rational approach to the causal frame problem, formally articulated at Marr's algorithmic level of analysis. Concretely, the main contributions of Chapter 3 are as follows:

- Substantiating the claim that PL may bear on how *time* is encoded in the mind.
- Showing that PLIF is consistent with a wide rage of findings in the literature.
- Demonstrating how the old concept of imprecise probabilities naturally emerges out of Simon's (1957) bounded rationality.
- PLIF is *not* from a "god's eye" point of view, and can handle any inference algorithm, including sample-based inference methods widely advocated in Bayesian models of cognition.
- Consistent with a growing acknowledgment in the literature that, not only time and causality are intimately linked, but that they *mutually constrain* each other in human cognition (see Buehner, 2014), PLIF formally shows how time can guide and constrain causal reasoning.

▷ **Chapter 4:** Chapter 4 pursues the notion of minimality in the key context of action and control. Concretely, the main contributions of Chapter 4 are as follows:

- Proposing an algorithm $\mathcal{C}^*$ which serves as the first rational, process-level account of how human adults devise their intervention strategies (by selecting on which intervenable variables to intervene) to control the state of a target node.
- The main prediction of $\mathcal{C}^*$, termed *the proximity principle*, is supported by experimental findings.
- $\mathcal{C}^*$'s output, $\boldsymbol{\mathcal{X}}^*$, serves as a distinctive pedagogical cue helping young children, not yet having developed elaborate intuitive theories, learn about the causal structure of their environment.

- $\mathcal{C}^*$ is the first rational, process-level account of a curious behavior demonstrated by young children called *overimitation*, generally taken as evidence for children's irrationality.

▷ **Chapter 5:** Chapter 5 revisits the fundamental notion of conditional probabilistic independence as the core concept which gives rise to minimality in probabilistic settings. Concretely, the main contributions of Chapter 5 are as follows:

- The proposed algorithm $\mathcal{D}^*$ serves as the first rational, *distributed*, process-level account of how humans handle probabilistic independence.
- $\mathcal{D}^*$ permits the implementation of $d$-separation in an asynchronous, distributed, message-passing fashion—a property consistent with the brain's computational machinery (see McClelland, 1989; Chater et al., 2006, *inter alia*) and fully in the spirit of the celebrated parallel-distributed-processing (PDP) research program in brain and cognitive sciences.
- $\mathcal{D}^*$'s use of BN links as a medium for inference is supported by recent work in neuroscience investigating possible implementation of BNs at the neural level.
- $\mathcal{D}^*$ demonstrates a peculiar tendency toward quick detection of NO-instance $d$-separation queries, which can be normatively-justified.

▷ **Chapter 6:** Chapter 6, for the first time in the literature, proposes a *neurally-plausible* and *computationally-efficient* framework which allows to transform any deterministic, discriminative neural network (e.g., deep convolutional neural networks and multilayer perceptron) into a probabilistic, generative model. Concretely, the main contributions of Chapter 6 are as follows:

- Converting cascade-correlation neural networks (CCNNs)—a class of self-organized, deterministic, discriminative models which have been successful in accounting for a variety of psychological phenomena—into probabilistic generative models, thereby enabling CCNNs to probabilistically generate exemplars from a category of interest.
- Our proposed framework gives rise to *self-organized* generative models: generative models possessing the self-constructive property of CCNNs. Such self-organized generative models could provide a wealth of developmental hypotheses as to how

the imaginative capacities of children change over development, and models with quantitative predictions to compare against.

- In accord with the maxim of Occam's razor, the proposed framework suggests that, in order to account for human generative abilities, one need not adhere to an encoder-decoder-type architecture (involving a forward model (encoder) and a fully separate inverse model (decoder)), but a single forward model, upon which MCMC operates, might suffice—a more parsimonious design.
- Importantly, the proposed framework: (1) suggests a modular account of human imagination which is supported by studies on learning and imaginative abilities of hippocampal amnesic patients as well as a growing body of brain imaging studies showing that perception and imagery share neural representation, (2) bridges computational, algorithmic, and implementational levels of analysis, (3) strongly suggests that, contrary to a widely-held view, the boundary between discriminative and generative models is blurry, and finally, (4) connects two dominant schools of thought in cognitive sciences, namely, connectionism and Bayesian cognition.

▷ **Chapter 7:** Chapter 7 formally articulate a new mode of enquiry, called Rational Minimalist Program (RMP), as a principled, rational methodology for studying cognition at Marr's algorithmic level of analysis. Importantly, Chapter 7 shows that the line of work pursued in Chapters 2 to 6 can all be viewed as instantiations of this methodology. RMP aims to bridge TCS and cognitive psychology, allowing these two fields to communicate, to benefit from each other's history as well as advances, and importantly, to engender new advances in each other through synergistic interactions.

# Part I:
# On Minimality in Reasoning

**Preface.** In Part I, the key role that the notion of minimality plays in the context of *probabilistic reasoning under partial knowledge* (Chapter 2) as well as *causal reasoning* (Chapter 3) is explored. Drawing on the fundamental understanding that a reasoner's attention span/scope is inevitably limited—as a manifestation of Simon's bounded rationality—Chapter 2 explores the question of how the reasoner whose probabilistic knowledge of a domain is acquired under such constraints can go about answering a probability of interest (called query) by merely entertaining those pieces of knowledge deemed relevant to the said task. In this light, the line of work pursued in Chapter 2 can be perceived as an adaptation of the well-known frame problem (FP)—a deep epistemological puzzle—to the realm of reasoning under partial knowledge. Chapter 3 explores the causal variant of the FP, the causal frame problem (CFP). Intriguingly, the line of work explored in Chapter 3 suggests that a satisfying algorithmic-level account of the CFP is intimately linked to another equally puzzling question: How *time* is encoded in the mind? Chapter 3 introduces a notion called potential level (PL) which bears on the aforesaid link.

# Chapter 2

# Multi-Context Models for Reasoning under Partial Knowledge*

*"Partial Knowledge is more triumphant than complete knowledge; it takes things to be simpler than they are, and so makes its theory more popular and convincing."*

— Friedrich Nietzsche, Human, All Too Human

## 2.1 Introduction

At an abstract level, an individual (also referred to as a reasoner) is faced with a domain where by "domain" we simply mean a collection of propositions or concepts which are mathematically encoded as random variables (RVs). Arriving at the complete probabilistic knowledge of the domain, i.e., to learn how all RVs in the domain probabilistically interact with one another, is indeed a demanding task, a task inevitably hindered by an individual's bounded rationality (Simon, 1957). In reality, a reasoner is often faced with a domain of which she merely possesses *partial* knowledge, that is, she only knows how *some* (not all) RVs in the domain interact. To make the setting under study more tangible, consider the following case. Suppose that the probabilistic knowledge of a domain is represented by a

*The material presented in Chapter 2 is partly based on "A. S. Nobandegani & I. N. Psaromiligkos; **Multi-Context Models for Reasoning under Partial Knowledge: Generative Process and Inference Grammar**, In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015."

probabilistic graphical model (PGM) $\mathcal{B}$, e.g., a Bayesian network (BN). Then the reasoner comes across a new RV, say $\psi$, and would like to incorporate it into $\mathcal{B}$ so as to achieve the complete probabilistic knowledge of the new domain (which now also includes $\psi$). However, incorporation of $\psi$ into $\mathcal{B}$ would require knowledge of how $\psi$ is probabilistically related to all the RVs already present in $\mathcal{B}$, a knowledge which may be, quite plausibly, unavailable to the reasoner. An interesting question that immediately arises is how to handle situations where only partial knowledge as to how $\psi$ is probabilistically related to $\mathcal{B}$ is available. An example would be when the reasoner merely knows how $\psi$ interacts probabilistically with only one RV, say $\phi$, in $\mathcal{B}$.

In this chapter, a novel graphical model, namely, the multi-context model (MCM) is put forward to represent the setting in which only partial probabilistic knowledge of a domain is available to the reasoner. More specifically, MCM is a graphical language to represent settings in which the joint probability distribution (JPD) over all RVs is not available, but what is available instead is the JPDs over a collection of subsets of RVs of the domain (referred to as sub-domains or *contexts*). These contexts are potentially overlapping, i.e., they could share some RVs. As pointed out elegantly by Pearl (1990), *"this state of partial knowledge is more common, because we often begin thinking about a problem through isolated frames, paying no attention to interdependencies."* Along the same line of thought, it is plausible to assume that the probabilistic knowledge of the domain at the early primitive stage consists of a collection of disjoint contexts and as the reasoner acquires more knowledge as to how the variables in the model are related to one another and thus probabilistically interact, contexts gradually go through a process very much like an evolution: contexts start to share some variables, overlaps begin to emerge and, once enough knowledge is obtained, a number of contexts could merge thereby giving rise to bigger contexts. This naturally raises the following fundamental question: How could a collection of consistent, probabilistically sound, and potentially overlapping contexts emerge *gradually* over the course of time? In an attempt to answer this question we present a generative process of constructing a contradiction-free MCM. Finally, we would like to note that the special case where the whole domain is modeled as a single context corresponds to the conventional way of modeling the probabilistic knowledge of a domain using a single PGM, e.g., by some BN.

Another yet crucial question which we address in this chapter—which is another motivation behind the development of the MCM—is how the task of inference (i.e., the evaluation of some probability of interest which is hereafter referred to as *query*) should be carried out

in a domain which is modeled according to some MCM. A query does not necessarily belong to any one of the contexts in particular and, in fact, may involve RVs from different contexts.

The chapter is structured as follows. After introducing the notation in Sec. 2.2, we define in Sec. 2.3 the MCM and, drawing on the notion of probabilistic conditioning, a generative process of constructing a contradiction-free MCM is discussed. Then, in Sec. 2.4 we elaborate on the problem of inference in a multi-context setting, i.e., in a domain whose probabilistic knowledge is encoded as an MCM. In Sec. 2.5 we discuss the relevant past work, comment on the proposed model, and discuss the implications of the proposed formalism for psychology. Finally, Sec. 2.7 concludes the chapter.

## 2.2 Terminology and Notation

In this section we present the mathematical notation and the terminology employed in this chapter. Random quantities are denoted by bold-faced letters; their realizations are denoted by the same letter but non-bold. More specifically, RVs are denoted by lower-case bold-faced letters, e.g., $\mathbf{x}$, while random vectors are denoted by upper-case bold letters, e.g., $\mathbf{X}$. $\mathit{Val}(\cdot)$ denotes the set of values a random quantity can take, e.g., $\mathit{Val}(\mathbf{x})$ is the set of all possible realizations of the RV $\mathbf{x}$. In this chapter, we assume that all random quantities are discrete.

The JPD over the RVs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is denoted by $\mathbb{P}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$; when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ comprise a vector $\mathbf{X}$ then $\mathbb{P}(\mathbf{X}) := \mathbb{P}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. We will use the notation $\mathbf{x}_{1:n}$ to denote the sequence of $n$ RVs $\mathbf{x}_1, \ldots, \mathbf{x}_n$. To simplify presentation and to prevent our expressions from becoming cumbersome, we incur the following abuse of notation: We denote the probability $\mathbb{P}(\mathbf{x} = x)$ by $\mathbb{P}(x)$ for some RV $\mathbf{x}$ and its realization $x \in \mathit{Val}(\mathbf{x})$. Also, $\mathbb{P}(\bar{x}) := \mathbb{P}(\mathbf{x} \neq x) = 1 - \mathbb{P}(x)$ for some $x \in \mathit{Val}(\mathbf{x})$, i.e., $\mathbb{P}(\bar{x})$ is the probability that $\mathbf{x}$ takes on any value other than $x$. For conditional probabilities we will use the notation $\mathbb{P}(x|y)$ instead of $\mathbb{P}(\mathbf{x} = x|\mathbf{y} = y)$. Similar notations will be used for the case of random vectors, i.e., $\mathbb{P}(X) := \mathbb{P}(\mathbf{X} = X)$, $\mathbb{P}(\bar{X}) := \mathbb{P}(\mathbf{X} \neq X) = 1 - \mathbb{P}(\mathbf{X} = X) = 1 - \mathbb{P}(X)$, and $\mathbb{P}(X|Y) := \mathbb{P}(\mathbf{X} = X|\mathbf{Y} = Y)$.

The subscript $\downarrow$ on a probability, e.g., $\mathbb{P}(x|y)_{\downarrow}$, denotes the minimum value the probability can take subject to the constraints induced by the available probabilistic knowledge. Likewise, the subscript $\uparrow$ on a probability denotes the maximum value the probability can take. Finally, the operator $[\cdot]^+$ gives the positive part of its argument, i.e., $[a]^+ := \max\{0, a\}$ for any real-valued $a$.

## 2.3 Multi-Context Model

As explained earlier, a *domain* is simply the set of all random variables (RVs) at hand. A *context* comprises a collection of RVs for which their JPD is precisely known, see Fig. 2.1(a). In general, two contexts could be disjoint (Fig. 2.1(b)) or overlapping (Fig. 2.1(c)).

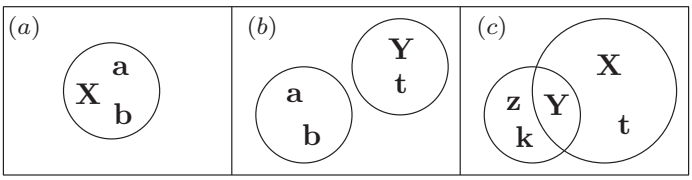



**Fig. 2.1** Graphical representation of contexts: **(a)** Context associated to $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{X})$. **(b)** Two disjoint contexts associated to $\mathbb{P}(\mathbf{a}, \mathbf{b})$ and $\mathbb{P}(\mathbf{Y}, \mathbf{t})$. **(c)** Two overlapping contexts associated to $\mathbb{P}(\mathbf{X}, \mathbf{Y}, \mathbf{t})$ and $\mathbb{P}(\mathbf{Y}, \mathbf{z}, \mathbf{k})$. The random vector $\mathbf{Y}$ is referred to as the *induced* part in Sec. 2.3.

A *multi-context model (MCM)* encodes the probabilistic knowledge of a domain as a collection of possibly overlapping contexts. This enables the handling of situations in which comprehensive knowledge of a domain is not available, but partial information is, in the form of JPDs of some subsets of the domain. Let us first motivate the proposed MCM by entertaining a simple yet enlightening example.

### 2.3.1 Motivating Example

Consider a domain consisting of the RVs $\mathbf{y}, \mathbf{z}$ in addition to a set of $n$ RVs, $\mathbf{x}_{1:n}$. A reasoner has formed a partial belief as to the probabilistic connections between the variables of the domain. More specifically, the reasoner knows precisely the JPDs $\mathbb{P}(\mathbf{y}, \mathbf{z})$ and $\mathbb{P}(\mathbf{x}_{1:n})$ but not the JPD $\mathbb{P}(\mathbf{y}, \mathbf{z}, \mathbf{x}_{1:n})$. This setting is described by an MCM that consists of two disjoint contexts, one associated to RVs $\mathbf{y}, \mathbf{z}$ and the other to $\mathbf{x}_{1:n}$, as shown in Fig. 2.2.

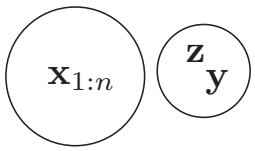



**Fig. 2.2** Problem statement as an MCM.

Assume that the following query is posed: Given the available information, what could be said about $\mathbb{P}(y|x_i)$ for some $i = 1, \cdots, n$? The RVs $\mathbf{y}$ and $\mathbf{x}_i$ belong to different contexts,

therefore, the JPD of $\mathbf{y}$ and $\mathbf{x}_i$, $\mathbb{P}(\mathbf{x}_i, \mathbf{y})$, is not available. The best one can hope for is to derive the range within which $\mathbb{P}(y|x_i)$ varies, namely, $[\mathbb{P}(y|x_i)_{\downarrow}, \mathbb{P}(y|x_i)_{\uparrow}]$. Let us for the moment assume the objective is to find $\mathbb{P}(y|x_i)_{\downarrow}$. Based on the conventional methodology, i.e., the approach adopted by past work (see Andersen and Hooker, 1990, 1994; Hansen et al., 1995, and references therein) one has to write down *all* the information as a list of linear equations and solve it as a linear program (LP). The main drawback of the conventional approach is that it cannot distinguish between what information is relevant and what is irrelevant for the posed query, and hence what needs to and what need not be considered in answering the query. The price for this is that the number of parameters required to merely formulate the query as an LP is exponential in $n$.

The key point, however, is that what information is relevant (or irrelevant) depends directly on the posed query, i.e., it is query-dependent. The main advantage of the proposed MCM over previous approaches is that it enables answering a query in a computationally efficient manner by distinguishing the relevant information from the irrelevant for the given query. This is realized thorough adopting the notion of *inference grammar*, a concept which will be systematically defined later. For our example, following the inference rule we will provide in Sec. 2.4.2, one can easily get $\mathbb{P}(y|x_i)_{\downarrow} = [\frac{\mathbb{P}(y)-\mathbb{P}(\bar{x}_i)}{\mathbb{P}(x_i)}]^+$.

The task of inference in an MCM is carried out on two different levels, which makes the task more computationally efficient:

(i) High-Level Reasoning: At this level, through the use of inference grammar, the relevant quantities are identified (e.g., $\mathbb{P}(y)$ and $\mathbb{P}(\bar{x}_i)$ in the case of our example).

(ii) Low-Level Reasoning: The relevant quantities, identified in (i), can be then computed by employing inference algorithms which take advantage of the potentially rich independence structure governing the contexts. For example, it could very well be the case that for the JPD associated to $\mathbf{x}_{1:n}$ a large number of conditional independence relations hold. In that case, stating the derivation of $\mathbb{P}(\bar{x}_i)$ (i.e., $1-\mathbb{P}(x_i)$) as an LP would be computationally inefficient[1] but unnecessary. Indeed, the task of finding $\mathbb{P}(\bar{x}_i)$ could be accomplished in a computationally efficient way using one of the many inference methods developed for probabilistic graphical models, a key point that the previous approaches do not take advantage of.

[1]The number of parameters required just to state the problem as an LP is exponential in $n$.

As a final step, in order to derive the lower/upper bound to the posed query, the quantities identified in (i) and subsequently calculated in (ii) are stated and solved as an LP.

The idea behind "high-level reasoning" will be explained and clarified further in Sec. 2.4.2 and 2.4.3, while the concept of "low-level reasoning" will be discussed in Sec. 2.4.1.

### 2.3.2 Generative Process of Contradiction-Free MCMs

The objective of the generative process we describe in this section is to provide a way to consistently[2] construct contexts, in a sequential manner, over a set of RVs. The act of constructing a context, i.e., of assigning a JPD to a subset of RVs, corresponds to forming a *subjective*[3] belief over those RVs. In this light, the act of constructing multiple contexts corresponds to *gradually* forming subjective beliefs over a number of subsets of variables in the domain; hence every context symbolizes an established belief over the RVs involved in that context.

We introduce this problem by considering a simple case shown in Fig. 2.3(a). Suppose

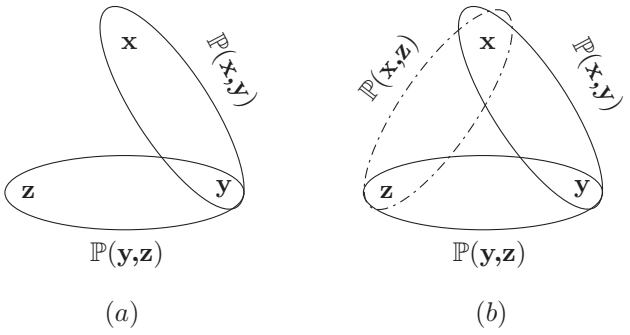



**Fig. 2.3** Generative process for contradiction-free Multi-Context Model. The dash-dotted contexts cannot be freely assigned.

there are three RVs, namely, $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$, present in the domain and let us consider the following question: Could one assign $\mathbb{P}(\mathbf{x}, \mathbf{y})$ and $\mathbb{P}(\mathbf{y}, \mathbf{z})$, *freely* and *gradually* in a consistent manner, over the three variables without introducing any sort of contradiction? It is easy to verify that the answer is positive. Indeed, one could start off by assigning $\mathbb{P}(\mathbf{x}, \mathbf{y})$. This assignment would, of course, induce the marginal $\mathbb{P}(\mathbf{y})$ and one can write $\mathbb{P}(\mathbf{y}, \mathbf{z}) = \mathbb{P}(\mathbf{y})\mathbb{P}(\mathbf{z}|\mathbf{y})$.

[2]That is, without introducing any form of contradictory result with respect to any probability assignment.

[3]One must not interpret the subjectivity of belief as "total disconnectivity from the reality." Thus, we adopt the Bayesian interpretation of probability in this section. The avid reader is referred to Chalmers (2013). An adherent to the frequentist interpretation of probability could think of contexts as being empirically constructed from a collection of data and thus skip Sec. 2.3.2 and proceed directly to the next section.

Then, to complete this task, one would just need to proceed with assigning $\mathbb{P}(\mathbf{z}|\mathbf{y})$. This process could be referred to as a *generative* process of the assignment of $\mathbb{P}(\mathbf{x}, \mathbf{y})$ and $\mathbb{P}(\mathbf{y}, \mathbf{z})$ over $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ without introducing any inconsistencies, in a gradual manner. Indeed, free-assignment refers to the act of freely assigning the non-induced, e.g., $P(\mathbf{z}|\mathbf{y})$, part of the *to-be-formed* belief, e.g., $P(\mathbf{y}, \mathbf{z})$. In other words, free-assignment signifies the observation that the already-formed belief does not impose any constraints on the non-induced part of the to-be-formed belief.

Let us now consider the case shown in Fig. 2.3(b). Could one assign $\mathbb{P}(\mathbf{x}, \mathbf{y}), \mathbb{P}(\mathbf{y}, \mathbf{z})$, and $\mathbb{P}(\mathbf{x}, \mathbf{z})$ freely and gradually in a consistent manner over the three variables without introducing any sort of contradiction? After some investigation, one can see that the answer is negative (Pearl, 1985). Not surprisingly, the reason for this has to do with the existence of a loop in the model: Once $\mathbb{P}(\mathbf{x}, \mathbf{y})$ and $\mathbb{P}(\mathbf{y}, \mathbf{z}) = \mathbb{P}(\mathbf{y})\mathbb{P}(\mathbf{z}|\mathbf{y})$ are assigned,[4] then $\mathbb{P}(\mathbf{x}, \mathbf{z})$ cannot be assigned freely. This is due to the fact that $\mathbb{P}(\mathbf{x}, \mathbf{z})$ has to satisfy some non-trivial conditions imposed by the already assigned contexts $\mathbb{P}(\mathbf{x}, \mathbf{y})$ and $\mathbb{P}(\mathbf{y}, \mathbf{z})$ (Pearl, 1985).

In summary, whenever it comes to generating a new context, the JPD associated to that context has to be separated into two parts: (i) the part induced by the already existing contexts, and (ii) the part containing new variables which have never been so far associated to any context (i.e., non-induced part). The key point in the generation of contradiction-free MCMs is that the former part has to be induced by some context which, itself, is already present in the domain. That is, all the induced parts have to be already contained within some context. Otherwise, to include the induced parts—each constrained by the context it is already in—in a new context, the newly created context would have to satisfy some nontrivial constraints and therefore could not be *freely* assigned.

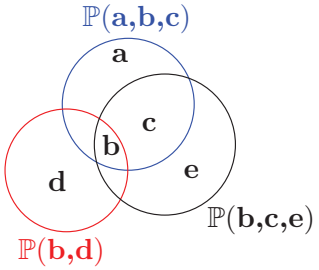



**Fig. 2.4** MCM for $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{c}), \mathbb{P}(\mathbf{b}, \mathbf{d})$, and $\mathbb{P}(\mathbf{b}, \mathbf{c}, \mathbf{e})$.

Let us discuss one final case to further clarify the process. Consider the multi-context

[4] $\mathbb{P}(\mathbf{y})$ is induced by the assignment of $\mathbb{P}(\mathbf{x}, \mathbf{y})$.

model in Fig. 2.4. Could this model be constructed freely and gradually in a probabilistically consistent manner? The answer is positive. We first assign $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{c})$, then we assign $\mathbb{P}(\mathbf{b}, \mathbf{c}, \mathbf{e}) = \mathbb{P}(\mathbf{b}, \mathbf{c})\mathbb{P}(\mathbf{e}|\mathbf{b}, \mathbf{c})$ where $\mathbb{P}(\mathbf{b}, \mathbf{c})$ is induced by our first assignment of $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{c})$. Finally, we assign $\mathbb{P}(\mathbf{b}, \mathbf{d}) = \mathbb{P}(\mathbf{b})\mathbb{P}(\mathbf{d}|\mathbf{b})$ where $\mathbb{P}(\mathbf{b})$ is induced by our first assignment of $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{c})$. A closer look reveals that this is not the only way we can gradually construct a contradiction-free model in this case; we could have performed the assignments in a different order.[5] Of course, the only thing which would have been different would be the induced probabilities. That is, if one does the assignment in the following order: (1) $\mathbb{P}(\mathbf{b}, \mathbf{d})$, (2) $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{c})$, (3) $\mathbb{P}(\mathbf{b}, \mathbf{c}, \mathbf{e})$ then the first assignment of $\mathbb{P}(\mathbf{b}, \mathbf{d})$ will induce $\mathbb{P}(\mathbf{b})$ for the second assignment of $\mathbb{P}(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbb{P}(\mathbf{b})\mathbb{P}(\mathbf{a}, \mathbf{c}|\mathbf{b})$ and the second assignment will induce $\mathbb{P}(\mathbf{b}, \mathbf{c})$ for the third assignment $\mathbb{P}(\mathbf{b}, \mathbf{c}, \mathbf{e}) = \mathbb{P}(\mathbf{b}, \mathbf{c})\mathbb{P}(\mathbf{e}|\mathbf{b}, \mathbf{c})$.

## 2.4 Inference in MCMs

In this section we consider *evidential* inference problems in multi-context settings. The objective is to evaluate (to the extent possible) a probability of the form $\mathbb{P}(\mathbf{O} = O|\mathbf{E} = E)$, called a *query*, where $\mathbf{O}$ and $\mathbf{E}$ are two mutually exclusive sets of RVs. The set $\mathbf{E}$ is the set of evidence variables and $\mathbf{O}$ is the set of RVs for which we are interested in knowing with what probability they take on the value $O$, upon the observation of $\mathbf{E} = E$. In multi-context settings, inference problems can be categorized into two broad classes:

- Intra-Contextual Inference Problems: For which the sets $\mathbf{E}$ and $\mathbf{O}$ both belong to the same context.
- Inter-Contextual Inference Problems: For which the sets $\mathbf{E}$ and $\mathbf{O}$ do not belong to a single context and, therefore, more than one context is involved in the inference problem.

In what follows, we will elaborate on these two cases.

[5] Yet, this is not always the case: suppose there are four RVs in the domain, namely, $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ and we would like to assign $\mathbb{P}(\mathbf{a}, \mathbf{b}), \mathbb{P}(\mathbf{b}, \mathbf{c})$, and $\mathbb{P}(\mathbf{c}, \mathbf{d})$. Performing the assignments in the order (1) $\mathbb{P}(\mathbf{a}, \mathbf{b})$, (2) $\mathbb{P}(\mathbf{b}, \mathbf{c})$, (3) $\mathbb{P}(\mathbf{c}, \mathbf{d})$ would not introduce any inconsistencies, in contrast to using the order (1) $\mathbb{P}(\mathbf{a}, \mathbf{b})$, (2) $\mathbb{P}(\mathbf{c}, \mathbf{d})$, (3) $\mathbb{P}(\mathbf{b}, \mathbf{c})$.

### 2.4.1 Intra-Contextual Inference Problem

One advantage of MCMs is that, once an inference problem is found to be an intra-contextual inference problem, one can take advantage of the rich independence structure potentially governing the context to accomplish the task of inference in a computationally efficient way. For instance, if the probabilistic knowledge of a context is presented in a form of a BN, then one can benefit from a variety of exact or approximate methods already developed for BNs. For a comprehensive study of such methods, the reader is referred to (Koller and Friedman, 2009). Hence, it is of great interest to have contexts whose probabilistic knowledge can be represented in some form of a PGM with sufficiently rich independence structure for which inference problems can be solved in a computationally efficient way. For example, if the probabilistic knowledge of a context is to be modeled according to some BN, we would like that BN to be as sparsely connected as possible and enjoy low tree-width to ensure computational efficiency for the task of inference (Chandrasekaran et al., 2012).

### 2.4.2 Inter-Contextual Inference Problem: Inference Grammar

In this section, we turn our attention to the task of inter-contextual inference. The RVs involved in the query for the inter-contextual inference problem do not belong to a single context. For this reason, the answer to the query is inevitably in the form of an interval indicating a lower and upper bound for the query. Since $\mathbb{P}(E|O) + \mathbb{P}(\bar{E}|O) = 1$ we have $\mathbb{P}(E|O)_{\uparrow} = 1 - \mathbb{P}(\bar{E}|O)_{\downarrow}$. Therefore, we can focus our attention on the minimization problem (i.e., identifying a lower bound to the probability of interest) realizing that any maximization problem (i.e., identifying an upper bound to the probability of interest) could be cast as a minimization problem and vice versa.

First, we are going to consider some simple queries which are posed to some example MCMs. These MCMs are depicted in Fig. 2.5(a-c). The goal here is to develop some insight as to which variables are indeed relevant and which are deemed irrelevant for a given query and the corresponding MCM.

We begin by considering a simple case: the disjoint MCM shown in Fig. 2.5(a). The rule to evaluate $\mathbb{P}(X|Y)_{\downarrow}$ is also given in Fig. 2.5(a). Interestingly enough, the expression only requires the intra-contextual quantities $\mathbb{P}(X)$ and $\mathbb{P}(Y)$ and it does not depend on any other RV present in the domain. In other words, as far as $\mathbb{P}(X|Y)_{\downarrow}$ is concerned, the MCM shown in Fig. 2.5(a) is equivalent to a much simpler MCM: the one corresponding to

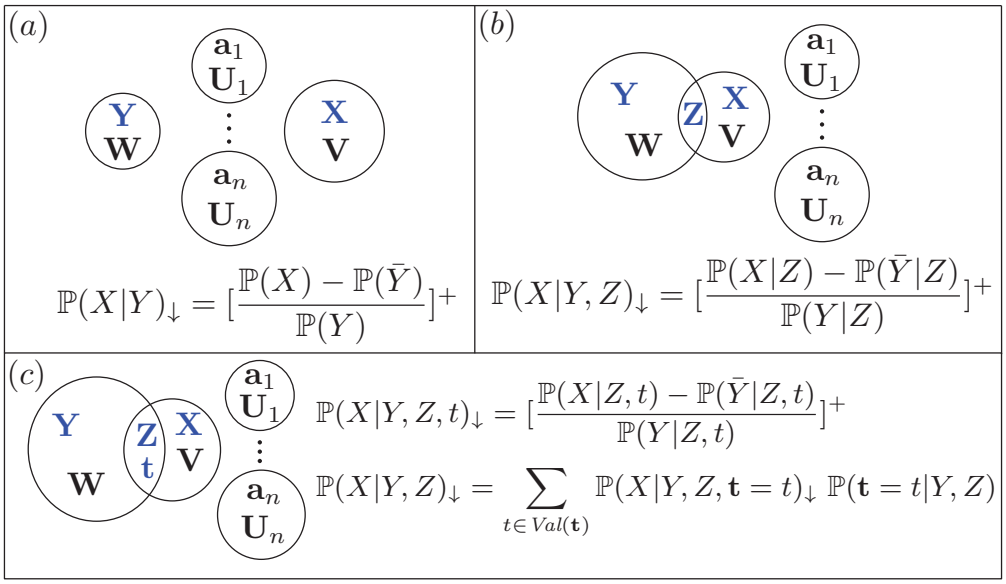



**Fig. 2.5** Sample inference rules given for some inter-contextual inference problems. The RVs involved in the query are shown in blue.

having only two disjoint contexts described by $\mathbb{P}(\mathbf{X})$ and $\mathbb{P}(\mathbf{Y})$. Next, we take the MCM given in Fig. 2.5(b) where there is an overlap between the context containing $\mathbf{X}$ and the one containing $\mathbf{Y}$. The overlapping part consists of the random vector $\mathbf{Z}$. The rule to evaluate $\mathbb{P}(X|Y,Z)_{\downarrow}$ is given in Fig. 2.5(b). Now, consider the MCM shown in Fig. 2.5(c) where we have the same setting we had in previous case but a new random variable $\mathbf{t}$ is added in the overlapping region. Notice that the expression for $\mathbb{P}(X|Y,Z,t)_{\downarrow}$ given in Fig. 2.5(c) is the same expression given for $\mathbb{P}(X|Y,Z)_{\downarrow}$ in Fig. 2.5(b) with the substitution of $Z,t$ instead of $Z$. That is, $Z$ in Fig. 2.5(b) and $Z,t$ in Fig. 2.5(c) are representing the same thing, namely, "all the variables in the overlapping region," and in that respect, they are ultimately the same. The rules are very much like sentences in predicate logic for which variables merely serve as place-holders.

The derivation of the rules given in Fig. 2.5(a-c) is not presented here. However, using the proof presented in Sec. A-II of Appendix A (to identify the relevant variables) and subsequently following the methodology outlined in Sec. A-III of Appendix A (to visualize the partitions and reason out the extent they overlap) it should be straightforward to derive the presented rules.

The sample set of rules presented is by no means exhaustive, nonetheless, due to the idea of context transformation that will be discussed in Sec. 2.4.3, they can be applied to a wide

range of interesting inter-contextual inference problems. We would like to clarify that our ultimate objective is *not* to compute and provide the complete set of rules that can answer all possible queries and for all possible MCMs, since simply, the set is infinite in size. What we need, therefore, is an algorithm, let us call it $\mathcal{I}^*$, that can provide the answer to the posed query being given an MCM as an input. The presented rules provide insights, and hints to the nature of $\mathcal{I}^*$ which needs to be devised to ideally handle *any* arbitrary query posed to *any*[6] MCM. In a sense, we can get a glimpse of the nature of $\mathcal{I}^*$ through analyzing the presented rules. In other words, the derived rules serve as a lens through which one can study $\mathcal{I}^*$. The reader is referred to Appendix A wherein the algorithm $\mathcal{I}^*$ (for handling arbitrary MCMs) is outlined.

The motivation behind giving this sample set of rules can now be summarized in the following.

1. To shed light on the general nature of a rule (which reflects on the nature of $\mathcal{I}^*$). More specifically, to illustrate that a rule enjoys two key properties, namely: (i) scale-invariance, (ii) resemblance to sentences in predicate logic, in that in both cases, variables are mere place-holders. For this resemblance, we refer to $\mathcal{I}^*$ as *inference grammar*.

2. To demonstrate that a rule is telling us which intra-contextual quantities are essential and which are irrelevant for a particular inter-contextual query.

3. To emphasize the key property that a rule derived under a specific MCM remains valid for and can be applied to infinitely many other MCMs all of which are linked through the notions of nestedness and transformation; hence generalization is achieved.

4. To lay down the foundation of *transformation* and *nestedness* which both play crucial roles in understanding the underlying machinery behind $\mathcal{I}^*$.

Next, we discuss another key property of the inference rules, namely, that of *scale-invariance*. Consider once again the case in Fig. 2.2. Now let us derive $\mathbb{P}(x_i|y)_\downarrow$, and $\mathbb{P}(X|y)_\downarrow$ where $\mathbf{X} \triangleq \mathbf{x}_{1:n}$. Using the rule given in Fig. 2.5(a), one arrives at the following results: $\mathbb{P}(x_i|y)_\downarrow = [\frac{\mathbb{P}(x_i)-\mathbb{P}(\bar{y})}{\mathbb{P}(y)}]^+$, and $\mathbb{P}(X|y)_\downarrow = [\frac{\mathbb{P}(X)-\mathbb{P}(\bar{y})}{\mathbb{P}(y)}]^+$. In other words, the expressions remain the same, regardless of the dimension of the quantity of interest, i.e., be it a single RV or be it a

[6] Although we believe that the MCMs generated through the generative process outlined in Sec. 2.3.2 are more cognitively plausible, nonetheless, from a pure mathematical point of view, it would be of interest to find an algorithm which could handle *any* MCM.

random vector comprised of many RVs. In this respect, once again, the inference rules resemble expressions in predicate logic. The intuition on the scale invariance is provided in Sec. A-III of Appendix A.

It is worth noting that $\mathcal{I}^*$ formulates the inter-contextual inference problem as a linear programming (LP) optimization (see Sec. A-I of Appendix A). The key issues to consider are: (i) what RVs have to be included in the LP, and (ii) the abstraction level $\mathcal{I}^*$ should choose to encode the RVs identified in step (i) for the LP, i.e., the parameterization of RVs identified in step (i) for the LP. In what follows, the concepts of nestedness and transformation are put forth. Once the two are introduced, one could apply a single rule (e.g., one in Fig. 2.5(a)) to a much larger number of MCMs; in fact to infinitely many MCMs.

### 2.4.3 Inter-Contextual Inference Problem: Nestedness and Transformation

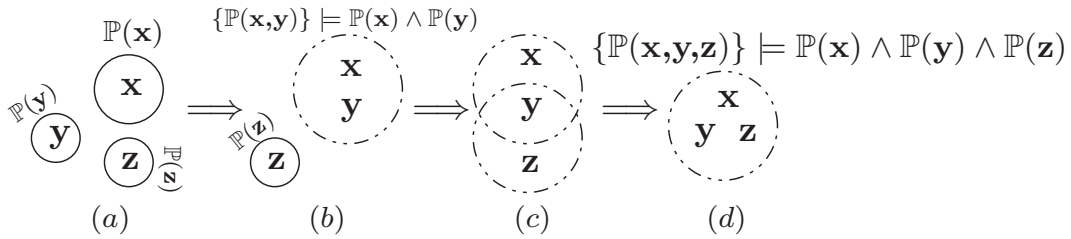



**Fig. 2.6** Inter-Contextual Inference Problem: Transformation and hierarchical construct. As one proceeds from the left to the right, a more comprehensive knowledge of domain is assumed to be available, of course hypothetically.

The nested property, or *nestedness*, refers to the fact that every MCM can be considered as an element of a family of MCMs. That family contains all MCMs which through marginalization can produce the original MCM. In such a case we simply say that the nested property holds between the original MCM and the family. The process of going from the original MCM to one of the members of the family is referred to as *transformation*. For example, the MCM containing three contexts $\{\mathbf{x}\}$, $\{\mathbf{y}\}$, and $\{\mathbf{z}\}$ shown in Fig. 2.6(a) is a member of a family of MCMs containing two contexts $\{\mathbf{x}, \mathbf{y}\}$ and $\{\mathbf{z}\}$, shown in Fig. 2.6(b), one of which is associated to a *family* of JPDs over $\mathbf{x}$ and $\mathbf{y}$ (the dash-dotted circle in Fig. 2.6(b)) which, if marginalized, produces the same $\mathbb{P}(\mathbf{x})$ and $\mathbb{P}(\mathbf{y})$ in the original MCM (left-most MCM). Mathematically, the set of all JPDs over RVs $\mathbf{x}$ and $\mathbf{y}$ which, if marginalized, produce specific marginal probability distributions $\mathbb{P}(\mathbf{x})$ and $\mathbb{P}(\mathbf{y})$ is denoted by $\{\mathbb{P}(\mathbf{x}, \mathbf{y})\} \models \mathbb{P}(\mathbf{x}) \wedge \mathbb{P}(\mathbf{y})$. The notion of the nested property enables us to look at one MCM as a subset of another

larger MCM. The nested property, furthermore, enables one to sort MCMs in a hierarchical construct as illustrated in Fig. 2.6 where moving from the left to the right corresponds to moving from lower levels of hierarchy to higher levels.

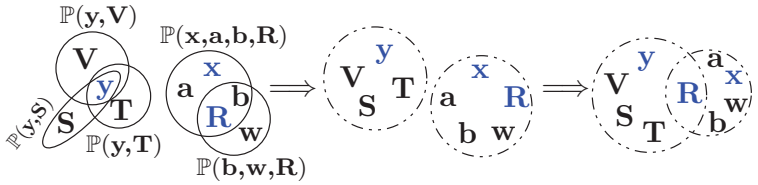



**Fig. 2.7** Transformation: Sample case.

To convey the idea, consider the case illustrated in Fig. 2.7. Suppose the query of interest is $\mathbb{P}(x|y,R)_{\downarrow}$. Then, one can first transform the original (left-most) MCM into the MCM shown in the middle, and subsequently into the right-most MCM. Hence, using the right-most MCM and the rule given in Fig. 2.5(b), one can write $\mathbb{P}(x|y,R)_{\downarrow} = [\frac{\mathbb{P}(x|R)-\mathbb{P}(\bar{y}|R)}{\mathbb{P}(y|R)}]^{+} = [\frac{\mathbb{P}(x|R)-1+\mathbb{P}(y|R)}{\mathbb{P}(y|R)}]^{+}$. If we had the knowledge of $\mathbb{P}(y|R)$ then the expression given above would have been sufficient to derive $\mathbb{P}(x|y,R)_{\downarrow}$. However, since $\mathbb{P}(y|R)$ is *not* known, we need to go through one more step. This is precisely due to, and emphasizes, the fact that by working on the right-most MCM we implicitly presumed that we were equipped with more knowledge than we really had. Using the middle MCM and the rule given in Fig. 2.5(a), one can conclude $\mathbb{P}(y|R)_{\downarrow} = [\frac{\mathbb{P}(y)-\mathbb{P}(\bar{R})}{\mathbb{P}(R)}]^{+}$. Altogether,[7] $\mathbb{P}(x|y,R)_{\downarrow} = \big([\frac{\mathbb{P}(x|R)-1+\mathbb{P}(y|R)}{\mathbb{P}(y|R)}]^{+}\big)_{\downarrow} = [\frac{\mathbb{P}(x|R)-1+\mathbb{P}(y|R)_{\downarrow}}{\mathbb{P}(y|R)_{\downarrow}}]^{+}$. It is worth noting that the same rule would apply if instead of the random vector $\mathbf{R}$ we were dealing with the random variable $\mathbf{a}$, i.e., to find $\mathbb{P}(x|y,a)_{\downarrow}$ one could use the same expression given for $\mathbb{P}(x|y,R)_{\downarrow}$ by substituting $a$ in place of $R$ in all the expressions. Arguments of this kind are made possible due to the idea of transformation which enables us to analyze the transformed MCM (e.g., the middle one in Fig. 2.7) rather than the original MCM (the left-most one in Fig. 2.7). Furthermore, the concept of transformation highlights a key idea: If a piece of information (i.e., an intra-contextual quantity) is irrelevant in the transformed MCM for the posed query, it must have been irrelevant in the original MCM in the first place. This statement, once again, sheds light on what intra-contextual quantities are relevant or irrelevant to derive a posed inter-contextual query on a given MCM.

[7]This is due to the observation that for function $f(y) = (\frac{k+y}{y})$ when $k < 0$, $\min_{1\geq y\geq t>0} f(y) = (\frac{k+t}{t})$.

## 2.5 Discussion

We will now discuss related work so as to build a connection between ours and previous attempts to incorporate partial probabilistic knowledge of a domain in the task of inference.

Attempting to combine Probabilistic Logic and BNs, Andersen and Hooker (1990, 1994) formulate the inference problem as an optimization problem subject to non-linear constraints so as to incorporate the conditional independence relations embedded in the BN. However, in our proposed framework, the issue of dealing with conditional independence relations does not arise at all, because these relations are dealt with during the derivation process of intra-contextual probabilities.

Hansen et al. (1995) point out that one could avoid non-linear optimization when the value for a conditional probability is at least imprecisely known. For example, the constraint $\mathbb{P}(a|b) = \mathbb{P}(a)$, if the value for $\mathbb{P}(a)$ is known either precisely or imprecisely within some interval $[\alpha, \beta]$, can be written as

$$\frac{\mathbb{P}(a,b)}{\mathbb{P}(b)} = \mathbb{P}(a) \in [\alpha, \beta] \Leftrightarrow \begin{cases} \mathbb{P}(a,b) - \alpha\mathbb{P}(b) > 0, \\ \mathbb{P}(a,b) - \beta\mathbb{P}(b) < 0. \end{cases}$$

Hence, the independence $\mathbb{P}(a|b) = \mathbb{P}(a)$ can be formulated as a number of linear constraints. However, the main drawback of this approach is that encoding a conditional independence relation such as $\mathbb{P}(\mathbf{x}|\mathbf{y}, \mathbf{a}_1, \cdots, \mathbf{a}_n) = \mathbb{P}(\mathbf{x}|\mathbf{y})$ requires a number of linear equations that is exponential in $n$ to be introduced into the optimization problem (Andersen and Hooker, 1994).

Drawing on the idea of context-specific independence (CSI) (Boutilier et al., 1996), Geiger and Heckerman (1991) propose the Bayesian multinet model which aims at taking advantage of the existing CSIs to perform inference, by modeling a single BN as multiple context-specific BNs. Translated into our multi-context setting, the Bayesian multinet model corresponds to the case where the whole domain is modeled as a single BN, i.e., a single-context MCM, that can be decomposed into multiple BNs each being valid for a specific instantiation of some RVs in the domain.

Thöne et al. (1992) point out the same concerns which led us to propose MCM, namely: (i) If unverified (in)dependencies are imposed between the variables in the domain then implausible results may arise; (ii) PGMs require one to have complete probabilistic knowledge of a domain which may not be available. Motivated by these, Thöne et al. (1992) give a

collection of rules to carry out inference in a domain. Very broadly speaking, this work is similar to ours in spirit with the main distinction being the level of abstraction chosen to perform inference. In Thöne et al. (1992), inference is performed in a very local and rule-based fashion and conditional independence relations are dealt with directly which complicates the task at hand, a task which is futile when it comes to dealing with domains of many variables. In our case, by introducing the notion of context and encoding conditional independence relations within contexts, we avoid having to contemplate the intra-contextual inference problem and leave this task for the corresponding context. This way, we can take advantage of the possibly rich independence structure governing the context and carry out the intra-contextual inference problem in a computationally efficient manner.

Finally, let us discuss some interesting aspects of the proposed model.

The degree of belief is encoded mathematically in the form of a probability distribution over the variables contained within the context. Furthermore, in the process of partial belief formation (which leads to the formation of contexts) the reasoner is ignorant as to how various contexts probabilistically interact (are related), except that, some contexts may in fact share a number of variables in between and hence overlap. Later on, in the process of the derivation of the query posed to the reasoner, this ignorance manifests in the uncertainty region represented by the min/max values for the inter-contextual query of interest. In other words, if the reasoner incurs ignorance as to the (in)dependency structure governing the variables present in the domain, then later on, in the process of derivation of the posed query, the reasoner has to pay the price by merely arriving at a *probability interval* rather than a point probability as an answer to the query of interest. Yet, the knowledge of the underlying dependency structure is a fundamental knowledge whose availability to the reasoner should *not* be postulated as an inevitability, but as an advantaged position.

The evolutionary process of MCM does not enforce a specific gradual expansion path, for the claim of MCM is merely that *any partial belief formation as to the domain can be modeled in the framework depicted by MCM*. That is, the reasoner may arrive at different MCMs, depending on the order in which the reasoner encounters different concepts and also depending on her background knowledge as to the nature of the potential connections between a collection of variables. Simply put, the order according to which the reasoner comes about knowing the concepts or propositions of the domain does matter (see the discussion on the order of belief formation in Sec. 2.3.2).

MCM enables one to carry out inference without having to commit to any unjustified

independence assumptions. In light of this, contexts symbolize the regions of the domain over which an (in)dependence structure is presumed and hence, the growth and merging of contexts indicates the formation of new (in)dependence structures over some parts of the domain which previously were unstructured. In short, MCM is meant to be invoked in circumstances where the observations and the a priori knowledge combined are not sufficient for the reasoner to form the full JPD over all of the domain variables and yet, quite crucially, the reasoner is reluctant to submit to any unjustified assumptions to compensate for such inadequacy of knowledge.

## 2.6 On the Implications of Multi-Context Model for Cognitive and Developmental Psychology

Attention is a well-explored subject in human psychology and neurophysiology (e.g., Broadbent, 1965; Kastner and Ungerleider, 2000; Pashler et al., 2001; Treue, 2001). Causes and effects of limited attention in humans as well as non-human animals are also investigated in the literature (e.g., Dukas and Kamil, 2001; Desimone and Duncan, 1995; Clark and Dukas, 2003; Pashler et al., 2001; Treue, 2001; Moran and Desimone, 1985), highlighting the significance of *what* a reasoner directs her attention toward, when faced with a task. Drawing on a reasoner's limited attention span and scope, which is yet another manifestations of Simon's (1957) bounded rationality, we put forth a novel graphical model, MCM, to formally represent a reasoner's state of partial knowledge of a domain. Furthermore, given the prominent role of BN in Bayesian models of cognition (Gopnik et al., 2004, *inter alia*), our proposed model generalizes BN to the realm of partial knowledge, where the term 'partial' signifies a reasoner's complete lack of knowledge as to parts of the underlying dependency structure of the domain. In particular, this generalization is achieved by drawing on the quintessence of BNs, i.e., the concept of *conditioning* (see Sec. 2.3).

Importantly, MCM could also significantly contribute to an influential line of work in developmental psychology concerning the computational modeling of one of the Younger and Cohen's key discoveries in infant information processing: the developmental shift from learning about visual stimulus features to learning about correlations between these features, in infants between four and ten months of age (see Oakes et al., 2011, for a review of psychological evidence). Concretely, MCM allows to: (1) computationally model a state of knowledge consisting of isolated features, without capturing the correlations between

them (4-month olds, see Fig. 2.8(a)), and (2) computationally model a state of knowledge capturing features as well as their correlations (10-month olds, see Fig. 2.8(c)). Furthermore, MCM allows to computationally model a set of intermediate stages in infant's knowledge representation (see Fig. 2.8(b)). An intriguing line of future work could be to investigate whether any of the intermediate stages shown in Fig. 2.8(b) can be experimentally confirmed, or that the transition from the representation invoked by 4-month olds to that of 10-month olds tends to be rather abrupt, leaving no room for any intermediate stages.

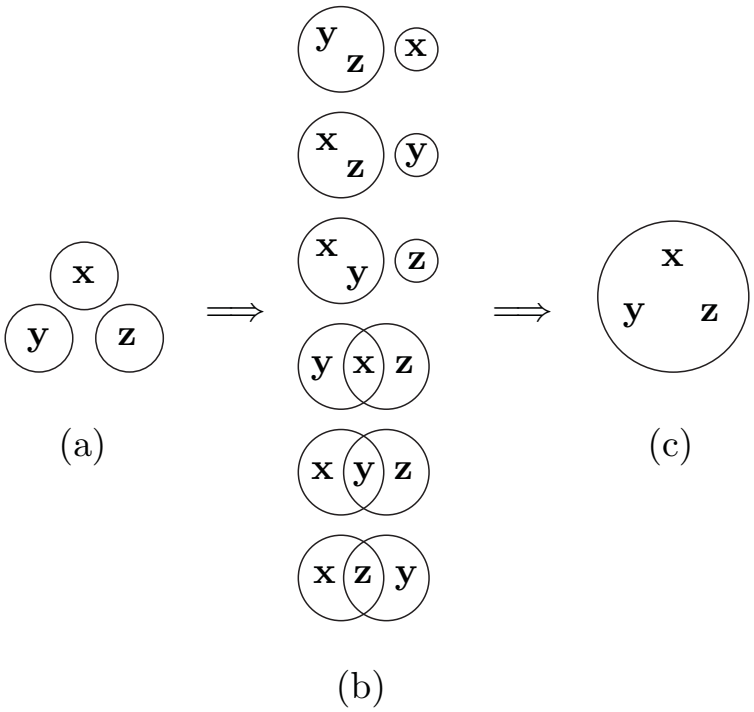



**Fig. 2.8** Computational modeling of the transition in infant's knowledge representation (Younger and Cohen, 1983, 1986), starting from a state of knowledge consisting of isolated features **(a)**, en route to attaining the state of complete knowledge capturing the correlations among those features **(c)**. MCM, furthermore, allows to formally capture possible intermediate stages in infant's knowledge representation **(b)**. Future work should investigate whether any of the intermediate stages shown in **(b)** can be experimentally confirmed, or that the transition from **(a)** to **(c)** tends to be rather developmentally abrupt, leaving no room for any intermediate stages.

MCM serves as a normative, representational-level model, with the normativity claim following from two statements: (1) MCM adheres to the maxim that the state of knowledge ought to progress from partial to complete. According to this maxim, a reasoner should

strive for minimizing his ignorance, aiming for achieving the state of knowledge with least uncertainty, or, more formally, with least *entropy*.[8] (2) MCM respects the maxim that, for reasoning, a reasoner should only consider what he knows, and plausibly, should make no *informed* assumptions about propositions of which he utterly knows nothing and has no prior knowledge of.

Before concluding this chapter, we would like to sketch the main predictions that follow from MCM modeling. It is worth reiterating that, in a domain modeled by MCM, intra-contextual queries result in probability intervals rather than precise probability values. Assuming that $[\alpha, \beta]$ denotes the probability interval implied by MCM for a posed evidential query, two main predictions follow. (1) Subjects' estimates for the posed query should tend to be close to the midpoint of the probability interval, $\frac{\alpha+\beta}{2}$, under the plausible assumption of the subjects' lack of preference for any particular value within that probability interval (computationally, the said lack of preference amounts to assigning an uninformative, uniform distribution over that probability interval). (2) For two queries with the same midpoint value, subjects' confidence ratings for the query with wider probability interval should be lower, under the plausible assumption that a wider probability interval should be construed by the subjects as indications for the existence of more fuzziness as to the query value, hence implying lower confidence ratings. Future work should investigate if these predictions are borne out by behavior data, empirically characterizing the extent to which human probabilistic judgment is rational under the state of partial knowledge captured by MCM (as a normative representational-level model).

## 2.7 Conclusion

In an attempt to establish a middle ground between Bayesian Logic and Probabilistic Logic (Andersen and Hooker, 1990, 1994), on one side, and PGMs[9] on the other, we proposed the Multi-Context Model to represent the state of partial knowledge regarding a domain. The generative process for a gradual construction of contradiction-free MCMs was discussed. The task of Inference for MCM was studied and, along the path, the notions of inference grammar, nestedness, and transformation were introduced. Finally, we elaborated on the implications of our proposed model for psychology, and discussed how it could significantly

[8] To a familiar reader, this should sound analogous to Karl Friston's *minimum free energy* principle.

[9] For instance, Bayesian Networks (Pearl, 1986), Markov Networks (Koller and Friedman, 2009), and Chain Graphs (Buntine, 1995).

contribute to the area of developmental psychology concerning infant information processing. To our knowledge, MCM is the first *normative*, *parsimonious*, representational-level model for capturing the state of partial knowledge, where the term partial signifies a reasoner's *complete* lack of knowledge as to parts of the underlying dependency structure of the domain. Last but not least, the algorithm $\mathcal{I}^*$ can be viewed as a rational process model (Griffiths et al., 2012) for inference under partial knowledge, in a domain modeled by an MCM.

# Chapter 3

# The Causal Frame Problem: An Algorithmic Perspective*

*"The frame problem goes very deep; it goes as deep as the analysis of rationality."*

Jerry Fodor (1987)

## 3.1 Introduction

At the core of any decision-making or reasoning task, resides an innocent-looking yet challenging question: Given an inconceivably large body of knowledge available to the reasoner, what constitutes the relevant for the task and what the irrelevant? The question, as it is posed, echoes the well-known frame problem (FP) in epistemology and philosophy of mind, articulated by Glymour (1987) as follows: "Given an enormous amount of stuff, and some task to be done using some of the stuff, what is the relevant stuff for the task?"

The question posed above perfectly captures what is really at the core of the FP, yet, it may suggest an unsatisfying approach to the FP at the algorithmic level of analysis (Marr, 1982). Indeed, the question may suggest the following two-step methodology: In the first step, out of all the body of knowledge available to the reasoner (termed, the model), she

*The material presented in Chapter 3 is partly based on "A. S. Nobandegani & I. N. Psaromiligkos; **The Causal Frame Problem: An Algorithmic Perspective**, In *Proceedings of the* $39^{th}$ *Annual Conference of the Cognitive Science Society (CogSci)*, 2017."

has to identify what is relevant to the task (termed, the relevant submodel); it is *only then* that she advances to the second step by performing *reasoning* or *inference* on the identified submodel. There is something fundamentally wrong with this methodology (which we term, *sequential* approach to reasoning) which bears on the following understanding: The relevant submodel, i.e., the portion of the reasoner's knowledge deemed relevant to the task, oftentimes is so enormous (or even infinitely large) that the reasoner—inevitably bounded in time and computational resources—would never get to the second step, had she adhered to such a methodology. In other words, in line with the notion of bounded rationality (Simon, 1957), a boundedly-rational reasoner must have the option, if need be, to merely consult a fraction of the potentially large—if not infinitely so—relevant submodel.

Icard and Goodman (2015) elegantly promote this insight when they write: "Somehow the mind must focus in on some 'submodel' of the 'full' model (including all possibly relevant variables) that suffices for the task at hand and is not too costly to use."[1] They then ask the following question: "what kind of simpler model should a reasoner consult for a given task?" This is an inspiring question, hinting to an interesting line of inquiry as to how to formally articulate a boundedly-rational approach to the FP, at Marr's (1982) algorithmic level of analysis.

In this chapter, we focus on the causal variant of the FP, the causal frame problem (CFP), stated as follows: Upon being presented with a causal query, how does the reasoner manage to attend to her causal knowledge relevant to the derivation of the query while rightfully dismissing the irrelevant? We adopt causal Bayesian networks (CBNs) (Pearl, 1988; Gopnik et al., 2004, *inter alia*) as a normative model to represent how the reasoner's internal causal model of the world is structured (i.e., reasoner's mental model). First, we introduce the notion of potential level (PL). PL, in essence, encodes the relative position of a node (representing a propositional variable or a concept) with respect to its neighbors in a CBN. Drawing on the psychological literature on causal judgment, we substantiate the claim that PL may bear on how *time* is encoded in the mind. Equipped with PL, we embark on investigating the CFP at Marr's algorithmic level of analysis. We propose an inference framework, termed PL-based inference framework (PLIF), which aims at empowering the boundedly-rational reasoner to consult (or retrieve[2]) parts of the underlying CBN deemed

[1] In an informative example on Hidden Markov Models (HMMs), Icard and Goodman (2015) present a setting wherein the relevant submodel is infinitely large—an example which highlights what is wrong with the sequential approach stated earlier.

[2] The terms "consult" and "retrieve" will be used interchangeably. We elaborate on the rationale behind

relevant for the derivation of the posed query (the relevant submodel) in a *local*, *bottom-up* fashion until the submodel is fully retrieved. PLIF allows the reasoner to carry out inference at *intermediate* stages of the retrieval process over the thus-far retrieved parts, thereby obtaining lower and upper bounds on the posed causal query. We show, in the Discussion section, that our proposed framework, PLIF, is consistent with a wide range of findings in the causal judgment literature, and that PL and PLIF make a number of predictions, some of which are already supported by the findings in the psychology literature.

In their work, Icard and Goodman (2015) articulate a boundedly-rational approach to the CFP at Marr's computational level of analysis, which, as they point out, is from a "god's eye" point of view. In sharp contrast, our proposed framework PLIF is *not* from a "god's eye" point of view and hence could be regarded, potentially, as a psychologically plausible proposal at Marr's algorithmic level of analysis as to how the mind both retrieves and, at the same time, carries out inference over the retrieved submodel to derive bounds on a causal query. We term this *concurrent* approach to reasoning, as opposed to the flawed sequential approach stated earlier.[3] The retrieval process progresses in a local, bottom-up fashion, hence the submodel is retrieved *incrementally*, in a *nested* manner.[4] Our analysis (Sec. 3.4.3) confirms Icard and Goodman's (2015) insight that even in the extreme case of having an infinitely large relevant submodel, the portion of which the reasoner has to consult so as to obtain a "sufficiently good" answer to a query could indeed be very small.

## 3.2 Potential Level and Time

Before proceeding further, let us introduce some preliminary notations. Random variables (RVs) are denoted by lower-case bold-faced letters, e.g., $\mathbf{x}$, and their realizations by non-bold lower-case letters, e.g., $x$. Likewise, sets of RVs are denoted by upper-case bold-faced letters, e.g., $\mathbf{X}$, and their corresponding realizations by upper-case non-bold letters, e.g., $X$. $Val(\cdot)$ denotes the set of possible values a random quantity can take on. Random quantities are assumed to be discrete unless stated otherwise. The joint probability distribution over $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is denoted by $\mathbb{P}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. We will use the notation $\mathbf{x}_{1:n}$ to denote the sequence of $n$ RVs $\mathbf{x}_1, \cdots, \mathbf{x}_n$, hence $\mathbb{P}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \mathbb{P}(\mathbf{x}_{1:n})$. The terms "node" and "variable" will

that in Sec. 3.5, where we connect our work to Long Term Memory and Working Memory.

[3] We elaborate more on this in the Discussion section.

[4] The term "nested" implies that the thus-far retrieved submodel is subsumed by every later submodel (provided that the reasoner proceeds with the retrieval process).

be used interchangeably. To simplify presentation, we adopt the following notation: We denote the probability $\mathbb{P}(\mathbf{x} = x)$ by $\mathbb{P}(x)$ for some RV $\mathbf{x}$ and its realization $x \in \mathit{Val}(\mathbf{x})$. For conditional probabilities, we will use the notation $\mathbb{P}(x|y)$ instead of $\mathbb{P}(\mathbf{x} = x|\mathbf{y} = y)$. Likewise, $\mathbb{P}(X|Y) = \mathbb{P}(\mathbf{X} = X|\mathbf{Y} = Y)$ for $X \in \mathit{Val}(\mathbf{X})$ and $Y \in \mathit{Val}(\mathbf{Y})$. A generic conditional independence relationship is denoted by $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})$ where $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ represent three mutually disjoint sets of variables belonging to a CBN. Furthermore, throughout this chapter, we assume that $\epsilon$ is some negligibly small positive real-valued quantity. Whenever we subtract $\epsilon$ from a quantity, we simply imply a quantity less than but arbitrarily close to the original quantity. The rationale behind adopting such a notation will become clearer in Sec. 3.4.

Before formally introducing the notion of PL, we articulate in simple terms what the idea behind PL is. PL simply induces a *chronological order* on the nodes of a CBN, allowing the reasoner to encode the timing between cause and effect.[5] As we will see, PL plays an important role in guiding the retrieval process used in our proposed framework. Next, PL is formally defined, followed by two clarifying examples.

**Def. 3.1. (Potential Level (PL))** Let $\mathit{par}(\mathbf{x})$ and $\mathit{child}(\mathbf{x})$ denote, respectively, the sets of parents (i.e., immediate causes) and children (i.e., immediate effects) of $\mathbf{x}$. Also let $T_0 \in \mathbb{R} \cup \{-\infty\}$. The PL of $\mathbf{x}$, denoted by $p_l(\mathbf{x})$, is defined as follows: (i) If $\mathit{par}(\mathbf{x}) = \varnothing$, $p_l(\mathbf{x}) = T_0$, and (ii) If $\mathit{par}(\mathbf{x}) \neq \varnothing$, $p_l(\mathbf{x})$ is a real-valued quantity selected from the interval $(\max_{\mathbf{y} \in \mathit{par}(\mathbf{x})} p_l(\mathbf{y}), \min_{\mathbf{z} \in \mathit{child}(\mathbf{x})} p_l(\mathbf{z}))$ such that $p_l(\mathbf{x}) - \max_{\mathbf{y} \in \mathit{par}(\mathbf{x})} p_l(\mathbf{y})$ indicates the amount of time which elapses between intervening simultaneously on all the RVs in $\mathit{par}(\mathbf{x})$ (i.e., $\mathit{do}(\mathit{par}(\mathbf{x}) = \mathit{par}_x)$) and $\mathbf{x}$ taking its value $x$ in accord with the distribution $\mathbb{P}(x|\mathit{par}_x)$. If $\mathit{child}(\mathbf{x}) = \varnothing$, substitute the upper bound of the given interval by $+\infty$. ■

Parameter $T_0$ symbolizes the origin of time, as perceived by the reasoner. $T_0 = 0$ is a natural choice, unless the reasoner believes that time continues unboundedly into the past, in which case $T_0 = -\infty$. The next two examples further clarify the idea behind PL. In both examples we assume $T_0 = 0$.

For the first example, let us consider the CBN depicted in Fig. 3.1(a) containing the RVs $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ with $p_l(\mathbf{x}) = 4, p_l(\mathbf{y}) = 4.7$, and $p_l(\mathbf{z}) = 5$. According to Def. 3.1, the given PLs can be construed in terms of the relative time between the occurrence of cause and effect as articulated next. Upon intervening on $\mathbf{x}$ (i.e., $\mathit{do}(\mathbf{x} = x)$), after the elapse of $p_l(\mathbf{y}) - p_l(\mathbf{x}) =$

[5] More precisely, PL induces a topological order on the nodes of a CBN, with temporal interpretations suggested in Def. 3.1.

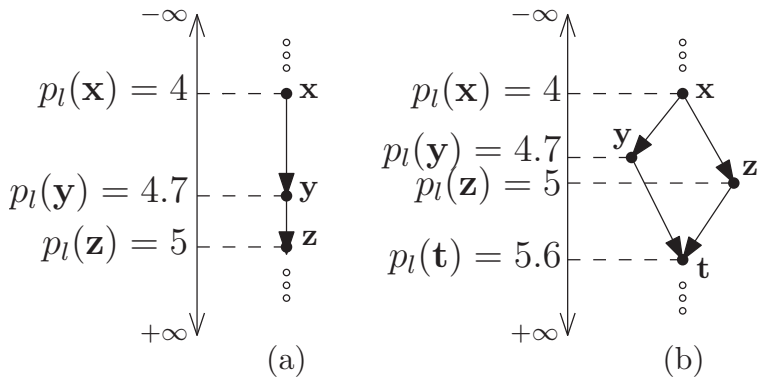



**Fig. 3.1** The relation between PL and time. Three hollow dots signify that the depicted CBNs extend into the past and future.

0.7 units of time, the RV $\mathbf{y}$ takes its value $y$ in accord with the distribution $\mathbb{P}(y|x)$. Likewise, upon intervening on $\mathbf{y}$ (i.e., $do(\mathbf{y} = y)$), after the elapse of $p_l(\mathbf{z}) - p_l(\mathbf{y}) = 0.3$ units of time, $\mathbf{z}$ takes its value $z$ according to $\mathbb{P}(z|y)$.

For the second example, consider the CBN depicted in Fig. 3.1(b) containing the RVs $\mathbf{x}, \mathbf{y}, \mathbf{z}$, and $\mathbf{t}$ with $p_l(\mathbf{x}) = 4, p_l(\mathbf{y}) = 4.7, p_l(\mathbf{z}) = 5$, and $p_l(\mathbf{t}) = 5.6$. Upon intervening on $\mathbf{x}$ (i.e., $do(\mathbf{x} = x)$) the following happens: (i) after the elapse of $p_l(\mathbf{y}) - p_l(\mathbf{x}) = 0.7$ units of time, $\mathbf{y}$ takes its value $y$ according to $\mathbb{P}(y|x)$, and (ii) after the elapse of $p_l(\mathbf{z}) - p_l(\mathbf{x}) = 1$ unit of time, $\mathbf{z}$ takes its value $z$ according to $\mathbb{P}(z|x)$. Also, upon intervening simultaneously on RVs $\mathbf{y}, \mathbf{z}$ (i.e., $do(\mathbf{y} = y, \mathbf{z} = z)$), after the elapse of $p_l(\mathbf{t}) - \max_{\mathbf{r} \in par(\mathbf{t})} p_l(\mathbf{r}) = 0.6$ units of time, $\mathbf{t}$ takes its value $t$ according to $\mathbb{P}(t|y, z)$.

In sum, the notion of PL bears on the underlying time-grid upon which a CBN is constructed, and adheres to Hume's principle of temporal precedence of cause to effect (Hume, 1975). A growing body of work in psychology literature corroborates Hume's centuries-old insight, suggesting that the timing and temporal order between events strongly influences how humans induce causal structure over them (Bramley et al., 2014; Lagnado and Sloman, 2006). The introduced notion of PL is based on the following hypothesis: When learning the underlying causal structure of a domain, humans may as well encode the temporal patterns (or some estimates thereof) on which they rely to infer the causal structure. This hypothesis is supported by recent findings suggesting that people have expectations about the delay length between cause and effect (Greville and Buehner, 2010; Buehner and May, 2004; Schlottmann, 1999). It is worth noting that we could have defined PL in terms of relative *expected time* between cause and effect, rather than relative absolute time. Under

such an interpretation, the time which elapses between the intervention on a cause and the occurrence of its effect would be modeled by a probability distribution, and PL would be defined in terms of the expected value of that distribution. Our proposed framework, PLIF, is indifferent as to whether PL should be construed in terms of absolute or expected time. Greville and Buehner (2010) show that causal relations with fixed temporal intervals are consistently judged as stronger compared to those with variable temporal intervals. This finding, therefore, seems to suggest that people expect, to a greater extent, fixed temporal intervals between cause and effect, rather than variable ones—an interpretation which, at least to a first approximation, favors construing PL in terms of relative absolute time (see Def. 3.1).[6]

## 3.3 Informative Example

To develop our intuition, and before formally articulating our proposed framework, let us present a simple yet informative example which demonstrates: (i) how the retrieval process can be carried out in a local, bottom-up fashion, allowing for retrieving the relevant submodel incrementally, and (ii) how adopting PL allows the reasoner to obtain bounds on a given causal query at intermediate stages of the retrieval process.

Let us assume that the posed causal query is $\mathbb{P}(x|y)$ where $\mathbf{x}, \mathbf{y}$ are two RVs in the CBN depicted in Fig. 3.2(a) with PLs $p_l(\mathbf{x}), p_l(\mathbf{y})$, and let $p_l(\mathbf{x}) > p_l(\mathbf{y})$. The relevant information for the derivation of the posed query (i.e., the relevant submodel) is depicted in Fig. 3.2(e).

Starting from the target RV $\mathbf{x}$ in the original CBN (Fig. 3.2(a)) and moving one step backwards,[7] $\mathbf{t}_1$ is reached (Fig. 3.2(b)). Since $p_l(\mathbf{y}) < p_l(\mathbf{t}_1)$, $\mathbf{y}$ must be a non-descendant of $\mathbf{t}_1$, and therefore, of $\mathbf{x}$. Hence, conditioning on $\mathbf{t}_1$ $d$-separates $\mathbf{x}$ from $\mathbf{y}$ (Pearl, 1988), yielding $(\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{t}_1)$. Thus $\mathbb{P}(x|y) = \sum_{t_1 \in \mathbf{V}al(t_1)} \mathbb{P}(x|y,t_1)\mathbb{P}(t_1|y) = \sum_{t_1 \in \mathbf{V}al(t_1)} \mathbb{P}(x|t_1)\mathbb{P}(t_1|y)$ implying: $\min_{t_1 \in Val(\mathbf{t}_1)} \mathbb{P}(x|t_1) \leq \mathbb{P}(x|y) \leq \max_{t_1 \in Val(\mathbf{t}_1)} \mathbb{P}(x|t_1)$. It is crucial to note that the given bounds can be computed using the information thus-far retrieved, i.e., the information encoded in the submodel shown in Fig. 3.2(b). Taking a step backwards from $\mathbf{t}_1$, $\mathbf{t}_2$ is reached (Fig. 3.2(c)). Using a similar line of reasoning to the one presented for $\mathbf{t}_1$, having $p_l(\mathbf{y}) < p_l(\mathbf{t}_2)$ ensures $(\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{t}_2)$. Therefore, the following bounds on the posed query

[6] There are cases, however, that, despite the precedence of cause to effect, quantifying the amount of time between their occurrences may bear no meaning, e.g., when dealing with hypothetical constructs. In such cases, PL should be simply construed as a topological ordering. From a purely computational perspective, PL is a generalization of *topological sorting* in computer science.

[7] Taking one step backwards from variable $\mathbf{q}$ amounts to retrieving all the parents of $\mathbf{q}$.

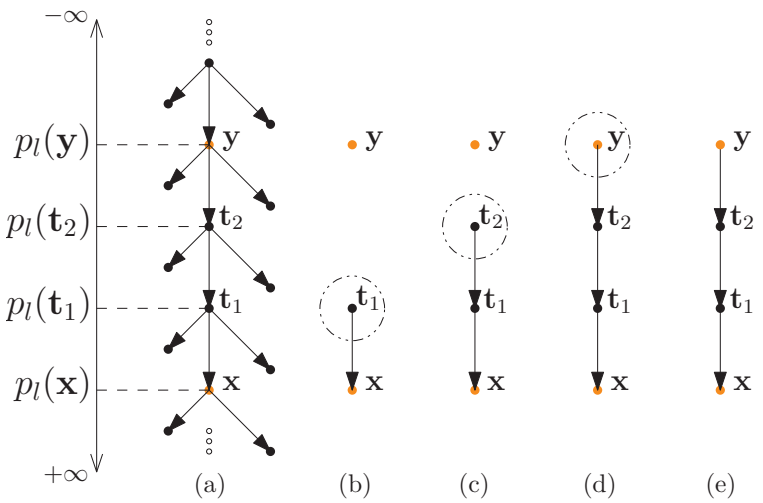



**Fig. 3.2** Example. Query variables are shown in orange.

can be derived, which, crucially, can be computed using the information thus-far retrieved: $\min_{t_2 \in Val(\mathbf{t}_2)} \mathbb{P}(x|t_2) \leq \mathbb{P}(x|y) \leq \max_{t_2 \in Val(\mathbf{t}_2)} \mathbb{P}(x|t_2)$. It is straightforward to show that the bounds derived in terms of $\mathbf{t}_2$ are equally tight or tighter than the bounds derived in terms of $\mathbf{t}_1$. Finally, taking one step backward from $\mathbf{t}_2$, $\mathbf{y}$ is reached (Fig. 3.2(d)) and the exact value for $\mathbb{P}(x|y)$ can be derived, again using the submodel thus-far retrieved (Fig. 3.2(d)).

We are now well-positioned to present our proposed framework.

## 3.4 PL-based Inference Framework (PLIF)

In this section, we intend to elaborate on how, equipped with the notion of PL, a generic causal query of the form[8] $\mathbb{P}(\mathbf{O} = O|\mathbf{E} = E)$ can be derived where $\mathbf{O}$ and $\mathbf{E}$ denote, respectively, the disjoint sets of target (or *objective*) and observed (or *evidence*) variables. In other words, we intend to formalize how inference over a CBN whose nodes are endowed with PL as an attribute should be carried out. Before we present the main result, a few definitions are in order.

**Def. 3.2. (Critical Potential Level (CPL))** The target variable with the least PL is denoted by $\mathbf{o}^*$ and its PL is referred to as the CPL. More formally, $p_l^* :\triangleq \min_{\mathbf{o} \in \mathbf{O}} p_l(\mathbf{o})$ and $\mathbf{o}^* :\triangleq \arg\min_{\mathbf{o} \in \mathbf{O}} p_l(\mathbf{o})$. E.g., for the setting given in Fig. 3.2(a), $\mathbf{o}^* = \mathbf{x}$, and $p_l^* = p_l(\mathbf{x})$.

[8] We do not consider interventions in this work. However, with some modifications, the presented analysis/results can be extended to handle a generic causal query of the form $\mathbb{P}(\mathbf{O} = O|\mathbf{E} = E, do(\mathbf{Z} = Z))$ where $\mathbf{Z}$ denotes the set of intervened variables.

Viewed through the lens of time, $\mathbf{o}^*$ is the furthest target variable into the past, with PL $p_l^*$.

There are two possibilities: (a) $p_l^* > T_0$, or (b) $p_l^* = T_0$, with $T_0$ denoting the origin of time (cf. Sec. 3.2). In the sequel, we assume that (a) holds.[9]

**Def. 3.3. (Inference Threshold (IT) and IT Root Set (IT-RS))** To any real-valued quantity, $\mathcal{T}$, corresponds a unique set, $\mathbf{R}_{\mathcal{T}}$, obtained as follows: Start at every variable $\mathbf{x} \in \mathbf{O} \cup \mathbf{E}$ with PL $\geq \mathcal{T}$ and backtrack along all paths terminating at $\mathbf{x}$. Backtracking along each path stops as soon as a node with PL less than $\mathcal{T}$ is encountered. Such nodes, together, compose the set $\mathbf{R}_{\mathcal{T}}$. It follows that: $\max_{\mathbf{t}\in\mathbf{R}_{\mathcal{T}}} p_l(\mathbf{t}) < \mathcal{T}$. $\mathcal{T}$ and $\mathbf{R}_{\mathcal{T}}$ are termed, respectively, inference threshold (IT) and the IT root set (IT-RS) for $\mathcal{T}$.

For example, the set of variables circled at the stages depicted in Figs. 3.2(b-d) are the IT-RSs for $\mathcal{T} = p_l(\mathbf{x}) - \epsilon$, $\mathcal{T} = p_l(\mathbf{t}_1) - \epsilon$, and $\mathcal{T} = p_l(\mathbf{t}_2) - \epsilon$, respectively. Note that instead of saying $\mathcal{T} = p_l(\mathbf{x}) - \epsilon$ we could have said: for any $\mathcal{T} \in (p_l(\mathbf{t}_1), p_l(\mathbf{x}))$. However, expressing ITs in terms of $\epsilon$ liberates us from having to express them in terms of intervals, thereby simplifying the exposition. We would like to emphasize that the adopted notation should not be construed as implying that the assignment of values to ITs is such a sensitive task that everything would have collapsed, had IT not been chosen in such a fine-tuned manner. To recap, in simple terms, $\mathcal{T}$ bears on how far into the past a reasoner is consulting her mental model in the process of answering a query, and $\mathbf{R}_{\mathcal{T}}$ characterizes the furthest-into-the-past concepts entertained by the reasoner in that process.

Next, we formally present the main idea behind PLIF, followed by its interpretation in simple terms.

**Proposition 3.1.** *Let $\mathbb{P}(O|E)$ denote the posed causal query, with $\boldsymbol{O}$ and $\boldsymbol{E}$ denoting, respectively, the disjoint sets of target and observed variables. For any chosen IT $\mathcal{T} < p_l^*$ and its corresponding $\boldsymbol{R}_{\mathcal{T}}$, define $\boldsymbol{S} :\triangleq \boldsymbol{R}_{\mathcal{T}} \setminus \boldsymbol{E}$. Then the following holds:*

$$\min_{S\in Val(\boldsymbol{S})} \mathbb{P}(O|S,E) \leq \mathbb{P}(O|E) \leq \max_{S\in Val(\boldsymbol{S})} \mathbb{P}(O|S,E). \tag{3.1}$$

*Crucially, the provided bounds can be computed using the information encoded in the submodel retrieved in the very process of obtaining the $\boldsymbol{R}_{\mathcal{T}}$.* □

The message of Proposition 3.1 is simple: For any chosen inference threshold $\mathcal{T}$ which is further into the past than $\mathbf{o}^*$, Proposition 3.1 ensures that the reasoner can condition on $\mathbf{S}$ and obtain the reported lower and upper bounds on the query by using *only* the information

[9]Under Case (b), to derive $\mathbb{P}(O|E)$, the set of all the ancestors of variables in $\mathbf{O} \cup \mathbf{E}$ should be retrieved and then inference should be carried out on the retrieved submodel.

encoded in the retrieved submodel.

It is natural to ask under what conditions the exact value to the posed query can be derived using the thus-far retrieved submodel, i.e., the submodel obtained during the identification of $\mathbf{R}_{\mathcal{T}}$. The following remark bears on that.

**Remark 1.** If for IT $\mathcal{T}$, $\mathbf{R}_{\mathcal{T}}$ satisfies either: (i) $\mathbf{R}_{\mathcal{T}} \subseteq \mathbf{E}$, or (ii) for all $\mathbf{r} \in \mathbf{R}_{\mathcal{T}}$, $p_l(\mathbf{r}) = T_0$, and $\min_{\mathbf{e}\in\mathbf{E}} p_l(\mathbf{e}) > \mathcal{T}$, or (iii) the lower and upper bound given in Proposition 3.1 are identical, then the exact value of the posed query can be derived using the submodel retrieved in the process of obtaining $\mathbf{R}_{\mathcal{T}}$. Fig. 3.2(d) shows a setting wherein (i) and (iii) are both met.

**Rationale behind Remark 1**. Case (i) and Case (iii) immediately follow from Proposition 3.1. Case (ii) implies that all the ancestors of variables in $\mathbf{O}\cup\mathbf{E}$ are retrieved, hence the sufficiency of the retrieved submodel for the exact derivation of the query; see also footnote 9.

### 3.4.1 Proof of Proposition 3.1

In this section, formal proof of Proposition 3.1 is presented.

Simple use of the total probability lemma yields:

$$\mathbb{P}(O|E) = \sum_{S\in Val(\mathbf{S})} \mathbb{P}(O|S,E)\mathbb{P}(S|E). \tag{3.2}$$

Equation (3.2) immediately reveals a simple fact, namely, that $\mathbb{P}(O|E)$ is a linear combination of the members of the set $\{\mathbb{P}(O|S,E)\}_{S\in Val(\mathbf{S})}$, an observation which grants the validity of the expression given in (3.1).

The key point which is left to be shown is the following: (Q.1) Why can the bounds given in (3.1) be computed using the submodel retrieved in the process of obtaining the corresponding $\mathbf{R}_{\mathcal{T}}$ for the adopted IT $\mathcal{T} < p_l^*$? This is where the notion of PL comes into play. To articulate the intended line of reasoning, let us introduce some notations first. According to Def. 3.3, any chosen IT $\mathcal{T}$ induces an IT-RS $\mathbf{R}_{\mathcal{T}}$. Let us partition the set of evidence variables $\mathbf{E}$ into three mutually disjoint sets $\mathbf{E}_T^+, \mathbf{E}_T$, and $\mathbf{E}_T^-$, where $\mathbf{E}_T$ denotes the set of variables in $\mathbf{E}$ which belong to the IT-RS $\mathbf{R}_{\mathcal{T}}$ (i.e., $\mathbf{E}_T :\triangleq \mathbf{E}\cap\mathbf{R}_{\mathcal{T}}$), $\mathbf{E}_T^+$ denotes the set of variables in $\mathbf{E}$ with PLs $\geq \mathcal{T}$, and finally, $\mathbf{E}_T^-$ denotes the set of variables in $\mathbf{E}$ which are neither in $\mathbf{E}_T$ nor in $\mathbf{E}_T^+$ (i.e., $\mathbf{E}_T^- :\triangleq \mathbf{E}\setminus(\mathbf{E}_T\cup\mathbf{E}_T^+)$). Note that, by construction, the PLs of the variables in $\mathbf{E}_T^-$ are less than the adopted IT $\mathcal{T}$, hence the adopted notation. For example, for the setting depicted in Fig. 3.2(b) (corresponding to the IT $\mathcal{T} = p_l(\mathbf{x}) - \epsilon$),

$\mathbf{E}_T = \varnothing, \mathbf{E}_T^+ = \varnothing$, and $\mathbf{E}_T^- = \{\mathbf{y}\}$. Also, for the setting depicted in Fig. 3.2(d) (corresponding to the IT $\mathcal{T} = p_l(\mathbf{t}_2) - \epsilon$), $\mathbf{E}_T = \{\mathbf{y}\}, \mathbf{E}_T^+ = \varnothing$, and $\mathbf{E}_T^- = \varnothing$. Next, we present a key result as a lemma.

**Lemma 3.1.** *Let $\mathbb{P}(O|E)$ denote the posed causal query. For any chosen IT $\mathcal{T} < p_l^*$ and its corresponding IT-RS $\boldsymbol{R}_\mathcal{T}$, the following conditional independence relation holds:*

$$(\mathbf{O} \perp\!\!\!\perp \mathbf{E}_T^- | \mathbf{R}_\mathcal{T} \cup \mathbf{E}_\mathcal{T}^+). \tag{3.3}$$

***Proof.*** The relations between the PLs of the variables involved in the statement (3.3) ensures that, according to $d$-separation criterion (Pearl, 1988), conditioning on the variables in $\mathbf{R}_\mathcal{T} \cup \mathbf{E}_\mathcal{T}^+$ blocks all the paths between the variables in $\mathbf{O}$ and $\mathbf{E}_T^-$, hence follows (3.3).

The following two-part argument responds to the question posed in (Q.1) in the affirmative. First, notice that:

$$\begin{aligned} \mathbb{P}(O|S,E) = &\ \mathbb{P}(O|S, E_\mathcal{T}, E_\mathcal{T}^-, E_\mathcal{T}^+) \\ = &\ \mathbb{P}(O|R_\mathcal{T}, E_\mathcal{T}^-, E_\mathcal{T}^+) \\ \overset{(3.3)}{=} &\ \mathbb{P}(O|R_\mathcal{T}, E_\mathcal{T}^+). \end{aligned} \tag{3.4}$$

Second, note that the process of obtaining $\mathbf{R}_\mathcal{T}$, namely, moving backwards from the variables in $\mathbf{O} \cup \mathbf{E}_\mathcal{T}^+$ until $\mathbf{R}_\mathcal{T}$ is reached, ensures that the submodel retrieved in this process suffices for the derivations of $\mathbb{P}(O|R_\mathcal{T}, E_\mathcal{T}^+)$. Using the approach introduced in (Geiger et al., 1989) for identifying the relevant information for the derivation of a query in a Bayesian network, this follows from the following fact: Conditioned on $\mathbf{R}_\mathcal{T} \cup \mathbf{E}_\mathcal{T}^+$, the set $\mathbf{O}$ is $d$-separated from all the nodes in the set $An(\mathbf{O} \cup \mathbf{E}) \setminus \mathbf{R}_\mathcal{T}$ whose PLs are less than the adopted IT $\mathcal{T}$. Note that $An(\mathbf{O} \cup \mathbf{E})$ denotes the ancestral graph for the nodes in $\mathbf{O} \cup \mathbf{E}$. This completes the proof. ■

### 3.4.2 How Tight the Bounds Given in Proposition 3.1 Really Are? On Maximally-Informative Bounds

Proposition 3.1 grants the validity of the following two statements: (1) A given query $\mathbb{P}(\mathbf{O}|\mathbf{E})$ is guaranteed to fall within the interval provided in (3.1), and (2) The upper and lower bounds provided in (3.1), can be exactly computed from the retrieved submodel the construction of which is delineated in Proposition 3.1. Now, an important question immediately present

itself: How tight the bounds provided in (3.1) really are? Put differently, could one drive bounds tighter than the ones given in (3.1), using the information encoded in the retrieved submodel the construction of which is delineated in Proposition 3.1? In what follows, we show that the bounds given in (3.1) are the "best" one can hope for to derive, using the information encoded in the retrieved submodel the construction of which is formally articulated in Proposition 3.1. We formally characterize this property of the bounds given in (3.1), under a notion which we term *maximally-informativeness*, defined as follows.

**Def. 3.4. (Maximally-Informativeness)** Let $\mathbb{P}(O|E)$ denote the posed causal query, with $\mathbf{O}$ and $\mathbf{E}$ denoting, respectively, the disjoint sets of target and observed variables. A probability interval $[\alpha, \beta]$ is called maximally-informative with respect to a submodel $\mathcal{M}$ iff the tightest probability interval which can be derived for $\mathbb{P}(O|E)$ using the information encoded in $\mathcal{M}$ is $[\alpha, \beta]$.

The following result then holds.

**Proposition 3.2.** *The probability interval provided in Proposition 3.1 is maximally-informative w.r.t. the submodel the construction of which is delineated in Proposition 3.1.*

***Proof.*** We present a constructive proof. We adopt the same notation used in Proposition 3.1. Let $\mathcal{T}^\dagger$ denote a chosen IT. Let us assume that the submodel which would be retrieved in the process of obtaining the corresponding $\mathbf{R}_{\mathcal{T}^\dagger}$ (for the adopted IT $\mathcal{T}^\dagger$) is constructed; the construction procedure is formally given in Def. 3.3. According to Proposition 3.1, the following holds: $\mathbb{P}(O|E) \in I :\triangleq [\min_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E), \max_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E)]$. To prove the claim of Proposition 3.2, it suffices to show that, for any $x \in I$, there exists a CBN $\mathcal{B}_x$ (consistent with the submodel already retrieved) for which $\mathbb{P}(O|E) = x$. It, therefore, suffices to formally delineate the procedure for the construction of $\mathcal{B}_x$. To construct $\mathcal{B}_x$, one simply needs to add to the already constructed submodel, a single binary common-cause node, $\mathbf{c}^\dagger$, for variables in $\mathbf{S}$ (see Proposition 3.1 for the definition of $\mathbf{S}$). Without loss of generality, heretofore we assume that all the variables in $\mathbf{S}$ are binary, with all-one assignment (for $\mathbf{S}$) yielding $\max_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E)$ and all-zero assignment yielding $\min_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E)$. The binary variable $\mathbf{c}_0$ has no parents, and its prior is defined as follows: $\mathbb{P}(\mathbf{c}^\dagger = 1) := \frac{x - \min_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E)}{\max_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E) - \min_{S \in Val(\mathbf{S})} \mathbb{P}(O|S,E)}$. For each variable $\mathbf{s} \in \mathbf{S}$, the conditional $\mathbb{P}(\mathbf{s}|\mathbf{c}^\dagger)$ is parameterized as follows: $\mathbb{P}(\mathbf{s} = 1|\mathbf{c}^\dagger = 1) = 1$ and $\mathbb{P}(\mathbf{s} = 0|\mathbf{c}^\dagger = 0) = 1$. In $\mathcal{B}_x$, variables in $\mathbf{E}$ which fall into the set $\mathbf{R}_{\mathcal{T}^\dagger}$ (i.e., $\mathbf{E} \cap \mathbf{R}_{\mathcal{T}^\dagger}$) are assumed to have no parents, and they a priori take on $E$ with probability one. This completes the construction of $\mathcal{B}_x$. It is easy to show that, in $\mathcal{B}_x$, $\mathbb{P}(O|E) = x$. This concludes the proof. ■

### 3.4.3 Case Study

Next, we intend to cast the hidden Markov model (HMM) studied in (Icard & Goodman, 2015, p. 2) into our framework. The setting is shown in Fig. 3.3(left). We adhere to the

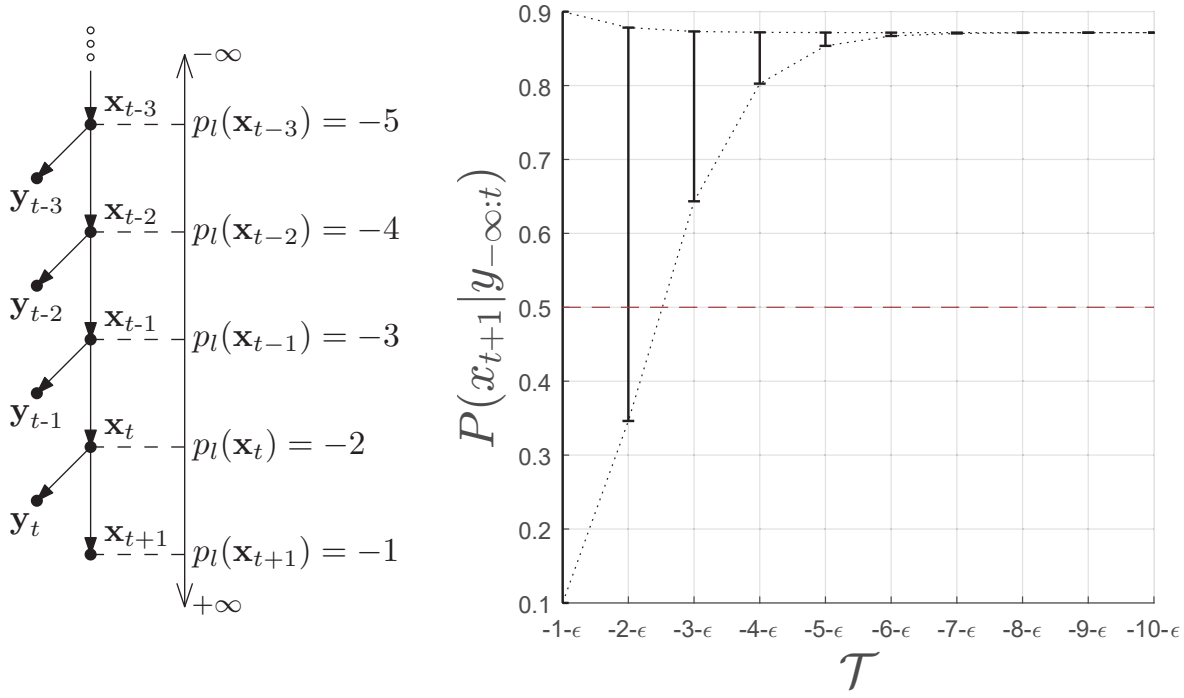



**Fig. 3.3** Left: The infinite-sized HMM discussed in (Icard & Goodman, 2015) with parameterization: $\mathbb{P}(x_{t+1}|x_t) = \mathbb{P}(\bar{x}_{t+1}|\bar{x}_t) = 0.9$, and $\mathbb{P}(y_t|x_t) = \mathbb{P}(\bar{y}_t|\bar{x}_t) = 0.8$. Right: Applying PLIF on the HMM shown in left. Vertical and horizontal axes denote, respectively, the value of the posed query $\mathbb{P}(x_{t+1}|y_{-\infty:t})$ and the adopted IT $\mathcal{T}$. The vertical bars depict the intervals within which the query lies due to Proposition 3.1. The dotted curves—which connect the lower and upper bounds of the intervals—show how the intervals shrink as IT $\mathcal{T}$ decreases.

same parameterization and query adopted therein. All RVs in this section are binary, taking on values from the set $\{0, 1\}$; $\mathbf{x} = x$ indicates the event wherein $\mathbf{x}$ takes the value 1, and $\mathbf{x} = \bar{x}$ implies the event wherein $\mathbf{x}$ takes the value 0. We assume $p_l(\mathbf{x}_{t+i}) = i - 2$.[10] We should note that the assignment of the PLs for the variables in $\{\mathbf{y}_{t-i}\}_{i=0}^{+\infty}$ does not affect the presented results in any way. The query of interest is $\mathbb{P}(x_{t+1}|y_{-\infty:t})$. Notice that after performing three steps of the sort discussed in the example presented in Sec. 3.3 (for the IT

[10]Note that the trend of the upper- and lower-bound curves as well as the size of the intervals shown in Fig. 3.3(right) are insensitive with regard to the choice of PLs for variables $\{\mathbf{x}_{t-i}\}_{i=-1}^{+\infty}$.

$\mathcal{T} = -3 - \epsilon$), the lower bound on the posed query exceeds 0.5 (shown by the red dashed line in Fig. 3.3(right)). This observation has the following intriguing implication. Assume, for the sake of argument, that we were presented with the following Maximum A-Posterior (MAP) inference problem: Upon observing all the variables in $\{\mathbf{y}_{t-i}\}_{i=0}^{+\infty}$ taking on the value 1, what would be the most likely state for the variable $\mathbf{x}_{t+1}$? Interestingly, we would be able to answer this MAP inference problem simply after three backward moves (corresponding to the IT $\mathcal{T} = -3 - \epsilon$). In Fig. 3.3(right), the intervals within which the posed query falls (due to Proposition 3.1) in terms of the adopted IT $\mathcal{T}$ are depicted.

Our analysis confirms Icard and Goodman's (2015) insight that even in the extreme case of having infinite-sized relevant submodel (Fig. 3.3(left)), the portion of which the reasoner has to consult so as to obtain a "sufficiently good" answer to the posed query could happen to be very small (Fig. 3.3(right)).

## 3.5 General Discussion

To our knowledge, PLIF is the first inference framework that capitalizes on *time* to constrain the scope of causal reasoning over CBNs, where the term scope refers to the portion of a CBN on which inference is carried out. PLIF does not restrict itself to any particular inference scheme. The claim of PLIF is that inference should be confined within and carried out over retrieved submodels of the kind suggested by Proposition 3.1 so as to obtain the reported bounds therein. In this light, PLIF can accommodate any inference scheme, including Belief Propagation (BP), and sample-based inference methods using Markov chain Monte Carlo (MCMC), as two prominent classes of inference schemes. MCMC-based methods have been successful in simulating important aspects of a wide range of cognitive phenomena and accounting for many cognitive biases (see Sanborn and Chater, 2016). Also, work in theoretical neuroscience has suggested mechanisms for how BP and MCMC-based methods could be realized in neural circuits (see Gershman and Beck, 2017; Lochmann and Deneve, 2011). For example, to cast BP into PLIF amounts to restricting BP's message-passing within submodels of the kind suggested by Proposition 3.1. In other words, assuming that BP is to be adopted as the inference scheme, upon being presented with a causal query, an IT according to Proposition 3.1 will be selected—at the *meta-level*—by the reasoner and the corresponding submodel, as suggested by Proposition 3.1, will be retrieved, over which inference will be carried out using BP. This will lead to obtaining lower and upper bounds

on the query, as reported in Proposition 3.1. If time permits, the reasoner builds up incrementally on the thus-far retrieved submodel so as to obtain tighter bounds on the query.[11] MCMC-based inference methods can be cast into PLIF in a similar fashion.

A growing body of work suggests that, for computational efficiency, humans flexibly reuse past inferences when faced with new but related queries, a strategy called *amortized inference* (Dasgupta et al., 2017; Stuhlmüller et al., 2013; Gershman and Goodman, 2014). Since PLIF's retrieval process progresses in a local, bottom-up fashion with the submodel being retrieved in an incremental, nested manner, PLIF naturally lends itself to this strategy. For instance, in the example discussed in Sec. 3.3, each member of the set $\{\mathbb{P}(x|t_2)\}_{t_2 \in \mathbf{V}al(t_2)}$ (maximum and minimum of which specify the bounds derived in terms of $\mathbf{t}_2$) can be efficiently computed in terms of the members of the set $\{\mathbb{P}(x|t_1)\}_{t_1 \in \mathbf{V}al(t_1)}$ (maximum and minimum of which specify the bounds derived in terms of $\mathbf{t}_1$) computed at an earlier stage of inference, since the following holds:

$$\mathbb{P}(x|t_2) = \sum_{t_1 \in \mathbf{V}al(t_1)} \mathbb{P}(x|t_1, t_2)\mathbb{P}(t_1|t_2) \overset{(\mathbf{x} \perp\!\!\!\perp \mathbf{t}_2 | \mathbf{t}_1)}{=} \sum_{t_1 \in \mathbf{V}al(t_1)} \mathbb{P}(x|t_1)\mathbb{P}(t_1|t_2).$$

The problem of what parts of a CBN are relevant and what are irrelevant for a given query, according to Geiger, Verma, and Pearl (1989), was first addressed by Shachter (1988). The approaches proposed for identifying the relevant submodel for a given query fall into two broad categories (cf. Mahoney & Laskey, 1998, and references therein): (i) top-down approaches, and (ii) bottom-up approaches. Top-down approaches start with the full knowledge of the underlying CBN and, depending on the posed query, gradually *prune* the irrelevant parts of the CBN. In this respect, top-down approaches are inevitably from "god's eye" point of view—a characteristic which undermines their cognitive-plausibility. Bottom-up approaches, on the other hand, incrementally construct a submodel (by moving backwards from the query variables), using which the posed query can be computed. It is crucial to note that bottom-up approaches cannot stop at *intermediate* steps during the backward move and run inference on the thus-far constructed submodel without running the risk of compromising some of the (in)dependence relations structurally encoded in the CBN, which would yield erroneous inferences. This observation is due to the fact that there exists no local signal revealing how the thus-far retrieved nodes are positioned relative to each other and to the

[11] The very property that the submodel gets constructed incrementally in a nested fashion guarantees that the obtained lower and upper bounds get tighter as the reasoner adopts smaller ITs; see Fig. 3.3(left).

to-be-retrieved nodes—a shortcoming circumvented in the case of PLIF by introducing PL. It is worth reiterating again that PLIF subscribes to what we call the concurrent approach to reasoning (as opposed to the flawed sequential approach mentioned earlier), whereby retrieval and inference take place *in tandem*. The HMM example analyzed in Sec. 3.4.3, with infinitely large relevant submodel, stresses the importance and shows the efficacy of the concurrent approach.

Work on causal judgment provides support for the so-called alternative neglect, according to which subjects tend to neglect alternative causes to a much greater extent in predictive reasoning than in diagnostic reasoning (Fernbach and Rehder, 2013; Fernbach et al., 2011). Alternative neglect, therefore, implies that subjects would tend to ignore parts of the relevant submodel while constructing it. Recent findings, however, seem to cast doubt on alternative neglect (Cummins, 2014; Meder et al., 2014). Meder et al.'s (2014), Experiment 1 demonstrates that subjects appropriately take into account alternative causes in predictive reasoning. Also, Cummins (2014) substantiates a two-part explanation of alternative neglect according to which: (i) subjects interpret predictive queries as requests to estimate the probability of the effect when only the focal cause is present, an interpretation which renders alternative causes irrelevant, and (ii) the influence of inhabitory causes (i.e., disablers) on predictive judgment is underestimated, and this underestimation is incorrectly interpreted as neglecting of alternative causes. Cummins' (2014) Experiment 2 shows that when predictive inference is queried in a manner that more accurately expresses the meaning of noisy-OR Bayes net (i.e., the normative model adopted by Fernbach et al. (2011)) likelihood estimates approached normative estimates. Cummins' (2014) Experiment 4 shows that the impact of disablers on predictive judgments is far greater than that of alternative causes, while having little impact on diagnostic judgments. PLIF commits to the retrieval of enablers as well as disablers. As mentioned earlier, PLIF abstracts away from the inference scheme operating on the retrieved submodel, and, hence, leaves it to the inference scheme to decide how the retrieved enablers and disablers should be weighted and subsequently integrated. In this light, PLIF is consistent with the results of Experiment 4 in Cummins (2014).

In an attempt to explain violations of screening-off reported in the literature, Park and Sloman (2013) find strong support for the contradiction hypothesis followed by the mediating mechanism hypothesis, and finally conclude that people do conform to screening-off once the causal structure they are using is correctly specified. PLIF is consistent with these accounts, as it adheres to the assumption that reasoners carry out inference on their *internal*

causal model (including all possible mediating variables and disablers), not the potentially incomplete one presented in the cover story (see also Rehder and Waldmann, 2017; Sloman and Lagnado, 2015).

Experiment 5 in Cummins (2014), consistent with Fernbach and Rehder (2013), shows that causal judgments are strongly influenced by memory retrieval/activation processes, and that both number of disablers and order of disabler retrieval matter in causal judgments. These findings suggest that the CFP and memory retrieval/activation are intimately linked. In that light, next, we intend to elaborate on the rationale behind adopting the term "retrieve" and using it interchangeably with the term "consult" throughout this chapter; this is where we relate PLIF to the concepts of Long Term Memory (LTM) and Working Memory (WM) in psychology and neurophysiology. Next, we elaborate on how PLIF could be interpreted through the lenses of two influential models of WM, namely, Baddeley and Hitch's (1974) Multi-component model of WM (M-WM) and Ericsson and Kintsch's (1995) Long-term Working Memory (LTWM) model. The M-WM postulates that "long-term information is downloaded into a separate temporary store, rather than simply activated in LTM," a mechanism which permits WM to "manipulate and create new representations, rather than simply activating old memories" (Baddeley, 2003). Interpreting PLIF through the lens of the M-WM model amounts to the value for IT being chosen (and, if time permits, updated so as to obtain tighter bounds) by the central executive in the M-WM and the submodel being incrementally "retrieved" from LTM into M-WM's episodic buffer. Interpreting PLIF through the lens of the LTWM model amounts to having no retrieval from LTM into WM and the submodel suggested by Proposition 3.1 being merely "activated in LTM" and, in that sense, being simply "consulted" in LTM. In sum, PLIF is compatible with both of the narratives provided by the M-WM and LTWM models.

A number of predictions follow from PL and PLIF. For instance, PLIF makes the following prediction: Prompted with a predictive or a diagnostic query (i.e., $\mathbb{P}(\mathbf{e}|\mathbf{c})$ and $\mathbb{P}(\mathbf{c}|\mathbf{e})$, respectively), subjects should not retrieve any of the effects of $\mathbf{e}$. Introspectively, this prediction seems plausible, and can be tested, using a similar approach to (Cummins, 2014; De Neys et al., 2003), by asking subjects to "think aloud" while engaged in predictive or diagnostic reasoning. Also, PL yields the following prediction: Upon intervening on cause $\mathbf{c}$, subjects should be sensitive to *when* effect $\mathbf{e}$ will occur, even in settings where they are not particularly instructed to attend to such temporal patterns. Recent findings suggesting that people have expectations about the delay length between cause and effect already provide

some supporting evidence for this prediction (Greville and Buehner, 2010; Buehner and May, 2004).

There is a growing acknowledgment in the literature that, not only time and causality are intimately linked, but that they *mutually constrain* each other in human cognition (see Buehner, 2014). In line with this view, we see our work also as an attempt to formally articulate how time could guide and constrain causal reasoning. While many questions remain open, we hope to have made some progress towards better understanding of the CFP at the algorithmic level of analysis.

# Part II:
# On Minimality in Action

**Preface.** Our daily experience suggests that we humans are both efficient and, perhaps, more intriguingly, thrifty in devising our interventions to achieve our desired goals. Among potentially enumerable variables of the environment amenable to intervention, we magically zero in on few pivotal variables to exercise our interventions on; but how? Formalization of this curious phenomenon is the aim of Part II. The results of Part II have important implications for a line of work in developmental psychology concerning causal learning by young children in pedagogical settings. Furthermore, the formalism developed in Part II establishes, for the first time in the literature, a *rational*, *algorithmic-level* account of a peculiar behavior demonstrated by young children in pedagogical settings (and generally taken as evidence for children's irrationality), namely, *overimitation*: children's persistently reproducing the adult's unnecessary actions.

# Chapter 4

# Probabilistic Structural Controllability in Causal Bayesian Networks

*"In the mind there is no absolute or free will; but the mind is determined to wish this or that by a cause, which has also been determined by another cause, and this last by another cause, and so on to infinity."*

— Baruch Spinoza, Ethics

## 4.1 Introduction

The aptitude to perceive causation plays a central role in human cognition, and *intervention* is the sole means of actively (in contrast with the passive mode of being a mere observer) interacting with a world governed by causal structures. Among possible intentions behind exerting intervention, the notion of "control" is a notable one—that is, informally speaking, to manipulate some variables[1] (also called driver variables) of a system to, either directly or indirectly, "guide" or "control" variables of the system which are of interest.

In this chapter, the problem of targeted probabilistic structural controllability (TPS-controllability) in the context of causal Bayesian networks (CBNs) is introduced and for-

[1]The terms "node" and "variable" will be used interchangeably throughout.

malized. The term "structural" signifies the condition wherein the agent is equipped merely with the causal structure of the domain under study. The term "targeted," on the other hand, emphasizes that the agent is interested in controlling the behavior of a specific subset (or all) of variables in the domain called target variables. Finally, the term "probabilistic" highlights the probabilistic nature of the problem under study.

At a high level, we define the problem of *probabilistic controllability* in the context of CBNs as follows: *How an agent, provided with the knowledge of the set of intervenable variables, should devise her intervention, i.e., (Q.1) "which" variables to intervene on, and (Q.2) "how" to intervene on those, so as to "control" the behavior of some particular variable(s) of interest in the domain (represented by a CBN), that is, to maximize or minimize the probability of the occurrence of a state of interest for a set of target variables.*

The problem of probabilistic *structural* controllability in the context of CBNs is then accordingly defined as that of probabilistic controllability—as stated above—with one crucial additional constraint on the agent's part: The agent is solely equipped with the knowledge of the underlying causal *structure* of the domain (i.e., the CBN's topology) and is uninformed of the parameterization thereof. In this work, we aim at identifying the minimal set of intervenable variables sufficient for TPS-controllability of an *arbitrary* CBN. Particularly, we devise an algorithm, $\mathcal{C}^*$, which identifies a sufficient set of intervenable variables for the purpose of TPS-controllability of a generic CBN. We also elaborate on the nature of minimality that the identified set satisfies.

The question of interest to this chapter has significant ramifications for studies on strategic planning and policy making. Equally importantly, the problem under study has notable connections to how humans, at the computational level of analysis (Marr, 1982) and in line with the rational analysis approach (Anderson, 1990), should devise their interventions to increase the odds of attaining their desired goals while faced with their uncertain environment. We will extensively elaborate on the implications of the work presented in this chapter for cognitive psychology and development psychology in Sec. 4.9.

## 4.2 Notation and Terminology

In this section, we present some preliminary notations and terminologies which will be adopted in this chapter. Random quantities are denoted by bold-faced letters; their realizations are denoted by the same letter but non-bold. More specifically, random variables

(RVs) are denoted by bold-faced lower-case letters, e.g., $\mathbf{x}$, and their realizations by non-bold lower-case letters, e.g., $x$. Likewise, sets of RVs are denoted by bold-faced calligraphic letters, e.g., $\boldsymbol{\mathcal{X}}$, and their corresponding realizations by non-bold calligraphic letters, e.g., $\mathcal{X}$. $Val(\cdot)$ denotes the set of possible values a random quantity can take on. To simplify presentation, we incur the following abuse of notation: We denote the probability $\mathbb{P}(\mathbf{x} = x)$ by $\mathbb{P}(x)$ for some RV $\mathbf{x}$ and its realization $x \in Val(\mathbf{x})$. For conditional probabilities, we will use the notation $\mathbb{P}(x|y)$ instead of $\mathbb{P}(\mathbf{x} = x|\mathbf{y} = y)$. Likewise, $\mathbb{P}(\mathcal{X}|\mathcal{Y}) := \mathbb{P}(\boldsymbol{\mathcal{X}} = \mathcal{X}|\boldsymbol{\mathcal{Y}} = \mathcal{Y})$ for $\mathcal{X} \in Val(\boldsymbol{\mathcal{X}})$ and $\mathcal{Y} \in Val(\boldsymbol{\mathcal{Y}})$. Random quantities are assumed to be discrete unless stated otherwise.

Throughout this chaper, the directed acyclic graph (DAG) $G = (V, E)$ characterizes the non-intervened causal structure of the domain where $V$ denotes the set of nodes/variables and $E$ denotes the set of edges. We adopt Pearl's notation $do(x) := do(\mathbf{x} = x)$ to denote an atomic intervention on $\mathbf{x}$ so as to force it to take on the value $x$. Also, $ip(\mathbf{x})$ denotes the intervention policy to be adopted for $\mathbf{x}$ the meaning of which will be clarified in the subsequent section; informally intervention policy refers to how the agent decides to manipulate the intervened variable (see Pearl, 2000, Sec. 4.2). Intervention policy may or may not functionally depend on other variables of the domain. As we will see later, intervention policy in its most generic form is nothing but a conditional probability distribution (CPD). Also, backward chaining (BC) on a variable refers to the simple process of identifying its parents (i.e., immediate causes) and the parents of the parents and so forth until the boundaries of the CBN are reached. Finally, $\delta(\cdot)$ denotes the Kronecker delta function.

Before proceeding further, let us formally define two key notions, namely, *subsumability* and *domination*.

**Def. 4.1 (Subsumability):** DAG $G_1 = (V_1, E_1)$ *subsumes* DAG $G_2 = (V_2, E_2)$, denoted in short by $G_1 \supseteq G_2$, iff $V_1 = V_2$ and $E_2 \subseteq E_1$. We refer to the set $E_1 \setminus E_2$, as the *surplus* of $G_1$ with respect to $G_2$.

**Def. 4.2 (Domination):** DAG $G_1 = (V_1, E_1)$ *dominates* DAG $G_2 = (V_2, E_2)$, denoted by $G_1 \succcurlyeq G_2$, iff there *exists* a parameterization of $G_1$ which yields a result for the objective of interest that is no worse than what is achievable by *any* parameterization of $G_2$. For instance, if the objective of interest is to maximize the probability of some event of interest, say $\mathbf{r} = r$ for some $r \in Val(\mathbf{r})$, then we write $G_1 \succcurlyeq G_2$ iff there exists a parameterization of $G_1$ which yields some value for the probability of interest, $\mathbb{P}(r)$, which is greater than or equal to what is achievable by any parameterization of $G_2$.

**Lemma 4.1. (Domination vs Subsumability):** Let $G_1, G_2$ be DAGs. Then, $G_1 \supseteq G_2 \Rightarrow G_1 \succcurlyeq G_2$.

***Proof.*** The proof is straightforward once we realize that one can very well take advantage of the extra edges of $G_1$ with respect to $G_2$ (i.e., the surplus of $G_1$ with respect to $G_2$) which gives one more "degrees of freedom" to entertain and hence to achieve a result which is equally good or better than what is achievable by any parameterization of $G_2$ in terms of the objective of interest. ■

In subsequent sections where we introduce our objectives of interest, the above statements will become clearer.

## 4.3 Motivating Examples

To develop some intuition as to the problem under study, we present in this section a series of informative examples.

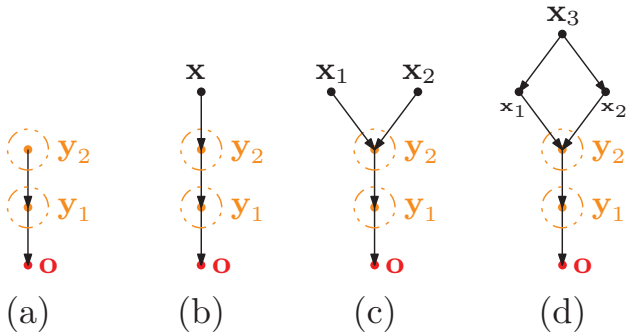



**Fig. 4.1** Motivating example.

Let us first consider the CBN depicted in Fig. 4.1(a). Variables $\mathbf{y}_1, \mathbf{y}_2$ are amenable to intervention (or, in short, *intervenable*). The objective is to make the occurrence of the event $\mathbf{o} = o \in Val(\mathbf{o})$ as likely as possible through intervening on a subset of variables $\{\mathbf{y}_1, \mathbf{y}_2\}$ (or to choose not to intervene at all which corresponds to choosing the empty set). The key question is how, by mere investigation of the *structure* of the CBN depicted in Fig. 4.1(a), to decide: (i) on which (intervenable) variables to intervene, and (ii) how the intervention should be exercised (a notion referred to as intervention policy (IP)). It is easy to come to the conclusion that, to make the occurrence of $\mathbf{o} = o$ as likely as possible, one needs to just intervene on $\mathbf{y}_1$ and force it into the state $\mathbf{y}_1 = y_1^*$ where $y_1^*$ is the realization for $\mathbf{y}_1$ conditioned on which the probability of event $\mathbf{o} = o$ is maximum, i.e., $y_1^* = \arg\max_{y_1 \in Val(\mathbf{y}_1)} \mathbb{P}(o|y_1)$.

It is crucial to realize that, due to the *structure* of the CBN depicted in Fig. 4.1(a), and *regardless* of its parameterization, it suffices for the agent to solely intervene on $\mathbf{y}_1$ for the purpose of TPS-controllability of $\mathbf{o} = o$ (the answer to (i)). Furthermore, since the agent is assumed to be equipped merely with the structure of the underlying CBN and not the parameterization thereof, based on the above argument on $\mathbf{y}_1$'s IP, the agent can just arrive at the conclusion that $\mathbf{y}_1$'s IP has the functional (or structural) form of[2] $\mathbb{P}(\mathbf{y}_1)$—that is, merely the *non-parametric* form of the IP (the answer to (ii)). Altogether, a solution to the problem of TPS-controllability of $\mathbf{o} = o$ is $\{\mathbf{y}_1\}$ (which is a sufficient set of variables to be intervened) along with $\mathbb{P}(\mathbf{y}_1)$ which is the *functional* form of $\mathbf{y}_1$'s IP. Following the same line of reasoning for the CBNs depicted in Figs. 1(b-d), it is straightforward to argue that $\{\mathbf{y}_1\}$ is a sufficient set for TPS-controllability of the target variable $\mathbf{o}$, and $\mathbf{y}_1$'s IP has the functional form of $\mathbb{P}(\mathbf{y}_1)$ akin to what we had for Fig. 4.1(a).

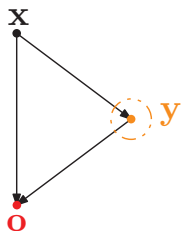



**Fig. 4.2** Motivating example.

Let us consider another example that highlights a key idea, namely, that we may need to broaden our understanding of the notion of intervention (see Pearl, 2000, Sec. 4.2). Consider the CBN depicted in Fig. 4.2. This time, only $\mathbf{y}$ is intervenable. Assume that (only for this particular example), all the variables are binary-valued; the prior probability on $\mathbf{x}$ is $\mathbb{P}(\mathbf{x})$ (which is assumed to be non-degenerate), $\mathbf{y} := \neg\mathbf{x}$, and $\mathbf{o} := \mathbf{x} \oplus \mathbf{y}$ where $\neg$ and $\oplus$ denote the logical connectives *not* and *xor*, respectively. It is easy to verify that the event $\mathbf{o} = 1$ occurs with probability one, regardless of the choice of $\mathbb{P}(\mathbf{x})$. For the problem of TPS-controllability of $\mathbf{o} = 1$, the agent has to decide whether or not to intervene on $\mathbf{y}$. Imagine an intervention were to be exercised on $\mathbf{y}$. Whether the agent would set $\mathbf{y} = 0$, or $\mathbf{y} = 1$, the objective event of $\mathbf{o} = 1$ would become less likely to happen compared to that of the (non-intervened) original model. At first glance, the fact that exercising intervention makes

[2]In fact, the agent can reason out one step further and come to the conclusion that $\mathbf{y}_1$'s IP must have the functional form of $\mathbb{P}(\mathbf{y}_1) = \delta(\mathbf{y}_1 = y^*)$, however, the agent cannot identify/specify the value of $y^*$—due to the lack of knowledge about the parameterization of the CBN.

the situation worse in terms of the objective of interest seems rather counter-intuitive. How could it be that having the freedom to manipulate variable $\mathbf{y}$ (which even happens to be one of the parents of the objective node) whatever way we like does not allow us to outperform the (non-intervened) original model? The answer lies in developing a better understanding of the term "whatever way we like." For that purpose, we need to broaden our conception of the notion of intervention and go beyond practicing merely a primitive atomic form of intervention denoted by $do(\mathbf{y} = y)$ in the literature. A more advanced form of intervention is to pick the state to which we want to force the intervened variable as a function of the states of some other variables of the domain. That is, IP may depend functionally on a collection of other variables in the domain. In this example, choosing $\mathbf{y}$'s IP, denoted by $ip(\mathbf{y})$, to functionally depend on $\mathbf{x}$ ensures that, the outcome achieved by exerting such intervention, is equally good or better (i.e., no worse) than that of the (non-intervened) original model. In such a setting, we simply adopt the following terminology/notation: The intervention pair $(\mathbf{y}, ip(\mathbf{y}) := \mathbb{P}(\mathbf{y}|\mathbf{x}))$ (comprising, in order, the set of intervened variables and their corresponding IPs) is equally good or better than both: (i) the intervention pair $(\mathbf{y}, ip(\mathbf{y}) := \mathbb{P}(\mathbf{y}) = \delta(\mathbf{y} = y))$ corresponding to the simplistic atomic form of intervention on $\mathbf{y}$ discussed above which does not depend on the state of $\mathbf{x}$ and, likewise, (ii) the intervention pair $(\varnothing, \varnothing)$ corresponding to the original model (without intervention).

The last example shows how our commonsense about the TPS-controllability problem is, at best, only *partially* correct: (†) Commonsense suggests that we should intervene on the intervenable ancestors of target variable(s) that are somewhat "closest" to the target variable(s). However, it might remain silent or, even worse, might mislead us in deciding about what functional form the IPs should possess (e.g., deciding between exercising $(\mathbf{y}, ip(\mathbf{y}) := \delta(\mathbf{y} = y))$ or $(\mathbf{y}, ip(\mathbf{y}) := \mathbb{P}(\mathbf{y}|\mathbf{x}))$ in the given example). In this light, the aim of this chapter is to formally articulate the TPS-controllability problem, to formalize our intuition about it, and importantly, to shed light on non-trivial aspects of the problem.

## 4.4 Intervention Policy

The notion of IP delineates how an intervened variable should be "manipulated." More specifically, IP indicates whether other variables play any role or not (and if so, how) in devising how the manipulation on a to-be-intervened variable is to be practiced. It is intuitive that the more variables we are allowed to functionally depend on while devising the IP of a

to-be-intervened variable, the more "degrees of freedom" we have in controlling the behavior of the to-be-intervened variable. In what follows, we will present a graphical representation for IP which is an adaptation of Tian's manipulated graph to our context (Tian, 2008). As we will see, this graphical representation, along with the idea of *subsumability*, allows us to formalize the above intuition. By allowing a larger number of variables for the IP of an intervened variable to *functionally* depend on, we also show that IPs can be organized in a hierarchical construct wherein moving up in the hierarchy amounts to empowering the agent to exercise more sophisticated forms of intervention.

### 4.4.1 Hierarchical Construct

Before proceeding further let us make a definition: The scope of an IP is the set of variables, with the exception of the intervened variable itself, that the IP functionally depends on. For instance, for $ip(\mathbf{y}) := \mathbb{P}(\mathbf{y}|\mathbf{s})$, the scope is comprised of the variable $\mathbf{s}$. In short, the idea of organizing IPs into a hierarchical construct is inspired by the simple realization that, by delimiting the set of variables the agent is allowed to incorporate into the scope of the intervened variable's IP (functionally represented by a conditional probability distribution), we impose a constraint on the expressive power of the IP.

The following notation henceforth will be employed to refer to different IP classes:

• IP *class*-0: This class refers to the set of IPs where the scope of each is the empty set, i.e., a setting wherein the IP(s) of the intervened variable(s) is not allowed to incorporate any variables into its scope. That is, if variable $\mathbf{x}$ is decided to be intervened and the agent is only permitted to adopt *class*-0 IPs, then, the agent is just allowed to place IP of the functional form $\mathbb{P}(\mathbf{x})$ on $\mathbf{x}$ to exercise her intervention. It is crucial to note that the agent is allowed to parameterize $\mathbb{P}(\mathbf{x})$ as she wishes, yet, the functional form of the IP is constrained.

• IP *class*-$j, \forall j \geq 1$: This class refers to the set of IPs where the scope of each is the ancestors of the corresponding intervened variable up to $i^{\text{th}}$ level. For instance, for the case of IP *class*-2, IP(s) of the intervened variable(s) is solely allowed to take into account the state of (i) the immediate causes, and (2) the immediate causes of the variables in (i), thereby, altogether functionally depending on all the ancestors up to the $2^{\text{nd}}$ level.

• IP *class*-$\infty$: This class refers to the set of IPs where the scope of each is *all* the ancestors of the corresponding intervened variable. Note that the complete set of ancestors of a variable $\mathbf{x}$ can be found by instantiating BC on $\mathbf{x}$.

Finally, it is crucial to notice the following. For an IP to be in a particular class amounts

to imposing a constraint solely on the *functional* form of the IP; the agent is free to choose any parameterization for the IP as she may wish. Therefore, $ip(\mathbf{x}) \in$ *class-i* simply means that the *functional* form of $ip(\mathbf{x})$ is constrained in accord with the definition of IP *class-i* given above, yet, it could be arbitrarily parameterized. Also, assuming that $\boldsymbol{\mathcal{X}} = \{\mathbf{x}_i\}_{i=1}^m$, the notation $ip(\boldsymbol{\mathcal{X}}) \in$ *class-j* will be adopted as a shorthand for the following: $ip(\mathbf{x}_i) \in$ *class-j*, $\forall i = 1, \ldots, m$.

### 4.4.2 Graphical Representation

In this section, we discuss a way of visualizing IPs which is an adaptation of Tian's manipulated graph to our context (Tian, 2008). If the IP of a to-be-intervened variable $\mathbf{x}$ functionally depends on $\mathbf{y}$, then we show this by a directed dash-dotted arrow emanating from $\mathbf{y}$ and pointing towards $\mathbf{x}$.[3] To ensure that any practice of intervention is fully expressed by such edges we introduce the following convention: For DAG $G = (V, E)$, a clamped variable $\mathfrak{C}$ is added to $V$.[4] Then, intervening on a variable $\mathbf{a}$ which has no parents and exerting $ip(\mathbf{a}) = \mathbb{P}(\mathbf{a})$ will be illustrated graphically by a dash-dotted edge emanating from $\mathfrak{C}$ towards $\mathbf{a}$. In general, upon $\mathbf{y}$ taking on the state $y$, the agent may decide to set the value of $\mathbf{x}$ to a fixed value $x$ (deterministic IP), or to set the value of $\mathbf{x}$ probabilistically (stochastic IPs), i.e., $\mathbf{x}$ takes on values from $Val(\mathbf{x})$ according to some conditional probability distribution $\mathbb{P}(\mathbf{x}|\mathbf{y})$. In both cases, $ip(\mathbf{x})$ is said to be *functionally* dependent on $\mathbf{y}$. Simply put, in devising the intervention policy of $\mathbf{x}$, namely, $ip(\mathbf{x})$, the state of $\mathbf{y}$ is taken into account. The notion of probabilistic IP is discussed in (Pearl, 2000, pp. 113-114) under the title of stochastic policy.

Let us first give some definitions which will prove useful in the subsequent sections. For the given definitions, DAG $G = (V, E)$ represents the (non-intervened) causal structure of the domain.

**Def. 4.3 (Intervention Pair):** A set of intervened variables $\boldsymbol{\mathcal{K}} \subseteq V$ along with their corresponding IPs comprise a pair, called an *intervention pair*, which is denoted by $(\boldsymbol{\mathcal{K}}, ip(\boldsymbol{\mathcal{K}}))$.

**Def. 4.4 (Intervention DAG ($i$-DAG)):** Every intervention pair $(\boldsymbol{\mathcal{K}}, ip(\boldsymbol{\mathcal{K}}))$ for $\boldsymbol{\mathcal{K}} \subseteq V$ uniquely specifies a DAG (as described above) which we refer to as the $i$-DAG associated to that intervention pair. The $i$-DAG associated to $(\boldsymbol{\mathcal{K}}, ip(\boldsymbol{\mathcal{K}}))$ is denoted by $(\boldsymbol{\mathcal{K}}, ip(\boldsymbol{\mathcal{K}}))_G$.

[3] Notice that according to Pearl (2000), upon intervening on $\mathbf{x}$, all the (pre-intervention) incoming edges into $\mathbf{x}$ should first be removed.

[4] It is implicitly assumed throughout this paper that the variable $\mathfrak{C}$ has been added to DAG $G$ *a priori*. Also, we will not depict $\mathfrak{C}$ in figures unless needed.

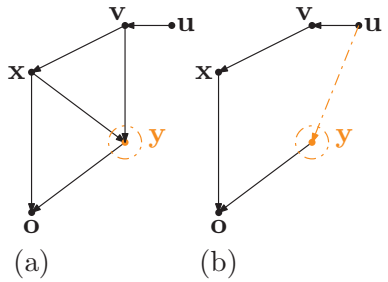



**Fig. 4.3** Sample case. **(a):** Original CBN. Variable $\mathbf{y}$ is to be intervened according to $ip(\mathbf{y}) := \mathbb{P}(\mathbf{y}|\mathbf{u})$. **(b):** The graphical representation of intervening on $\mathbf{y}$ with $ip(\mathbf{y}) := \mathbb{P}(\mathbf{y}|\mathbf{u})$. The figure simply illustrates the fact that the state of $\mathbf{y}$ gets decided (potentially probabilistically) according to the state of $\mathbf{u}$.

**Def. 4.5 ($i$-Subsumability):** For $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \subseteq V$, $i$-DAG $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G$ $i$-subsumes $i$-DAG $(\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$, denoted in short by $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G \supseteq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$, iff (i) $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G \supseteq (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$, (ii) the set of the dash-dotted edges in $(\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$ is a subset of the set of the dash-dotted edges in $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G$, and (iii) the surplus of $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G$ with respect to $(\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$ is solely comprised of dash-dotted edges.

**Def. 4.6 ($i$-Domination):** For $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \subseteq V$, $i$-DAG $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G$ $i$-dominates $i$-DAG $(\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$, denoted by $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G \succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$ for short, iff there exist a parameterization for the dash-dotted edges (see Fig. 4.3) in $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G$ which yields a result for the objective of interest that is no worse than what is achievable by *any* parameterization of the dash-dotted edges in $i$-DAG $(\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$.

**Lemma 4.2. ($i$-Domination vs $i$-Subsumability):** *Let DAG $G = (V, E)$ characterize the causal structure of the domain. Let $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G$ and $(\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$ be two* i*-DAGs for some $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \subseteq V$. Then, the following holds: $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G \supseteq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G \Rightarrow (\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}))_G \succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}))_G$.*

***Proof.*** The rationale is similar to the one presented for Lemma 4.1. ■

The following corollary immediately follows from Lemma 4.2.

**Corollary 4.1.** For $G = (V, E)$ and $\forall \boldsymbol{\mathcal{K}} \subseteq V$, the following holds true: $\forall j \geq m$, $(\boldsymbol{\mathcal{K}}, ip(\boldsymbol{\mathcal{K}}) \in \textit{class-}j)_G \succcurlyeq_i (\boldsymbol{\mathcal{K}}, ip(\boldsymbol{\mathcal{K}}) \in \textit{class-}m)_G$.

## 4.5 TPS-Controllability of CBNs: Formalization

Let DAG $G = (V, E)$ characterize the causal structure of the domain. Let $V = V_i \cup \bar{V}_i$ where $V_i$ denotes the set of nodes/variables amenable to intervention (or, in short, *intervenable*), and

$\bar{V}_i$ be the complement of $V_i$, i.e., $\bar{V}_i = V \setminus V_i$. The probability of interest, in its generic form takes the form[5] $\mathbb{P}(\boldsymbol{\mathcal{O}} = \mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\boldsymbol{\mathcal{X}}) = ip(\mathcal{X})])$, or in short $\mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})])$, where $\boldsymbol{\mathcal{O}}$ denotes the set of target variables, $\mathcal{O}$ denotes the realization of interest, $\boldsymbol{\mathcal{X}}$ denotes the set of intervened variables, $ip(\boldsymbol{\mathcal{X}})$ denotes the *functional* form (i.e., non-parametric representation) of the to-be-adopted intervention policy, and $ip(\mathcal{X})$ denotes a specific parameterization[6] of $ip(\boldsymbol{\mathcal{X}})$. Also, $do[\boldsymbol{\mathcal{X}}; ip(\boldsymbol{\mathcal{X}})]$ denotes the setting wherein variables $\boldsymbol{\mathcal{X}}$ are intervened according to $ip(\boldsymbol{\mathcal{X}})$ (functional form) and, likewise, $do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]$ denotes the setting wherein variables $\boldsymbol{\mathcal{X}}$ are intervened according to $ip(\boldsymbol{\mathcal{X}}) = ip(\mathcal{X})$. One can write, $\boldsymbol{\mathcal{O}} = \boldsymbol{\mathcal{O}}_i \cup \bar{\boldsymbol{\mathcal{O}}}_i$ where $\boldsymbol{\mathcal{O}}_i \subseteq V_i$ and $\bar{\boldsymbol{\mathcal{O}}}_i \subseteq \bar{V}_i$. Objectives of interest could have any of the following forms:

$$\max_{\boldsymbol{\mathcal{X}} \subseteq V_i} \left( \max_{ip(\mathcal{X}) \in \textit{class-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \right), \tag{4.1}$$

$$\min_{\boldsymbol{\mathcal{X}} \subseteq V_i} \left( \min_{ip(\mathcal{X}) \in \textit{class-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \right), \tag{4.2}$$

and,

$$\max_{\boldsymbol{\mathcal{X}} \subseteq V_i} \left( \min_{ip(\mathcal{X}) \in \textit{class-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \right), \tag{4.3}$$

$$\min_{\boldsymbol{\mathcal{X}} \subseteq V_i} \left( \max_{ip(\mathcal{X}) \in \textit{class-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \right). \tag{4.4}$$

In the sequel, we focus on objectives (4.1) and (4.2).[7] Therefore, whenever we use the statement "objective of interest" we are specifically referring to both of objectives (4.1) and (4.2) unless stated otherwise.

Next, we devise an algorithm, $\mathcal{C}^*$, for the problem of TPS-controllability of CBNs. $\mathcal{C}^*$ outputs a set of intervenable variables, $\boldsymbol{\mathcal{X}}^*$, which is "optimal" with respect to objectives (4.1) and (4.2). In other words, $\boldsymbol{\mathcal{X}}^*$ is a sufficient choice of variables to intervene on (according to IP *class-*$\infty$) to satisfy objectives (4.1) and (4.2). Formally put, for $\boldsymbol{\mathcal{X}}^*$ the following holds:

[5]The connection to Pearl's notation for *do*-calculus is as follows (see Pearl, 2000, p. 114): $\mathbb{P}(y)|_{\mathbb{P}^*(\mathbf{x}|\mathbf{z})} = \mathbb{P}(y|do[\mathbf{x}; \mathbb{P}^*(\mathbf{x}|\mathbf{z})])$.

[6]Which is equivalent to a specific parameterization of the dash-dotted edges representing the intervention policy exercised on variables $\boldsymbol{\mathcal{X}}$ in the corresponding $i$-DAG (see Fig. 4.3).

[7]The solution to both objectives (4.3) and (4.4) is the empty set, i.e., to intervene on none of the intervenable variables at all. In fact, a more general result can be established for minimax and maximin objectives: Subject to the constraint that the IPs of the to-be-intervened variables have to belong to IP *class-j*, the solution to both minimax and maximin objectives is the empty set, for all $j \geq 1$. For details, the reader is referred to Sec. B-I of Appendix B.

$\forall \boldsymbol{\mathcal{Y}} \subseteq V_i$,

$$(\boldsymbol{\mathcal{X}}^*, ip(\boldsymbol{\mathcal{X}}^*) \in \textit{class-}\infty)_G \succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}) \in \textit{class-}\infty)_G.$$

For the proof, the reader is referred to Sec. B-II of Appendix B.

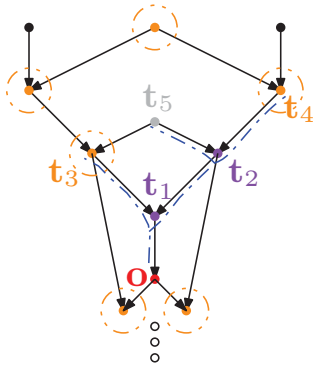



**Fig. 4.4** Sample Case: Variable **o** (depicted in red) is the target variable. The intervenable variables (i.e., members of $V_i$) are circled. BC execution paths are colored in blue and illustrated by dash-dotted lines. Upon initiating BC at the target variable **o**, we arrive at $\mathbf{t}_1$ (depicted in purple) located at the junction. Next, we arrive at $\mathbf{t}_2$ and $\mathbf{t}_3$. Since $\mathbf{t}_3 \in V_i$, BC terminates at $\mathbf{t}_3$. On the other hand, since $\mathbf{t}_2 \notin V_i$, BC continues. Having performed BC on $\mathbf{t}_2$, we arrive at $\mathbf{t}_4$ and $\mathbf{t}_5$. Since $\mathbf{t}_4 \in V_i$, BC terminates on $\mathbf{t}_4$. At the end, since $\mathbf{t}_5$ (depicted in grey) has no parents (immediate causes), BC terminates at $\mathbf{t}_5$ as well. Therefore, by mere investigation of the structure, $\mathcal{C}^*$ outputs the set $\boldsymbol{\mathcal{X}}^* = \{\mathbf{t}_3, \mathbf{t}_4\}$ as a solution to the objectives (4.1) and (4.2) for this particular setting.

### 4.5.1 Algorithm $\mathcal{C}^*$

Let us explain simply how $\mathcal{C}^*$ works. BC has to be initiated on nodes in $\boldsymbol{\mathcal{O}}$. Upon reaching any node in $V_i$, the BC execution path terminates at that node. This procedure continues until, for all of the BC execution paths, either: (i) The BC execution path gets terminated at some node belonging to $V_i$, or (ii) a node with no parents is reached. The set of intervenable variables at which BC terminates constitute $\mathcal{C}^*$'s output denoted by $\boldsymbol{\mathcal{X}}^*$. Algorithm $\mathcal{C}^*$ nicely captures our commonsense with respect to the TPS-Controllability problem, as alluded to in (†) in the last paragraph of Sec. 4.3. Fig. 4.4 depicts a sample execution of $\mathcal{C}^*$. The worst-case running time of $\mathcal{C}^*$ is $O(|E| + |V|)$, i.e., linear in the size of $G$.[8] It is crucial to

[8]For more on this, the reader is referred to Sec. B-V of Appendix B.

note that despite the simplicity of $\mathcal{C}^*$, the proof of its correctness is far from trivial. For the proof, the reader is referred to Sec. B-II of Appendix B.

## 4.6 Reducing the Scope of IPs: Toward Minimal Scopes

So far we have shown that $(\boldsymbol{\mathcal{X}}^*, ip(\boldsymbol{\mathcal{X}}^*) \in \textit{class-}\infty)_G \succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}) \in \textit{class-}\infty)_G, \forall \boldsymbol{\mathcal{Y}} \subseteq V_i$. From Lemma 4.2, it immediately follows that $(\boldsymbol{\mathcal{X}}^*, ip(\boldsymbol{\mathcal{X}}^*) \in \textit{class-}\infty)_G \succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}) \in \textit{class-}j)_G$, $\forall \boldsymbol{\mathcal{Y}} \subseteq V_i$, $\forall j \in \mathbb{N} \cup \{0, \infty\}$, where $\mathbb{N}$ denotes the set of natural numbers. A key question now arises: Could the condition of *class-*$\infty$ be relaxed for $\boldsymbol{\mathcal{X}}^*$ without jeopardizing its optimality? That is, are all the ancestors of $\boldsymbol{\mathcal{X}}^*$, as dictated by *class-*$\infty$, always necessary? Formally put, are there any settings for which we can replace *class-*$\infty$ in the expression $(\boldsymbol{\mathcal{X}}^*, ip(\boldsymbol{\mathcal{X}}^*) \in \textit{class-}\infty)_G \succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}) \in \textit{class-}j)_G$, $\forall \boldsymbol{\mathcal{Y}} \subseteq V_i$, $\forall j \in \mathbb{N} \cup \{0, \infty\}$ with something less demanding and yet preserve the validity of the expression? In what follows we will show, through a series of examples, that there indeed exist settings for which the *class-*$\infty$ condition can be safely relaxed. In this respect, the examples serve as a "proof-of-concept" in the hope that they provide some intuition as to the question being posed. We will conclude this section by presenting a lemma and a conjecture; the validity or falsity of the conjecture is left to be shown in future work. The conjecture has hitherto defied a proof, yet all our attempts to refute it has thus-far been futile.

### 4.6.1 Motivating Examples

Recall that the term "objective of interest" refers to objectives (1) and (2).

Let us consider the CBN depicted in Fig. 4.5(a). Notice the resemblance of the CBN shown in Fig. 4.5(a) and the one presented in Fig. 4.1(a). Variable $\mathbf{y}$ is the only variable in the system which is intervenable and the target variable is $\mathbf{o}$. Execution of $\mathcal{C}^*$ results in $\boldsymbol{\mathcal{X}}^* = \{\mathbf{y}\}$. To exercise an IP *class-*$\infty$ on $\mathbf{y}$ amounts to having $ip(\mathbf{y}) = \mathbb{P}(\mathbf{y}|\mathbf{x})$. However, it is easy to come to the conclusion that simply exerting an atomic intervention on $\mathbf{y}$ (i.e., IP *class-*0) suffices for achieving the objective of interest and $\mathbf{x}$ need not be incorporated into the scope of $\mathbf{y}$'s IP; see footnote 2. The same line of reasoning holds true for the CBN depicted in Figs. 4.5(b).

Now let us consider the CBN shown in Fig. 4.5(c) which is one of the CBNs explicated in Sec. 4.3. As was the case in Sec. 4.3, $\mathbf{y}$ is the only variable in the system which is intervenable and the target variable is $\mathbf{o}$. Execution of $\mathcal{C}^*$ results in $\boldsymbol{\mathcal{X}}^* = \{\mathbf{y}\}$. As discussed in Sec. 4.3,

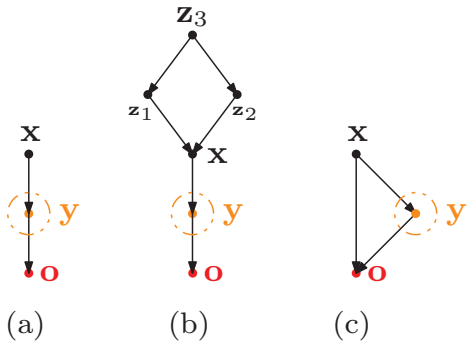



**Fig. 4.5** Motivating example.

depending on the parameterization of the CBN, $\mathbf{x}$ may need to be incorporated into the scope of $\mathbf{y}$'s IP. In Sec. 4.3 we provided a parameterization which indeed necessitated the incorporation of $\mathbf{x}$ into the scope of $\mathbf{y}$'s IP in order to achieve the objective of interest. A comparison between the CBNs shown in Fig. 4.5(a) and Fig. 4.5(c) reveals the following: For the CBN shown in Fig. 4.5(a), exercising atomic intervention on $\mathbf{y}$ renders $\mathbf{x}$ $d$-separated from $\mathbf{o}$, whereas, for the CBN depicted in Fig. 4.5(c), exercising atomic intervention on $\mathbf{y}$ does not render $\mathbf{x}$ $d$-separated from $\mathbf{o}$. We will return to this observation at the end of this section.

Consider now the CBN depicted in Fig. 4.6(a). Variables $\mathbf{y}_1, \mathbf{y}_2$ are intervenable and the target variable is $\mathbf{o}$. Execution of $\mathcal{C}^*$ results in $\boldsymbol{\mathcal{X}}^* = \{\mathbf{y}_1, \mathbf{y}_2\}$. Exercising IP *class*-$\infty$ on $\mathbf{y}_1$ and $\mathbf{y}_2$ amounts to having $ip(\mathbf{y}_1) = \mathbb{P}(\mathbf{y}_1|\mathbf{x})$ and $ip(\mathbf{y}_2) = \mathbb{P}(\mathbf{y}_2|\mathbf{x})$. It is straightforward to show that simply exercising atomic interventions on $\mathbf{y}_1$ and $\mathbf{y}_2$ suffices for achieving the objective of interest and $\mathbf{x}$ need not be incorporated into the scopes of $\mathbf{y}_1$ and $\mathbf{y}_2$'s IPs, hence IP *class*-0 for $\mathbf{y}_1$ and $\mathbf{y}_2$ suffices. More specifically, $\mathbf{y}_1$ and $\mathbf{y}_2$ must be forced to take, respectively, the values $y_1^*$ and $y_2^*$ where $(y_1^*, y_2^*) = \arg\max_{(y_1,y_2)\in Val(\mathbf{y}_1)\times Val(\mathbf{y}_2)} \mathbb{P}(o|y_1, y_2)$ and $(y_1^*, y_2^*) = \arg\min_{(y_1,y_2)\in Val(\mathbf{y}_1)\times Val(\mathbf{y}_2)} \mathbb{P}(o|y_1, y_2)$ to satisfy the maximax and minimin objectives, respectively. Notice once again that exercising atomic interventions on $\{\mathbf{y}_1, \mathbf{y}_2\}$ renders $\mathbf{x}$ $d$-separated from $\mathbf{o}$. We will revisit this observation at the end of this section.

Finally, let us consider the CBN depicted in Fig. 4.7(a). Variables $\mathbf{y}$ is the only variable in the system which is intervenable and the target variable is $\mathbf{o}$. Execution of $\mathcal{C}^*$ results in $\boldsymbol{\mathcal{X}}^* = \{\mathbf{y}\}$. Exercising IP *class*-$\infty$ on $\mathbf{y}$ amounts to having $ip(\mathbf{y}) = \mathbb{P}(\mathbf{y}|\mathbf{x}, \mathbf{z})$. Nonetheless, simple calculations, as presented below, reveal that the mere incorporation of $\mathbf{z}$ into the scope of $ip(\mathbf{y})$ suffices for achieving the objective of interest and $\mathbf{x}$ need not be incorporated into the scope of $ip(\mathbf{y})$. As we will see, the aforesaid result is an immediate implication of

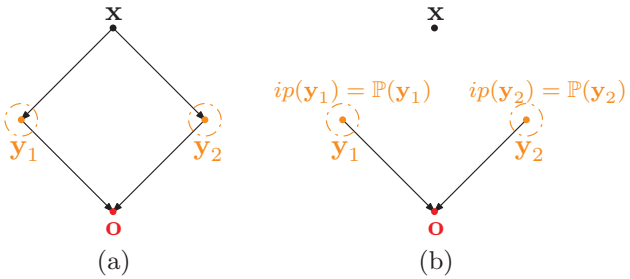



**Fig. 4.6** **(a)** Executing $\mathcal{C}^*$ yields $\boldsymbol{\mathcal{X}}^* = \{\mathbf{y}_1, \mathbf{y}_2\}$. **(b)** Graphical representation of $ip(\mathbf{y}_i) = \mathbb{P}(\mathbf{y}_i)$ for $i = 1, 2$.

Rule 1 of Pearl's *do*-calculus in our context: In DAG $G_{\overline{\mathbf{y}}}$, conditioning on $\mathbf{x}$ $d$-separates $\mathbf{z}$ from the target variable $\mathbf{o}$. Let $\mathbb{P}^*(\mathbf{y}|\mathbf{x}, \mathbf{z})$ denote the optimal *class*-$\infty$ IP which should be exercised on $\mathbf{y}$ so as to achieve the objective of interest.

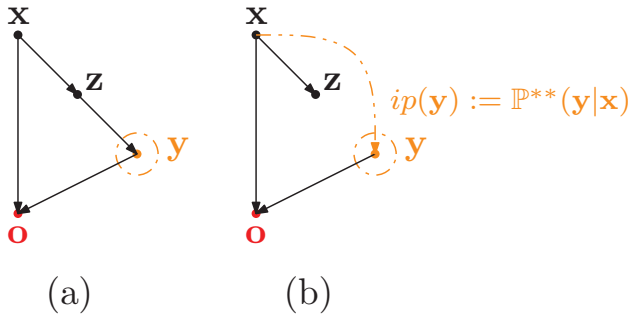



**Fig. 4.7** **(a)** Executing $\mathcal{C}^*$ yields $\boldsymbol{\mathcal{X}}^* = \{\mathbf{y}\}$. **(b)** Graphical representation of $ip(\mathbf{y}) = \mathbb{P}^{**}(\mathbf{y}|\mathbf{x})$.

Through simple calculations, we show that the aforesaid IP can be "replaced" with an IP, $\mathbb{P}^{**}(\mathbf{y}|\mathbf{x})$, into the scope of which $\mathbf{z}$ need not be incorporated, thereby "relaxing" the IP *class*-$\infty$ condition. $\mathbb{P}(o|do[\mathbf{y}, \mathbb{P}^*(\mathbf{y}|\mathbf{x}, \mathbf{z})])$ can be expressed as follows.[9]

$$\mathbb{P}(o|do[\mathbf{y}, \mathbb{P}^*(\mathbf{y}|\mathbf{x}, \mathbf{z})])$$

[9]Stochastic policies can be expressed in terms of atomic interventions as explained in (Pearl, 2000, pp. 113-114) and (Pearl, 1995, p. 684).

$$
\begin{aligned}
&= \sum_{x,y,z} \mathbb{P}(x)\mathbb{P}(z|x)\mathbb{P}^*(y|x,z)\mathbb{P}(o|do(y),x,z) \\
&\overset{R.1}{=} \sum_{x,y,z} \mathbb{P}(x)\mathbb{P}(z|x)\mathbb{P}^*(y|x,z)\mathbb{P}(o|do(y),x) \\
&= \sum_{x,y} \mathbb{P}(x)\mathbb{P}(o|do(y),x) \sum_{z} \mathbb{P}(z|x)\mathbb{P}^*(y|x,z) \\
&= \sum_{x,y} \mathbb{P}(x)\mathbb{P}(o|do(y),x)\Big(\mathbb{P}^{**}(y|x)\Big).
\end{aligned}
$$

Rule 1 of *do*-calculus is applied at the second step denoted by $R.1$ and simple marginalization is carried at the third step whose result is written between parenthesis in the final expression.

If the reader follows the approach adopted for the last example also for the examples depicted in Fig. 4.5(a) and Fig. 4.6(a), she will realize that indeed Rule 1 of *do*-calculus plays a critical role in deciding what variables suffice to be included into the scopes of the intervened variables in those cases as well. In fact, the observations we made earlier in this section grant the applicability of Rule 1 of *do*-calculus in examples given in Fig. 4.5(a) and Fig. 4.6(a). In conclusion, it appears that Rule 1 of Pearl's *do*-calculus plays an important role in relaxing IP *class*-$\infty$ requirement; future work will hopefully shed light on this matter. Next we formally state our results as a lemma, followed by a conjecture.

**Lemma 4.3.** *Let DAG $G = (V, E)$ characterize the causal structure of the domain. To preserve optimality with respect to the objectives given in (1) and (2), the following variables need not be incorporated into the scope of $ip(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}}^*$: Any $\boldsymbol{y} \in V_i \setminus \boldsymbol{\mathcal{X}}^*$ which is an ancestor of $\boldsymbol{x}$.*

**Conjecture 4.1.** *Let DAG $G = (V, E)$ characterize the causal structure of the domain, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}^*$ be a to-be-intervened variable. Let ip-scope$(\boldsymbol{x})$ denote the scope of $ip(\boldsymbol{x})$ which is selected according to the following procedure:*

> *Assume that all the ancestors of $\boldsymbol{x}$ are intervenable in addition to the variables in the set $V_i$. Run the algorithm $\mathcal{C}^*$. Let $\boldsymbol{\mathcal{W}}^*$ denote the set of variables that $\mathcal{C}^*$ outputs. Set ip-scope$(\boldsymbol{x}) := \boldsymbol{\mathcal{W}}^* \setminus \boldsymbol{\mathcal{X}}^*$.*

*Let $ip(\boldsymbol{\mathcal{X}}^*|\boldsymbol{\mathcal{S}}^*)$ denote the set of IPs to be exercised on $\boldsymbol{\mathcal{X}}^*$ whose scopes (i.e., elements of $\boldsymbol{\mathcal{S}}^*$) are selected according to the above procedure. Then, the following holds: $(\boldsymbol{\mathcal{X}}^*; ip(\boldsymbol{\mathcal{X}}^*|\boldsymbol{\mathcal{S}}^*))_G \succcurlyeq_i (\boldsymbol{\mathcal{X}}^*; ip(\boldsymbol{\mathcal{X}}^*) \in$ class-$\infty)_G$.*

## 4.7 On the Minimality of $\mathcal{C}^*$'s Output

Let us present two definitions and two propositions which bear on the minimality of $\mathcal{C}^*$'s output $\boldsymbol{\mathcal{X}}^*$.

**Def. 4.7 (Locally Structurally Minimal (LSM)):** $\mathcal{C}^*$'s output, $\boldsymbol{\mathcal{X}}^*$, is LSM with respect to (the input) DAG $G$ iff there exists a parameterization of $G$ such that *no* proper subset of $\boldsymbol{\mathcal{X}}^*$, namely, $\boldsymbol{\mathcal{X}}^{**}$, exists for which the following holds: $(\boldsymbol{\mathcal{X}}^{**}, ip(\boldsymbol{\mathcal{X}}^{**}) \in \textit{class-}\infty)_G \succcurlyeq_i (\boldsymbol{\mathcal{X}}^*, ip(\boldsymbol{\mathcal{X}}^*) \in \textit{class-}\infty)_G$.

**Def. 4.8 (Uniformly Structurally Minimal (USM)):** $\mathcal{C}^*$ is called USM iff, for any (input) DAG $G$, $\mathcal{C}^*$'s output, $\boldsymbol{\mathcal{X}}^*$, is LSM with respect to that DAG $G$.

**Proposition 4.1.** *$\mathcal{C}^*$ is USM for the maximax objective given in (4.1).*

***Proof.*** The proof is constructive. The objective of interest is maximax given in (4.1). Let us assume, without loss of generality, that all the RVs are binary-valued and the desired state is for all the target variables to take on value one. Our goal is to parameterize an *arbitrary* $G$ in such a way that: (i) the desired state happens with probability one if variables $\boldsymbol{\mathcal{X}}^*$ are all set to one through exerting atomic interventions, and (ii) the desired state happens with probability zero otherwise. Start at $\boldsymbol{\mathcal{X}}^*$. Parameterize the CPD of each $\mathbf{x}^* \in \boldsymbol{\mathcal{X}}^*$ such that it always takes on the value zero. Moving along the BC execution paths terminated at $\boldsymbol{\mathcal{X}}^*$, proceed towards the target variables which are descendants[10] of $\boldsymbol{\mathcal{X}}^*$. Along the way, parameterize the CPD associated to any variable $\mathbf{k}$ such that, conditioned on $\mathbf{k}$'s parents which are descendants of $\boldsymbol{\mathcal{X}}^*$ (denoted by $par_{\boldsymbol{\mathcal{X}}^*}(\mathbf{k})$), $\mathbf{k}$ takes on value one iff all $par_{\boldsymbol{\mathcal{X}}^*}(\mathbf{k})$ take value one.[11] In other words, intermediate variables like $\mathbf{k}$ work as an *and* logical gate. Proceed in the aforementioned manner until all the target variables (which are descendants of $\boldsymbol{\mathcal{X}}^*$) are reached. It is easy to verify that indeed the desired state happens with probability one iff all the variables $\boldsymbol{\mathcal{X}}^*$ are set to one and, furthermore, intervening on any proper subset of $\boldsymbol{\mathcal{X}}^*$ in any way does not yield such an outcome. This concludes the proof. ■

**Proposition 4.2.** *$\mathcal{C}^*$ is USM for the minimin objective given in (4.2) when the set of target variables $\boldsymbol{\mathcal{O}}$ is a singleton.*

***Proof.*** The proof is constructive. Let us assume, without loss of generality, that all the RVs are binary-valued and the undesired state for the target variable (i.e., the state whose

[10] For any target variable $\mathbf{q}$ which is not a descendant of $\boldsymbol{\mathcal{X}}^*$, parameterize $\mathbb{P}(\mathbf{q}|par(\mathbf{q}))$ such that $\mathbf{q}$ takes the value 1 with probability one.

[11] In other words, the CPD of $\mathbb{P}(\mathbf{k}|par(\mathbf{k}))$ is parameterized in such a manner that the parents of $\mathbf{k}$ which are *not* descendants of $\boldsymbol{\mathcal{X}}^*$ are rendered ineffective.

probability is to be minimized) is the zero state. Our goal is to parameterize an *arbitrary* $G$ in such a way that: (i) the target variable takes the zero state with probability one unless *all* the variables in $\boldsymbol{\mathcal{X}}^*$ are set to one, and (ii) once all the variables in $\boldsymbol{\mathcal{X}}^*$ are intervened to take the value one, the target variable takes the value zero with probability zero. Start at $\boldsymbol{\mathcal{X}}^*$. Parameterize the CPD of each $\mathbf{x}^* \in \boldsymbol{\mathcal{X}}^*$ such that it always takes on the value zero. Moving along the BC execution paths terminated at $\boldsymbol{\mathcal{X}}^*$, proceed towards the target variable. Along the way, parameterize the CPD associated to any variable $\mathbf{k}$ such that, conditioned on $\mathbf{k}$'s parents which are descendants of $\boldsymbol{\mathcal{X}}^*$ (denoted by $par_{\boldsymbol{\mathcal{X}}^*}(\mathbf{k})$), $\mathbf{k}$ takes on value one iff all $par_{\boldsymbol{\mathcal{X}}^*}(\mathbf{k})$ take the value one.[12] In other words, intermediate variables like $\mathbf{k}$ serve as an *and* logical gate. Proceed in the aforesaid manner until the target variable is reached. It is easy to verify that indeed the undesired state for the target variable occurs with probability zero iff *all* the variables in $\boldsymbol{\mathcal{X}}^*$ are set to one. This concludes the proof. ■

It is intriguing to see that the output of $\mathcal{C}^*$ (with $\mathcal{C}^*$ formally capturing our intuition with respect to the TPS-Controllability problem) satisfies the said non-trivial minimality property.

## 4.8 Optimal Intervention Policy: Computational Complexity

In what follows, we elaborate on the computational complexity of finding the Optimal Intervention Policy (OIP) to be exercised on $\boldsymbol{\mathcal{X}}^*$. Let us formally define the problems the complexity of which are of concern, namely, OIP-MAXMAX-FP and OIP-MINMIN-FP.

**Def. 4.9** (OIP-MAXMAX-FP): Given a CBN $\mathcal{B}$ with causal structure $G$, parameterized by distribution $\mathbb{P}$ (which factorizes over $G$), and the corresponding set $\boldsymbol{\mathcal{X}}^*$, output the OIP to be exercised on $\boldsymbol{\mathcal{X}}^*$, that is the IP which is optimal with respect to maximax objective given in (4.1), i.e., $\arg\max_{ip(\mathcal{X}^*)\in class\text{-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}^*; ip(\mathcal{X}^*)])$.

**Def. 4.10** (OIP-MINMIN-FP): Given a CBN $\mathcal{B}$ with causal structure $G$, parameterized by distribution $\mathbb{P}$ (which factorizes over $G$), and the corresponding set $\boldsymbol{\mathcal{X}}^*$, output the OIP to be exercised on $\boldsymbol{\mathcal{X}}^*$, that is the IP which is optimal with respect to minimin objective given in (4.2), i.e., $\arg\min_{ip(\mathcal{X}^*)\in class\text{-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}^*; ip(\mathcal{X}^*)])$.

**Proposition 4.3.** OIP-MAXMAX-FP *and* OIP-MINMIN-FP *are both NP-hard.*

The reader is referred to Sec. B-IV of Appendix B for the proof of Proposition 4.3.

[12] In other words, the CPD of $\mathbb{P}(\mathbf{k}|par(\mathbf{k}))$ is paremeterized in such a manner that the parents of $\mathbf{k}$ which are *not* descendants of $\boldsymbol{\mathcal{X}}^*$ are rendered ineffective.

Interestingly, the NP-hardness results for OIP-MAXMAX-FP and OIP-MINMIN-FP are established using a special class of (degenerate) CBNs for which any Exact Inference or Maximum A-Posterior (MAP) query can be answered in poly-time (hence, tractable).

## 4.9 On the Connections to Cognitive Psychology

A defining characteristic of Algorithm $\mathcal{C}^*$ is its strong tendency to select the closest intervenable nodes to target node(s) for intervention.[13] That is, Algorithm $\mathcal{C}^*$ endorses the following dictum: "The closer a to-be-intervened node is to the target node(s), the better," which we refer to as the *proximity principle*. It is worth nothing that the proclaimed proximity principle follows from the machinery of $\mathcal{C}^*$ (i.e., by $\mathcal{C}^*$'s starting at target nodes and moving backwards toward intervenable nodes). Interestingly, the said proximity principle is a direct implication of White's "dissipation effect" in causal settings (White, 1997, 2000; Edwards et al., 2015). Concretely, Edwards et al. (2015) suggest that, in settings wherein causal relations are probabilistic rather than deterministic, human subjects are more inclined to intervene on immediate causes of the target variable. Through a series of experiments, Edwards et al. (2015) provide support for a general preference of human subjects to intervene on immediate causes rather than intermediate ones (see Edwards et al., 2015, p. 1921), fully consistent with the proximity principle entailed by $\mathcal{C}^*$. In that light, Algorithm $\mathcal{C}^*$ serves as the first rational, process-level account of how subjects devise their intervention strategies (by selecting on which intervenable variables to intervene) to "control" the state of a target node.

Edwards et al.'s (2015) experiments also provide support for a tendency to intervene on root-causes when subjects were asked to decide where they would prefer to intervene so as to bring about long-term goals (as opposed to short-term goals). The concepts of long-term and short-term goals inevitably involve the notion of time, and particularly, how long it may take for the effect of an intervention to "reach" the target variable(s). In that light, subjects being presented with such tasks will be essentially concerned with the following two criteria together: (i) maximizing/minimizing the probability of the desired/undesired state for the target variable, and (ii) the amount of time it takes for the effect of an intervention to reach target variable(s). The existence of the second criterion inevitably makes the task lie outside the scope of the problem addressed in this chapter. In Chapter 3, we formally

[13] The extreme case being to intervene right on the target node(s) and set it to the desired value.

articulated how time can be integrated into the formalism of CBNs, by introducing the notion of potential level (PL). Future work should investigate how PL and the formalism of probabilistic structural controllability can be brought together to account for psychological findings concerning attaining long-term vs. short-term goals such as Edwards et al.'s (2015).

### 4.9.1 Implications for Developmental Psychology: Overimitation and Causal Learning

Imitation is an effective learning strategy, allowing for sophisticated forms of cultural transmission (Tomasello, 2009; Whiten et al., 2011; Nielsen, 2012). Children are surprisingly prolific imitators, but there are also times when their copying of others' actions appears to be obviously irrational, inducing major errors in reasoning (Lyons et al., 2007; Whiten et al., 1996). The terms *overimitation* particularly refers to this children's persistently reproducing the adult's unnecessary actions. Several studies have documented this behavior, suggesting that it is most likely uniquely human (Horner and Whiten, 2005), that it exists across various cultures (Nielsen and Tomaselli, 2010), that it emerges in early ages and increases with age (McGuigan et al., 2007; Nielsen and Tomaselli, 2010; McGuigan and Whiten, 2009), and that it occurs despite children's ability to distinguish relevant actions from irrelevant actions (Nielsen and Tomaselli, 2010). Surprisingly, children have been observed to engage in overimitation in various contexts (Call et al., 2005; Carpenter et al., 2002; Horner and Whiten, 2005; McGuigan et al., 2007; Nagell et al., 1993; Want and Harris, 2002; Whiten et al., 1996), even in settings where chimpanzees correctly ignore the unnecessary, irrelevant steps (Horner and Whiten, 2005; Nagell et al., 1993; Want and Harris, 2002; Whiten et al., 1996).

In a series of studies, Lyons et al. (2007) provide evidence for children having a strong tendency to encode all of an adult's purposeful actions as causally necessary, automatically revising their causal beliefs about the object accordingly. Lyons et al. (2007) show that children are frequently unable to avoid reproducing the adult's irrelevant actions despite countervailing task demands, time pressure, and even direct warnings, thereby providing strong evidence for their automatic causal encoding (ACE) account. The resulting distortions in children's causal beliefs, as Lyons et al. (2007) show, are the actual cause of the overimitation, not implicit social demands (Horner and Whiten, 2005; Nielsen, 2006; Uzgiris, 1981) or imitative habit (McGuigan et al., 2007; Whiten et al., 1996) as previously believed. Drawing on the fact that learning the casual structure of a domain is indeed computationally

hard, Lyons et al. (2007) argue that adults and children alike rely on the intentional manipulations/interventions of knowledgeable people to learn causally important operations, with adults doing it deliberately whereas children automatically (which leads them to overimitate). Fully consistent with this view, our proposed algorithm $\mathcal{C}^*$ formally shows, having observed an adult's purposeful actions, how children should revise their (mental) causal model, represented by a CBN. As prescribed by $\mathcal{C}^*$, the revised causal model (of a child) should be such that $\mathcal{C}^*$'s run on the revised model outputs the adult's purposeful set of actions (i.e., $\boldsymbol{\mathcal{X}}^*$). The aforesaid condition severely constraints how children should go about revising their causal model and considerably limits the set of causal models they possibly need to entertain, thereby reducing the computational complexity of the task—in line with Lyons et al.'s (2007) view. Hence, $\boldsymbol{\mathcal{X}}^*$ serves as a distinctive pedagogical cue helping young children, not yet having developed elaborate intuitive theories, learn about the causal structure of their environment. Furthermore, consistent with Lyons et al.'s (2007) ACE account of overimitation, our proposed algorithm $\mathcal{C}^*$ suggests that $\boldsymbol{\mathcal{X}}^*$—which simultaneously satisfies optimality and minimality criteria—is indeed a rational choice for children to imitate, a behavior which will be perceived by an external observer (i.e., the experimenter) as an instance of *over*imitation. (The observation that in Lyons et al.'s (2007) experiments children frequently copied all the actions demonstrated by the experimenter suggests that children deemed unnecessary to add extra actions to and/or remove any action from the set of actions presented by the experimenter, with the former strongly supporting the sufficiency of the set and the latter the minimality of the set.) In this light, our proposed algorithm $\mathcal{C}^*$ serves as a rational, algorithmic-level account of overimitation under single-demonstration condition (i.e., where children are presented with a single demonstration of how to generate the outcome by the experimenter). To our knowledge, $\mathcal{C}^*$ is the first rational, algorithmic-level account of this curious behavior.[14]

### On Asynchronous Distributed Implementation of $\mathcal{C}^*$

We would like to elaborate on an asynchronous distributed implementation of $\mathcal{C}^*$ which is of interest for Marr's algorithmic level of analysis (Marr, 1982). Let us assume that

[14]Here, we are particularly focusing on single-demonstration condition, aka one-shot learning setting (e.g., Lyons, Young, & Keil, 2007; Whiten, Custance, Gomez, Teixidor, & Bard, 1996). Some recent studies have focused on *repeated*-demonstration condition which inevitably requires children to combine various statistical information and perform statistical inference (e.g., Buchsbaum, Gopnik, Griffiths, & Shafto, 2011). In that light, Repeated-demonstration condition falls outside the scope of our work.

nodes symbolize computational units which can communicate with their immediate neighbors (i.e., their children and parents) through the edges of the underlying DAG—symbolizing communication channels. The distributed algorithm begins by each node in $\boldsymbol{\mathcal{O}}_i$ sending tokens to its parents.[15] Node $\mathbf{x}$, upon receiving a token from any of its children proceeds as follows: (i) if $\mathbf{x}$ is *not* intervenable (i.e., $\mathbf{x} \notin V_i$), upon receipt of the first token, it propagates the token to all its parents, and simply ignores all subsequent tokens received from any of its children.[16] (ii) if $\mathbf{x}$ is intervenable (i.e., $\mathbf{x} \in V_i$) then it absorbs the token and does not communicate any tokens to their parents. Upon termination of the distributed algorithm, the set of all absorbing nodes, together, form the set $\boldsymbol{\mathcal{X}}^*$. Time-complexity of the above distributed algorithm is $O(l_d)$ where $l_d$ denotes the length of the longest directed path in the underlying DAG.

### Bounded Rationality and $\mathcal{C}^*$

In this section, we elaborate on how $\mathcal{C}^*$ should be construed within the context of bounded rationality (Simon, 1957). The reasoner—inevitably bounded in time and computational resources—executes $\mathcal{C}^*$ within the time frame available to her. The set of nodes/variables that, within the available time frame, she gets the chance to identify should be understood as a boundedly-rational approximation to $\boldsymbol{\mathcal{X}}^*$.

## 4.10 Related Work and Conclusion

Finally, we give an overview of the ideas explored in the literature which are, in spirit, related to the problem under study in this work. The idea of Structure Control Theory (SCT) proposed by Lin (1974) in the context of Linear Time-Invariant (LTI) systems governed by first-order differential equations (a.k.a. state equations) perhaps comes closest to our problem. In such domains, all variables are deterministic and the states of variables change in time according to the dynamics represented by state equations.

Liu et al. (2011), drawing on the idea of SCT proposed by Lin (1974), aimed at identifying the minimal set of variables which are sufficient for the purpose of structural controllability of a generic large-scale LTI system.[17] In a subsequent work, Gao et al. (2014), relaxed the

[15] Variables in $\bar{\boldsymbol{\mathcal{O}}}_i$ absorb their tokens immediately and do not send any tokens to their parents.

[16] In this light, informally speaking, a non-intervenable variable $\mathbf{x}$ remains 'invisible' to the flow of tokens.

[17] Note that, intervening on $V_i$ (according to IP *class*-$\infty$) is the trivial solution to the TPS-controllability problem. Similar, in spirit, to the contribution of (Liu et al., 2011), using the sole topology of $G$, the proposed

objective of structural controllability of the system in whole, to merely that of a particular set of desired variables called target variables. This line of thought was motivated by the understanding that in large-scale systems, it may neither be attainable nor required to control the full system but, rather, to merely control a subset of the variables of the system (analogous to target variables in our problem) which are deemed pivotal for the realization of the task at hand. In that light, Gao et al. (2014) were concerned with the same question underlying our work, yet, perused it in a radically different setting. In (Liu et al., 2011; Gao et al., 2014), both variables and their inter-connections are deterministic in nature whereas, in our case, both have probabilistic natures, a point of departure which leads to a substantially different line of work—both semantically and syntactically.

It is crucial to note that the line of work investigated in this chapter does not lend itself to influence diagram (ID) modeling, for the following important feature of IDs: In IDs, the set of nodes on which intervention should be exercised is *pre-specified* at the outset, leaving no room for the agent to decide upon which of intervenable nodes intervention should be exercised. A defining property of TPS-controllability problem is that the agent gets to decide, out of $2^{|V_i|}$ possible choices, on which subset of intervenable nodes intervention should be applied (recall (Q1) from Sec. 4.1, as a defining feature of probabilistic controllability problem). Also, it is easy to show that the very choice of intervenable nodes for intervention modulates the IP class which is optimal for exercising on them. These features make the line of work pursued in this chapter radically distinct from, and fall outside the scope of ID formalism.

To conclude, the problem of TPS-controllability in the context of CBNs was introduced and formalized in this work. A linear-time algorithm, $\mathcal{C}^*$, was devised to identify a sufficient set of intervenable variables for the purpose of TPS-controllability of a generic CBN; the minimality of $\mathcal{C}^*$'s output was also characterized. We also elaborated on the computational complexity of the task and presented a lemma as well as a conjecture on reducing the scope of IPs. The implications of this work for psychology were also extensively investigated. Concretely, we highlighted the implications of our results for a line of work in developmental psychology concerning causal learning by young children in pedagogical settings, and established, for the first time in the literature, a rational, algorithmic-level account of a curious behavior demonstrated by young children called overimitation, generally taken as evidence for children's irrationality. The provided results are also of importance to studies on strate-

linear-time algorithm $\mathcal{C}^*$ outputs—among $2^{|V_i|}$ potential subsets—a *non-trivial* solution to the problem, $\mathcal{X}^*$, which is both minimal (see Sec. 4.7) and optimal with respect to the objectives given in (4.1) and (4.2).

gic planning and policy making concerned with efficient practice of intervention in order to maximize (minimize) the odds of desired (undesired) outcomes.

# Part III: Conditional Independence, $d$-Separation, and Minimality

**Preface.** It was Judea Pearl's great moment of insight when he recognized the fundamental role of (conditional) probabilistic independence in how human's probabilistic/causal knowledge is structured. Baffled with how humans oftentimes judge probabilistic independence with "clarity, conviction and consistency" in spite of their incompetence in giving estimates of the probabilities involved, Pearl put forth an elegant graph-theoretic notion, called $d$-separation, for the purpose of judging independencies by merely entertaining the topology of the BN representing how the knowledge of the reasoner is structured. Motivated by human cognition and particularly the distributed machinery of the brain, in his seminal work in 1986, Pearl put forth an asynchronous, distributed, message-passing algorithm called Belief Propagation (BP) to address how inference could be efficiently carried out on BNs.[1] Despite recognizing the fundamental role that the notion of $d$-separation plays in human cognition, Pearl left one question unanswered: How could $d$-separation be implemented in an asynchronous, distributed, message-passing manner akin, in spirit, to the BP scheme? Part III puts forward such an algorithm for implementing $d$-separation. Furthermore, through the introduction of a key graph-theoretic notion, termed minimal-refutation module, Part III shows how the notion of minimality manifest itself in a distributed, message-passing implementation of $d$-separation.

[1] BP went on to become not only an extremely prominent inference algorithm for PGMs, but also it turned out to be unreasonably successful for a series of problems which, on the face of it, had nothing to do with the original motivation behind proposing BP, namely, decoding the low-density parity-check (LDPC) codes as well as Turbo codes, and solving the Boolean satisfiability problem.

# Chapter 5

# Asynchronous, Distributed Algorithm for $d$-Separation: Towards Cognitively Plausible Implementations

## 5.1 Introduction

In his 1986 paper, Pearl put forth a graph-theoretic notion called $d$-separation, allowing for reading off probabilistic independence relations from the mere structure of a Bayesian network (BN) (Pearl, 1986).[1] Ever since its inception, $d$-separation has proved fundamental in a variety of domains, e.g., probabilistic reasoning (Pearl, 1988), causal reasoning (Pearl, 2000), decision making (Shachter, 1998; Koller and Friedman, 2009), and has played important roles in a broad range of areas, e.g., handling missing data (Mohan and Pearl, 2014), extrapolation across populations (Pearl and Bareinboim, 2014), and deep learning (Goodfellow et al., 2016).

In this chapter, we put forth the Three-Color Algorithm, denoted by $\mathcal{D}^*$, which permits implementing Pearl's $d$-separation criterion. That is, $\mathcal{D}^*$ allows to decide, based on the sole topology of a BN, if an arbitrary conditional independence relation holds in that BN. The algorithm $\mathcal{D}^*$ is distributed and asynchronous, and outperforms previously proposed algo-

[1]The edges of a BN may have causal semantics, in which case the formalism of causal BN should be invoked. The graph-theoretic notion of $d$-separation remains valid regardless of whether the edges enjoy causal interpretations or not.

rithms for implementing $d$-separation in terms of worst-case running time. We provide a comprehensive analysis of the computational properties of $\mathcal{D}^*$, along with several refined time-complexity bounds. A detailed comparison between $\mathcal{D}^*$ and previously proposed algorithms is provided in the Discussion section, later in this chapter. We will also elaborate on the implications of the work presented in this chapter for neuroscience and psychology in Sec. 5.6.

## 5.2 Preliminaries and Notations

Let us introduce the notation adopted in this chapter. Lower bold-faced letters (e.g., $\mathbf{x}$) denote random variables and upper bold-faced letters (e.g., $\mathbf{X}$) represent sets of random variables. A generic $d$-separation relation is denoted by $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ with $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ representing three mutually disjoint sets of variables belonging to the DAG $G$ where $G$ represents the topology of the underlying BN. Read $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ as follows: $\mathbf{C}$ $d$-separates $\mathbf{A}$ from $\mathbf{B}$ in DAG $G$. Similarly, $(\mathbf{A} \not\perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ denotes that $\mathbf{C}$ does not $d$-separate $\mathbf{A}$ from $\mathbf{B}$ in DAG $G$. For ease of notation, we use $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ to denote both a $d$-separation relation (i.e., $\mathbf{C}$ $d$-separates $\mathbf{A}$ from $\mathbf{B}$ in DAG $G$) and to denote a $d$-separation query (i.e., does $\mathbf{C}$ $d$-separate $\mathbf{A}$ from $\mathbf{B}$ in DAG $G$?); the distinction should be clear from the context. Let also $G_{An(\mathbf{K})}$ denote the ancestral graph for the variables in set $\mathbf{K}$ belonging to the underlying DAG $G$ (Lauritzen et al., 1990), i.e., the set of nodes for $G_{An(\mathbf{K})}$ comprises the nodes in $\mathbf{K}$ and all the ancestors of the nodes in $\mathbf{K}$ (hence, $G_{An(\mathbf{K})}$ is a subgraph of the underlying DAG $G$).

Next, a notion called refutation-module is introduced; this will be used later in our formal analysis of the propose algorithm.

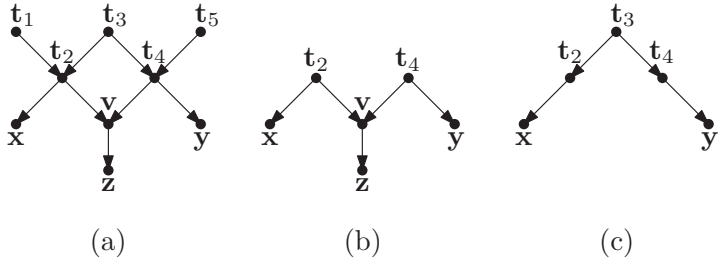



**Fig. 5.1** Examples for refutation modules. (a) The underlying DAG $G$ is depicted, for which $(\mathbf{x} \not\perp\!\!\!\perp \mathbf{y}|\mathbf{z})_G$. (b,c) Two refutation-modules for the $d$-separation query $(\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z})_G$ are depicted.

**Definition 5.1. (Refutation-Module)** Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three mutually disjoint sets belonging to a DAG $G$. Let also $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$. A connected subgraph of $G$, $\mathcal{M}_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$, serves as a refutation-module for the $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, iff $\mathcal{M}_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ satisfies the following two conditions: (1) $\mathcal{M}_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ contains an active path $P$ between a node $\mathbf{x} \in \mathbf{X}$ and a node $\mathbf{y} \in \mathbf{Y}$, and (2) for every head-to-head node $\mathbf{v}$ on $P$, $\mathcal{M}_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ contains a directed path between $\mathbf{v}$ and a node $\mathbf{c} \in \mathbf{C}$. See Fig. 5.1 for examples on refutation-module.

**Lemma 1.** *Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be three mutually disjoint sets belonging to a DAG $G$. Let also $(\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$. The following two statements hold: (1) $G$ must contain at least one refutation-module for the d-separation query $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$, (2) $G$ may contain more than one refutation-module for the d-separation query $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$, hence the non-uniqueness of the refutation-module.*

***Proof.*** Claim (1) immediately follows from the definition of $d$-separation; see (Pearl, 1986). Claim (2) follows from the examples depicted in Fig. 5.1.

**Definition 5.2. (Minimal Refutation-Module)** Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ be three disjoint sets of nodes belonging to a DAG $G$. Also, let $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$. Let $\mathcal{M}^*_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ denote the refutation-module for the $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$ which possesses the smallest number of edges. We refer to $\mathcal{M}^*_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ as the *minimal* refutation-module in $G$ for the $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$.

It is easy to prove by construction that the minimal refutation-module $\mathcal{M}^*_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ need not be unique.

## 5.3 The Three-Color Algorithm $\mathcal{D}^*$

In this section, we show how the proposed algorithm $\mathcal{D}^*$ allows to decide if a generic $d$-separation query of the form $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ holds in a DAG $G$; $\mathcal{D}^*$ is an asynchronous, distributed, message-passing algorithm. More specifically, in $\mathcal{D}^*$, nodes of the underlying DAG $G$—symbolizing computational units—autonomously engage in communicating messages to their immediate neighbors via the edges of the DAG $G$—symbolizing communication channels. We assume that communication channels are reliable, bidirectional, and first-in first-out (FIFO) (Lynch, 1996).

The proposed algorithm $\mathcal{D}^*$ is outlined next. Throughout an execution of $\mathcal{D}^*$, variables in $\mathbf{C}$ ignore any message received from any of their children, and also do not send any message to any of their children. The variables in the sets $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ initially activate in the states

represented by colors (●), red (●), and white (○), respectively. Following the prescriptions of the original Belief Propagation algorithm (Pearl, 1986, Sections 1.3 and 2.2.3), we assume that the variables in the sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$ acquire their initial states in a *self-activated* manner.[2] $\mathcal{D}^*$ begins with nodes in $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ sending their colors as messages to their parents.[3] Node $\mathbf{x}$, upon receiving a message, follows two simple steps in the following order:

(i) If $\mathbf{x}$'s current color differs from that of the received message, $\mathbf{x}$ replies by sending back its own color as a message to the transmitter node. If $\mathbf{x}$ is in the state of having no color (denoted by $\varnothing$) prior to the receipt of the message, it does not send back any message to the transmitter node.

(ii) $\mathbf{x}$ updates its color in accord with the following primitive rules, altogether composing the Color Update Grammar (CUG):

$$(\varnothing, \bullet) \to \bullet, (\varnothing, \bullet) \to \bullet, (\varnothing, \circ) \to \circ,$$
$$(\bullet, \bullet) \to \bullet, (\bullet, \bullet) \to \bullet, (\circ, \circ) \to \circ,$$
$$(\circ, \bullet) \to \bullet, (\circ, \bullet) \to \bullet,$$
$$(\bullet, \circ) \to \bullet, (\bullet, \circ) \to \bullet,$$
$$(\bullet, \bullet) \to \text{clash}, (\bullet, \bullet) \to \text{clash},$$

where the syntax is: ($\mathbf{x}$'s current color, received message) $\to$ $\mathbf{x}$'s new color. If $\mathbf{x}$'s new color turns out to be different from its old color, with the exception of the transmitter node in (i), $\mathbf{x}$ sends its new color as a message to (a) all its parents, and (b) only those children of $\mathbf{x}$ with which $\mathbf{x}$ has communicated before.

The rules given in the first row of the CUG correspond to white-, green-, and red-colored nodes sending their colors to their yet-unvisited parents. Rules in the second row ensure that the colors of white-, green-, and red-colored nodes persist upon interacting with nodes of the same color. Rules stated in the third row bear on the key understanding that the white color functions as a mere place-holder getting "replaced" by interacting with green-,

[2] Alternatively, we provide an asynchronous, distributed, $O(l)$-time message-passing algorithm in Sec. S-VI of Appendix C, which permits a predesignated source node to disseminate information regarding the initial states of the nodes through the graph $G$, where $l$ denotes the length of the longest undirected path in $G$.

[3] This very act will initiate the propagation of colors in a backwards manner throughout the network. This becomes clearer when we present an example in Sec. 5.4.

or red-colored nodes. Rules in the fourth row, guarantee the persistence of colors green and red upon interacting with white. Finally, rules given in the last row correspond to the clash event the implication of which is discussed in Remark 1.

**Remark 1:** A clash between colors green (●) and red (●) at a node, anytime throughout an execution of $\mathcal{D}^*$, signals the falsity of the input $d$-separation query, upon which $\mathcal{D}^*$ decides that $(\mathbf{A} \not\perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$. Note that "clash" is a *termination state* for a node.

Note that asynchrony of $\mathcal{D}^*$ stems from the fact that there exists no *global clock* for the system and hence any node, upon receiving a message, follows Steps (i) and (ii) *autonomously*, i.e., informally, without having to attend to what computations other nodes in $G$ are performing.[4]

Some of the computational properties of the proposed algorithm $\mathcal{D}^*$ are formally articulated in Proposition 5.1 below.

**Proposition 5.1.** *The following statements hold for $\mathcal{D}^*$.*

*(1) For a given d-separation query $(\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C})_G$ and DAG $G$,*

*"$\boldsymbol{C}$ does not d-separate $\boldsymbol{A}$ from $\boldsymbol{B}$ in $G$" $\iff$ "Clash takes place during $\mathcal{D}^*$'s execution".*

*(2) $\mathcal{D}^*$'s message-passing is confined within the ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$.*

*(3) During $\mathcal{D}^*$'s execution, either a clash between colors red (●) and green (●) takes place (see Remark 1) upon which $\mathcal{D}^*$ decides that $(\boldsymbol{A} \not\perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C})$, or a state of equilibrium will be reached in $O(l_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})})$ time where $l_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ denotes the length of the longest undirected path in the ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$.*

*(4) Message-passing terminates in $O(1)$ time after reaching the state of equilibrium, thereby guaranteeing the termination of $\mathcal{D}^*$.*

*(5) Message-complexity of $\mathcal{D}^*$ is $O(|E_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}|)$ where $E_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ is the set of the edges of the ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$.*

*(6) Communication-complexity of $\mathcal{D}^*$ is $O(|E_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}|)$ bits where $E_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ is the set of the edges of the ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$.*

The reader is referred to Sec. C-VI of Appendix C for the proof of Proposition 5.1.

[4]That is, using the formalism of *state transition systems*, without having to attend to the states other nodes in the distributed system are in.

### 5.3.1 High-Level Understanding of $\mathcal{D}^*$

$\mathcal{D}^*$ has a simple machinery as we informally discuss here. Upon variables in $\mathbf{A}\cup\mathbf{B}\cup\mathbf{C}$ sending their colors to their parents, colors white ($\circ$), green ($\bullet$), and red ($\bullet$) begin to propagate in a *backwards* manner throughout the network. In the midst of this process, white-color nodes which have a node colored either red ($\bullet$) or green ($\bullet$) in their neighborhood, change their color to that of their neighbors, and if ever a clash occurs between colors red and green, $\mathcal{D}^*$ decides that the input $d$-separation query is false (i.e., a NO-instance $d$-separation query). The proof of correctness for $\mathcal{D}^*$ is presented in Sec. S-I of Appendix C.

### 5.3.2 A Note On The Termination of $\mathcal{D}^*$

According to Proposition 5.1, if the input $d$-separation query presented to $\mathcal{D}^*$ is true (i.e., a YES-instance $d$-separation query), the system reaches a state of equilibrium in $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ time and message-passing is guaranteed to terminate in $O(1)$ time after that. However, due to its local view, a node cannot know if such a global state has been reached. This is a fairly standard situation for an asynchronous distributed algorithm to find itself in (Mattern, 1987; Tel, 2000), leading to the introduction of the fundamental concept of Termination-Detection (DT) in distributed systems literature; see (Tel, 2000, Ch. 8). There exist a variety of DT-algorithms in the literature (e.g., Dijkstra et al., 1983; Mattern, 1987; Mittal et al., 2004, 2007).[5]

## 5.4 $\mathcal{D}^*$ in Action: A Case Study

In this section, we present an example to illustrate an execution and highlight the simplicity of $\mathcal{D}^*$. Let us consider the BN depicted in Fig. 5.2(a). Let the posed $d$-separation query be $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$ where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2\}$, and $\mathbf{Z} = \{\mathbf{z}\}$. According to $d$-separation criterion (Pearl, 1988), observation of $\mathbf{z}$ activates the path $\mathbf{x}_1 \leftarrow \mathbf{t}_1 \leftarrow \mathbf{t}_2 \leftarrow \mathbf{t}_3 \rightarrow \mathbf{t}_4 \leftarrow \mathbf{t}_5 \rightarrow \mathbf{t}_6 \rightarrow \mathbf{t}_7 \rightarrow \mathbf{y}_1$, thereby yielding the falsity of the $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$ (hence, the input is a NO-instance query); see Fig. 5.2(a). An execution of $\mathcal{D}^*$ is illustrated using successive *snapshots* shown in Figs. 5.2(b-f) with each figure depicting the global state of

[5]For example, Mittal et al. (2004) propose two DT-algorithms, each having detection latency of $O(D)$ where $D$ is the diameter of the underlying graph $G$, and $G$ is allowed to have an arbitrary topology.

the system (i.e., nodes' colors) at some instance in global time (aka system's *configuration*).[6] As depicted in Fig. 5.2(b), variables in sets $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ initially self-activate in the states represented by colors green (●), red (●), and white (○), respectively. Also recall that, as explicated in Sec. 5.3, variables in $\mathbf{Z}$ ignore any message received from any of their children, and also do not send any message to any of their children—depicting the downlinks of the variables in $\mathbf{Z}$ in a dash-dotted format simply illustrates this statement pictorially in Fig. 5.2(b). The colors green (●), red (●), and white (○) propagate in a backwards manner (Figs. 5.2(c-d)) and white gets replaced by green or red once it becomes a neighbor of a node with such colors (Figs. 5.2(d-f)). Eventually, in the configuration depicted in Fig. 5.2(f), a clash takes place between colors green and red at a node (circled node in Fig. 5.2(f)), upon which $\mathcal{D}^*$ decides that $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$.[7]

Notice that, since $\mathbf{w}$ is unobserved (Fig. 5.2(a)), the path $\mathbf{x}_2 \to \mathbf{w} \leftarrow \mathbf{y}_2$ indeed remains blocked; this is nicely captured by the machinery of $\mathcal{D}^*$. Algorithm $\mathcal{D}^*$ prevents $\mathbf{x}_2$ and $\mathbf{y}_2$ from sending their colors in the forward direction (i.e., along the edges pointing to $\mathbf{w}$), thereby guaranteeing the occurrence of no clash along the blocked path $\mathbf{x}_2 \to \mathbf{w} \leftarrow \mathbf{y}_2$. Also notice that, since $\mathbf{z}$ is observed (Fig. 5.2(a)), the path $\mathbf{x}_2 \leftarrow \mathbf{z} \to \mathbf{y}_2$ is blocked as well. Once again the machinery of $\mathcal{D}^*$, due to $\mathbf{z}$ refraining from engaging in message-exchange with its children, ensures that no clash takes place due to the blocked path $\mathbf{x}_2 \leftarrow \mathbf{z} \to \mathbf{y}_2$.

## 5.5 Discussion

A number of algorithms for the implementation of $d$-separation are proposed in the literature; see (Geiger et al., 1989; Lauritzen et al., 1990; Shachter, 1998; Koller and Friedman, 2009; Butz et al., 2016). Assuming $|E| \geq |V|$, to decide if $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ holds in $G$, the worst-case running time of Geiger et al.'s, Koller and Friedman's, Shachter's, and Butz et al.'s is $O(|E|)$ and that of Lauritzen et al.'s algorithm[8] is $O(|V|^2)$ where $|V|$ and $|E|$ denote the number of the nodes and the edges of the underling DAG $G$, respectively. Note that, since for any DAG

[6] Cast into Lamport's *space-time diagram* (Lamport, 1978), each figure depicts the global state of the system which corresponds to a vertical *time-cut* positioned at a global time (see (Mattern, 1987)) and the time-cuts corresponding to Figs. 5.2(b-f) are successively ordered.

[7] The order according to which the circled node receives the incoming red- and green-colored massages in the configuration depicted in Fig. 5.2(f) is irrelevant; either way a clash will take place. This property can be formalized in a wider sense under the notion of *order-invariance* which will be discussed in Sec. 5.5.

[8] The reader is referred to (Geiger et al., 1989) for a detailed analysis of the running-time of Lauritzen et al.'s algorithm.

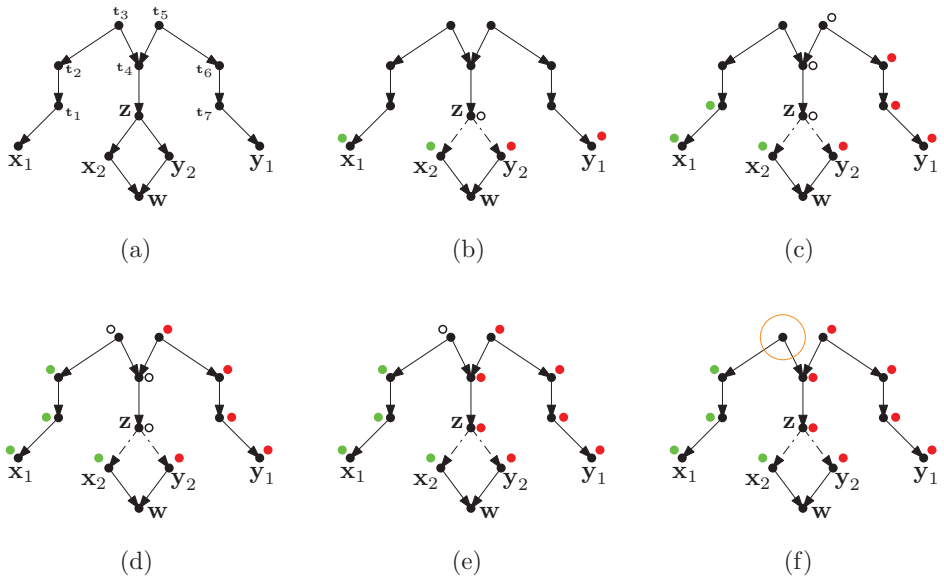



**Fig. 5.2** Illustrative example. The underlying DAG $G$ is shown in (a). The initial configuration of the system is portrayed in (b), wherein variables in sets $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ self-activate in the states represented by colored green (●), red (●), and white (○), respectively. Depicting the downlinks of the variables in $\mathbf{Z}$ in a dash-dotted format simply symbolizes that the variables in $\mathbf{Z}$ ignore any message received from any of their children, and also do not send any message to any of their children. $\mathcal{D}^*$ begins by nodes in $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ sending their colors as messages to their parents and proceeds as shown in (c-f) with each figure depicting a snapshot of the global state of the system (i.e., nodes' colors) at some instance in global time. Eventually, upon occurrence of a clash between colors green and red (at the circled node in (f)), $\mathcal{D}^*$ decides that $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$.

$G$, $|E| \leq |V|^2$, an $O(|E|)$-time algorithm (e.g., Geiger et al.'s) outperforms an $O(|V|^2)$-time algorithm (e.g., Lauritzen et al.'s) in terms of worst-case runtime[9] (see Geiger et al., 1989, for more discussions on this). According to Proposition 5.1, the time-complexity of the proposed algorithm $\mathcal{D}^*$ is $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ where $l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$ denotes the length of the longest undirected path in the ancestral graph $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$.[10] Since, for any DAG $G$, $l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} \leq |E| \leq |V|^2$, the proposed algorithm $\mathcal{D}^*$ outperforms all the previously proposed algorithms in terms of the worst-case running time.[11] Particularly, the gain is significant in dense DAGs. Note that,

[9]The gain in particularly significant in sparse graphs, where $|E| = O(|V|)$.

[10]The reader is referred to Sec. C-II of Appendix C where the time-complexity analysis of $\mathcal{D}^*$ is presented.

[11]According to Proposition 5.1, a NO-instance $d$-separation query can be decided by $\mathcal{D}^*$ in time $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$; see also Lemma A.1 in Sec. A-I of the Appendix. The upper-bound $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ is an improvement over the worst-case runtime of all the previously proposed algorithms. Also note that, adopting a DT-algorithm with detection latency of $O(D)$ (see (Mittal et al., 2004, 2007) for such DT-algorithms), a

in the limit as the underlying DAG $G$ gets denser, the worst-case runtime performances of the previously proposed algorithms become identical, i.e., $O(|V|^2)$.

The proposed algorithm $\mathcal{D}^*$ restricts its exploration solely in the ancestral graph $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, as formalized by Statement (2) of Proposition 5.1. The idea of exploring the ancestral graph is at the core of Lauritzen et al.'s algorithm for $d$-separation (Lauritzen et al., 1990). However, in sharp contrast to Lauritzen et al.'s algorithm, the proposed algorithm $\mathcal{D}^*$ need not *moralize* the ancestral graph $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$. As Geiger et al. (1989) point out, the moralization step of Lauritzen et al.'s algorithm requires $O(|V|^2)$ time in the worst-case.

Another noteworthy property of $\mathcal{D}^*$ is that it is tailored towards quick detection of false $d$-separation queries (i.e., NO-instance queries), manifested in an occurrence of a clash according to Remark 1. For a NO-instance $d$-separation query, Proposition 5.2, below, gives a more refined upper-bound on the time required for an occurrence of a clash, thereby formalizing the said claim. The reader is referred to Sec. C-III of Appendix C for the proof of Proposition 5.2.

**Proposition 5.2.** *Let $\boldsymbol{A} = \{\boldsymbol{a}_i\}_i$, $\boldsymbol{B} = \{\boldsymbol{b}_j\}_j$, $\boldsymbol{C} = \{\boldsymbol{c}_k\}_k$ be three disjoint sets of nodes belonging to a DAG $G$. Let $l^d_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ denote the length of the longest directed path in the ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$, and $l^{ij}_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ the length of the shortest unblocked path between the nodes $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ in $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$. As a convention, if all paths between $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ are blocked, $l^{ij}_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})} = \infty$. If $(\boldsymbol{A} \not\perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C})_G$ (hence, a NO-instance d-separation query) then a clash between colors green (●) and red (●) occurs in time $O\big(l^d_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})} + \min_{i,j} l^{ij}_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}\big)$, upon which $\mathcal{D}^*$ decides that $(\boldsymbol{A} \not\perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C})_G$.*

In Sec. 5.2, we formally defined a notion called refutation-module (see Definition 5.1). In the language of computational complexity and theorem-proving, a refutation-module $\mathcal{M}_{(\mathbf{X}\not\perp\!\!\!\perp\mathbf{Y}|\mathbf{Z})_G}$ can serve as a *certificate* (or *witness*) for disproving a $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$. This interpretation is related to the verifier-based definition of the complexity class *coNP*. Next, in Proposition 5.3, we provide an even more refined upper-bound on the time required for an occurrence of a clash, thereby strengthening our claim as to $\mathcal{D}^*$ being tailored toward quick detection of false $d$-separation queries. The reader is referred to Sec. C-IV of Appendix C for the proof of Proposition 5.3.

**Proposition 5.3.** *Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be three disjoint sets of nodes belonging to a DAG $G$. Also, let $(\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$ (hence, a NO-instance d-separation query). Let $\mathcal{M}_{(\boldsymbol{X}\not\perp\!\!\!\perp\boldsymbol{Y}|\boldsymbol{Z})_G}$*

YES-instance $d$-separation query can be decided by $\mathcal{D}^*$ in time $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + D)$ where $D$ is the diameter of $G$. Once again, since $l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} \leq |E|, D \leq |E|, |E| \leq |V|^2$, the upper-bound $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + D)$ is an improvement over the worst-case runtime of all the previously proposed algorithms. (Notice that, for any DAG $G$, $\frac{1}{2}(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + D) \leq |E|$, hence follows $|E| = \Omega(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + D)$.)

*denote a refutation-module for the query* $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$ *with* $l^d_{\mathcal{M}}$ *and* $|P_{\mathcal{M}}|$ *denoting the length of the longest directed path and the shortest unblocked path in* $\mathcal{M}_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}$*, respectively. Finally, let* $\mathcal{M}^*_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}$ *denote the minimal refutation-module for the query* $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$*, with* $E_{\mathcal{M}^*_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}}$ *denoting the set of the edges of* $\mathcal{M}^*_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}$*. Then the following statement holds true: A clash between colors green (●) and red (●) occurs in time* $O(\min_{\mathcal{M}_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}}\{l^d_{\mathcal{M}} + |P_{\mathcal{M}}|\}) \leq O(|E_{\mathcal{M}^*_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}}|)$*, upon which* $\mathcal{D}^*$ *decides that* $(\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$*.*

To further highlight the significance of Proposition 5.3, let us consider the following *nondeterministic* algorithm $\mathcal{A}$. Algorithm $\mathcal{A}$ takes as input a DAG $G$ along with a $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, and outputs YES or NO depending on whether the input query is a YES-instance or a NO-instance query, respectively.

(i) Nondeterministically guess (1) the minimal refutation-module $\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}$ in $G$ for the $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$ (by definition, $\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}$ contains an active path, $P^*$, between a node $\mathbf{x}^* \in \mathbf{X}$ and a node $\mathbf{y}^* \in \mathbf{Y}$, and also contains a set of observed variables $\mathbf{Z}^* \subseteq \mathbf{Z}$)[12], and (2) the corresponding nodes $\mathbf{x}^*, \mathbf{y}^*, \mathbf{Z}^*$ belonging to $\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}$.

(ii) Verify that (1) $\mathbf{x}^* \in \mathbf{X}$, $\mathbf{y}^* \in \mathbf{Y}$, and $\mathbf{Z}^* \subseteq \mathbf{Z}$ (this can be straightforwardly verified in $O(|\mathbf{X}|+|\mathbf{Y}|+|\mathbf{Z}^*||\mathbf{Z}|)$ time), (2) $\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}$ is a subgraph of $G$ (this can be straightforwardly verified in $O(|E_{\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}}|)$ time), and (3) $d$-separation relation $(\mathbf{x}^* \perp\!\!\!\perp \mathbf{y}^*|\mathbf{Z}^*)$ does not hold in DAG $\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}$ (this can be verified in $O(|E_{\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}}|)$ time, using Geiger et al.'s algorithm (Geiger et al., 1989)). If all the verification steps (1)-(3) pass, output NO; otherwise, output YES.

Altogether, presented with a NO-instance $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, algorithm $\mathcal{A}$ outputs NO in $O(|E_{\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}}| + |\mathbf{X}| + |\mathbf{Y}| + |\mathbf{Z}^*||\mathbf{Z}|)$ nondeterministic time. Interestingly according to Proposition 5.3, presented with a NO-instance $d$-separation query $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, the machinery of $\mathcal{D}^*$ ensures that a clash between colors green (●) and red (●) occurs within at most $O(|E_{\mathcal{M}^*_{(\mathbf{X} \not\perp \mathbf{Y}|\mathbf{Z})_G}}|)$ time, upon which $\mathcal{D}^*$ decides that $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$. It is crucial to note that the presented argument solely concerns NO-instance $d$-separation queries.

Proposition 5.4, given below, further strengthens the claim of Proposition 5.3. The reader is referred to Sec. C-V of Appendix C for the proof of Proposition 5.4.

**Proposition 5.4.** *The upper-bound* $O(\min_{\mathcal{M}_{(\boldsymbol{X} \not\perp \boldsymbol{Y}|\boldsymbol{Z})_G}}\{l^d_{\mathcal{M}}+|P_{\mathcal{M}}|\})$ *given in Proposition 5.3 is tighter than the one given in Proposition 5.2.*

[12]Note that if $P^*$ does not contain any head-to-head node, then $\mathbf{Z}^* = \varnothing$

Finally, we would like to point out an interesting property of the CUG, referred to as *order-invariance*, which is characterized informally as follows: *The order according to which nodes in the network receive their messages is irrelevant.* More formally, the order-invariance property can be stated as follows: Assume that a node $\mathbf{x}$ is at state $S_i^{\mathbf{x}}$ and upon receiving the sequence of messages $M_1, M_2, \cdots, M_n$ ends up in state $S_f^{\mathbf{x}}$. Then the following holds true for the node $\mathbf{x}$. For any permutation $\pi$ defined on the set $\{1, 2, \cdots, n\}$, the node $\mathbf{x}$, starting at state $S_i^{\mathbf{x}}$, would end up in the state $S_f^{\mathbf{x}}$ upon receiving the sequence of messages $M_{\pi(1)}, M_{\pi(2)}, \cdots, M_{\pi(n)}$. The reader is referred to Sec. C-VIII of Appendix C for a formal treatment of the order-invariance property and its proof.

## 5.6 On the Implications for Psychology and Neuroscience

It is inconceivable how chaotic the world would seem to humans, faced with innumerable decisions a day to be made under uncertainty, had they been lacking the very capacity to distinguish the relevant from the irrelevant—a capacity which, computationally, amounts to handling probabilistic independence relations efficiently. As Pearl (1986) elegantly puts it: "Whereas a person may show reluctance to giving a numerical estimate for a conditional probability $P(\mathbf{x}_i|\mathbf{x}_j)$, that person can usually state with ease whether $\mathbf{x}_i$ and $\mathbf{x}_j$ are dependent or independent, namely, whether or not knowing the truth of $\mathbf{x}_i$ will alter the belief in $\mathbf{x}_j$." He then continues: "Likewise, people tend to judge the three-place relationships of conditional dependency (i.e., $\mathbf{x}_i$ influences $\mathbf{x}_j$ given $\mathbf{x}_k$) with clarity, conviction, and consistency. This suggests that the notions of dependence and conditional dependence are more basic to human reasoning than are the numerical values attached to probability judgments." Some literature in cognitive psychology, however, does not fully embrace the statement "with clarity, conviction, and consistency" as Pearl put it. For example, the experimental work done by Rehder (2014) suggests that adults exhibit deviations from the Markov condition. In contrast, drawing on the experimental studies of Park and Sloman (2013), Sloman and Lagnado (2015) conclude that people indeed uphold the Markov condition and the reason behind the observed deviations is that, under experimental conditions, people may not solely adhere to the information provided by the experimenter and may bring their own background knowledge into the experiment (see also Rehder and Waldmann, 2017). Specifically, Park and Sloman (2013) found strong support for their contradiction hypothesis followed by the mediating mechanism hypothesis, and finally concluded that people do conform to Markov

condition once the causal structure people are using is correctly specified (i.e., people's mental causal models).

All the algorithms proposed in the literature for the implementation of $d$-separation (Geiger et al., 1989; Lauritzen et al., 1990; Shachter, 1998; Koller and Friedman, 2009; Butz et al., 2016) have been so far sequential, i.e., coordinated and executed by a supervisory unit, without any concurrency or parallelism in computation—a characteristic which undermines their cognitive plausibility. The proposed algorithm $\mathcal{D}^*$ permits the implementation of $d$-separation in an asynchronous, distributed, message-passing fashion—a property consistent with the brain's computational machinery (see McClelland, 1989; Chater et al., 2006) and fully in the spirit of the celebrated parallel-distributed-processing (PDP) research program in brain and cognitive sciences.

The Algorithm $\mathcal{D}^*$, in the spirit of Pearl's Belief Propagation scheme, employs the edges of the underlying BN as the medium through which message-passing between nodes takes place. The latter echos Pearl's insight (1986) when he advocated the idea that a BN must not be viewed "merely as a passive parsimonious code for storing factual knowledge but also as a computational architecture for reasoning about that knowledge." $\mathcal{D}^*$ adheres to this idea. Recent literature in neuroscience investigating possible implementation of BNs at the neural level supports Pearl's idea (see Lochmann and Deneve, 2011; Gershman and Beck, 2017). Lochmann and Deneve (2011) advocate the idea that a BN's node can be represented at the neural level by a single (Deneve, 2008a,b) or a population of neurons (Ma et al., 2006) with the neural network resembling a "mirror image" of the BN it implements—though sometimes not a 'perfect' mirror—and the links of the neural network providing the medium for inference to be carried out—either in a form of Belief Propagation or Sample-based methods like Gibbs sampling.[13]

Interestingly, the peculiar tendency of $\mathcal{D}^*$ toward quick detection of NO-instance $d$-separation queries is consistent with our pre-theoretical intuition that humans tend to detect possible dependencies between concepts and propositions rather swiftly, once such dependencies do exist. The following two questions then immediately present themselves: (Q1) Why should people have such a tendency in the first place? (Q2) Could this tendency be supported based on any rational grounds? Despite its form, (Q1) may not merit any real answer: Assuming that the mind is implementing $d$-separation by a process akin to $\mathcal{D}^*$, the existence

[13]For more on how probability distributions can be encoded at the neural level, the reader is referred to Lochmann and Deneve (2011).

of the said tendency is merely a logical implication of the very process by which the mind is implementing $d$-separation (i.e., it emerges out of the very machinery of the underlying psychological processes at work).[14] In that light, $\mathcal{D}^*$ can be viewed as a rational, process-level account of the said tendency. Question (Q2), however, is more subtle and indeed demands a real answer. In what follows we provide two arguments supporting the rationality of the foregoing tendency. The first argument is based on the assumption that propositions/concepts are dependent more often than not, i.e., the majority of the $d$-separation instances that the mind encounters are NO-instances—an assumption which a priori appears to be neither plausible nor implausible, and therefore requires empirical investigations. It then follows that the foregoing tendency is simply a consequence of the mind acting as a rational *optimizer*, trying to attain good performance in terms of *expected* runtime (i.e., average-case analysis). The second argument relies on the assumption that, abstractly speaking, the mind incurs a higher rate of loss (defined as incurred cost per unit of time) for discovering a dependency when one does exist, compared to the condition wherein one does not exist and the mind recognizes that. Once again, the aforesaid tendency simply follows from the mind acting as a rational optimizer, attempting to minimize the total cumulative loss. But why should the rate of loss under the condition wherein a dependency does exist be higher? That is, informally put, why should the mind be so hasty in detecting dependencies under that condition? One possible explanation is that it is crucial for the mind to swiftly detect dependencies under that condition, with the rationale being that delay in detecting those dependencies could be harmful to the reasoner and potentially jeopardize her life, hence important from an evolutionary standpoint. Furthermore, given the prominent role that explanation and inference play in human cognition (see Lombrozo, 2016), it is crucial for the mind to promptly detect those factors deemed (probabilistically) relevant to the task faced by the reasoner.

Let us more formally characterize the condition (which was alluded to above) under which the aforesaid tendency can be given rational basis. Let $\mathbf{T}_A$ denote the runtime of an algorithm $A$ implementing $d$-separation criterion, $\pi_{\text{YES}}$ and $\pi_{\text{NO}}$ denote the prior probability of the input being a YES-instance and NO-instance $d$-separation query, respectively. Let also $\mathbf{T}_A^{\text{YES}}$ and $\mathbf{T}_A^{\text{NO}}$ denote the worst-case runtime of $A$ on YES-instance and NO-instance $d$-separation queries, respectively. Finally, let $\mathcal{L}_{\text{YES}} \in \mathbb{R}^{>0}$ and $\mathcal{L}_{\text{NO}} \in \mathbb{R}^{>0}$ denote the cost per

[14] Question (Q1) is somewhat analogous to asking "why" Fermat's Last Theorem is true? As Edward Witten points out, aside from proving the correctness of Fermat's Last Theorem, this question does not mean much since Fermat's Last Theorem is after all an inevitable consequence of natural numbers, i.e., natural numbers, by virtue of their existence, naturally give rise to Fermat's Last Theorem.

unit of time incurred by $A$ for delay in detecting a YES-instance and NO-instance $d$-separation query, respectively. Then, for any underlying DAG $G$, the following holds true:

$$\mathbb{E}[\mathbf{T}_A] \leq \frac{\mathcal{L}_{\text{YES}}}{\mathcal{L}_{\text{YES}} + \mathcal{L}_{\text{NO}}} \mathbf{T}_A^{\text{YES}} \pi_{\text{YES}} + \frac{\mathcal{L}_{\text{NO}}}{\mathcal{L}_{\text{YES}} + \mathcal{L}_{\text{NO}}} \mathbf{T}_A^{\text{NO}} \pi_{\text{NO}},$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to the (unknown) distribution on the set of all $d$-separation queries. Then, under the condition

$$\mathcal{L}_{\text{NO}} \pi_{\text{NO}} \geq \mathcal{L}_{\text{YES}} \pi_{\text{YES}}, \tag{5.1}$$

it is rational for the mind to demonstrate the aforementioned tendency toward quick detection of NO-instance $d$-separation queries. The two arguments presented above in support of the rationality of the said tendency are just special cases of Condition (5.1): The first argument corresponds to Condition (5.1) subject to the assumptions $\pi_{\text{NO}} \geq \pi_{\text{YES}}$ and $\mathcal{L}_{\text{NO}} = \mathcal{L}_{\text{YES}}$. The second argument corresponds to Condition (5.1) subject to the assumptions $\pi_{\text{NO}} = \pi_{\text{YES}}$ and $\mathcal{L}_{\text{NO}} \geq \mathcal{L}_{\text{YES}}$.

From among the arguments presented above, unless the validity of $\pi_{\text{NO}} \geq \pi_{\text{YES}}$ is empirically confirmed, our second argument appears to provide the firmest rational basis for the foregoing tendency. Future work should investigate if humans demonstrate the forgoing (apparently) normatively-justified tendency in probabilistic independence judgment tasks, or that, on the contrary, they systematically deviate from this behavior.

## 5.7 Conclusion

We presented a new algorithm, $\mathcal{D}^*$, for implementing $d$-separation, which outperforms previously proposed algorithms in terms of worst-case runtime; the gain is particularly significant in the case of dense graphs. A detailed analysis of the proposed algorithm, including its message- and communication-complexity, along with several refined time-complexity bounds were presented. The introduction of a new graph-theoretic concept, refutation module, permitted a formal characterization of the curious tendency of the proposed algorithm towards quick detection of NO-instance $d$-separation queries. Along the way, important connections were made to the verifier-based definition of the complexity class *coNP*. The work presented in this chapter enhances our understanding of the computational properties of $d$-separation, and crucially, highlights subtle, previously unknown algorithmic properties of $d$-separation

in an unexplored territory: the distributed computing setting. In addition to theoretical contributions, current work might have important implications for how $d$-separation can be implemented in neural circuits. Being similar, in spirit, to Pearl's Belief Propagation (which has played important roles in the theoretical neuroscience literature (see e.g., Gershman and Beck, 2017; George and Hawkins, 2009; Litvak and Ullman, 2009; Rao, 2004; Lochmann and Deneve, 2011)), the asynchronous, distributed nature of the proposed algorithm positions it as a plausible candidate, at Marr's (1982) algorithmic level of analysis—contrary to all the previously proposed algorithms whose sequential nature severely undermines their cognitive plausibility. In that light, $\mathcal{D}^*$ can be then viewed as the first rational, *distributed*, process-level account of how humans handle probabilistic independence (see Griffiths et al., 2009, 2012). Last but not least, the simplicity of $\mathcal{D}^*$ potentially makes it a good candidate for pedagogical purposes.

# Part IV: On Minimality in Learning and Imagination

**Preface.** Capitalizing on the notion of minimality in the context of learning, Fahlman and Lebiere (1989) proposed an influential class of self-organized neural network called cascade-correlation neural networks (CCNNs), which recruited hidden units one at a time *as needed*, thereby constructing an architecture *sufficient* for capturing the regularities in the training set, with good generalization performances. However, CCNNs, by virtue of recruiting only new hidden units during the course of learning, dismissed a crucial aspect of human learning, i.e., relying on the knowledge acquired in the past to unravel new learning tasks as they come about. To address this shortcoming of CCNNs, Shultz and Rivest (2001) introduced a new type of CCNNs called knowledge-based cascade-correlation (KBCC), which was allowed to recruit previously learned neural networks as well as single hidden units during learning. In that light, CCNNs and KBCC both capitalized on the notion of minimality in learning, with the former ignoring past knowledge while the latter exploiting it. Humans are not only adept in recognizing what class an input instance belongs to (i.e., classification task), but perhaps more remarkably, they can imagine (i.e., *generate*) plausible instances of a desired class with ease, when prompted. In computational terms, the notion of generating examples from a desired class can be formalized in terms of sampling from some underlying probability distribution. Despite their appeal from a learning perspective as well as their success in accounting for a variety of psychological phenomena, it remained an open question if CCNNs and/or KBCC could be enabled to probabilistically generate samples, mimicking human imaginative capacity. Addressing this open problem is the topic of Chapter 6. Chapter 6, for the first time in the literature, proposes a *neurally-plausible* and *computationally-efficient* framework, allowing to transform any deterministic, discriminative neural network (e.g., deep convolutional neural networks and multilayer perceptron) into a probabilistic, generative model. Using this framework, Chapter 6 shows, as a proof-of-concept, how CCNNs (and KBCC alike) can be converted into probabilistic generative models, thereby enabling CCNNs

to probabilistically generate samples from a category of interest. Concretely, the proposed framework: (1) suggests a modular account of human imagination which is supported by studies on learning and imaginative abilities of hippocampal amnesic patients as well as a growing body of brain imaging studies showing that perception and imagery share neural representation, (2) gives rise to *self-organized* generative models, (3) strongly suggests that, contrary to a widely-held view, the boundary between discriminative and generative models is blurry, (4) bridges computational, algorithmic, and implementational levels of analysis, and finally, (5) connects two dominant schools of thought in cognitive sciences, namely, connectionism and Bayesian cognition. The framework presented in Chapter 6 views imagination as a collaborative effort of two separate modules, with one responsible for sampling from a distribution induced on the input-output mapping learned by the other module, suggesting that a two-module architecture is *sufficient* to account for human imaginative ability manifested in generating new samples from a desired class. In accord with the maxim of Occam's razor, the proposed framework suggests that, in order to account for human generative abilities, one need not adhere to an encoder-decoder-type architecture (involving a forward model (encoder) and a fully separate inverse model (decoder)), but a single forward model, upon which MCMC operates, might suffice—a more parsimonious design.

# Chapter 6

# Converting Deterministic, Discriminative Neural Networks into Probabilistic, Generative Models: A Case Study of Cascade-Correlation Neural Nets*

*"Everything you can imagine is real."* — Pablo Picasso

A green-striped elephant! Probably no one has seen such a thing—no surprise. But what is a surprise is our ability to easily imagine one. Humans are not only adept in recognizing what class an input instance belongs to (i.e., classification task), but more remarkably, they can imagine (i.e., *generate*) plausible instances of a desired class, when prompted. In fact, humans can generate instances of a desired class, say, elephant, that they have never encountered before, like, a green-striped elephant.[1] In this sense, humans' generative

*The material presented in Chapter 6 is partly based on "A. S. Nobandegani & T. R. Shultz ; **Converting Cascade-Correlation Neural Nets into Probabilistic Generative Models**, In *Proceedings of the* $39^{th}$ *Annual Conference of the Cognitive Science Society (CogSci)*, 2017."

[1]In *counterfactual* terms: Had a human seen a green-striped elephant, s/he would have yet recognized it

capacity goes beyond merely retrieving from memory. In computational terms, the notion of generating examples from a desired class can be formalized in terms of *sampling* from some underlying probability distribution, and has been extensively studied in machine learning under the rubric of probabilistic generative models.

Cascade-Correlation Neural Networks (CCNNs) (Fahlman and Lebiere, 1989) are a well-known class of discriminative (as opposed to generative) models that have been successful in simulating a variety of phenomena in the developmental literature, e.g., infant learning of word-stress patterns in artificial languages (Shultz and Bale, 2006), syllable boundaries (Shultz and Bale, 2006), visual concepts (Shultz, 2006), and have also been successful in capturing important developmental regularities in a variety of tasks, e.g., the balance-scale task (Shultz et al., 1994; Shultz and Takane, 2007), transitivity (Shultz and Vogel, 2004), conservation (Shultz, 1998), and seriation (Mareschal and Shultz, 1999). Also, CCNNs exhibit several similarities with known brain functions: distributed representation, self-organization of network topology, layered hierarchical topologies, both cascaded and direct pathways, an S-shaped activation function, activation modulation via integration of neural inputs, long-term potentiation, growth at the newer end of the network via synaptogenesis or neurogenesis, pruning, and weight freezing (Westermann et al., 2006). Nonetheless, in virtue of being deterministic and discriminative, CCNNs have so far lacked the capacity to probabilistically generate examples from a category of interest.

In this work, we propose a framework which allows transforming CCNNs into probabilistic generative models, thereby enabling CCNNs to generate samples from a category. Our proposed framework is based on a Markov Chain Monte Carlo (MCMC) method, called the Metropolis-Adjusted Langevin (MAL) algorithm, which employs the gradient of the target distribution to guide its explorations towards regions of high probability, thereby significantly reducing the undesirable random walk often observed at the beginning of an MCMC run (a.k.a. the burn-in period). MCMC methods are a family of algorithms for sampling from a desired probability distribution, and have been successful in simulating important aspects of a wide range of cognitive phenomena, e.g., temporal dynamics of multistable perception (Gershman et al., 2012; Moreno-Bote et al., 2011), developmental changes in cognition (Bonawitz et al., 2014b), category learning (Sanborn et al., 2010), causal reasoning in children (Bonawitz et al., 2014a), and accounting for many cognitive biases (Dasgupta et al., 2016).

as an elephant. Geoffrey Hinton once told a similar story about a pink elephant!

Furthermore, work in theoretical neuroscience has shed light on possible mechanisms according to which MCMC methods could be realized in generic cortical circuits (Buesing et al., 2011; Moreno-Bote et al., 2011; Pecevski et al., 2011; Gershman and Beck, 2017). In particular, Moreno-Bote et al. (2011) showed how an attractor neural network implementing MAL can account for multistable perception of drifting gratings, and Savin and Deneve (2014) showed how a network of leaky integrate-and-fire neurons can implement MAL in a biologically-realistic manner.

## 6.1 Cascade-Correlation Neural Networks

CCNNs are a special class of deterministic artificial neural networks, which construct their topology in an autonomous fashion—an appealing property simulating developmental phenomena (Westermann et al., 2006) and other cases where networks need to be constructed. CCNN training starts with a two-layer network (i.e., the input and the output layer) with no hidden units, and proceeds by recruiting hidden units one at a time, as needed. Each new hidden unit is trained to maximally correlate with residual error in the network built so far, and is recruited into a hidden layer of its own, giving rise to a deep network with as many hidden layers as the number of recruited hidden units. CCNNs use sum-of-squared error as an objective function, and typically use symmetric sigmoidal activation functions with range $-0.5$ to $+0.5$ for hidden and output units.[2] Some variants have been proposed: Sibling-Descendant Cascade-Correlation (SDCC) (Baluja and Fahlman, 1994) and Knowledge-Based Cascade-Correlation (KBCC) (Shultz and Rivest, 2001). Although in this chapter we focus on standard CCNNs, our proposed framework can handle SDCC and KBCC as well.

## 6.2 The Metropolis-Adjusted Langevin Algorithm

MAL (Roberts and Tweedie, 1996) is a special type of MCMC method, which employs the gradient of the target distribution to guide its explorations towards regions of high probability, thereby reducing the burn-in period. More specifically, MAL combines the two concepts of Langevin dynamics (a random walk guided by the gradient of the target distribution), and the Metropolis-Hastings algorithm (an accept/reject mechanism for generating a sequence of samples the distribution of which asymptotically converges to the target distribution).

[2]Fahlman and Lebiere (1989) also suggest linear, Gaussian, and asymmetric sigmoidal (with range 0 to +1) activation functions as alternatives. Our proposed framework can be straightforwardly adapted to handle all such activation functions.

**Algorithm 1** The Metropolis-Adjusted Langevin Algorithm

**Input**: Target distribution $\pi(\mathbf{X})$, parameter $\tau \in \mathbb{R}_+$, number of samples $N$.
**Output**: Samples $\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(N-1)}$.

Pick $\mathbf{X}^{(0)}$ arbitrarily.
**for** $i = 0, \ldots, N-1$ **do**
Sample $\mathbf{u} \sim$ Uniform[0,1]
Sample $\mathbf{X}^* \sim q(\mathbf{X}^*|\mathbf{X}^{(i)}) = \mathcal{N}(\mathbf{X}^{(i)} + \tau \nabla \log \pi(\mathbf{X}^{(i)}), 2\tau\mathbb{I})$
    **if** $\mathbf{u} < \min\{1, \frac{\pi(\mathbf{X}^*)q(\mathbf{X}^{(i)}|\mathbf{X}^*)}{\pi(\mathbf{X}^{(i)})q(\mathbf{X}^*|\mathbf{X}^{(i)})}\}$ **then**
        $\mathbf{X}^{(i+1)} \leftarrow \mathbf{X}^*$
    **else**
        $\mathbf{X}^{(i+1)} \leftarrow \mathbf{X}^{(i)}$
    **end if**
**end for**
**return** $\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(N-1)}$

We denote random variables with small bold-faced letters, random vectors by capital bold-faced letters, and their corresponding realizations by non-bold-faced letter. The MAL algorithm is outlined in Algorithm 1 wherein $\pi(\mathbf{X})$ denotes the target probability distribution, $\tau$ is a positive real-valued parameter specifying the time-step used in the Euler-Maruyama approximation of the underlying Langevin dynamics, $N$ denotes the number of samples generated by the MAL algorithm, $q$ denotes the proposal distribution (a.k.a. transition kernel), $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$, and $\mathbb{I}$ denotes the identity matrix. The sequence of samples generated by the MAL algorithm, $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots$, is guaranteed to converge in distribution to $\pi(\mathbf{X})$ (Robert and Casella, 2013). It is worth noting that work in theoretical neuroscience has shown that MAL, outlined in Algorithm 1, can be implemented in a neurally-plausible manner (Savin and Deneve, 2014; Moreno-Bote et al., 2011).[3] In the following section, we propose a target distribution $\pi(\mathbf{X})$, allowing CCNNs to generate samples from a category of interest.

[3]More precisely, it has been shown how the continuous-time version of MAL, Langevin dynamics, can be implemented in a neurally-plausible manner. But note that MAL amounts to sampling from the underlying Langevin dynamics.

## 6.3 The Proposed Framework

In what follows, we propose a framework which transforms CCNNs into probabilistic generative models, thereby enabling them to generate samples from a category of interest. The proposed framework is based on the MAL algorithm given in Sec. 6.2. Let $f(X; W^*)$ denote the input-output mapping learned by a CCNN, and $W^*$ denote the set of weights for a CCNN after training.[4] Upon termination of training, presented with input $X$, a CCNN outputs $f(X; W^*)$. Note that, in case a CCNN possesses multiple output units, $f(X; W^*)$ will be a vector rather than a scalar. To convert a CCNN into a probabilistic generative model, we use the MAL algorithm with its target distribution $\pi(\mathbf{X})$ being set as follows:

$$\begin{aligned} \tilde{\pi}(\mathbf{X}) &\triangleq p(\mathbf{X}|\mathbf{Y} = L_j) \\ &= \frac{1}{Z} \exp(-\beta ||L_j - f(\mathbf{X}; W^*)||_2^2), \end{aligned} \tag{6.1}$$

where $||\cdot||_2$ denotes the $l_2$-norm, $\beta \in \mathbb{R}_+$ is a *damping factor*, $Z$ is the normalizing constant, and $L_j$ is a vector whose element corresponding to the desired class is $+0.5$ (i.e., its $j^{\text{th}}$ element) and the rest of its elements are $-0.5$s. The intuition behind Eq. (6.1) can be articulated as follows: For an input instance $\mathbf{X} = X$ belonging to the desired class $j$,[5] the output of the network $f(X; W^*)$ is expected to be close to $L_j$ in $l_2$-norm sense. In this light, Eq. (6.1) is adjusting the likelihood of input instance $X$ to be inversely proportional to the (base-$e$) exponentiation of the said $l_2$ distance.

For a reader familiar with probabilistic graphical models, the expression in Eq. (6.1) looks similar to the expression for the joint probability distribution of Markov random fields and probabilistic energy-based models, e.g., Restricted Boltzman Machines and Deep Boltzman Machines. However, there is a crucial distinction: The normalizing constant $Z$, the computation of which is intractable in general, renders learning in those models computationally intractable.[6] The appropriate way to interpret Eq. (6.1) is to see it as a Gibbs distribution for a non-probabilistic energy-based model whose energy is defined as the square of the prediction error (LeCun et al., 2006). Section 1.3 of LeCun et al. (2006) discusses the topic of

[4] Formally, $f(\cdot; W^*) : \prod_{i=1}^{n} D_i \rightarrow \prod_{j=1}^{m} R_j$ where $D_i$ and $R_j$ denote the set of values that input unit $i$ and output unit $j$ can take on, respectively.

[5] In counterfactual terms, this is equivalent to saying: Had input instance $X$ been presented to the network, it would have classified $X$ in class $j$.

[6] More specifically, $Z$ renders the computation of the gradient of the log-likelihood for those models intractable.

Gibbs distribution for non-probabilistic energy-based models in the context of discriminitive learning, computationally modeled by $p(\mathbf{Y}|\mathbf{X})$ (i.e., to predict a class given an input), and raises the same issue that we highlighted above regarding the intractability of computing the normalizing constant $Z$ in general. In sharp contrast to LeCun et al. (2006), our framework is proposed for the purpose of generating examples from a desired class, as evidenced by Eq. (6.1) being defined in terms of $p(\mathbf{X}|\mathbf{Y})$. Also crucially, the intractability of computing $Z$ raises no issue for our proposed framework due to an intriguing property of the MAL algorithm according to which the normalizing constant $Z$ need not be computed at all.[7]

Due to Line 4 of Algorithm 1, MAL's proposal distribution $q$ requires the computation of $\nabla \log \tilde{\pi}(\mathbf{X}^{(i)})$, which essentially involves computing $\nabla f(\mathbf{X}^{(i)}; W^*)$ (note that the gradient is operating on $\mathbf{X}^{(i)}$, and $W^*$ is treated as a set of fixed parameters). The multi-layer structure of CCNN ensures that $\nabla f(\mathbf{X}^{(i)}; W^*)$ can be efficiently computed using Backpropagation. Alternatively, in settings where CCNNs recruit a small number of input units (hence, the cardinality of $\mathbf{X}^{(i)}$ is small), $\nabla f(\mathbf{X}^{(i)}; W^*)$ can be obtained by introducing negligible perturbation to a component of input signal $\mathbf{X}^{(i)}$, dividing the resulting change in the network's outputs by the introduced perturbation, and repeating this process for all components of input signal $\mathbf{X}^{(i)}$. It is worth noting that although the idea of computing gradients through introducing small perturbations would lead to a computationally inefficient approach for *learning* CCNNs, it leads to a computationally efficient approach for *generation*, as the number of input units are typically much fewer than the number of weights in CCNNs (and artificial neural networks in general). It is crucial to note that the normalizing constant $Z$ plays no role in the computation of $\nabla \log \tilde{\pi}(\mathbf{X}^{(i)})$.

It is worth noting that the target distribution given in Eq. (6.1) is applicable to any deterministic, discriminative NN (whose input-output mapping is denoted by $f(\mathbf{X}^{(i)}; W^*)$), with CCNNs being simply a particular class of such models. In other word, the target distribution given in Eq. (6.1) is not concerned with what class of deterministic, discriminative NN is responsible for the implementation of the input-output mapping $f(\mathbf{X}^{(i)}; W^*)$. In that light, the proposed framework allows to transform any deterministic, discriminative NN into a probabilistic, generative model.[8]

[7]The MAL algorithm inherits this property from the Metropolis-Hasting algorithm, which it uses as a subroutine.

[8]We should note that the vector $L_j$ in Eq. (6.1) needs to be adapted to the deterministic, discriminative NN which is to be converted to a probabilistic, generative model. For example, if the components of the output vector fall within zero and one (as is most often the case due to adopting a softmax unit), $L_j$ should be set as follows: the element of $L_j$ corresponding to the desired class should be set to 1 (i.e., the $j^{\text{th}}$ element

## 6.4 Simulations

In this section we demonstrate the efficacy of our proposed framework through simulations. We particularly focus on learning which can be accomplished by two input and one output units. This permits visualization of the input-output space, which lies in $\mathbb{R}^3$. Note that our proposed framework can handle arbitrary number of input and output units; this restriction is solely for ease of visualization.

### 6.4.1 Continuous-XOR Problem

In this section, we show how our proposed framework allows a CCNN, trained on the continuous-XOR classification task, to generate examples from a category of interest. The output unit has a symmetric sigmoidal activation function with range $-0.5$ and $+0.5$. The training set consists of 100 samples in the unit-square $[0, 1]^2$, paired with their corresponding labels. More specifically, the training set is comprised of all the ordered-pairs starting from $(0.1, 0.1)$ and going up to $(1, 1)$ with equal steps of size 0.1, paired with their corresponding labels (i.e., $+0.5$ for positive samples and $-0.5$ for negative samples); see Fig. 6.1(a). After training, a CCNN with 6 hidden layers is obtained whose input-output mapping, $f(x_1, x_2; W^*)$, is shown in Fig. 6.1(b).[9]

Fig. 6.2 shows the efficacy of our proposed framework in enabling CCNNs to generate samples from a category of interest, under various choices for MAL parameter $\tau$ (see Algorithm 1) and damping factor $\beta$ (see Eq. (6.1)); generated samples are depicted by red dots. For the results shown in Fig. 6.2, the category of interest is the category of positive examples, i.e., the category of input patterns which, upon being presented to the (learned) network, would be classified as positive by the network. Because $\tau$ controls the amount of jump between consecutive proposals made by MAL, the following behavior is expected: For small $\tau$ (Fig. 6.2(a)) consecutive proposals are very close to one another, leading to a slow exploration of the input domain. As $\tau$ increases, bigger jumps are made by MAL (Fig. 6.2(b)).[10] Parameter $\beta$ controls how severely deviations from the desired class label

of $L_j$) and the rest of the elements of $L_j$ should be set to 0.

[9]Due to the inherent randomness in CCNN construction, training could lead to networks with different structures. However, since in this chapter we are solely concerned with generating examples using CCNNs rather than how well CCNNs could learn a given discriminitive task, we arbitrarily pick a learned network. Note that our proposed framework can handle CCNNs with arbitrary structures; in that light, the choice of network is without loss of generality.

[10]Yet, too large a $\beta$ is not good either, leading to a sparse and coarse-grained exploration of the input

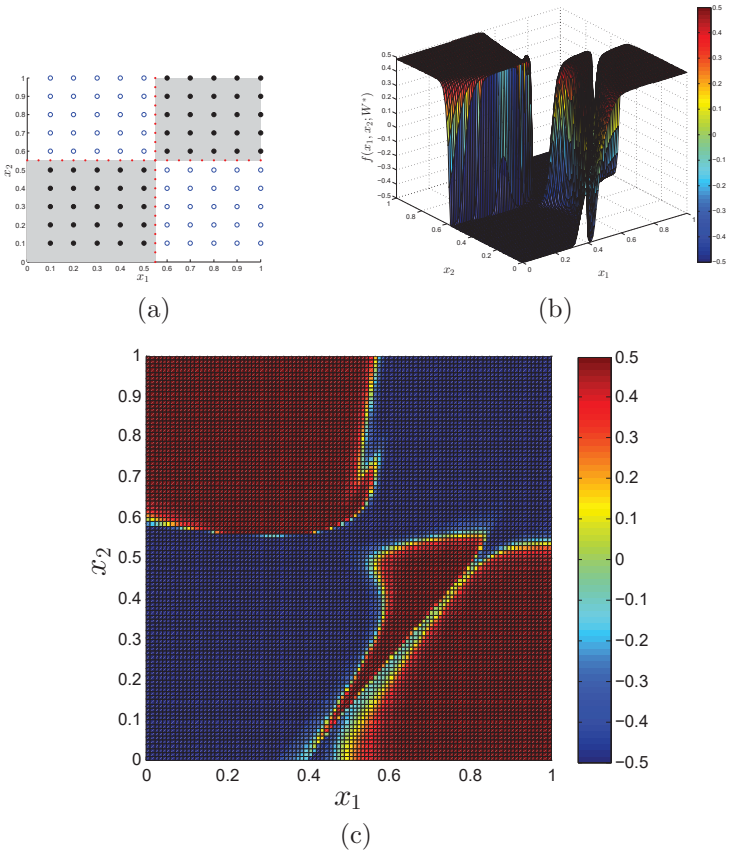
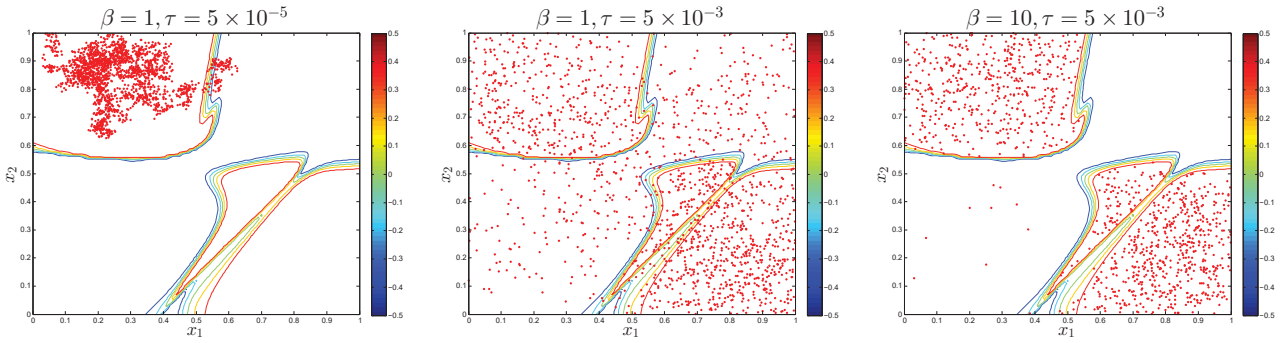



**Fig. 6.1** A CCNN trained on the continuous-XOR classification task. **(a)** Training patterns. All the patterns in the gray quadrants are negative examples with label $-0.5$, and all the patterns in the white quadrants are positive examples with label $+0.5$. Red dotted lines depict the boundaries. **(b)** The input-output mapping, $f(x_1, x_2; W^*)$, learned by a CCNN, along with a colorbar. **(c)** The top-down view of the curve depicted in **(b)**, along with a colorbar.

(here, $+0.5$) are penalized. The larger the parameter $\beta$, the more severely such deviations are penalized and the less likely MAL moves toward such regions of input space. Acceptance Rate (AR), defined as the number of accepted moves divided by the total number of suggested moves, is also presented for the results shown in Fig. 6.2. Fig. 6.2(c) shows that for $\tau = 5 \times 10^{-3}$ and $\beta = 10$, our proposed framework demonstrates desirable performance: virtually all of the generated samples fall within the desired input regions (i.e., the regions associated with hot colors, signaling the closeness of network's output to $+0.5$ in those re-

space. Some measures have been proposed in computational statistics for properly choosing $\tau$ (see Roberts and Rosenthal, 1998).




(a) $N = 2000$, $AR = 99.55\%$ (b) $N = 2000$, $AR = 75.25\%$ (c) $N = 2000$, $AR = 57.85\%$

**Fig. 6.2** Generating example for the positive category, under various choices for MAL parameter $\tau$ and damping factor $\beta$. Contour-plot of the learned mapping, $f(x_1, x_2; W^*)$, along with its corresponding colorbar is shown in each sub-figure. Generated samples are depicted by red dots. $N$ denotes the total number of samples generated by MAL, and $AR$ denotes the corresponding acceptance rate. **(a)** $\tau = 5 \times 10^{-5}$ leads to a very slow exploration of the input space. **(b)** $\tau = 5 \times 10^{-3}$ leads to an adequate exploration of the input space, however, $\beta = 1$ is not penalizing undesirable input regions severely enough. **(c)** A desirable performance is achieved by $\tau = 5 \times 10^{-3}$ and $\beta = 10$.

gions; see Fig. 6.1(c)) and the desired regions are adequately explored (i.e., all hot-colored input regions being visited and almost evenly explored).

Fig. 6.2 depicts all the first $N = 2000$ samples generated by MAL, without excluding the so-called burn-in period. In that light, the result shown in Fig. 6.2(c) nicely demonstrates how MAL—by directing its suggestions toward the direction of gradient and therefore moving toward regions with high likelihood—could alleviate the need for discarding a (potentially large) number of samples generated at the beginning of an MCMC which are assumed to be unrepresentative of equilibrium state, a.k.a. the burn-in period. Fig. 6.3 shows the performance of our framework in enabling the learned CCNN to generate from the category of negative examples, with $\tau = 5 \times 10^{-3}$ and $\beta = 10$.

### 6.4.2 Two-Spirals Problem

Next, we show how our proposed framework allows a CCNN, trained on the famously difficult two-spirals classification task (Fig. 6.4), to generate examples from a category of interest. The output unit has a symmetric sigmoidal activation function with range $-0.5$ and $+0.5$. The training set consists of 194 samples (97 samples per spiral), in the square $[-6.5, 6.5]^2$, paired with their corresponding labels ($+0.5$ and $-0.5$ for positive and negative samples,

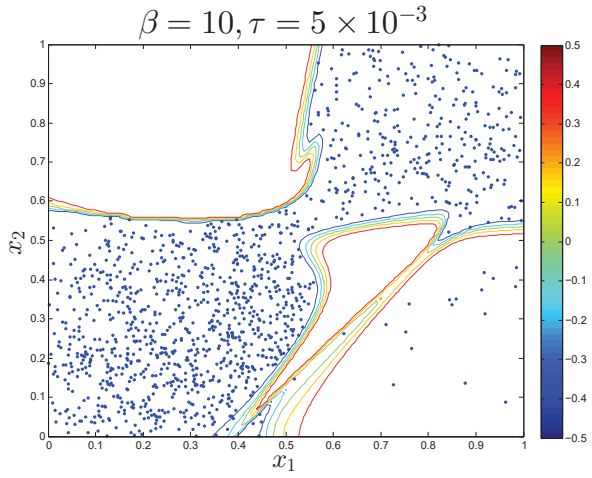



**Fig. 6.3** Generating example for the negative category, with $\tau = 5 \times 10^{-3}, \beta = 10$. Generated samples are shown by blue dots. Total number of samples generated is $N = 2000$, with $AR = 65.13\%$.

respectively). The training patterns are shown in Fig. 6.4(a) (see Chalup and Wiklendt, 2007, for details). After training, a CCNN with 14 hidden layers is obtained whose input-output mapping, $f(x_1, x_2; W^*)$, is depicted in Fig. 6.4(b).

Fig. 6.5(left) and Fig. 6.5(right) show the efficacy of our proposed framework in enabling CCNNs to generate samples from the positive and negative categories, respectively. Although similar patterns of behavior observed in Sec. 6.4.1 due to increasing/decreasing $\beta$ and $\tau$ are observed here as well, due to the lack of space such results are omitted. The results in Fig. 6.5 depict all the first $N = 15000$ samples generated by MAL, without excluding the burn-in period. In that light, these results again demonstrate the efficacy of MAL in alleviating the need for discarding a (potentially large) number samples generated at the beginning of an MCMC run.

Interestingly, our proposed framework also allows CCNNs to generate samples subject to some forms of constraints. For example, Fig. 6.6 demonstrates how our proposed framework enables a CCNN, trained on the continuous-XOR classification task (see Sec. 6.4.1), to generate examples from the positive category, under the following constraint: Generated samples must lie on the curve $x_2 = 0.25 \sin(8\pi x_1) + 0.5$. To generate samples from the positive category while satisfying this constraint, MAL adopts our proposed target distribution given in Eq. (6.1), and treats $x_1$ as an independent and $x_2$ as a dependent variable.

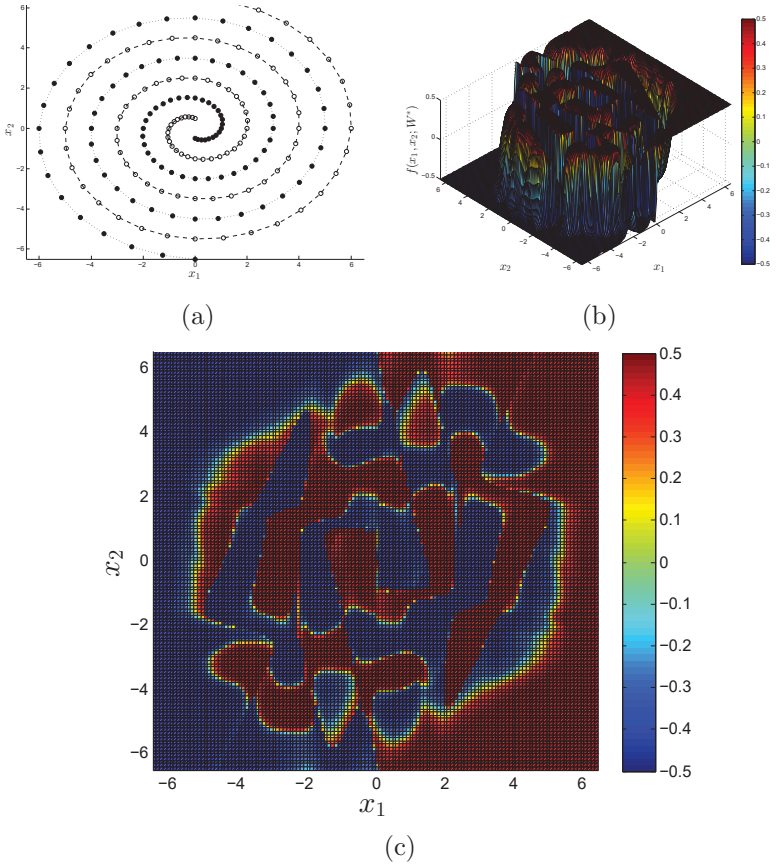



**Fig. 6.4** A CCNN trained on the two-spirals classification task. **(a)** Training patterns. Positive patterns (associated with label +0.5) are shown by hollow circles, and negative patterns (associated with label −0.5) by black circles. Positive spiral is depicted by a dashed line, and negative spiral by a dotted line. **(b)** The input-output mapping, $f(x_1, x_2; W^*)$, learned by a CCNN, along with a colorbar. **(c)** The top-down view of the curve depicted in **(b)**, along with a colorbar.

## 6.5 General Discussion

Although we focused on CCNNs as a case study, our proposed framework allows to transform any deterministic, discriminative neural network (e.g., multilayer perceptron and deep convolutional neural networks) into a probabilistic, generative model. Importantly, our framework is both neurally-plausible and computationally-efficient. The neural-plausibility of our framework stems from the fact that MAL can be implemented in a neurally-plausible manner (Savin and Deneve, 2014; Moreno-Bote et al., 2011). Furthermore, the computational-efficiency of our framework follows from the following three statements: (1) our framework

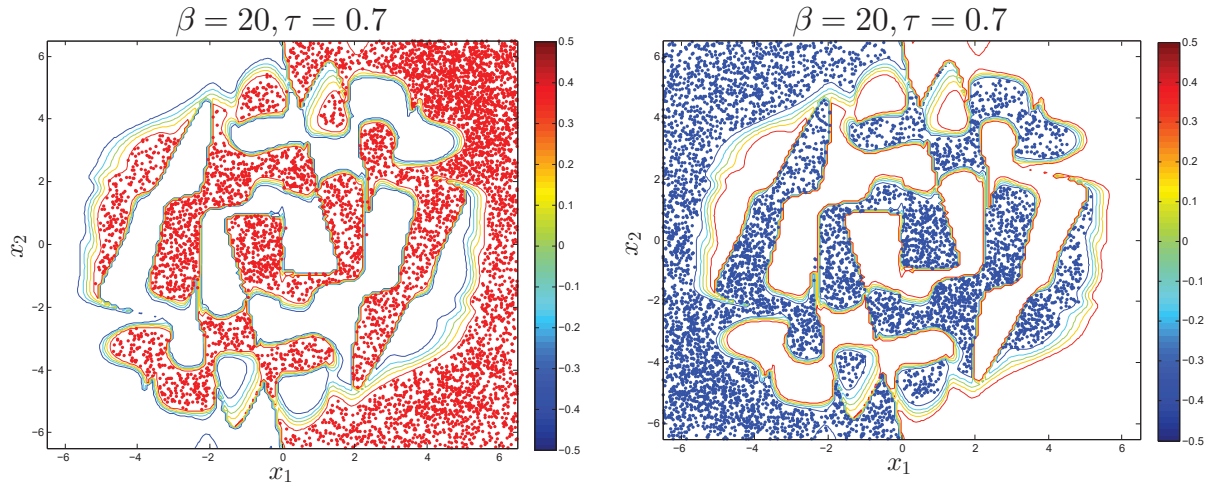



**Fig. 6.5** Generating example for the positive and negative categories, with $\beta = 20$ and $\tau = 0.7$. Contour-plot of the learned mapping, $f(x_1, x_2; W^*)$, along with its corresponding colorbar is shown in each sub-figure. $N$ denotes the total number of samples generated by MAL, and $AR$ denotes the corresponding acceptance rate. Left: Generated example for the positive category, with $N = 15000$ and $AR = 40.69\%$; generated samples are depicted by red dots. Right: Generated example for the negative category, with $N = 15000$ and $AR = 40.28\%$; generated samples are depicted by blue dots.

leverages the gradient of the target distribution (Eq. 6.1) to *guide* its search through the input space, (2) the gradient of the target distribution can be efficiently computed using backpropagation (see Sec. 6.3), and finally (3) our framework does not require that the normalizing constant of the target distribution (Eq. 6.1) be computed at all.[11] Furthermore, our

[11]Recently, Jern and Kemp (2013) advocated a two-step account of exemplar generation: (1) using the training set, a joint probability distribution $\mathbb{P}(\mathbf{X}, \mathbf{Y})$ should be learned over input-output pair $(\mathbf{X}, \mathbf{Y})$, (2) generating exemplars then amounts to drawing samples from $\mathbb{P}(\mathbf{X}|\mathbf{Y}) = \mathbb{P}(\mathbf{X}, \mathbf{Y})/\mathbb{P}(\mathbf{X})$. Our proposed framework is fully consistent with Jern and Kemp's account. Aside from the fact that our framework focuses on neural networks while Jern and Kemp's account is mainly directed toward learning probability distributions, a distinctive feature of our framework is that it substitutes the computationally intractable Step (1) of Jern and Kemp's account (which involves learning a joint distribution, possibly consisting of hidden variables) with a step which can be carried out in a much more computationally efficient manner, namely, learning a discriminative, deterministic NN; our framework subsequently *induces* (instead of learning) a probability distribution on the input-output mapping leaned by the discriminative, deterministic NN (see Eq. 6.1). Also, in multiple occasions, Jern and Kemp (2013) advocate the idea that a *purely* trial-and-error-style approach to exemplar generation could be very inadequate (particularly, for their *randomly-sample-and-score* account, which builds on discriminative models of classification, and as Jern and Kemp point out, seems to be psychologically implausible), implicitly suggesting that more informed mechanisms for exemplar generation might be needed. Crucially, our framework leverages the gradient of the target distribution to inform its search through the input space. We believe that gradient-based MCMCs (e.g., MAL) may be good candidates

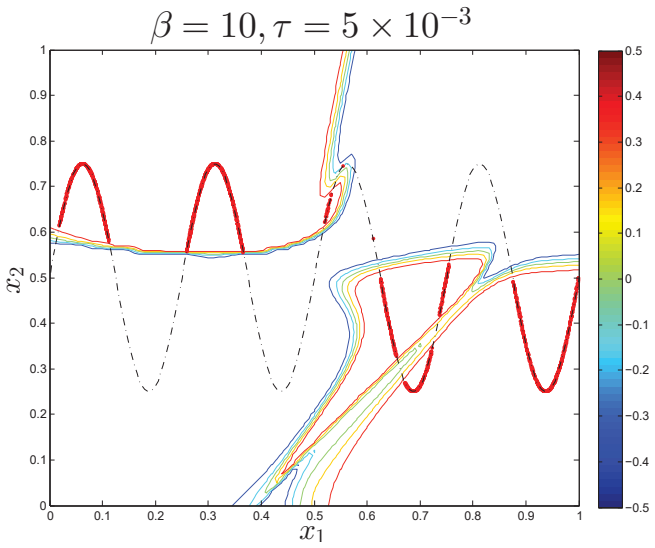



**Fig. 6.6** Generating examples for the positive category, under constraint $x_2 = 0.25\sin(8\pi x_1) + 0.5$ (dash-dotted curve), with $N = 5000$ and $AR = 39.82\%$. Contour-plot of the learned mapping, $f(x_1, x_2; W^*)$, along with its corresponding colorbar is depicted. Generated samples are shown by red dots, which appear mainly as solid red curves due to high density.

proposed framework, together with recent work in theoretical neuroscience showing possible neurally-plausible implementations of MAL (Savin and Deneve, 2014; Moreno-Bote et al., 2011), suggests an intriguing modular hypothesis according to which generation could result from two separate modules interacting with each other (in our case, a CCNN and a neural network implementing MAL). This hypothesis yields the following prediction: There should be some brain impairments which lead to a marked decline in a subject's performance in generative tasks (i.e., tasks involving imagery, or imaginative tasks in general) but leave

for addressing the sensible concern alluded to by Jern and Kemp. Also, the recent surge of interest in computational statistics on gradient-based MCMCs for dealing with high-dimensional distributions (Barp, Briol, Kennedy, & Girolami, 2017) lends further credibility to this idea. It is worth noting that according to Jern and Kemp (2013), their *sample-by-parts* account (as an instantiation of their *sampling* account) cannot handle cases wherein exemplars comprise correlated parts, while our framework in principle can. (This is due to the fact that possible correlations between different parts of an input $\mathbf{X} = X$ is expected to be captured by a discriminative neural network, though training. For example, when a convolutional neural net is trained on a face recognition task, the correlation between various constituent parts of a face, e.g., eyes, ears, nose, lip, etc. will be captured by the trained network.) Furthermore, it is easy to formally show that our proposed target distribution (Eq. 6.1) nicely captures the idea that some exemplars of a category are more likely than others, a property that people are sensitive to when generating exemplars (Jern and Kemp, 2013). Finally, since consecutive samples generated by MCMCs tend to be correlated (see Sanborn and Chater, 2016), our framework can potentially explain the violations of independence effect documented by Jern and Kemp (2013).

the subject's learning abilities (nearly) intact. Studies on learning and imaginative abilities of hippocampal amnesic patients already provide some supporting evidence for this idea (Hassabis et al., 2007; Spiers et al., 2001; Brooks and Baddeley, 1976).

According to Line 4 of Algorithm 1, to generate the $i^{\text{th}}$ sample, MAL requires access to a fine-tuned, Gaussian noise with mean $\mathbf{X}^{(i)} + \tau \nabla \log \pi(\mathbf{X}^{(i)})$ for its proposal distribution $q$. Recently Savin and Deneve (2014) showed how a network of leaky integrate-and-fire neurons can implement MAL in a neurally-plausible manner. However, as Gershman and Beck (2017) point out, Savin and Deneve leave unanswered what the source of that fine-tuned Gaussian noise could be. Our proposed framework may provide an explanation, not for the source of Gaussian noise, but for its fine-tuned mean value. According to our modular account, the main component of the mean value, which is $\nabla \log \pi(\mathbf{X}^{(i)})$, may come from another module (in our case, a CCNN) which has learned some input-output mapping $f(X; W^*)$, based on which the target distribution $\pi(\mathbf{X}^{(i)})$ is defined (see Eq. (1)).

In accord with the maxim of Occam's razor, the proposed framework suggests that, in order to account for human generative abilities, one need *not* adhere to an encoder-decoder-type architecture (involving a forward model (encoder), and a fully separate inverse model (decoder)), but a single forward model, upon which MCMC operates, might suffice—a more parsimonious design. Next, we articulate three propositions which cast doubt on the plausibility of an encoder-decoder-type architectures as an account of human imaginative abilities. Firstly, encoder-decoder-type architectures predict that the time-complexity of discrimination and generation should be predominately comparable. However, our daily experience strongly suggests to the contrary, with generation often appearing to be more effortful (see Jern and Kemp, 2013). Secondly, encoder-decoder-type architectures predict that discriminative and generative abilities are fully dissociated, with one functionality being completely independent of the other. According to our modular account, however, discriminative abilities can be preserved while generative abilities are impaired, but no the other way around. This follows from the fact that our modular account posits that generation *piggybacks* on discrimination, but not vice-versa. The fact that there have been no reports of any subjects with impaired discriminative abilities but spared imaginative abilities calls into question the validity of the encoder-decoder-type architectures prediction mentioned above. A growing body of work in brain imaging suggests that perception and imagery share neural representations, corroborating the view that the phenomenological similarity between imagery and perception is mirrored in similar neural representations (e.g., O'Craven and Kanwisher, 2000;

Cichy et al., 2012; Ishai et al., 2000; Grill-Spector and Malach, 2004; Reddy and Kanwisher, 2007; De Beeck et al., 2008). For example, in a recent study using a combination of functional magnetic resonance imaging (fMRI) and multivariate pattern classification, Cichy et al. (2012) provide evidence supporting the view that perception and imagery systematically share representations of both content (i.e., the category of object seen by a subject) and location (i.e., where the object is seen to be) in ventral visual cortex. These findings lend further support to our modular account, and, at the same time, cast considerable doubt on encoder-decoder-type architectures as an account of human imagery, as they maintain the view that perception and imagery operate on fully dissociated neural circuits. Finally, having separate modules for discrimination and generation, as maintained by encoder-decoder-type architectures, ensues some nontrivial problems as to how one module should be updated in light of the other's update. For example, suppose after the training of an encoder-decoder architecture, the discriminative module (i.e., the encoder) is presented with more data, and in order to accommodate the new data which have come to light, the discriminative module has to go through an update (i.e., further training). The question that immediately arises is how the decoder should be modified in response to the newly introduced update to the encoder—a computationally nontrivial question. However, in the case of our modular account, this key question simply never arises in the fist place.

The idea of sample generation under constraints could be an interesting line of future work. Humans clearly have the capacity to engage in imaginative tasks under a variety of constraints, e.g., when given incomplete sentences or fragments of a picture people can generate possible completions (Sanborn and Chater, 2016). Also, our proposed framework can be used to let a CCNN generate samples from a category of interest at any stage during CCNN construction. In that light, our proposed framework, along with a neurally-plausible implementation of MAL, gives rise to a *self-organized generative model*: a generative model possessing the self-constructive property of CCNNs. Such self-organized generative models could provide a wealth of developmental hypotheses as to how the imaginative capacities of children change over development, and models with quantitative predictions to compare against. We see our work as a step towards such models. Last but not least, our framework strongly suggests that, contrary to a widely-held view, the boundary between discriminative and generative models is blurry—perhaps they are just two sides of the same coin!

# Chapter 7

# Epilogue

*"Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity..."*

— Isaac Newton, Pricipia Mathematica

## 7.1 Paying Attention to Signs

In our attempt to uncover—arguably a glimpse of—how minimality and sufficiency play out in the contexts of learning, reasoning, action, and imagination, we encounter key notions and principles: *symmetry*, *invariance*, *scale-invariance* (aka *self-similarity*), *nestedness*, *locality*, *maximal-informativeness*, *anytime algorithms*, and, finally, the notion of *asynchronous, distributed mechanism*. Another intriguing property of fundamental notions like minimality is that their investigations oftentimes invoke other fundamental notions—birds of a feather flock together.

The central argument that we would like to make is the following: *In the quest for articulating a plausible, algorithmic-level account of cognition, the aforesaid notions—and arguably many others that we either did not touch on or are simply not known awaiting to be discovered—ought to be actively sought for, as guiding principles.* That is, it is not merely enough for them to incidentally manifest themselves in our algorithmic-level accounts; but we are to force them, consciously, into the very fabrics of the formulation of the problem under study as well as its solution.

The next section, perhaps, sheds some light on the subject matter.

### 7.1.1 A Quick Lesson from Physics

According to Feynman[1], it was Poincaré who first realized the significant role symmetry plays in the characteristics of physical laws, inviting physicists to pay attention to it. Steven Weinberg[2] points out that, at the time of his graduate studies, having no idea of the nature of the forces, all that an elementary-particle physicist could possibly do was to work out the symmetries the fundamental forces acting on the particles would have to satisfy. It turned out, ultimately, that the nature of the forces were in fact dictated by those symmetries—satisfying those symmetries, serving as a set of constraints, would simply leave no room for the forces to take any other forms. Weinberg goes on to point out that a story of the same kind also holds for Einstein's general relativity.

## 7.2 Minimality as a Guiding Principle

The ideas explored in this dissertation suggest that pursuing the key notion of minimality—and sufficiency as a relaxation of it—is a fruitful endeavor for understanding cognition at the computational and algorithmic levels. Inspired by this, next, we put forward a mode of enquiry, termed the Rational Minimalist Program (RMP), which brings together Anderson's (1990) rational analysis methodology and the key notion of minimality. In RMP, the term "rational"—echoing Anderson's rational analysis approach—emphasizes the *adaptive* nature of a cognitive system as well as its purposive attitude of striving for *optimality*,[3] and the term "minimalist" captures the cognitive system's very attempt toward attaining optimality, in the thriftiest manner possible (i.e., with *the least redundancy* in terms of resources).

### 7.2.1 A Formalization of the Notion of Minimality

Before articulating the RMP, let us present a formal characterization of the notion of minimality. Depending on the nature of the resource with respect to which the notion of minimality is invoked, minimality can be formalized in two ways, with the second way being a generalization of the first. The first way applies to resources such as time, space, exchanged

[1] *The Character of Physical Laws: Symmetry in Physical Law*, The messenger lectures, Cornell University, 1964.

[2] *Of Beauty and Consolation: Episode 6*

[3] According to Anderson's rational analysis methodology, human performance on task $\mathfrak{T}$ is successful insofar as it *approximates* the optimal solution to task $\mathfrak{T}$, assuming no computational limitations on the reasoner's mental faculties (Chater and Oaksford, 1999).

messages/bits, and samples, which we refer to as *non-set-theoretic* resources.[4] The second way, applies to resources such as knowledge-base and its elements, and neural modules with specific functionalities, which we refer to as *set-theoretic* resources. Like before, we use the notion of sufficiency as a proxy (and a relaxation) for the notion of minimality.

**Def 7.1 (sufficiency, non-set-theoretic resources)** For a non-set-theoretic resource $R$, a $t_R \in \mathbb{R}^{\geq 0}$ amount of $R$ is said to be *sufficient* for task $\mathfrak{T}$ iff $\mathfrak{T}$ can be accomplished by $t_R$ amount of resource $R$.

**Def 7.2 (minimality, non-set-theoretic)** For a non-set-theoretic resource $R$, a $t_R \in \mathbb{R}^{\geq 0}$ amount of $R$ is said to be *minimal* for task $\mathfrak{T}$ iff (1) $t_R$ amount of $R$ is sufficient for task $\mathfrak{T}$, and (2) $\nexists\, t'_R < t_R$ s.t. $t'_R$ amount of $R$ is sufficient for task $\mathfrak{T}$.

**Def 7.3 (sufficiency, set-theoretic resources)** Set $S$ is said to be *sufficient* for task $\mathfrak{T}$ iff using the members of $S$, $\mathfrak{T}$ can be accomplished.

**Def 7.3 (minimality, set-theoretic resources)** Set $S$ is said to be *minimal* for task $\mathfrak{T}$ iff (1) $S$ is sufficient for task $\mathfrak{T}$, and (2) $\nexists\, S' \subsetneqq S$ s.t. $S'$ is sufficient for task $\mathfrak{T}$.

### 7.2.2 Rational Minimalist Program: Pursuing Rationality at the Algorithmic Level of Analysis

As mentioned earlier, RMP is a methodology which integrates the notion of minimality (as formalized in Sec. 7.2.1) and Anderson's (1990) rational analysis approach. RMP outlines a principled way to studying cognition at the algorithmic level. Concretely, by drawing on the concept of minimality in addition to that of optimality (with the latter being characteristic of Anderson's rational analysis), RMP outlines a *principled*, *algorithmic-level* methodology, paralleling Anderson's rational analysis approach which was devised for systematic investigation of cognition at the computational level. Crucially, RMP adds a new dimension to rationality, namely, that of minimality. In Anderson's rational analysis approach, rationality is characterized solely based on the concept of optimality, hence having only one dimension. In RMP, rationality has two dimensions: optimality and minimality. That is, according to

[4] In algorithmics and statistical learning theory, the implications of the aforesaid resources are characterized in terms of *time-complexity*, *space-complexity*, *communication-complexity*, and *sample-complexity*.

RMP, a cognitive system is rational (1) to the extent that it attains optimality, and (2) to the extent that it satisfies minimality while achieving (1). Simply put, in Anderson's rational analysis approach (which is a computational-level methodology) a cognitive system strives for optimality, while, in RMP (which is an algorithmic-level methodology) a cognitive system strives for *minimalist-optimality*—to attain optimality *with the least resources* required for doing so, or, formally, to attain optimality while satisfying minimality criterion (as formalized in Sec. 7.2.1). Hence, the term "minimalist" conveys the following message: with the *least redundancy* in terms of resources, highlighting the thrifty attitude of a cognitive system in striving for optimality. Informally speaking, RMP holds the view that the mind tends to just put in the *bare minimum* (of resources) to attain optimality—the stingy mind!

We are now well-positioned to delineate RMP in five steps, outlined below.

Rational Minimalist Program

1. Formally articulate the **objective** of the cognitive system, $\mathfrak{T}$.
2. Formally specify the cognitive system's **mental model** of its environment. (data structure)
3. Except for resource $R$, postulate no constraints on the cognitive system's computational/cognitive resources.
4. Given (1)-(3) and a **performance guarantee** on $\mathfrak{T}$, devise an algorithm $\mathcal{A}$ which attains **minimalist-optimality** (with optimality referring to fully satisfying the performance guarantee on $\mathfrak{T}$, and minimality being invoked w.r.t. resource $R$).
5. See if predictions of $\mathcal{A}$ are borne out empirically. If not, revise and re-evaluate.

In the following section, we elaborate on how various steps in RMP can be relaxed by acknowledging Simon's (1957) principle of bounded rationality, mimicking the natural move from perfect rationality/optimality toward bounded optimality (Russell, 1997).

### 7.2.3 Relaxing Rational Minimalist Program (RMP): From (Perfect) RMP to Bounded RMP

Analogous to dropping the unrealistic assumption of "unlimited cognitive/computational resources" made in Anderson's (1990) rational analysis approach (which counts as a move from perfect optimality to bounded optimality in Russell's (1997) terms), the restrictive role of various cognitive resources can be increasingly introduced into RMP, thereby increasingly incorporating Simon's (1957) principle of bounded rationality into the methodology of RMP. For example, Step 3 of RMP can be relaxed, straightforwardly, by imposing restrictions on other cognitive resources except $R$ and/or by imposing limits on the amount of computation which can be carried out by the cognitive system in a unit of time. Also, Step 4 of RMP can be relaxed in various ways. The said performance guarantee can be construed as attaining the optimal solution OPT (as is the case for *exact algorithms*), or as attaining merely an approximation to OPT (as is the case for *approximation algorithms*). We refer to such relaxations of RMP as bounded RMP. Therefore, RMP and its relaxed variant bounded RMP, together, form a mode of enquiry—which is parallel to Anderson's rational analysis approach and its relaxation bounded optimality (with the latter considering cognitive/computational limitations), which were introduced for studying cognition at the computational level[5]—outlining a *rational* approach to studying cognition at the *algorithmic* level of analysis (see Fig. 7.1).

### 7.2.4 Rational Minimalist Program's Implications

By focusing on a particular resource $R$ (see RMP's Step 3), RMP aims to formally characterize the computational significance of resources $R$ for cognition. That is, RMP aims to precisely characterize what a restriction imposed on resources $R$ implies for a cognitive system. For example, it could be that a certain constraint on resources $R$ renders the objective of the cognitive system outright unattainable, echoing the notions of *impossibility* and *inapproximability* results in theoretical computer science (TCS). The formulation of RMP, therefore, is very appealing from the vantage point of TCS, and makes contact with a wide

[5]However, we should note that Russel's (1997) formulation of bounded optimality in terms of value of computation (VOC) (Horvitz, 1990) allows for a principled, process-level methodology, targeted at problems which can be cast as a reinforcement learning problem (see Griffiths et al., 2015). Broadly speaking, nonetheless, Russel's (1997) bounded optimality paradigm is an *optimization-based* methodology, focusing on optimizing the VOC, whereas RMP (and, by extension, bounded RMP) is an *algorithm-design* methodology, which is after devising minimalist-optimal algorithms.

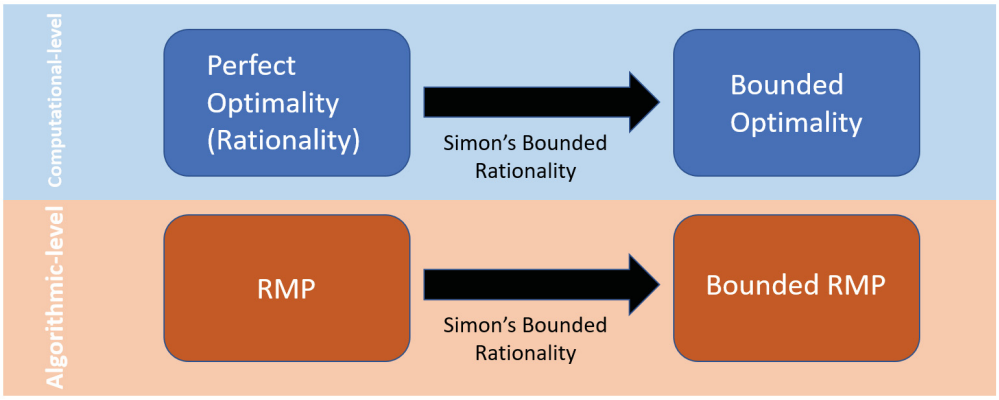



**Fig. 7.1** Rational minimalist program (RMP) and its relaxation bounded RMP, together, form a rational mode of inquiry for studying cognition at the algorithmic level, serving as a parallel research program to Anderson's (1990) rational analysis approach and its relaxation bounded optimality (with the latter considering cognitive/computational limitations) which were devised for studying cognition at the computational level of analysis.

range of topics in TCS: *exact* and/or *approximation algorithms*, *parameterized complexity*, *fixed-parameter tractability* (with the restriction imposed on $R$ serving as a natural parameter), and developing *inapproximability results*. When the task of the cognitive system is a decision problem, pursuing RMP naturally makes contact with the notions of *shortest proof* (for *NP*) and *shortest disproof* (for *coNP*) (for an example on this, see Sec. 5.2 where we introduce the notion of minimal refutation-module).

### 7.2.5 A Dual Interpretation of Rational Minimalist Program: D-RMP

In what follows, we develop a dual interpretation of RMP, termed D-RMP. Our objective for such a development is that, despite the mathematical equivalence of RMP and D-RMP, some problems/tasks lend themselves more naturally to D-RMP than RMP.[6] Under the assumption that having access to more of resource $R$ allows a cognitive system to attain a strictly better performance guarantee on a task $\mathfrak{T}$ of interest (which is equivalent to assuming that the quality of the cognitive system's performance on $\mathfrak{T}$ is strictly increasing in the size of $R$), the algorithmic solution to the following two problems is the same: (This claim can

[6]For the reader familiar with representational systems, this should come as no surprise. Given two equivalent representational systems (i.e., if statement $e$ can be represented in one system, it can be represented in the other as well, for all $e$), one representational system may allow for a simpler, more parsimonious representation of a proposition than the other; e.g., the tabular representation of a joint probability distribution vs. its representation by a Bayesian network.

be straightforwardly established using proof by contradiction.)

RATIONAL MINIMALIST PROGRAM (RMP) VS. ITS DUAL (D-RMP)

**(P1)** For a given performance guarantee on $\mathfrak{T}$, which algorithm uses the least amount of resource $R$ to attain that? (RMP version)

**(P2)** For a given amount of resource $R$, which algorithm yields the best performance guarantee on $\mathfrak{T}$? (D-RMP version)

Let us clarify the above equivalence by giving a concrete example. Assuming that the error incurred on a task $\mathfrak{T}$ is strictly decreasing in the amount of time spent by a cognitive system on $\mathfrak{T}$, the following algorithmic problems have the same solution:

**(1a)** Which algorithm uses the least amount of time to attain a given error rate on $\mathfrak{T}$?

**(1b)** For a given amount of time, which algorithm yields the least error rate on $\mathfrak{T}$?

Or, reiterating the above example when the resource of interest is the number of random samples used by the cognitive system (a setting which is of great interest to sample-based accounts, widely entertained in Bayesian models of cognition), the following equivalence follows:

**(2a)** Which algorithm uses the least number of samples to attain a given error rate on $\mathfrak{T}$?

**(2b)** For a given number of samples available to a cognitive system, which algorithm yields the least error rate on $\mathfrak{T}$?

Intriguingly, the well-known *speed-accuracy trade-off*, a very ubiquitous effect in the literature on judgment and decision-making, arises naturally out of the established duality between RMP and D-RAMP. Arguably, the fact that such an important effect naturally arose from this duality under extremely minimal assumptions about the nature of the decision-task faced by the cognitive system, elevates the credibility of RMP (and D-RMP, by duality) as a methodology, and highlights its robustness under a wide range of decision-making tasks' parameterizations.

### 7.2.6 On the Connection between RMP and Chomsky's Minimalist Program in Linguistics

A reader familiar with Chomsky's Minimalist Program (MP) in linguistics may rightfully wonder if there is any connection between MP and the proposed methodology RMP; after all, their names strongly suggest that they should have a great deal in common (otherwise, calling it 'Rational Minimalist Program' would have been a bad choice). Indeed, they do have a great deal in common. Simply put, RMP is, in essence, an extension of MP (which is solely concerned with the human language faculty, a.k.a. Universal Grammar) to studying human cognition, as a whole, at the algorithmic level of analysis. This understanding is eminent in the following characterization of MP:[7] "The minimalist program for linguistic theory adopts as its working hypothesis the idea that Universal Grammar is *'perfectly'* designed, that is, it contains *nothing more than* what follows from our best guesses regarding conceptual, biological, physical necessity" (Boeckx, 2006, emphasis added). That is, using the terminology of RMP, MP in linguistics seeks to show that the human language faculty is minimalist-optimal; in the above quote from Boeckx (2006), the terms 'perfectly' and 'no more than' correspond to optimality and minimality criteria, respectively. In that light, broadly speaking, linguistic MP can be construed as an instantiation of the proposed methodology RMP within the context of the human language faculty. This understanding has significant implications for RMP: Successes of linguistic MP count as successes of RMP, and, likewise, failures of MP count as that of RMP. There is substantial literature on linguistic MP which we do not get into; for a great exposition of MP and its historical, philosophical, biological, and empirical grounds see Boeckx (2006).

The two concepts of *economy* and *virtual conceptual necessity*, that occupy center stage in linguistic MP (Boeckx, 2006), eminently highlight two key features of RMP.[8] Economy, as a manifestation of least effort principle, chiefly echoes the minimalist aspect of the RMP's minimalist-optimality objective (RMP, Step 4), and virtual conceptual necessity captures the limitations, constraints, and conditions imposed by the mind's computational and cognitive

[7] Likewise, the following characterization of linguistic MP attests to the claim that MP and RMP share the same core concept: "Minimalism is animated by the belief that the old adage 'Least is best' is not only methodologically desirable but also true of the design of the language faculty" (Boeckx, 2006) (Boeckx, 2006).

[8] In Boeckx's (2006) view, the concept of *symmetry* also represent the third pillar of linguistic MP. Similar to economy, symmetry also bears on the the minimalist aspect of the RMP's minimalist-optimality objective (RMP, Step 4). Our emphasis on pursuing the notions of symmetry, self-symmetry, and invariance in Sec. 7.1 clearly echoes this point.

apparatus (RMP, Step 3).

In light of the said connections between MP and RMP, it is worth noting Chomsky's (2000) view on MP:

> "The minimalist goal of discovering how perfect/well-designed language is will inevitably meet with obstacles. Language seems to be full of imperfections, properties that do not seem to follow from economy, virtual conceptual necessity, or symmetry. When faced with some apparent property $P$ of language, the way to proceed is to find out whether:
>
> (i) $P$ is real, and an imperfection (i.e., a real problem for minimalism)
>
> (ii) $P$ is not real, contrary to what had been supposed
>
> (iii) $P$ is real, but not an imperfection; [once scrutinized, $P$ can be shown to be] part of a best way to meet design specification."

By the same logic, RMP, too, will inevitably meet with obstacles; and the way to proceed is what Chomsky (2000) suggests above. In short, what RMP seeks to understand is how much of human cognition meets the principle of minimalist-optimality. And we might be surprised that a large portion of human behaviors documented in the literature as evidence for human irrationality would ultimately turn out to be accounted for by appealing to RMP's minimalist-optimality criterion.

As Boeckx (2006) notes "[Linguistic] minimalism is worth pursuing because, to the extent that one can reach explanations by following minimalist guidelines, such explanations will have a deep and pleasing character;" the same obviously holds for any explanation provided by pursuing RMP, making the pursuit of RMP a pleasing endeavor.

Next, we show that the pursuit of RMP has already helped us develop a deeper understanding of important aspects of human cognition.

### 7.2.7 Instantiations of Rational Minimalist Program

In this section, we elaborate on how the line of work pursued in Chapters 2 to 6 can all be viewed as instantiations of RMP methodology.

In Chapter 2, inspired primarily by the reasoner's limited attention span and scope (as a manifestation of Simon's bounded rationality), we introduced a new graphical model, MCM, to represent the reasoner's partial knowledge of a domain. Concretely, MCM served

as a normative, probabilistic, representational-level model for capturing the state of partial knowledge of a domain. Due to the state of knowledge being partial, the best one could hope for was to derive optimal bounds given the available partial knowledge to the reasoner represented by an MCM. Hence, the task $\mathfrak{T}$ was finding lower and upper bounds on a posed query, and the performance guarantee became those upper and lower bounds being optimal, given the reasoner's state of partial knowledge represented by an MCM. We then proposed a computationally-efficient algorithm which could identify a *sufficient* set of variables for the aforementioned task, and output optimal lower and upper bound on the posed query. Also, MCM served as the first normative, probabilistic, representational-level account of an important developmental shift from features in isolation to correlations between those features, in infants between four an ten months of age. In that light, the line of work investigated in Chapter 2 is an instantiation of RMP with its Step 4 being relaxed (due to settling for sufficiency instead of minimality)—hence an instantiation of bounded RMP.

In our investigation of the causal frame problem in Chapter 3, inspired by Simon's bounded rationality, we substantiated the concurrence approach to reasoning. Guided (in retrospect) by the key notions of locality and nestedness, the task $\mathfrak{T}$ was deriving lower and upper bounds on a posed query given a retrieved submodel, requiring the retrieved submodel to be *sufficient* for the derivation of the proposed bounds on the posed query. We also formally showed that the reported lower and upper bounds on the query were the "best" one could possibly hope for, through the introduction of the key notion of maximally-informativeness—hence "maximally-informativeness" served as the performance guarantee. Furthermore, we showed that our proposed framework were consistent with a wide range of findings in the literature, and substantiated the claim that the introduced graph-theoretic notion of potential level (PL) might bear on how time is encoded in the mind. In that light, the line of work explored in Chapter 3 is an instantiation of RMP with its Step 4 being relaxed (due to settling for a sufficient submodel rather than the minimal submodel)—hence an instantiation of bounded RMP.[9]

The line of work investigated in Chapter 4 is an instantiation of RMP without any relax-

[9]Following Russell's (1997) bounded optimality methodology (which is, broadly, equivalent to the resource-rational methodology proposed by Griffiths et al. (2015)), Icard and Goodman (2015) presented a boundedly-rational approach to the CFP, formally articulated at Marr's computational level of analysis, but left unanswered how the CFP can be addressed at Marr's algorithmic level. However, pursuing bounded RMP allowed us to present a formal, boundedly-rational, *algorithmic-level* approach to the CFP. This apparently suggests that RMP (and, by extension, bounded RMP and D-RMP) is a more fruitful methodology for investigating cognition at the algorithmic level.

ations, with the task $\mathfrak{T}$ being TPS-controllability of a CBN formally captured by maximax, minimin, maximin, and minimax objectives, and the performance guarantee being to be optimal w.r.t. the said objectives (hence, no approximations). We proposed a linear-time algorithm which outputted a minimal set of intervenable nodes $\boldsymbol{\mathcal{X}}^*$ for the maximax and minimin objectives.[10] Also, the formalism developed in Chapter 4 established the first rational, algorithmic-level account of a curious behavior demonstrated by young children called *overimitation*, generally taken as evidence for children's irrationality.

The fundamental idea of asynchronous, distributed mechanisms came into play in Chapter 5, where we revisited Pearl's key notion of $d$-separation. The task $\mathfrak{T}$ was deciding if a posed $d$-separation query held in a BN, with the performance guarantee being that the output of the proposed algorithm be correct for every input instance (hence, no approximations). Through the introduction of the graph-theoretic notion of minimal refutation-module, we formally characterized how the proposed algorithm $\mathcal{D}^*$ explores the smallest subgraph of the BN, to disprove a given NO-instance $d$-separation query. Consistent with the brain's computational machinery (see McClelland, 1989; Chater et al., 2006, *inter alia*) and fully in the spirit of the celebrated parallel-distributed-processing (PDP) research program in brain and cognitive sciences, $\mathcal{D}^*$ permitted the implementation of $d$-separation in an asynchronous, distributed, message-passing fashion. Also, recent work in neuroscience investigating possible implementation of BNs at the neural level supports $\mathcal{D}^*$'s use of BN links as a medium for inference, and we also provided normative grounds for $\mathcal{D}^*$'s tendency toward quick detection of NO-instance $d$-separation queries. In that light, the line of work explored in Chapter 5 is an instantiation of RMP.

In Chapter 6, we proposed a neurally-plausible and computationally-efficient framework, allowing to transform any deterministic, discriminative neural network (e.g., deep convolutional neural networks and multilayer perceptron) into a probabilistic, generative model. The resource of interest in this chapter was the number of neural modules/units required for performing the tasks of learning and exemplar generation, thereby RMP making contact with the implementational level of analysis. The line of work investigated in Chapter 6, *prima facie*, seems to be detached from RMP. However, as we argue next, it indeed follows the spirit of RMP, which is a cognitive system striving for minimality. As far as learning was concerned, cascade-correlation neural network (CCNN), would strive for recruiting the least number of hidden units required for learning the underlying task—in line with the maxim

[10] We showed that the solution to maximin and minimax objectives is, in both case, the empty set.

of a cognitive system having to strive for minimality, according to RMP—with the task $\mathfrak{T}$ being to minimize the sum-of-squared error (objective function), and the performance guarantee being to be optimal w.r.t. the said objective function (or to stop the learning process altogether, if no improvement is achieved after certain number of epochs). As far as exemplar generation was concerned, The task $\mathfrak{T}$ was to draw samples from a target distribution induced on the input-output mapping leaned by a deterministic, discriminative NN (particularly, a CCNN), with the performance guarantee being the distribution of the sequence of generated samples converging to the target distribution asymptotically. In accord with the maxim of Occam's razor, the proposed framework suggested that, in order to account for human generative abilities, one need not adhere to an encoder-decoder-type architecture (which involves a forward model (encoder) and a fully separate inverse model (decoder)), but a single forward model, upon which MCMC operates, might suffice—a more parsimonious design.

### 7.2.8 A Principled, Rational Approach to Studying Cognition at the Algorithmic Level: What to Expect? What to Gain?

The mode of enquiry which was pursued, instantiated, advocated, and finally systematically outlined in this dissertation, i.e., RMP, portrays a perspective on understanding cognition at the process level that is strikingly different from the way in which cognitive psychologies predominantly have studied psychological processes to date, which involves devising *ad hoc* processes to account for experimental data. These ad hoc processes are often entertained with a deliberate disregard for, or little appeal to, rational grounds, i.e., what makes them normatively justified. But two fundamental questions immediately present themselves: (Q1) Why should we pay any attention to normative principles in going about uncovering psychological processes? (Q2) Isn't accounting for data *all* that matters? Let us first address Question (Q1). Appealing to normative principles, evidently, is not a necessity in any attempts toward discovering a psychological process; it perfectly suffices if you simply guess one which accounts for the data. But two key points should not be dismissed here. Firstly, the data almost never sufficiently constrain the problem of devising psychological processes. That is, as far as empirical data are concerned, the problem of devising psychological processes is highly under-constrained (the concept of overfitting should come to mind). Therefore, guiding principles like rationality, should be sought for in order to simply *regularize* this combinatorial search in the space of psychological processes. The point we just made is analogous to the

"poverty of stimuli" argument entertained in linguistics, and, more broadly, in philosophy of mind. Secondly, if one adheres to the view that a thorough understanding of the mind involves not only "*what* cognitive processes are carried out in the mind," but also, "*why* those processes are any good, and, why they make us 'smarter' than many other species, across various tasks and domains," then the answer to Question (Q1) should appear to be obvious. The "why" questions, therefore, is at least as important as the "what" questions—if not more! Hence, in short, adopting a principled, rational approach toward studying the mind at the process level, *simultaneously* serves two key purposes: alleviating the complexity of the combinatorial search problem in the space of cognitive processes, while guaranteeing that the outcome of the search, which is an algorithm, is "good," in a reasonable sense (in the case of RMP, the output is good in the sensible manner that it is *minimalist-optimal*). Addressing Question (Q2) is the subject of the next section.

## 7.3 Why Do We Need Guiding Principles?

In 1905, when Einstein put forward his theory of Special Relativity, there were experiments which were at odds with its predictions. In Murray Gell-Mann's[11] words, once Einstein was asked if he was worried about the fact that some experiments were disproving his newly proposed theory; "[t]he theory is so beautiful, it must be right," Gell-Mann quoting Einstein—implying that all those experiments must be wrong. Some decades later, Murray Gell-Mann and colleagues put forward a theory predicting the Weak Force despite knowing that seven important experiments were in disagreement with their proposal. Gell-Mann[12] reflecting on that memory, says that "the theory was so beautiful that it could not be wrong... it turned out that all seven experiments were wrong—every single one of them." How could a scientist be so sure of his/her proposed theory while, at the very same time, there exists a good number of experiments disproving it? In the case of Einstein or Gell-Mann, how did they know that it was the experiments, not their ideas, that were wrong? Evidently, it is far more feasible to perform well-controlled experiments in sciences like physics than in behavioral and social sciences. This understanding further highlights the crucial role of guiding principles in our attempts to formally investigate the inner-workings of the human

[11] *Beauty and Elegance in Physics*, The 2009 Harry Mullin Memorial Lecture.
[12] *Beauty and Elegance in Physics*, The 2009 Harry Mullin Memorial Lecture.

mind, with all its insurmountable perplexity.[13] They might well turn out to be our ultimate winning cards in our game against nature! The hide-and-seek game of uncovering the nature of the human mind!

[13]This echoes Chomsky's argument in favor of the Galilean style; for a thorough exposition, the reader is referred to Boeckx (2006).

# Appendix A

Appendix A is structured as follows. Sec. A-I presents a short version of $\mathcal{I}^*$, called $\mathcal{I}^*_{non-scale}$, which ignores the scale-invariance property discussed in Sec. 2.4.2 and solely aims to identify a sufficient set of RVs for inclusion into the LP. The correctness proof for $\mathcal{I}^*_{non-scale}$ is presented in Sec. A-II. Sec. A-III provides a proof for the example on scale-invariance property given in Sec. 2.4.2 and also introduces an intriguing way of *visualizing* the scale-invariance property. Finally, Sec. A-III (a) delves more deeply into the scale-invariance property and the intuition behind it, (b) highlights a key concept called *orthogonality* which is essential for understanding the scale-invariance property, and (c) presents the algorithm $\mathcal{I}^*_{scale}$ (which is the algorithm $\mathcal{I}^*$ alluded to above) as a variant of $\mathcal{I}^*_{non-scale}$, which strives to fully embrace the scale-invariance property.

## A-I $\mathcal{I}^*_{non-scale}$: A short version of $\mathcal{I}^*$ without scale-invariance property

$\mathcal{I}^*$ aims at minimally parameterizing the information contained in an MCM so that the posed inter-contextual query can be stated as an LP with the fewest number of parameters. As pointed out earlier in Sec. 2.4.2, $\mathcal{I}^*$ has to decide on the following: (i) what RVs have to be included in the LP, and (ii) the abstraction level required to efficiently encode the information on the RVs identified in step (i) for the LP, in our case, the parameterization of the identified RVs.

In what follows, a simple algorithm, $\mathcal{I}^*_{non-scale}$, is sketched which only performs (i) and ignores (ii). In other words, $\mathcal{I}^*_{non-scale}$ identifies the relevant RVs needed to derive the *exact* lower/upper bound for the inter-contextual query, however, it does not aim at minimally encoding them into the LP. In Sec. A-III, we provide the intuition behind the scale-invariance property presented in Sec. 2.4.2 and, ultimately, we present the algorithm $\mathcal{I}^*_{scale}$ as a variant

of $\mathcal{I}^*_{non-scale}$ which aims to accomplish (i) as well as (ii). $\mathcal{I}^*_{non-scale}$ consists of three steps:

(1) Identify all the RVs involved in the posed query (e.g., in $P(X|Y, z)$ these are the random vector $X$, random vector $Y$ and RV $z$).

(2a) If any two of the already identified RVs belong to two overlapping contexts, identify all the *overlapping* RVs between these two contexts (e.g., in Fig. 5(b) and for the query $P(X|Y)$ for which step (1) would identify $X$ and $Y$, random vector $Z$ in the overlapping region must be identified as well).

(2b) If any two of the already identified RVs belong to two contexts connected through a chain of overlapping contexts: identify all the RVs contained in all the *overlapping* regions of the chain of contexts.

(3) Parameterize only the identified RVs in steps (1), (2a), and (2b) (remove all the other RVs from the MCM—there is no need to encode the information on any other RVs not identified in steps (1), (2a), and (2b)).

■

It should be noted that whether the posed query involves minimization or maximization does not affect which RVs need to be identified by $\mathcal{I}^*_{non-scale}$. Finally, It is worth noting that with a minor modification to step (3) of $\mathcal{I}^*_{non-scale}$, the scale-invariance property could be achieved. The modification has to do with the question of how to *minimally* encode the information on each RV identified in steps (1), (2a), and (2b) of $\mathcal{I}^*_{non-scale}$.

To demonstrate the operation of $\mathcal{I}^*_{non-scale}$ on a more complicated MCM that involves loops, consider the following example sketched in Fig. 7.2(a). The query of interest is $\mathbb{P}(X|Y)_{\downarrow}$.

Next, we are going to sketch the proof for $\mathcal{I}^*_{non-scale}$. Let us first state the claim formally and then provide the proof.

## A-II Proof for $\mathcal{I}^*_{non-scale}$:

**Lemma**: *Given a posed query and an MCM, if all the information on the RVs identified in steps (1) to (2b) of $\mathcal{I}^*_{non-scale}$ is stated and then solved as an LP, the exact solution (i.e., a min or max) can be derived for the posed query; all the remaining information available in the MCM is deemed irrelevant to the derivation of the query, hence the sufficiency.*

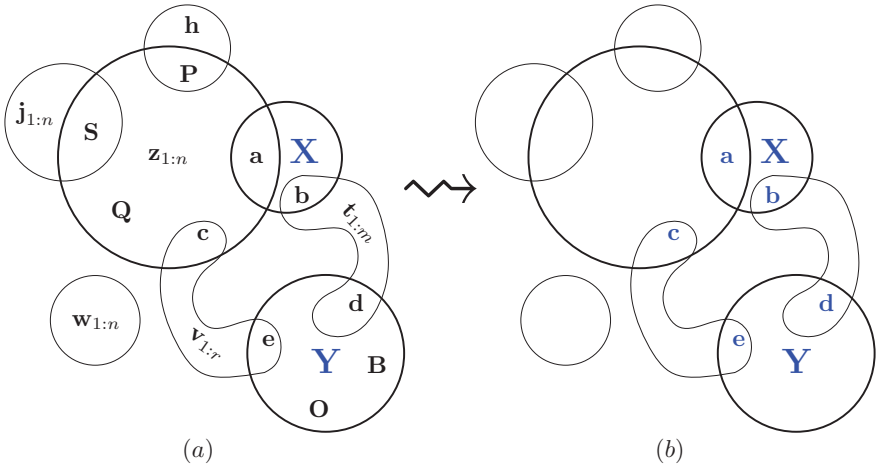



**Fig. 7.2** (a) Sample MCM. The RVs involved in the posed query are depicted in blue. (b) In Step (1) **X** and **Y** are identified; in step (2b) the RVs **b**, **d** as well as **a**, **c**, and **e** are identified. According to step (3) of $\mathcal{I}^{*}_{non-scale}$ all of the information as to the RVs **X**, **Y**, **b**, **d**, **a**, **c**, and **e** has to be stated as an LP to derive the query.

**Proof:** Our proof is constructive. In the proof we entertain two ideas, namely (i) the idea of generative process and, particularly, that of *conditioning* also used in Sec. 2.3.2, and (ii) the notion we refer to as the *locality of information*. Suppose that all the RVs discussed in steps (1) to (2b) of $\mathcal{I}^{*}_{non-scale}$ are identified. The key insight is that the information on how the remaining RVs probabilistically interact with each other is completely local in nature and, therefore, irrelevant to the derivation of the posed query. To see this, one can start off with the identified RVs and then in a gradual fashion add on[1] the rest of the RVs (through the idea of conditioning discussed in Sec. 2.3.2). Quite crucially, this very process of adding the non-identified RVs to the model can be done completely in a local fashion, i.e., without imposing any constraints on how the identified RVs probabilistically interact. The mere fact that those RVs can be added into the model: (i) subsequent to the identified ones, and (ii) without inducing any sort of constraints on the identified ones, deems them irrelevant to the derivation of the query. ■

[1]This is based on the fundamental property that a JPD can be expanded using the chain rule of probability in an arbitrary order.

## A-III Scale-Invariance Property: Intuition

Here, we will provide a proof for the example on scale-invariance property given in Sec. 2.4.2. Although the proof is provided for a special query, the methodology used in the proof provides an insightful way of *visualizing* an inference problem. The idea behind the proof is very simple and related to visualizing the connection of a RV to the underlying sample space using Venn diagrams. Without loss of generality, we assume that all the RVs present in the domain are binary[2]. Random vector $\mathbf{X} = \mathbf{x}_{1:n}$ partitions the sample space $\Omega$ into $2^n$ disjoint regions each of which corresponds to a realization of $\mathbf{X}$. If each realization of the random vector $\mathbf{x}_{1:n}$ corresponds to a binary number (i.e., binary-coding the realizations), then one can conclude $Val(\mathbf{X}) = \{0, 1, \cdots, 2^n - 1\}$. Let us index the partitions by their corresponding realization of $\mathbf{X}$. An illustrative example of an induced partitioning of the sample space $\Omega$ due to random vector $\mathbf{X} = \mathbf{x}_{1:n}$ is depicted in Fig. 7.3(a), and a partitioning induced by RVs $\mathbf{y}$ and $\mathbf{z}$ is sketched in Fig. 7.3(b). We note that the mere knowledge of the distribution function of a random quantity does not provide one with the knowledge of the underlying partitions. For this particular example, since the JPD over $\mathbf{X}, \mathbf{y}, \mathbf{z}$ is not available, the knowledge of how the partitions induced by $\mathbf{y}, \mathbf{z}$ (Fig. 7.3(b)) and the ones induced by $\mathbf{X}$ (Fig. 7.3(a)) interact, i.e., to what extent they overlap, remains unspecified. Therefore, since $\mathbb{P}(X|y) = \frac{\mathbb{P}(X,y)}{\mathbb{P}(y)}$, to minimize (maximize) $\mathbb{P}(X|y)$, the quantity $\mathbb{P}(X, y)$ has to be minimized (maximized). Pictorially, the minimization (maximization) of $\mathbb{P}(X, y)$ corresponds to the minimization (maximization) of the overlap between the partitions corresponding to the events $\{\mathbf{X} = X\}$ and $\{\mathbf{y} = y\}$; hence, very simply, $\mathbb{P}(X, y)_{\downarrow} = [\mathbb{P}(X) + \mathbb{P}(y) - 1]^+$ and $\mathbb{P}(X, y)_{\uparrow} = \min\{\mathbb{P}(X), \mathbb{P}(y)\}$. The key point, which yields the scale-invariance property, is that to derive the minimum (maximum) overlap between the partitions corresponding to the events $\{\mathbf{X} = X\}$ and $\{\mathbf{y} = y\}$ *the information as to how the other partitions—corresponding to the other realizations of the present RVs in the model—interact with one another neither needs to be known nor to be encoded into the LP*; a fact which results in not requiring to encode the information as to the other realizations. Hence the only pieces of information that are required to be encoded and then solved as an LP are $\mathbb{P}(X)$ and $\mathbb{P}(y)$. The same line of reasoning could be adopted for $\mathbb{P}(x_i|y)$. The idea of scale-invariance, therefore, aims to avoid the encoding of the information as to the partitions induced on $\Omega$ which are yet deemed to be irrelevant to the derivation of the posed query; hence one needs to encode

[2]The generalization of the argument to non-binary RVs is straightforward.

solely the relevant ones into the LP.

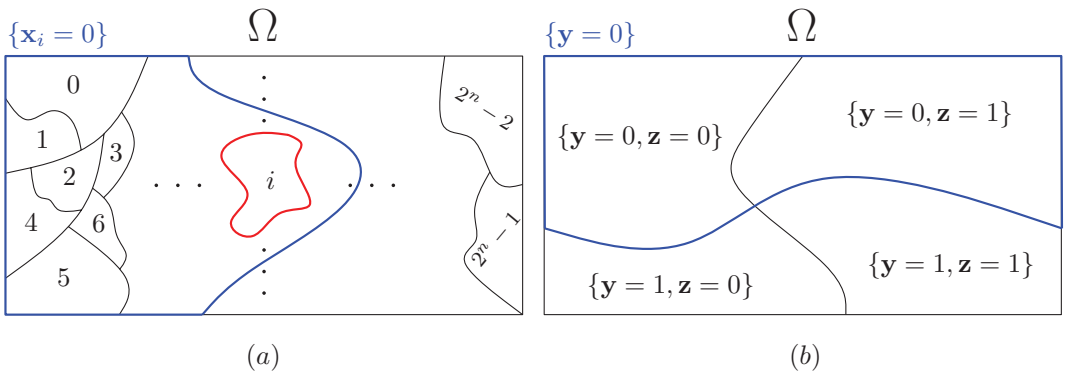



**Fig. 7.3** Sample Space: (a) Partitioning induced on $\Omega$ due to $\mathbf{X} = \mathbf{x}_{1:n}$. The blue region corresponds to the partition associated to the event $\{\mathbf{x}_i = 0\}$ and the red one to that of $\{\mathbf{X} = i\}$ where $i \in \mathit{Val}(\mathbf{X})$. (b) Partitioning induced on $\Omega$ due to RVs $\mathbf{y}$ and $\mathbf{z}$. The blue region corresponds to the partition associated to the event $\{\mathbf{y} = 0\}$.

## A-IV More on Scale-Invariance Property & Algorithm $\mathcal{I}^*_{scale}$

In what follows, drawing on the visualization methodology discussed in Sec. A-III, we present a series of examples to provide further intuition as to the scale-invariance property. As we well see in the following, the scale-invariance property is rooted in a key notion which we refer to as *orthogonality*. Finally, we present a variant of $\mathcal{I}^*_{non-scale}$, called $\mathcal{I}^*_{scale}$, which aims at minimally encoding the RVs identified by $\mathcal{I}^*_{non-scale}$ into the LP. As we will see, $\mathcal{I}^*_{scale}$ will use $\mathcal{I}^*_{non-scale}$ as a subroutine.

As our first example, Let us consider the MCM depicted in Fig. 7.4(a). The $\mathbb{P}(X, Y, Z)_{\downarrow}$ be the query of interest. Notice that RVs $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ will be identified by steps (1) to (2b) of $\mathcal{I}^*_{non-scale}$ for the aforesaid query. Adopting the same visualization method discussed in Sec. 7.3 (see Figs 7.4(b-c)), deriving $\mathbb{P}(X, Y, Z)_{\downarrow}$ amounts to minimizing the extent the two partitions $\{\mathbf{X} = X, \mathbf{Z} = Z\}$ and $\{\mathbf{Y} = Y, \mathbf{Z} = Z\}$ overlap. We refer to the aforesaid overlap as the solution region. Thus, $\mathbb{P}(X, Y, Z)_{\downarrow} = [\mathbb{P}(Z) - (\mathbb{P}(X, Z) + \mathbb{P}(Y, Z))]^+$.[3] The key understanding is that for the derivation of $\mathbb{P}(X, Y, Z)_{\downarrow}$ (and $\mathbb{P}(X, Y, Z)_{\uparrow}$), all the variables $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ can be treated as bi-valued variables where each either takes the value appearing in the posed query (e.g., $\mathbf{X} = X$) or the complement of it (e.g., $\mathbf{X} \neq X$ also denoted by

[3]Following the same line of reasoning, for $\mathbb{P}(X, Y, Z)_{\uparrow}$ one would get $\mathbb{P}(X, Y, Z)_{\uparrow} = \min\{\mathbb{P}(X, Z), \mathbb{P}(Y, Z)\}$.

$\mathbf{X} = \bar{X}$) thereby clumping all the other realizations together. It is crucial to notice that the solution region for $\mathbb{P}(X, Y, Z)_{\downarrow}$ has no intersection with any of the partitions corresponding to the aforesaid complements. We refer to this phenomenon as *orthogonality* defined as follows: a partition is orthogonal to the solution region iff it has zero intersection with the solution region. Knowing that the partitions associated to the aforesaid complements are orthogonal to the solution region for $\mathbb{P}(X, Y, Z)_{\downarrow}$ (and likewise $\mathbb{P}(X, Y, Z)_{\uparrow}$) allows us to safely encode $\mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ as bi-valued variables in stating the problem as an LP.

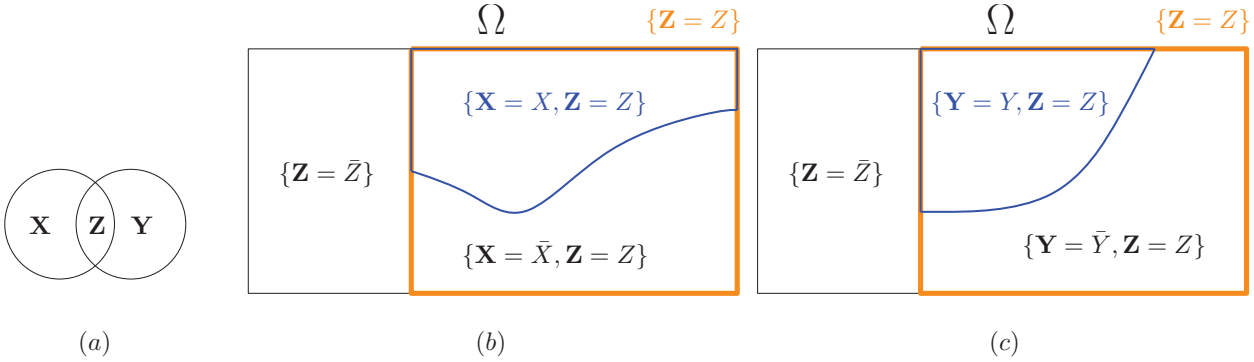



**Fig. 7.4** Sample Space: **(a)** MCM representing two overlapping contexts $\mathbb{P}(\mathbf{X}, \mathbf{Z})$ and $\mathbb{P}(\mathbf{Y}, \mathbf{Z})$. **(b)** Partitioning induced on $\Omega$ due to RVs $\mathbf{X}$ and $\mathbf{Z}$. **(c)** Partitioning induced on $\Omega$ due to RVs $\mathbf{X}$ and $\mathbf{Z}$. The orange region corresponds to the partition associated to the event $\{\mathbf{Z} = Z\}$.

Let us once again consider the MCM depicted in Fig. 7.4(a) but this time let $\mathbb{P}(X, Y)_{\downarrow}$ be the query of interest. Note that the only difference between the query under study here and that of the previous example is the omission of $\mathbf{Z}$. Notice that, similar to the previous example, RVs $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ will be identified by steps (1) to (2b) of $\mathcal{I}^{*}_{non-scale}$ for the query of interest. To answer $\mathbb{P}(X, Y)_{\downarrow}$, in every partition corresponding to a realization of $\mathbf{Z}$, we need to carry out the same line of reasoning adopted in the previous example. In other words, for every $Z \in \mathit{Val}(\mathbf{Z})$ and the partition thereof, the extent to which the two partitions $\{\mathbf{X} = X, \mathbf{Z} = Z\}$ and $\{\mathbf{Y} = Y, \mathbf{Z} = Z\}$ overlap needs to be minimized. Thus, $\mathbb{P}(X, Y)_{\downarrow} = \sum_{Z \in \mathit{Val}(\mathbf{Z})} \mathbb{P}(X, Y, Z)_{\downarrow} = \sum_{Z \in \mathit{Val}(\mathbf{Z})} [\mathbb{P}(Z) - (\mathbb{P}(X, Z) + \mathbb{P}(Y, Z))]^{+}$.[4] The key understanding which follows form this result is the following. To derive $\mathbb{P}(X, Y)_{\downarrow}$, similar to the previous example, variables $\mathbf{X}$ and $\mathbf{Y}$ can be treated as bi-valued. However, in contrast to the previous example, variable $\mathbf{Z}$ can no longer be treated as bi-valued; that is, in stating

[4] Following the same line of reasoning, for $\mathbb{P}(X, Y)_{\uparrow}$ one would get $\mathbb{P}(X, Y)_{\uparrow} = \sum_{Z \in \mathit{Val}(\mathbf{Z})} \min\{\mathbb{P}(X, Z), \mathbb{P}(Y, Z)\}$.

the problem as an LP, we cannot clump any of the realizations of $\mathbf{Z}$ together. This key understanding manifests itself in the very act of summing over all the realizations of $\mathbf{Z}$ in the expression given for $\mathbb{P}(X,Y)_{\downarrow}$ above. This should come as no surprise noticing that *none* of the partitions $\{\mathbf{Z}=Z\}, \forall Z \in \mathit{Val}(\mathbf{Z})$, is orthogonal to the solution region for $\mathbb{P}(X,Y)_{\downarrow}$ (and, likewise, for $\mathbb{P}(X,Y)_{\uparrow}$). Yet, notice that the partitions $\{\mathbf{X}=\bar{X}\}$ as well as $\{\mathbf{Y}=\bar{Y}\}$ are orthogonal to the solution region for $\mathbb{P}(X,Y)_{\downarrow}$ (and, likewise, $\mathbb{P}(X,Y)_{\uparrow}$), allowing us—akin to the previous example—to safely encode $\mathbf{X},\mathbf{Y}$ as bi-valued variables in stating the problem as an LP.

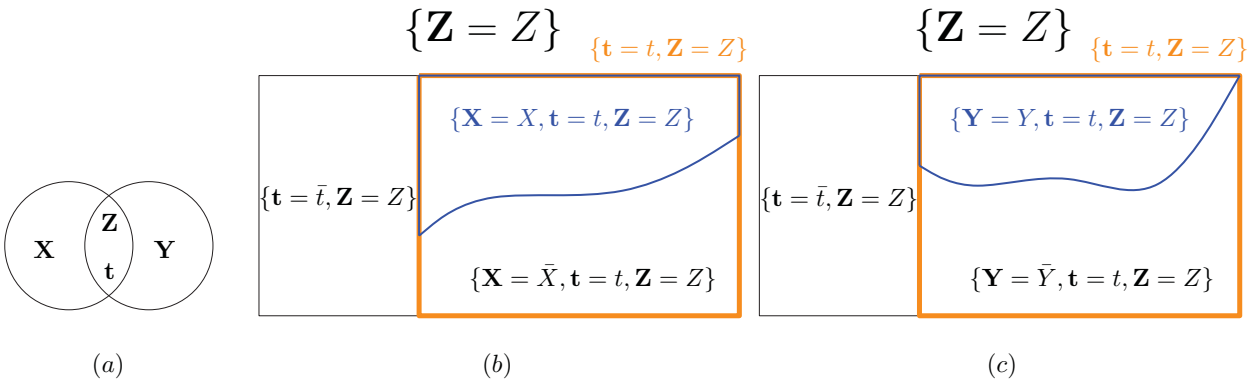



**Fig. 7.5** Sample Space: **(a)** MCM representing two overlapping contexts $\mathbb{P}(\mathbf{X},\mathbf{Z},\mathbf{t})$ and $\mathbb{P}(\mathbf{Y},\mathbf{Z},\mathbf{t})$. **(b)** Partitioning induced on $\{\mathbf{Z}=Z\}$ due to RVs $\mathbf{X}$ and $\mathbf{t}$. **(c)** Partitioning induced on $\{\mathbf{Z}=Z\}$ due to RVs $\mathbf{Y}$ and $\mathbf{t}$. The orange region corresponds to the partition associated to the event $\{\mathbf{t}=t,\mathbf{Z}=Z\}$.

As our final example, let us consider the MCM depicted in Fig. 7.5(a) and let $\mathbb{P}(X,Y,Z)_{\downarrow}$ be the query of interest. Note that the only difference between the MCM under study here and that of the previous example is the addition of RV $\mathbf{t}$ in the overlapping region between the two contexts. Notice also that RVs $\mathbf{X},\mathbf{Y},\mathbf{Z}$ and $\mathbf{t}$ will be all identified by steps (1) to (2b) of $\mathcal{I}^{*}_{non-scale}$ for the aforesaid query. Following the same line of reasoning adopted in the previous example yields: $\mathbb{P}(X,Y,Z)_{\downarrow} = \sum_{t\in \mathit{Val}(\mathbf{t})} \mathbb{P}(X,Y,Z,t)_{\downarrow} = \sum_{t\in \mathit{Val}(\mathbf{t})} [\mathbb{P}(Z,t) - (\mathbb{P}(X,Z,t) + \mathbb{P}(Y,Z,t))]^{+}$.[5] The key understanding which follows form this result is the following. To derive $\mathbb{P}(X,Y,Z)_{\downarrow}$, similar to the first example, RVs $\mathbf{X},\mathbf{Y}$ and $\mathbf{Z}$ (despite being in the overlapping region) can be treated as bi-valued, however, RV $\mathbf{t}$ cannot be treated as bi-valued in stating the problem as an LP. Indeed this understanding complies with the following key observation that all the partitions $\{\mathbf{X}=\bar{X}\}, \{\mathbf{Y}=\bar{Y}\}$, and $\{\mathbf{Z}=\bar{Z}\}$

[5] Following the same line of reasoning, for $\mathbb{P}(X,Y,Z)_{\uparrow}$ one would get $\mathbb{P}(X,Y,Z)_{\uparrow} = \sum_{t\in \mathit{Val}(\mathbf{t})} \mathbb{P}(X,Y,Z,t)_{\uparrow} = \sum_{t\in \mathit{Val}(\mathbf{t})} \min\{\mathbb{P}(X,Z,t), \mathbb{P}(Y,Z,t)\}$.

are orthogonal to the solution region whilst the partition $\{\mathbf{t} = \bar{t}\}$ does not meet this criterion.

We are now ready to present the algorithm $\mathcal{I}^*_{scale}$ as a variant of $\mathcal{I}^*_{non-scale}$ which enjoys the scale-invariance property. Before we proceed further, let us introduce the following notations: $id(\mathbb{P}(P|Q))$ denotes the set of RVs which would be identified by steps (1) to (2b) of $\mathcal{I}^*_{non-scale}$, had the query been $\mathbb{P}(P|Q)$. For query $\mathbb{P}(O|E)$, $\mathcal{I}^*_{scale}$ consists of two simple steps. The purpose the first step, drawing on the key notion of orthogonality, is to identify which RVs can be treated as bi-valued and which cannot in stating the problem as an LP, which takes place is step (2). $\mathcal{I}^*_{scale}$ is sketched bellow.

(1) Form the following sets:
   (1a) $\textsc{Expand}_1 := id(\mathbb{P}(O, E)) \setminus (\mathbf{O} \cup \mathbf{E})$.
   (1b) $\textsc{Expand}_2 := id(\mathbb{P}(E)) \setminus \mathbf{E}$.
   (1c) $\textsc{Expand}_{total} := \textsc{Expand}_1 \cup \textsc{Expand}_2$.
   (1d) $\textsc{ort}_{total} := id(\mathbb{P}(O|E)) \setminus \textsc{Expand}_{total}$.

(2) In stating the problem as an LP, parameterize RVs in $id(\mathbb{P}(O|E))$ in the following manner: Parameterize RVs in $\textsc{ort}_{total}$ as bi-valued and parameterize the rest of the identified RVs, $\textsc{Expand}_{total}$, as they are along with all their realizations—that is, without clumping any of their realizations together.

The justification for $\mathcal{I}^*_{scale}$ is elaborated next. Knowing that $\mathbb{P}(O|E) = \frac{\mathbb{P}(O,E)}{\mathbb{P}(E)}$, step (1a) identifies the RVs, $\textsc{Expand}_1$, which cannot be treated as bi-valued in stating the problem as an LP for the query $\mathbb{P}(O, E)$, and step (1b) identifies the RVs, $\textsc{Expand}_2$, which cannot be treated as bi-valued in stating the problem as an LP for the query $\mathbb{P}(E)$. Step (1c), by combining the identified variables in steps (1a) and (1b), identifies the RVs, $\textsc{Expand}_{total}$, which cannot be treated as bi-valued in stating the problem as an LP for the main query $\mathbb{P}(O|E)$. The rationale behind step (1c) is the following: If an RV, $\mathbf{L}$, is identifies in either step (1a) or (1b) as one which is not allowed to be treated as bi-valued, then $\mathbf{L}$ cannot be treated as bi-valued in stating the problem as an LP for the main query $\mathbb{P}(O|E)$. Finally, based on the result of step (1c) and the notion of orthogonality, step (1d) identifies the RVs, $\textsc{ort}_{total}$, which can be treated as bi-valued in stating the problem as an LP for the main query $\mathbb{P}(O|E)$.

**Remark A.1.** Let $\mathbb{P}(O|E)$ denote the posed query. If $\mathbb{P}(E)$ happens to be an intra-contextual quantity, then, since $\mathbb{P}(O|E) = \frac{\mathbb{P}(O,E)}{\mathbb{P}(E)}$, deriving the minimum (maximum) for

$\mathbb{P}(O|E)$ amounts to finding the minimum (maximum) for $\mathbb{P}(O, E)$ and subsequently dividing it by the quantity $\mathbb{P}(E)$.

# Appendix B

## B-I On Minimax and Maximin Objectives

In this section we claim that, subject to the constraint that the IP's of the to-be-intervened variables has to belong to IP *class-j*, the solution to both minimax and maximin problem is the empty set, for all $j \geq 1$. Let us present a lemma using which it is easy to justify the claim made above.

**Lemma B.1.** *Let DAG $G = (V, E)$ represent the causal structure of the domain. $\forall j \geq 1$ and $\forall \boldsymbol{\mathcal{X}} \subseteq V_i$, the following inequalities hold:*

$$\min_{ip(\mathcal{X}) \in \textit{class-}j} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \leq \mathbb{P}(\mathcal{O}),$$
$$\mathbb{P}(\mathcal{O}) \leq \max_{ip(\mathcal{X}) \in \textit{class-}j} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]).$$

***Proof.*** The proof for Lemma B.1 is straightforward due to the realization that, $\forall j \geq 1$, $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}) \in \textit{class-}j)_G$ $i$-subsumes the original CBN, thus $(\boldsymbol{\mathcal{X}}, ip(\boldsymbol{\mathcal{X}}) \in \textit{class-}j)_G \succcurlyeq_i (\varnothing, \varnothing)_G$, $\forall j \geq 1$. ■

Using Lemma B.1, the next inequalities follow: $\forall j \geq 1$,

$$\max_{\boldsymbol{\mathcal{X}} \subseteq V_i} \left( \min_{ip(\mathcal{X}) \in \textit{class-}j} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \right) \leq \mathbb{P}(\mathcal{O}),$$
$$\mathbb{P}(\mathcal{O}) \leq \min_{\boldsymbol{\mathcal{X}} \subseteq V_i} \left( \max_{ip(\mathcal{X}) \in \textit{class-}j} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})]) \right).$$

One can readily conclude from the above inequalities the claim made in the paper as to the solution to the minimax and maximin problems; the solution to both is the empty set.

## B-II Algorithm $\mathcal{C}^*$: Proof

**On the Sufficiency of $\boldsymbol{\mathcal{X}}^*$:** Next, we present the proof for sufficiency of $\boldsymbol{\mathcal{X}}^*$ where by sufficiency we mean the following: Intervening (according to IP *class*-$\infty$) on any variables in addition to $\boldsymbol{\mathcal{X}}^*$ does not yield any improvement upon what is achievable through merely intervening (according to IP *class*-$\infty$) on $\boldsymbol{\mathcal{X}}^*$. Let $V_i = V_i^{BC} \cup \bar{V}_i^{BC}$ where $V_i^{BC}$ is the set of intervenable variables which belong to the subgraph generated by executing BC on the target nodes $\boldsymbol{\mathcal{O}}$. It is obvious that intervening on any variable in $\bar{V}_i^{BC}$ is pointless due to the following argument[1]: ($\star$) variables in $\bar{V}_i^{BC}$ have no direct or indirect causal effect on any of the target nodes. Now, what is left to be shown is why, among all variables in $V_i^{BC}$, it suffices to intervene on $\boldsymbol{\mathcal{X}}^*$ (according to IP *class*-$\infty$) or, differently put, why intervening on any additional variables does not improve upon what is achievable through intervening merely on $\boldsymbol{\mathcal{X}}^*$ (according to IP *class*-$\infty$). Formally, the question is why the following holds: $\forall \boldsymbol{\mathcal{Y}} \subseteq V_i^{BC}$

$$(\boldsymbol{\mathcal{X}}^*, ip(\boldsymbol{\mathcal{X}}^*) \in \textit{class-}\infty)_G \not\succcurlyeq_i (\boldsymbol{\mathcal{Y}}, ip(\boldsymbol{\mathcal{Y}}) \in \textit{class-}\infty)_G.$$

Notice that, the extension to the case of $\forall \boldsymbol{\mathcal{Y}} \subseteq V_i$ immediately follows from argument ($\star$). The realization of the fact that variable $\mathbf{y} \in V_i^{BC}$ was not selected (for intervention) by $\mathcal{C}^*$ implies that $\mathbf{y}$'s causal effect on the target variables which are descendants[2] of $\mathbf{y}$ must have been mediated through some of the nodes selected by $\mathcal{C}^*$ say $\boldsymbol{\mathcal{Y}}^\dagger \subseteq \boldsymbol{\mathcal{X}}^*$ (otherwise, $\mathbf{y}$ would have been selected). The claim as to the redundancy of further intervening on $\mathbf{y}$ in addition to exerting intervention (according to IP *class*-$\infty$) on $\boldsymbol{\mathcal{Y}}^\dagger$ is supported as follows. Let $G = (V, E)$ be the DAG associated to the (non-intervened) underlying causal structure of the domain. First, notice that evaluating $\mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}; ip(\mathcal{X})])$ amounts to simply deriving $\mathbb{P}(\mathcal{O})$ in the $i$-DAG the dash-dotted edges of which are parameterized in accord with the IP $ip(\mathcal{X})$; cf. (Tian, 2008), (Pearl, 2000, pp. 113-114) and (Pearl, 1995, p. 684). Our goal is to show the following: (i) There exits a setting (characterized by an $i$-DAG) wherein $\mathbf{y}$ is *not* intervened but every variable $\mathbf{y}^\dagger \in \boldsymbol{\mathcal{Y}}^\dagger$ is intervened according to an IP into the scope of which neither $\mathbf{y}$ nor the set of $\mathbf{y}$'s descendants which lie on a (directed) path from $\mathbf{y}$ to $\mathbf{y}^\dagger$ (denoted by the set $\boldsymbol{\mathcal{M}}(\mathbf{y}, \mathbf{y}^\dagger)$) is included, and (ii) the $i$-DAG in (i) $i$-dominates an $i$-DAG in which variables in $\mathbf{y} \cup \boldsymbol{\mathcal{Y}}^\dagger$ are intervened according to an *arbitrarily*-parameterized

[1] This statement immediately follows from Rule 3 of Pearl's *do*-calculus (cf. Pearl, 2000, p. 95).

[2] Intervening on $\mathbf{y}$, obviously, could only influence $\mathbf{y}$'s descendants.

IP *class*-$\infty$. The fact that the above line of reasoning can be carried out for any such $\mathbf{y}$ and its corresponding set $\boldsymbol{\mathcal{Y}}^{\dagger}$ grants the sufficiency of $\boldsymbol{\mathcal{X}}^{*}$. In the argument presented next, we make use of the following result whose proof is provided in Sec. B-III: ($\star\star$) Among all optimal IPs which can be exercised on intervenable variable $\mathbf{x}$, there exists one which is of deterministic nature. Next, we present the ideas allowing us to arrive at the setting discussed above. Let us assume—hypothetically of course—that, in addition to $V_i$, all the variables in $\cup_{\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}}\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger})$ are also intervenable; we will return to the rationale behind this assumption (referred to by (A.1)) shortly. Using the result stated in ($\star\star$), it then follows that, $\forall\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}$, there exists an optimal IP to be exercised on $\mathbf{y}^{\dagger}$ (which is deterministic) into the scope of which neither $\mathbf{y}$ nor any of the variables in $\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger})$ is included—these exclusions are made possible due to the determinism of the IP.[3] Once the aforesaid IPs are exercised on variables in $\boldsymbol{\mathcal{Y}}^{\dagger}$, intervention (according to IP *class*-$\infty$) on any variable in $(\cup_{\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}}\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger}))\cup\mathbf{y}$ yields no effect on any target variables—in the $i$-DAG characterizing the very setting under study, none of the variables in $(\cup_{\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}}\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger}))\cup\mathbf{y}$ has any direct or indirect causal effect on any of the target variables. Hence, we can intervene on the variables in $(\cup_{\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}}\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger}))\cup\mathbf{y}$ in any way we may wish. Let us intervene on variables in $(\cup_{\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}}\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger}))\cup\mathbf{y}$ in such a manner that the IP exerted on each is identical to the corresponding CPD in the (non-intervened) CBN—hence returning to the state of exerting no interventions on variables in $(\cup_{\mathbf{y}^{\dagger}\in\boldsymbol{\mathcal{Y}}^{\dagger}}\boldsymbol{\mathcal{M}}(\mathbf{y},\mathbf{y}^{\dagger}))\cup\mathbf{y}$ from the point of view of the original (non-intervened) CBN. At this moment the setting discussed in (ii) is achieved. This concludes the proof once we realize that the assumption (A.1) amounts to having an $i$-DAG which, due to Lemma 4.2, $i$-dominates the $i$-DAG which corresponds to having only the variables in $V_i$ to be intervenable. ■

## B-III On Deterministic vs Stochastic Intervention Policies

Let $\mathbf{x}$ be an intervenable variable, $ip(\mathbf{x})$ be the IP to be exercised on $\mathbf{x}$, and set $\boldsymbol{\mathcal{S}}_{\mathbf{x}}$ denote the scope of $ip(\mathbf{x})$. For the case of a deterministic IP, given a realization of the variables involved in the scope, say $\boldsymbol{\mathcal{S}}_{\mathbf{x}}=\mathcal{S}_{\mathbf{x}}$, the state of $\mathbf{x}$ becomes fully determined (i.e., with probability one). Hence, for the case of exerting a deterministic IP on $\mathbf{x}$, the following holds: $ip(\mathbf{x})=g(\boldsymbol{\mathcal{S}}_{\mathbf{x}})$ where $g(\cdot)$ denotes the corresponding deterministic function; see (Pearl, 2000, p. 113). On the other hand, for the case of exercising an stochastic IP on $\mathbf{x}$, given a realization of the variables involved in the scope, say $\boldsymbol{\mathcal{S}}_{\mathbf{x}}=\mathcal{S}_{\mathbf{x}}$, the state of $\mathbf{x}$ is specified probabilistically

[3]For more elaboration on this, the reader is referred to the last paragraph of Sec. B-III.

according to $\mathbb{P}(\mathbf{x}|\boldsymbol{\mathcal{S}}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}})$.

Although deterministic IPs are a "special case" of stochastic IPs, as we will see in the next lemmas, for maximax and minimin objectives, adopting stochastic IPs does not yield any advantage over using only deterministic IPs. More formally, among all optimal IPs which can be exercised on intervenable variable $\mathbf{x}$, there exists one which is of deterministic nature.[4]

**Lemma B.2.** *Let $G = (V, E)$ denote the causal structure of the domain. Let $\boldsymbol{x}$ be an intervenable variable. Let $\mathbb{P}(\boldsymbol{x}|\boldsymbol{\mathcal{S}}_{\boldsymbol{x}})$ denote the IP to be exercised on $\boldsymbol{x}$ where $\boldsymbol{\mathcal{S}}_{\boldsymbol{x}}$ denote the scope of $\boldsymbol{x}$'s IP which is a subset of $\boldsymbol{x}$'s ancestors. Then, among all possible parameterizations of $\mathbb{P}(\boldsymbol{x}|\boldsymbol{\mathcal{S}}_{\boldsymbol{x}})$ which are optimal with respect to maximax objective, there exists one of deterministic nature. That is, given a realization of $\boldsymbol{\mathcal{S}}_{\boldsymbol{x}} = \mathcal{S}_{\boldsymbol{x}}$, the state of $\boldsymbol{x}$ becomes fully determined with probability one. Hence, $\boldsymbol{x} := g(\boldsymbol{\mathcal{S}}_{\boldsymbol{x}})$ where $g(\cdot)$ is some deterministic function.*

***Proof.*** First, notice that the effect of stochastic policies can be expressed in terms of atomic interventions as explained in (Pearl, 2000, pp. 113-114) and (Pearl, 1995, p. 684). The lemma then follows from a simple understanding that the query $\mathbb{P}(\boldsymbol{\mathcal{O}}|do[\mathbf{x}; \mathbb{P}(\mathbf{x}|\boldsymbol{\mathcal{S}}_{\mathbf{x}})])$ is a linear function of the parameters involved in the CPD $\mathbb{P}(\mathbf{x}|\boldsymbol{\mathcal{S}}_{\mathbf{x}})$. For more elaborations on the aforesaid linear functional dependence, the reader is referred to Chan (2005) where this idea is discussed under the topic of "sensitivity analysis of Bayesian networks" (see Chan and Darwiche, 2001; Castillo et al., 1997; Russell et al., 1995). ■

Similar result can be established for the minimin objective given in (2) as presented in the next lemma.

**Lemma B.3.** *Let $G = (V, E)$ denote the causal structure of the domain. Let $\boldsymbol{x}$ be an intervenable variable. Let $\mathbb{P}(\boldsymbol{x}|\boldsymbol{\mathcal{S}}_{\boldsymbol{x}})$ denote the IP to be exercised on $\boldsymbol{x}$ where $\boldsymbol{\mathcal{S}}_{\boldsymbol{x}}$ denote the scope of $\boldsymbol{x}$'s IP which is a subset of $\boldsymbol{x}$'s ancestors. Then, among all possible parameterizations of $\mathbb{P}(\boldsymbol{x}|\boldsymbol{\mathcal{S}}_{\boldsymbol{x}})$ which are optimal with respect to minimin objective, there exists one of deterministic nature. That is, given a realization of $\boldsymbol{\mathcal{S}}_{\boldsymbol{x}} = \mathcal{S}_{\boldsymbol{x}}$, the state of $\boldsymbol{x}$ becomes fully determined with probability one. Hence, $\boldsymbol{x} := h(\boldsymbol{\mathcal{S}}_{\boldsymbol{x}})$ where $h(\cdot)$ is some deterministic function.*

***Proof.*** The same line of reasoning provided for the proof of Lemma B.2 applies here as well. ■

The understanding captured in Lemmas B.2 and B.3 plays an important role in the argument provided on the sufficiency of $\boldsymbol{\mathcal{X}}^*$ in Sec. B-II. Using the notation adopted therein,

[4]For the reader familiar with game theory, there is an interesting analogy between the aforesaid statement and the following result in game theory (Polak, 2007): "If a mixed strategy is a best response then each of the pure strategies involved in the mix must itself be a best response."

there exists an optimal intervention (according to IP *class*-$\infty$) to be exercised on $\mathbf{y} \in V_i^{BC}$—which was not selected by $\mathcal{C}^*$—which amounts to having $\mathbf{y} := f(anc(\mathbf{y}))$ where $anc(\mathbf{y})$ denotes the set of ancestors of $\mathbf{y}$ and $f(\cdot)$ denote some deterministic function. Simply put, such optimal intervention on $\mathbf{y}$ makes $\mathbf{y}$ a deterministic function of its ancestors. Following the same line of reasoning, there exists an optimal intervention (according to IP *class*-$\infty$) to be exercised on each $\mathbf{y}^\dagger \in \boldsymbol{\mathcal{Y}}^\dagger$ which makes $\mathbf{y}^\dagger$ a deterministic function of its ancestors. Notice that $\mathbf{y}$ is an ancestor for any variable in $\boldsymbol{\mathcal{Y}}^\dagger$. Hence follows: $\forall \mathbf{y}^\dagger \in \boldsymbol{\mathcal{Y}}^\dagger, anc(\mathbf{y}) \subset anc(\mathbf{y}^\dagger)$. The latter immediately implies that the inclusion of $\mathbf{y}$ into the scope of IP of each variable $\mathbf{y}^\dagger \in \boldsymbol{\mathcal{Y}}^\dagger$ is pointless and hence $\mathbf{y}$ can be safely removed from the scope of any $\mathbf{y}^\dagger$'s IP.

## B-IV Optimal Intervention Policy: Computational Complexity

In what follows, we elaborate on the computational complexity of the problem of finding the Optimal Intervention Policy (OIP) for the purpose of probabilistic controllability (P-controllability) of CBNs. More specifically, we show that, under both maximax and minimin objectives the aforesaid problem is $\mathcal{NP}$-hard. This is accomplished by showing that OIP contains a subproblem that is $\mathcal{NP}$-complete under both maximax and minimin objectives. This subproblem is constructed by selecting a special class of CBNs which will be denoted by $\mathfrak{B}_{G_0}$. This proof technique is known as *proof by restriction*.

Let us formally define the function problems the complexity of which we are interested in investigating, namely, OIP-MAXMAX-FP and OIP-MINMIN-FP.

**Def. B.1.** (OIP-MAXMAX-FP): Given a CBN $\mathcal{B}$ with causal structure $G$, parameterized by distribution $\mathbb{P}$ (which factorizes over $G$), and the corresponding set $\boldsymbol{\mathcal{X}}^*$, output the OIP to be exercised on $\boldsymbol{\mathcal{X}}^*$, that is the IP which is optimal with respect to maximax objective given in (1) in the paper, i.e., $\arg\max_{ip(\mathcal{X}^*)\in class\text{-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}^*; ip(\mathcal{X}^*)])$.

**Def. B.2.** (OIP-MINMIN-FP): Given a CBN $\mathcal{B}$ with causal structure $G$, parameterized by distribution $\mathbb{P}$ (which factorizes over $G$), and the corresponding set $\boldsymbol{\mathcal{X}}^*$, output the OIP to be exercised on $\boldsymbol{\mathcal{X}}^*$, that is the IP which is optimal with respect to minimin objective given in (2) in the paper, i.e., $\arg\min_{ip(\mathcal{X}^*)\in class\text{-}\infty} \mathbb{P}(\mathcal{O}|do[\boldsymbol{\mathcal{X}}^*; ip(\mathcal{X}^*)])$.

In what follows, we restrict our attention to a class of CBNs denoted by $\mathfrak{B}_{G_0}$ which we finally use to prove $\mathcal{NP}$-hardness results for both OIP-MAXMAX-FP and OIP-MINMIN-FP. $\mathfrak{B}_{G_0}$ is formally defined next.

**Def. B.3. (Class $\mathfrak{B}_{G_0}$)**: Let $\mathfrak{B}_{G_0}$ be a class of CBNs defined over binary variables $\{\mathbf{X} =$

$\mathbf{x}_{1:n}, \mathbf{y}\}$ with causal structure $G_0$ and parameterized by a set of degenerate prior distributions on input variables $\mathbf{x}_i$'s such that $\forall i$, $\mathbb{P}(\mathbf{x}_i = 0) = 1$, and a Boolean formula $\phi$ with size $m$ where $\phi(X) = \mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X))$. Hence, $\mathbb{P}(X, y) = [1 - \phi(X)]^{(1-y)}[\phi(X)]^y \mathbb{1}(X \in \{0\}^n)$ where $\mathbb{1}(\cdot)$ is the indicator function. The DAG $G_0$ enjoys the $n$-to-1 topology depicted in Fig. 7.6(a). Every input variable $\mathbf{x}_i \in \mathbf{X}$ is intervenable. Variable $\mathbf{y}$ (also called the output variable) is not intervenable. A generic member of $\mathfrak{B}_{G_0}$ is depicted in Fig. 7.6(b).

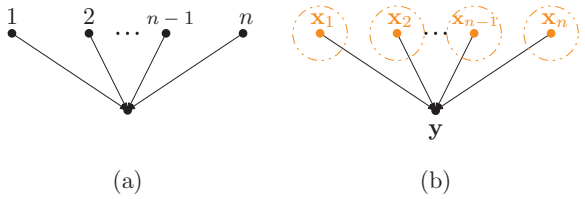



**Fig. 7.6** **(a)** DAG $G_0$ enjoys the $n$-to-1 topology as depicted where $n$ denotes the number of input variables $\mathbf{x}_i$'s. **(b)** A generic member of the class $\mathfrak{B}_{G_0}$. Circled variables are intervenable.

Note that every member of $\mathfrak{B}_{G_0}$, denoted by $(G_0, \phi) \in \mathfrak{B}_{G_0}$, is uniquely characterized by its corresponding Boolean formula $\phi$. Also note that, for any $(G_0, \phi) \in \mathfrak{B}_{G_0}$, algorithm $\mathcal{C}^*$ outputs the corresponding input variables $\mathbf{X}$ over which the CBN $(G_0, \phi)$ is defined; hence, $\boldsymbol{\mathcal{X}}^* = \mathbf{X}$.

### B-IV.I P-Controllability for $\mathfrak{B}_{G_0}$

In what follows, we elaborate on the problem of P-controllability under maximax and min-imin objectives for the class $\mathfrak{B}_{G_0}$.

#### B-IV.I.I P-Controllability under Maximax Objective for $\mathfrak{B}_{G_0}$

The problem of P-controllability under the maximax objective for a CBN $(G_0, \phi) \in \mathfrak{B}_{G_0}$ can be cast as follows:

$$\max_{X \in \{0,1\}^n} \mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X)) = \max_{X \in \{0,1\}^n} \phi(X).$$

Thus, the OIP under maximax objective for a CBN $(G_0, \phi) \in \mathfrak{B}_{G_0}$ can be stated as follows:

$$\arg\max_{X \in \{0,1\}^n} \mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X)) = \arg\max_{X \in \{0,1\}^n} \phi(X).$$

Next, we elaborate on the complexity of finding the OIP when the objective of interest is maximax. Let us formally define the corresponding decision problem, OIP-MAXMAX-DP$_{G_0}$, as follows.

**Def. B.4.** Given a CBN $(G_0, \phi) \in \mathfrak{B}_{G_0}$, decide whether there exists an atomic IP, $do(\mathbf{X} = X^*)$, such that $\mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X^*)) = 1$. Hence, let

$$\text{OIP-MAXMAX-DP}_{G_0} = \{(G_0, \phi) \mid \exists X^* \text{ s.t. } \mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X^*)) = 1\}.$$

Let SAT $= \{\phi \mid \phi$ is a satisfiable Boolean formula$\}$. Lemma B.4 then follows.

**Lemma B.4.** OIP-MAXMAX-DP$_{G_0}$ is polynomial-time equivalent to SAT.

***Proof.*** Recall that for any $(G_0, \phi) \in \mathfrak{B}_{G_0}$ holds $\mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X^*)) = \phi(X^*)$. Hence,

$$\begin{aligned} \text{OIP-MAXMAX-DP}_{G_0} &= \{(G_0, \phi) \mid \exists X^* \text{ s.t. } \phi(X^*) = 1\} \\ &= \{(G_0, \phi) \mid \phi \text{ is a satisfiable Boolean formula}\}. \end{aligned}$$

Note that given the $n$-to-1 topology of $G_0$ with $n$ denoting the size of the input variables, any instance $\phi \in$ SAT with size $m$ can be transformed into its corresponding instance, $(G_0, \phi) \in \mathfrak{B}_{G_0}$, in $O(n + m)$ time. The transformation of an instance $(G_0, \phi) \in \mathfrak{B}_{G_0}$ into its corresponding instance, $\phi \in$ SAT, can be accomplished in a straightforward manner in $O(m)$ time where $m$ denotes the size of the Boolean formula $\phi$. ■

**Corollary B.1.** From Lemma B.4 and $\mathcal{NP}$-completeness of SAT immediately follows that OIP-MAXMAX-DP$_{G_0}$ is $\mathcal{NP}$-complete.

Let us state the main result on the complexity of OIP-MAXMAX-FP as follows.

**Lemma B.5.** OIP-MAXMAX-FP is $\mathcal{NP}$-hard.

***Proof.*** The proof follows from Colloraly B.1 and the reduction of OIP-MAXMAX-DP$_{G_0}$ to OIP-MAXMAX-FP due the simple understanding that the class of CBNs, $\mathfrak{B}_{G_0}$, is a subset of the set of instances for which OIP-MAXMAX-FP is defined. ■

### B-IV.I.II P-Controllability under Minimin Objective for $\mathfrak{B}_{G_0}$

Likewise, the problem of P-controllability under minimin objective for $(G_0, \phi)$ can be cast as follows:

$$\min_{X \in \{0,1\}^n} \mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X)) = \min_{X \in \{0,1\}^n} \phi(X).$$

Thus, the OIP under minmin objective for $(G_0, \phi)$ can be stated as follows:

$$\arg\min_{X\in\{0,1\}^n} \mathbb{P}(\mathbf{y}=1|do(\mathbf{X}=X)) = \arg\min_{X\in\{0,1\}^n} \phi(X).$$

Next, we elaborate on the complexity of finding an OIP when the objective of interest is minimin. Let us formally define the corresponding decision problem, OIP-MINMIN-DP$_{G_0}$, as follows.

**Def. B.5.** Given a CBN $(G_0, \phi) \in \mathfrak{B}_{G_0}$, decide whether there exists an atomic intervention policy $do(\mathbf{X} = X^*)$ such that $\mathbb{P}(\mathbf{y} = 0|do(\mathbf{X} = X^*)) = 1$. Hence let

$$\text{OIP-MINMIN-DP}_{G_0} =$$
$$\{(G_0, \phi) \mid \exists X^* \text{ s.t. } \mathbb{P}(\mathbf{y}=1|do(\mathbf{X}=X^*)) = 0\}.$$

Also let us define the TAUTOLOGY decision problem, in its language representation, as follows.

**Def. B.6.** Let TAUTOLOGY $= \{\phi \mid \phi \text{ is always true}\}$. Hence,

$$\overline{\text{TAUTOLOGY}} = \{\phi \mid \exists X^* \text{ s.t. } \neg\phi(X^*) = 1\},$$

where $\overline{L}$ denotes the complement of the language $L$.

**Lemma B.6.** OIP-MINMIN-DP$_{G_0}$ is polynomial-time equivalent to $\overline{\text{TAUTOLOGY}}$.

***Proof.*** Recall that for any $(G_0, \phi) \in \mathfrak{B}_{G_0}$ holds $\mathbb{P}(\mathbf{y} = 1|do(\mathbf{X} = X^*)) = \phi(X^*)$. Hence,

$$\begin{aligned}
\text{OIP-MINMIN-DP}_{G_0} &= \{(G_0, \phi) \mid \exists X^* \text{ s.t. } \mathbb{P}(\mathbf{y}=1|do(\mathbf{X}=X^*)) = 0\} \\
&= \{\phi \mid \exists X^* \text{ s.t. } \phi(X^*) = 0\} \\
&= \{\phi \mid \exists X^* \text{ s.t. } \neg\phi(X^*) = 1\}.
\end{aligned}$$

The same argument provided in Lemma B.4 can be made as to the transformation of an instance of OIP-MINMIN-DP$_{G_0}$ to that of $\overline{\text{TAUTOLOGY}}$ and vice versa. ■

**Corollary B.2.** From Lemma B.6 and $\mathcal{NP}$-completeness[5] of $\overline{\text{TAUTOLOGY}}$, follows that OIP-MINMIN-DP$_{G_0}$ is $\mathcal{NP}$-complete.

Let us state the main result on the complexity of OIP-MINMIN-FP as follows.

**Lemma B.7.** OIP-MINMIN-FP is $\mathcal{NP}$-hard.

[5]Note that TAUTOLOGY is co$\mathcal{NP}$-complete (see Arora and Barak, 2009).

***Proof.*** The proof follows from Colloraly B.2 and the reduction of OIP-MINMIN-DP$_{G_0}$ to OIP-MINMIN-FP due the simple understanding that the class of CBNs, $\mathfrak{B}_{G_0}$, is a subset of the set of instances for which OIP-MINMIN-FP is defined. ■

## B-V Running-Time Analysis of $\mathcal{C}^*$

It is straightforward to implement $\mathcal{C}^*$ using the well-known Breadth First Search (BFS) algorithm: Simply scan the underlying DAG in a BFS fashion and, upon hitting an intervenable node $\mathbf{v} \in V_i$, identify it and do not scan the parents of $\mathbf{v}$. Finally, return the set of identified nodes as $\boldsymbol{\mathcal{X}}^*$. Hence, the worst-case running time of $\mathcal{C}^*$ is $O(|E| + |V|)$ where $|E|$ and $|V|$ denote, respectively, the number of edges and vertices of the underlying DAG (see Cormen et al., 2001).

# Appendix C

Throughout Appendix C, let $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ denote the posed $d$-separation query with DAG $G$ representing the topology of the underlying BN. Throughout the proofs and arguments to follow, it is assumed that communication channels are reliable, bidirectional, and first-in first-out (FIFO) (Lynch, 1996). For the time-complexity analysis of $\mathcal{D}^*$, we adhere to the same assumptions adopted in (Lynch, 1996). More specifically, we assume: (ASM-1) an upper-bound of $\alpha$ for a process to perform Steps (i) and (ii) upon receipt of a message, and (ASM-2) an upper-bound of $\beta$ on the delivery time for each message in a channel. Note that the parameters $\alpha$ and $\beta$ are arbitrary but finite constants. Also note that, as the number of messages exchanged by $\mathcal{D}^*$ on an edge is $O(1)$ (see Statement (5) of Proposition 5.1 in the main text), the effect of pileups (aka congestion) on a channel has been considered in Assumptions (ASM-1) and (ASM-2).[1]

## C-I $\mathcal{D}^*$: Proof of Correctness

In what follows, we prove three statements which, taken together, grant the correctness of $\mathcal{D}^*$. The three statements are given below.

(I) For a given $d$-separation query $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ and DAG $G$,

"$\mathbf{C}$ does not $d$-separate $\mathbf{A}$ from $\mathbf{B}$ in $G$" $\iff$ "Clash takes place during $\mathcal{D}^*$'s execution".

(II) During $\mathcal{D}^*$'s execution, either a clash between colors red (●) and green (●) takes place (cf. Remark 5.1 in the main text) upon which $\mathcal{D}^*$ decides that $(\mathbf{A} \not\perp\!\!\!\perp \mathbf{B}|\mathbf{C})$, or a state of equilibrium will be eventually reached.

[1] Since the number of messages exchanged by $\mathcal{D}^*$ on an edge is $O(1)$, the following holds: (a) the number of messages in any channel queue is at most $O(1)$, and (b) the number of messages awaiting in a process's send buffer is at most $O(1)$.

(III) Message-passing terminates in $O(1)$ time after reaching the state of equilibrium, thereby guaranteeing the termination of $\mathcal{D}^*$.

The proofs of Statements (I) to (III) are presented next.

### C-I.I Proof of Statement (I)

Next, the proof of Statement (I) is presented. First, the proof of the forward direction is outlined (Sec. C-I.I.I), followed by the proof of the backward direction (Sec. C-I.I.II).

#### C-I.I.I Proof of Statement (I): Forward Direction

We prove the forward direction of Statement (I) next. This is accomplished by proving the following: Conditioned on the set $\mathbf{C}$, if there exists an unblocked path between $\mathbf{a} \in \mathbf{A}$ and $\mathbf{b} \in \mathbf{B}$ (for any $\mathbf{a}, \mathbf{b}$), a clash of the kind stated in Remark 1 is unavoidable during $\mathcal{D}^*$'s execution. A path $l$ is said to be unblocked (Pearl, 1988) if and only if (a) for every collider node $\mathbf{n}$ on $l$, either $\mathbf{n}$ or some of $\mathbf{n}$'s descendants are in $\mathbf{C}$, and (b) for every non-collider node $\mathbf{m}$, $\mathbf{m} \notin \mathbf{C}$. The proof rests on a simple understanding that a generic unblocked path can be decomposed into v-structured and non-v-structured *modules* as illustrated in Fig. 7.7. Neighboring modules share a common vertex which we refer to as *joint vertex* (e.g., the nodes $\mathbf{j}_1, \mathbf{j}_2$ in Fig. 7.7(a)). The end-point vertex of a non-v-structured subpath which is not a joint vertex is termed *source vertex*; see Fig. 7.7(b1) and Fig. 7.7(b3). In principle, an unblocked path may have multiple v-structured modules. For ease of exposition, the unblocked path $p$ depicted in Fig. 7.7(a) possesses only one v-structured module. Note that the proof that follows does not make this restrictive assumption.

Next, we prove the inevitability of a clash for unblocked paths possessing non-v-structured as well as v-structured modules.[2] The proof comprises two parts. In Part I, we show the inevitability of a clash over such an unblocked path, $l^*$, provided that no message is destined from a node outside $l^*$ to a node belonging to $l^*$. Using the arguments provided in Part I, in Part II we show that *regardless* of the messages destined from nodes outside $l^*$ to the nodes belonging to $l^*$, the occurrence of a clash on $l^*$ is inevitable (i.e., eventually happens).

**Proof of Part I:** Rules $(\varnothing, \bullet) \rightarrow \bullet$, and $(\varnothing, \bullet) \rightarrow \bullet$ sketched in Step (ii) of $\mathcal{D}^*$, along with $\mathcal{D}^*$'s initialization phase wherein all the nodes in the sets $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ propagate their colors

[2]The adaptation of the argument for the other two cases where the unblocked path is solely comprised of either non-v-structured modules or v-structured modules is straightforward.

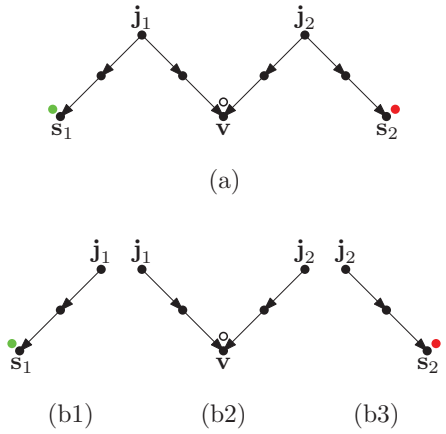



**Fig. 7.7** Decomposition of a generic unblocked path into v-structured and non-v-structured modules. **(a)** A generic unblocked path $p$ comprised of v-structured as well as non-v-structured modules. The nodes $\mathbf{s}_1$ and $\mathbf{s}_2$ are source vertices. The nodes $\mathbf{j}_1$ and $\mathbf{j}_2$ are joint vertices. The node $\mathbf{v}$ is a collider. Without loss of generality, $\mathbf{s}_1, \mathbf{s}_2$, and $\mathbf{v}$ are assumed to be initialized with colors green, red and white, respectively. **(b1)** A non-v-structured module of the unblocked path $p$ with the source vertex $\mathbf{s}_1$. **(b2)** The v-structured module of the unblocked path $p$. **(b3)** A non-v-structured module of the unblocked path $p$ with the source vertex $\mathbf{s}_2$.

to their parents, ensure that all non-v-structured modules are fully explored and, by the end of exploration, all the nodes within each non-v-structured module will be homogeneously colored consistent with that of the respective source vertex, except for the joint vertex which requires more careful consideration (†). The propagation of white (∘) through the DAG $G$ in a backward manner ensures that v-structured modules are fully explored and, by the end of exploration, all the nodes within each v-structured module will be homogeneously colored in white (∘), except for the joint vertices which require more careful consideration (‡). The consideration advised in (†) and (‡) is explicated next.[3] The joint vertex connecting a non-v-structured module to a v-structured module may first become white (∘) or whatever the color of the source vertex of the non-v-structured module is, depending on whether the joint vertex first receives a message from the non-v-structured module or the v-structured module, respectively. However, and quite importantly, its color eventually becomes that of the source vertex of the non-v-structured module and, according to Step (ii) of $\mathcal{D}^*$, it sends its color down the v-structured module. In short, any joint vertex $\mathbf{j}$ will eventually serve as a *relay*

[3]The analysis of the case for the joint vertex between two adjacent v-structured modules does not require special consideration since it will become white (∘) first and thereafter, according to the CUG, will function as a *relay*, transferring the color of one branch to the other.

transferring the color of one side to the other in one of the following two ways: (1) either **j** becomes white and then, upon receiving a red- or green-colored message from a neighbor on one side, **j** changes it color and sends its new color down the other side, or (2) **j** first becomes green or red (due to receiving, respectively, a green or red message from a neighbor on one side) and then receives a white-colored message (∘) from a neighbor **w** residing on the other side, upon which—in an act analogous to *handshaking* in communication networks—**j** sends back its color to **w** which, in turn, initiates a chain reaction thereby **w** and its white-colored neighbors alter their color to that of **j** and so do their white-colored neighbors and so forth. This key understanding that joint vertices, as just explained, essentially serve as a relay transferring the color of one module to the other neighboring module, in addition to the fact that the color of the two source vertices are different, together, grants the conclusion that a clash between the colors green and red along the unblocked path $l^*$ eventually takes place. This concludes the proof of Part I.

**Proof of Part II:** As stated earlier, Part II concerns with showing the following: A message received by a node belonging to an unblocked path $l^*$ which is sent from a node lying outside $l^*$ cannot prevent the clash from happening on $l^*$. That is, informally, the occurrence of a clash cannot be prevented by any message coming from a node residing outside $l^*$ to a one belonging to $l^*$, say $\mathbf{n}_{\text{in}}$. We consider all the possible scenarios (i.e., scenarios (c1) to (c6) listed below) and show that indeed the claim of Part II holds true. Before we proceed further, let us introduce a notation. Let $\langle \alpha, \beta \rangle$ denote the following: $\mathbf{n}_{\text{in}}$'s current color is $\alpha$ and the color of the message (coming from a node residing outside $l^*$) destined to $\mathbf{n}_{\text{in}}$ is $\beta$. For example, $\langle \circ, \bullet \rangle$ implies that $\mathbf{n}_{\text{in}}$'s current color is white and the incoming message is red.

(c1) For $\langle \circ, \circ \rangle$, $\langle \bullet, \bullet \rangle$, $\langle \bullet, \bullet \rangle$: According to the CUG, if $\mathbf{n}_{\text{in}}$ receives a message whose color is identical to its current color, $\mathbf{n}_{\text{in}}$'s current color persists. Hence, the claim of Part II remains true under such circumstances.

(c2) For $\langle \bullet, \bullet \rangle$, $\langle \bullet, \bullet \rangle$: According to the CUG, these cases immediately lead to the occurrence of a clash. Hence, the claim of Part II remains valid under such circumstances.

(c3) For $\langle \bullet, \circ \rangle$, $\langle \bullet, \circ \rangle$: According to the CUG, if $\mathbf{n}_{\text{in}}$'s current color is green or red, it preserves its color upon receiving a white-colored message. Hence, the claim of Part II holds true under such circumstances.

(c4) For $\langle \varnothing, \bullet \rangle$, $\langle \circ, \bullet \rangle$: According to the CUG, if $\mathbf{n}_{\text{in}}$'s current color is white or $\mathbf{n}_{\text{in}}$ does not currently have any color, upon receiving a green-colored message, $\mathbf{n}_{\text{in}}$'s color becomes green and thereafter it acts as a green-colored source vertex for $l^*$.[4] This can only expedite the occurrence of a clash on $l^*$.

(c5) For $\langle \varnothing, \bullet \rangle$, $\langle \circ, \bullet \rangle$: The line of reasoning is similar to the one given for (c4).

(c6) For $\langle \varnothing, \circ \rangle$: According to the CUG, this interaction changes $\mathbf{n}_{\text{in}}$'s color to white. However, due to the machinery of $\mathcal{D}^*$, color white merely acts as a placeholder awaiting to be replaced by green or red upon interacting with one of the kind. In this light, altering the state of $\mathbf{n}_{\text{in}}$ from $\varnothing$ to white cannot prevent a clash from happening on $l^*$.

This concludes the proof of Part II and, together with Part I, concludes the proof of the forward direction of Statement (I). ■

### C-I.I.II Proof of Statement (I): Backward Direction

We prove the backward direction of Statement (I) next, using proof by contraposition. That is, we prove: For a given $d$-separation query $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C})_G$ and DAG $G$,

"$\mathbf{C}$ $d$-separates $\mathbf{A}$ from $\mathbf{B}$ in $G$" $\Rightarrow$ "Clash does not take place during $\mathcal{D}^*$'s execution".

According to (Pearl, 1988), the statement "$\mathbf{C}$ $d$-separates $\mathbf{A}$ from $\mathbf{B}$ in $G$" is equivalent to the following: Every path between any $\mathbf{a} \in \mathbf{A}$ and any $\mathbf{b} \in \mathbf{B}$ is blocked. According to (Pearl, 1988), a path $l$ is said to be blocked if and only if at least one of the two statements holds: (a2) There exists a collider node $\mathbf{n}$ on $l$ where neither $n$ nor any of $\mathbf{n}$'s descendants is in $\mathbf{C}$, (b2) There exists a non-collider node $\mathbf{m}$ on $l$ where $\mathbf{m} \in \mathbf{C}$. Therefore, altogether, the statement "$\mathbf{C}$ $d$-separates $\mathbf{A}$ from $\mathbf{B}$ in $G$" is equivalent to the statement that every path connecting $\mathbf{a} \in \mathbf{A}$ and $\mathbf{b} \in \mathbf{B}$ has to at least contain a *subpath* of the type specified in (a2) and (b2). Hence, for a clash to take place on path $l$, one of the colors green or red has to pass through $l$'s corresponding subpath and collide with the other color. In what follows, we consider all such subpaths and show that, the very existence of such subpaths on every path connecting $\mathbf{a} \in \mathbf{A}$ and $\mathbf{b} \in \mathbf{B}$, grants the impossibility of an occurrence of

[4] More specifically, once $\mathbf{n}_{\text{in}}$ becomes green it acts as a green-colored source vertex for the two *subpaths* of $l^*$ which lie at the two sides of $\mathbf{n}_{\text{in}}$ and share $\mathbf{n}_{\text{in}}$ as their common node. For example, for the path $\mathbf{v}_1 \leftarrow \mathbf{v}_2 \leftarrow \mathbf{v}_3 \rightarrow \mathbf{n}_{\text{in}} \rightarrow \mathbf{v}_4 \leftarrow \mathbf{v}_5$, the two subpaths are $\mathbf{v}_1 \leftarrow \mathbf{v}_2 \leftarrow \mathbf{v}_3 \rightarrow \mathbf{n}_{\text{in}}$ and $\mathbf{n}_{\text{in}} \rightarrow \mathbf{v}_4 \leftarrow \mathbf{v}_5$.

a clash during $\mathcal{D}^*$'s execution. These subpaths can be of three types: (1) the green node and the red node are separated by a head-to-tail node which is observed (Fig. 7.8(a)), (2) the green node and the red node are separated by a common cause (aka confounder) which is observed (Fig. 7.8(b)), and finally (3) the green node and the red node are separated by a common effect (aka collider) which is neither itself nor any of its descendants is observed (Fig. 7.8(c)).

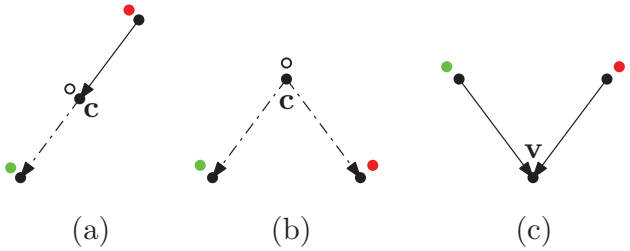



**Fig. 7.8** The three types of subpaths. Depicting the downlinks of a variable $\mathbf{c} \in \mathbf{C}$ in a dash-dotted format simply symbolizes a crucial property of $\mathcal{D}^*$ according to which $\mathbf{c}$ ignores any message received from any of its children, and also does not send any message to any of its children. **(a)** The green node and the red node are separated by a head-to-tail variable $\mathbf{c} \in \mathbf{C}$. **(b)** The green node and the red node are separated by a confounder $\mathbf{c} \in \mathbf{C}$. **(c)** The green node and the red node are separated by a collider $\mathbf{v}$ where neither $\mathbf{v}$ nor any of $\mathbf{v}$'s descendants is in the set $\mathbf{C}$.

Next, we consider each case at a time and prove that $\mathcal{D}^*$'s machinery prevents the occurrence of a clash along any of the aforesaid subpaths depicted in Figs 7.8(a-c). The proof for (1) and (2) immediately follows form the following crucial property of $\mathcal{D}^*$: Variables in $\mathbf{C}$ ignore any message received from any of their children, and also do not send any message to any of their children (depicting the outgoing edges from $\mathbf{c}$ in Figs 7.8(a-b) simply symbolizes this property). Case (3) requires more careful consideration. The only way for color green/red (on the one side) to reach color red/green (on the other side)—thereby generating a clash—was for the collider to be white-colored so that, by being replaced by either green or red, it would allow colors green and red to meet and hence a clash would occur. However, since (i) neither the collider nor any of its descendants is observed (and hence none of them are white), and also (ii) $\mathcal{D}^*$'s machinery dictates the propagation of the color white in a *backwards* manner through the corresponding ancestors of the white-colored nodes, altogether, the collider cannot become white during an execution of $\mathcal{D}^*$. This concludes the proof. ■

### C-I.II Proof of Statement (II)

The state transition diagram for $\mathcal{D}^*$ is given in Fig. 7.9. The states represent a node's color and the edges represent transitions due to receiving messages whose colors are depicted on the edges.

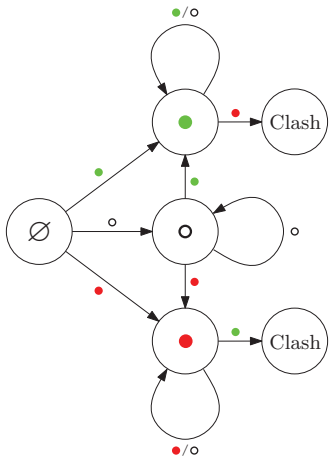



**Fig. 7.9** State transition diagram. The message which ought to be received for a transition to take place is depicted on the corresponding edge. In case multiple messages engender the same transition, they are all detailed on the corresponding edge separated by slashes.

A simple inspection of the diagram reveals that a node's color cannot alternate between any two states. This is due to the fact that the diagram has no cycles of length two or greater. This observation implies that either a clash takes place upon which $\mathcal{D}^*$ decides that the input $d$-separation query is false, or a state of equilibrium will eventually be reached. By definition, equilibrium is a global state of a network $G$ according to which none of the nodes in $G$ alters its state (i.e., its color) once that state is reached. This concludes the proof of Statement (II).

### C-I.III Proof of Statement (III)

In the analysis to follow, we adhere to Assumptions (ASM-1) and (ASM-2) presented in the first paragraph of Appendix C. We analyze all potential post-equilibrium, in-transit messages.[5] An in-transit message can be of three colors: (a1) green, (b1) red, or (c1) white.

[5]Cast into Lamport's *space-time diagram* (Lamport, 1978), these are the messages that cross a vertical *time-cut* positioned at a (global) time which is after the occurrence of the state of equilibrium; see (Mattern, 1987).

We consider each possibility next. Case (a1): If the state of equilibrium has indeed been reached, a green-colored in-transit message must be destined to a green-colored node. Indeed, if the green-colored in-transit message were destined to a red-, white-, or $\varnothing$-colored node, it would lead, respectively, to a clash, a change in the color of the destination node, and once again, a change in the color of the destination node—all of which are in contradiction with the assumption that the state of equilibrium has already been reached. According to $\mathcal{D}^*$, therefore, a green-colored in-transit message will be absorbed by the corresponding destination node (which is of the same color) in time at most $\beta$ leading to the generation of no new messages. The same line of reasoning can be adopted to conclude the following (Case (b1)): A red-colored in-transit message will be absorbed by the corresponding destination node (which is of the same color) in time at most $\beta$ leading to the generation of no new messages. Next, we consider the possibility of an in-transit message being white. A white-colored in-transit message could be destined to: (a2) a white-colored node, (b2) a green-colored node, or (c2) a red-colored node. (A white-colored in-transit cannot be destined to an $\varnothing$-colored node, as it would lead to a change in the color of the destination node—contradicting with the equilibrium assumption.) We consider each possibility in order. Case (a2): A white-colored in-transit message which is destined to a white-colored node reaches its destination in time at most $\beta$ and, according to $\mathcal{D}^*$, will be absorbed upon reception leading to the generation of no new messages. Case (b2): A white-colored in-transit message from node $\mathbf{x}$ to a green-colored node $\mathbf{g}$ reaches its destination, $\mathbf{g}$, in time at most $\beta$ and, according to Step (i) of $\mathcal{D}^*$, $\mathbf{g}$ replies, in time at most $\alpha$, by sending a green-colored message to $\mathbf{x}$ which, according to Case (a1), will be absorbed by $\mathbf{x}$ without generating any further new messages. (Note that, according to the CUG, the receipt of a white-colored message by a green-colored node does not lead to any color update, and hence $\mathbf{g}$ does not generate any messages due to Step (ii) of $\mathcal{D}^*$.) Case (c2) can be handled in the same manner as Case (b2). ■

## C-II Time-Complexity Analysis of $\mathcal{D}^*$

We present the results in the form of two lemmas as follows.

**Lemma C.1.** *For a given DAG $G$ and disjoint sets $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$, if $(\boldsymbol{A} \not\perp \boldsymbol{B}|\boldsymbol{C})_G$ (hence, a* NO*-instance d-separation query), then $\mathcal{D}^*$'s execution grants that a clash of the kind stated in Remark 1 occurs in $O(l_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})})$ time where $l_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ denotes the length of the longest undirected path in the ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$.*

**Proof.** The proof relies on the high-level understanding of $\mathcal{D}^*$'s machinery as discussed in Sec. 5.3.1, and Statements (1) and (2) of Proposition 5.1 (see Sec. C-VI of Appendix C for the proof). To obtain an upper bound on the time it takes for the clash to happen, we perform the propagation of colors through the DAG $G$ in two phases as follows. Phase-I: Starting at the nodes in $\mathbf{C}$, color white ($\circ$) propagates backwards through the DAG $G$. Phase-I ensures that all the nodes in $G$ which could potentially become white in the absence of colors red and green in the graph, indeed become white. Adopting (ASM-1) and (ASM-2) and the notation introduced therein, Phase-I is completed by time $(\alpha+\beta)l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$ where $l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$ denotes the longest directed path in $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$. Hence, Phase-I takes $O(l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ time. Phase-II: colors green ($\bullet$) and red ($\bullet$) (corresponding to the nodes in $\mathbf{A}$ and $\mathbf{B}$, respectively) will be introduced back into $G$ and begin to propagate through $G$ as dictated by the machinery of $\mathcal{D}^*$ until along some path between a node in $\mathbf{A}$ and a node in $\mathbf{B}$ a clash takes place.[6] Adopting (ASM-1) and (ASM-2) and the notation introduced therein, after the completion of Phase-I, within time $(\alpha+\beta)l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$ a clash takes place on a path between a node in $\mathbf{A}$ and a node in $\mathbf{B}$ where $l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$ denote the length of the longest undirected path in $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$. Hence, putting Phase-I and Phase-II together, by time $(\alpha+\beta)(l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}+l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ a clash takes place. Note that the parameters $\alpha$ and $\beta$ are arbitrary but finite constants. Since for any DAG $G$, $l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} \geq l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, the claimed upper bound $O(l_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ follows. ■

Using the above line of reasoning, we can prove the following lemma.

**Lemma C.2.** *For a given DAG $G$ and disjoint sets $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$, if $(\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C})_G$*

[6] Introducing colors green ($\bullet$) and red ($\bullet$) back into the DAG $G$ should be interpreted as follows: Through exerting external signals, the colors of the nodes in $\mathbf{A}$ and $\mathbf{B}$ are altered to green and red, respectively. The provided interpretation is equivalent to endowing each node $\mathbf{n} \in \mathbf{A}\cup\mathbf{B}$ with a dummy child $\mathbf{n}_{ext}$ and having $\mathbf{n}_{ext}$ colored (instead of $\mathbf{n}$) green (if $\mathbf{n} \in \mathbf{A}$) or red (if $\mathbf{n} \in \mathbf{B}$) in the initialization phase of $\mathcal{D}^*$, thereby making any node in $\mathbf{n} \in \mathbf{A}\cup\mathbf{B}$ inherit its red/green color from its newly introduced dummy child instead of being initialized by the corresponding color in the initialization phase of $\mathcal{D}^*$. By this construction, we purposefully delay the occurrence of a clash.

*(hence, a* YES*-instance query), then $\mathcal{D}^*$'s execution grants that a state of equilibrium will be reached in $O(l_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})})$ time where $l_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$ denotes the length of the longest undirected path in ancestral graph $G_{An(\boldsymbol{A}\cup\boldsymbol{B}\cup\boldsymbol{C})}$.*

Note that, in the context of Lemma C.2, the inevitability of equilibrium state follows from Statement (1) of Proposition 5.1 and Statement II given in Sec. C-I of Appendix C.

## C-III Proof of Proposition 5.2

The analysis presented next follows the same line of reasoning presented in the proof of Lemma C.1 in Sec. C-II of Appendix C. For the time-complexity analysis presented blow we adhere to Assumptions (ASM-1) and (ASM-2) outlined in the first paragraph of Appendix C.[7] Note that both the parameters $\alpha$ and $\beta$ are arbitrary but finite constants. The claimed upper-bound $O\big(l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + \min_{i,j} l^{ij}_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}\big)$ follows from the following two statements: (s1) By time $(\alpha+\beta)l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, all the nodes in $G$ which could potentially become white in the absence of colors red and green in the graph, become white, and (s2) After the completion of (s1), a clash takes place on the shortest unblocked path between $\mathbf{a}_i$ and $\mathbf{b}_j$ in the ancestral graph $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, within time $(\alpha+\beta)l^{ij}_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, $\forall i,j$ (note that, according to the proof of Statement (1) of Proposition 1, on any unblocked path between $\mathbf{a}_i$ and $\mathbf{b}_j$ a clash eventually takes place). Hence, by time $(\alpha+\beta)(l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + l^{ij}_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})})$ a clash will have occurred. Note that (s2) holds for all $i,j : 1 \leq i \leq |\mathbf{A}|, 1 \leq j \leq |\mathbf{B}|$. Also note that (s1) and (s2) correspond, respectively, to Phase-I and Phase-II presented in the proof of Lemma C.1 in Sec. C-II of Appendix C. From (s1) and (s2) follows the claimed upper-bound $O\big(l^d_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})} + \min_{i,j} l^{ij}_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}\big)$ in Proposition 5.2. ■

[7]Note that, as the number of messages exchanged by $\mathcal{D}^*$ on an edge is $O(1)$ (see Statement (5) of Proposition 1), the effect of pileups (aka congestion) on a channel has been considered in Assumptions (ASM-1) and (ASM-2).

## C-IV Proof of Proposition 5.3

The analysis presented next follows the same line of reasoning presented in the proof of Proposition 5.2 (see Sec. C-III for the proof). For the time-complexity analysis presented blow we adhere to Assumptions (ASM-1) and (ASM-2) outlined in the first paragraph of Appendix C; see also footnote 7. Let $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ denote a refutation-module for $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, with $E_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}$ denoting the set of the edges of $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$. Let $l^d_{\mathcal{M}}$ and $|P_{\mathcal{M}}|$ denote, respectively, the length of the longest directed path and the shortest unblocked path in $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$. (Recall that, according to Lemma 5.1 in the main text, DAG $G$ must contain at least one refutation-module for $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, and, due to Definition 5.1 in the main text, $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ must contain at least one unblocked path between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$.) Also, let $\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ denote the minimal refutation-module for $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_G$, with $E_{\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}$ denoting the set of the edges of $\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$, and $l^d_{\mathcal{M}^*}$ and $|P_{\mathcal{M}^*}|$ denoting the length of the longest directed path and the shortest unblocked path in $\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$, respectively. The claimed upper-bound $O(\min_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}\{l^d_{\mathcal{M}} + |P_{\mathcal{M}}|\})$ follows from the following two statements: (s1) By time $(\alpha+\beta)l^d_{\mathcal{M}}$, all the nodes in $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ which could potentially become white in the absence of colors red and green in the graph, become white, and (s2) After the completion of (s1), a clash takes place along the shortest unblocked path $P_{\mathcal{M}}$ in $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ within $(\alpha+\beta)|P_{\mathcal{M}}|$. (Note that (s1) and (s2) correspond, respectively, to Phase-I and Phase-II presented in the proof of Lemma C.1 in Sec. C-II of Appendix C.) Hence, taken together, a clash will have occurred along $P_{\mathcal{M}}$ by time $(\alpha+\beta)(l^d_{\mathcal{M}} + |P_{\mathcal{M}}|)$. Since the above argument holds for any arbitrary refutation-module $\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$, it follows that a clash will have occurred by time $(\alpha+\beta)\min_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}\{l^d_{\mathcal{M}} + |P_{\mathcal{M}}|\}$, hence the claimed upper-bound $O(\min_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}\{l^d_{\mathcal{M}} + |P_{\mathcal{M}}|\})$ in Proposition 5.3. Finally, since $\min_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}(l^d_{\mathcal{M}} + |P_{\mathcal{M}}|) \leq l^d_{\mathcal{M}^*} + |P_{\mathcal{M}^*}|$, $l^d_{\mathcal{M}^*} \leq |E_{\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}|$, and $|P_{\mathcal{M}^*}| \leq |E_{\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}|$, it follows that $\min_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}\{l^d_{\mathcal{M}} + |P_{\mathcal{M}}|\} \leq O(|E_{\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}|)$, hence the claimed upper-bound $O(|E_{\mathcal{M}^*_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}|)$ on the time for the occurrence of a clash. This concludes the proof. ■

## C-V Proof of Proposition 5.4

Consider any refutation-module $\mathcal{M}^{\dagger}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ satisfying the following condition: $\mathcal{M}^{\dagger}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}$ contains the unblocked path $\min_{i,j} l^{ij}_{An(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$. By definition, it immediately follows that $|P_{\mathcal{M}^{\dagger}}| \leq \min_{i,j} l^{ij}_{An(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$ and $l^{d}_{\mathcal{M}^{\dagger}} \leq l^{d}_{An(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$; for the notation, see Propositions 5.2 and 5.3 in the main text. Hence, $l^{d}_{\mathcal{M}^{\dagger}} + |P_{\mathcal{M}^{\dagger}}| \leq l^{d}_{An(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})} + \min_{i,j} l^{ij}_{An(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$. Since $\min_{\mathcal{M}_{(\mathbf{X}\not\perp\mathbf{Y}|\mathbf{Z})_G}}\{l^{d}_{\mathcal{M}} + |P_{\mathcal{M}}|\} \leq l^{d}_{\mathcal{M}^{\dagger}} + |P_{\mathcal{M}^{\dagger}}|$, the claim of Proposition 5.4 follows. ■

## C-VI Proof of Proposition 5.1

### Proof of Statement (1)

The reader is referred to Sec. C-I.I of Appendix C for the proof.

### Proof of Statement (2)

Let $\mathbf{I}_T$ denote the set of all nodes $\mathbf{i}$ in $G$ which have ever sent a message up to a (global) time $T$. Then, the validity of Statement (2) follows from the following recursion. If **(†)** Prior to applying Steps (i) and (ii) on the corresponding recipients of $\mathbf{i}$'s messages, $\mathbf{i} \in \mathbf{I}_T$, the set $\mathbf{I}_T$ satisfies the following two conditions: ($\star$) Any node in the set belongs to $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, and ($\star\star$) Any node to which a member of the set has ever sent a message belongs to $G_{An(\mathbf{A}\cup\mathbf{B}\cup\mathbf{C})}$, then **(‡)** After applying Steps (i) and (ii) on the corresponding recipients of $\mathbf{i}$'s messages (denoted by the set $recp(\mathbf{i})$) for all $\mathbf{i} \in \mathbf{I}_T$, the set $(\cup_{\mathbf{i}\in\mathbf{I}_T} recp(\mathbf{i})) \cup \mathbf{I}_T$ indeed satisfies ($\star$) and ($\star\star$). Note that at the initial configuration of $\mathcal{D}^*$, nodes in $\mathbf{A}\cup\mathbf{B}\cup\mathbf{C}$ send their corresponding colors to their parents. Hence, Statement (†) holds at the initial configuration. Let us now assume that Statement (†) holds for the set of all nodes $\mathbf{i}$ in $G$ that have ever sent a message up to a (global) time $T$, i.e., $\mathbf{i} \in \mathbf{I}_T$. According to Step (i) of $\mathcal{D}^*$, a node $\mathbf{x}$ which is the recipient of a message from $\mathbf{i}$ replies back by sending its own color to the sender. Note that, according to Statement (†), the sender must adhere to ($\star$) and ($\star\star$). According to Step (ii) of $\mathcal{D}^*$, the node $\mathbf{x}$ sends its updated color to (a) all its parents, and (b) those children of $\mathbf{x}$ with which $\mathbf{x}$ has communicated before. Based on the argument provided above regarding Step (i), $\mathbf{x}$ sending messages to its parents guarantees that the set $(\cup_{\mathbf{i}\in\mathbf{I}_T} recp(\mathbf{i})) \cup \mathbf{I}_T$ indeed satisfies ($\star$) and ($\star\star$). Also, due to the constraints "with which $\mathbf{x}$ has communicated before" in (b) and Statement ($\star\star$), $\mathbf{x}$ sending messages to the nodes prescribed in (b) guarantees that the set $(\cup_{\mathbf{i}\in\mathbf{I}_T} recp(\mathbf{i})) \cup \mathbf{I}_T$ indeed satisfies ($\star$) and ($\star\star$). The above argument establishes the validity of the recursion. The recursion given above, together with the fact that Statement (†) holds at the initial configuration of $\mathcal{D}^*$, grants the validity of Statement (2). This concludes the proof. ■

### Proof of Statement (3)

Statement (3) follows from Statement (I) (see Sec. C-I.I of Appendix C), Lemma C.1 and Lemma C.2 (see Sec. C-II of Appendix C).

### Proof of Statement (4)

The reader is referred to Sec. C-I.III of Appendix C for the proof.

### Proof of Statement (5)

In what follows we prove that the number of messages exchanged on an edge is bounded above by a constant which is independent of the size of the graph. Based on Step (ii) of $\mathcal{D}^*$, a node sends out a message upon updating its color and discovering that its new color is different from its $pre - update^T$ color. According to the state transition diagram depicted in Fig. 7.9, throughout an execution of $\mathcal{D}^*$ a node could pass through at most five states $\{\varnothing, \circ, \bullet, \bullet, clash\}$, and once a node changes its state it cannot go back to that state ever again (due to the nonexistence of any loop of the length at least two in the state transition diagram, see Fig. 7.9). Hence, Step (ii) of $\mathcal{D}^*$ results in $O(1)$ messages to be exchanged per channel. There are $\binom{5}{2} = 10$ possible ways of pairing nodes of different colors together. By inspection, Step (i) of $\mathcal{D}^*$ results in having the highest number of messages exchanged on the edge between a white-colored node and a green-colored node, which, as we show, results in $O(1)$ messages to be exchanged on that edge. When a (newly) white-colored node **p** sends a white-colored message to its green-colored neighbor **q** the following exchange of messages takes place due to Step (i): white-colored message (sent by the white-colored initiator node **p**) will be received by the green-colored node **q**; **q** will send back a green-colored message (due to Step (i)). Upon receipt of the green-colored message by **p**, **p** will send a new white-colored message to **q** (due to Step (i)) and also will updates its color to green based on the CUG. Upon receipt of the new white-colored message, **q** will send back a green-colored message (due to Step (i)) to **p**. However, since **p**'s color has been updated to green (hence, identical to that of the received message), this time **p** does not send any message to **q** due to Step (i). This completes the proof. ■

### Proof of Statement (6)

Throughout an execution of $\mathcal{D}^*$ there exist three types of messages which can be exchanged between nodes, namely, a white-, a green-, or a red-colored message. Therefore, to encode the said three types two bits are required. Statement (6) then follows from Statement (5) and the argument provided above. ■

## C-VII Alternative, Centralized Initialization of $\mathcal{D}^*$

Below, we explain how the initialization phase of $\mathcal{D}^*$ can be accomplished, in a distributed manner, in $O(l)$ time, where $l$ denotes the length of the longest undirected path in $G$.

First, using a combination of broadcast and convergecast, a pre-assigned initiator node, **s**, sends the initialization message $\langle$`INITIALIZE`$\rangle$ (containing the list of nodes in sets **A**, **B**, and **C**) to all the other nodes in $G$, and receives an acknowledgement that all nodes have received $\langle$`INITIALIZE`$\rangle$. Using the *AsynchSpanningTree* algorithm in (Lynch, 1996), this can be done in time $O(l)$; cf. (Lynch, 1996, p. 499, 2nd paragraph).

Then, **s** broadcasts the control message, $\langle \texttt{START}_{\mathcal{D}^*} \rangle$, to all the nodes in $G$. Using the *AsynchSpanningTree* algorithm, this can be done in time $O(D)$, where $D$ denotes the diameter of $G$. Upon receipt of $\langle \texttt{START}_{\mathcal{D}^*} \rangle$ by a node in $\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}$, it sends its color to its parents, as prescribed in the initialization phase of $\mathcal{D}^*$ outlined in Sec. 5.3 in the main text.

## C-VIII On the Order-Invariance Property

Before we state the result in a form of a lemma, let us introduce the following notation. We adopt the expression $S_i^{\mathbf{x}} \overset{M_1,\cdots,M_n}{\rightsquigarrow} S_f^{\mathbf{x}}$ to state the following: Starting at the state $S_i^{\mathbf{x}}$, node $\mathbf{x}$ transits to the state $S_f^{\mathbf{x}}$ upon receiving the sequence of messages $M_1, \cdots, M_n$ where $M_1, \cdots, M_n \in \{\circ, \bullet, \bullet\}$ and $S_i^{\mathbf{x}}, S_f^{\mathbf{x}} \in \{\varnothing, \circ, \bullet, \bullet, clash\}$ . Let us now formally state the result as a lemma.

**Lemma C.3.** *Let* $\mathbf{x}$ *be a node in the network. Then, the following holds:*

$$(S_i^{\mathbf{x}} \overset{M_1,M_2}{\rightsquigarrow} S_f^{\mathbf{x}}) \Rightarrow (S_i^{\mathbf{x}} \overset{M_2,M_1}{\rightsquigarrow} S_f^{\mathbf{x}}).$$

**Proof.** The proof can be straightforwardly accomplished by examining all the possible cases and showing that the statement holds true for all of them. To provide a sample case, consider $\circ \overset{\circ,\bullet}{\rightsquigarrow} \bullet$. Using the CUG given in Sec. 5.3, it is straightforward to check that $\circ \overset{\bullet,\circ}{\rightsquigarrow} \bullet$ holds true. ■

The order-invariance property is captured in the following lemma.

**Lemma C.4** *Let* $\mathbf{x}$ *be a node in the network and let* $\pi$ *be an arbitrary permutation defined on the set* $\{1, 2, \cdots, n\}$*. Then, the following holds:*

$$(S_i^{\mathbf{x}} \overset{M_1,\cdots,M_n}{\rightsquigarrow} S_f^{\mathbf{x}}) \Rightarrow (S_i^{\mathbf{x}} \overset{M_{\pi(1)},\cdots,M_{\pi(n)}}{\rightsquigarrow} S_f^{\mathbf{x}})$$

**Proof.** The proof follows from Lemma C.3 and the understanding that the sequence $M_{\pi(1)}, \cdots, M_{\pi(n)}$ (for an arbitrary permutation $\pi$) can be constructed from the original sequence $M_1, \cdots, M_n$ through a series of pairwise permutations.[8] ■

It is worth noting that the order-invariance property formalized above is analogous to the key notion of *exchangeability* in probability theory.

[8] This is essentially the idea behind the well-known sorting algorithm, Insertion Sort.

# Bibliography

Andersen, K. and Hooker, J. N. (1990). Probabilistic logic for belief nets. In *International Congress of Cybernetics and Systems, New York City*.

Andersen, K. A. and Hooker, J. N. (1994). Bayesian logic. *Decision Support Systems*, 11(2):191–210.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Psychology Press.

Aquinas, T. (1945). *Basic Writings of St. Thomas Aquinas,* A.C. Pegis (trans.). New York: Random House.

Arora, S. and Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Review Neuroscience*, 4(10):829–839.

Baddeley, A. D. and Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, 8:47–89.

Baker, A. (2016). Simplicity. *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*.

Baldea, M. and Daoutidis, P. (2006). Model reduction and control of reactor–heat exchanger networks. *Journal of Process Control*, 16(3):265–274.

Baluja, S. and Fahlman, S. E. (1994). Reducing network depth in the cascade-correlation learning architecture. Technical Report # CMU-CS-94-209, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Berwick, R. C. and Chomsky, N. (2015). *Why Only Us: Language and Evolution*. MIT press.

Boeckx, C. (2006). *Linguistic Minimalism: Origins, Concepts, Methods, and Aims*. Oxford University Press.

Bonawitz, E., Denison, S., Gopnik, A., and Griffiths, T. L. (2014a). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive Psychology*, 74:35–65.

Bonawitz, E., Denison, S., Griffiths, T. L., and Gopnik, A. (2014b). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, 18(10):497–500.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-specific independence in bayesian networks. In *Proceedings of the* $12^{th}$ *International Conference on Uncertainty in Artificial Intelligence*, pages 115–123. Morgan Kaufmann Publishers Inc.

Bramley, N. R., Gerstenberg, T., and Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 236–241.

Broadbent, D. (1965). Information processing in the nervous system. *Science*, 150(3695):457–462.

Brooks, D. and Baddeley, A. (1976). What can amnesic patients learn? *Neuropsychologia*, 14(1):111–122.

Buehner, M. J. (2014). Time and causality: editorial. *Frontiers in Psychology*, 5:228.

Buehner, M. J. and May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology Section B*, 57(2):179–191.

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7(11):e1002211.

Buntine, W. L. (1995). Chain graphs for learning. In *Proceedings of the* $11^{th}$ *Conference on Uncertainty in Artificial Intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc.

Butz, C. J., Dos Santos, A. E., and Oliveira, J. S. (2016). Relevant path separation: A faster method for testing independencies in bayesian networks. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models (PGM)*, pages 74–85.

Call, J., Carpenter, M., and Tomasello, M. (2005). Copying results and copying actions in the process of social learning: chimpanzees (pan troglodytes) and human children (homo sapiens). *Animal Cognition*, 8(3):151–163.

Carpenter, M., Call, J., and Tomasello, M. (2002). Understanding "prior intentions" enables two–year–olds to imitatively learn a complex task. *Child Development*, 73(5):1431–1441.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1997). Sensitivity analysis in discrete bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 27(4):412–423.

Chalmers, A. F. (2013). *What Is This Thing Called Science?* Hackett Publishing.

Chalup, S. K. and Wiklendt, L. (2007). Variations of the two-spiral task. *Connection Science*, 19(2):183–199.

Chan, H. (2005). *Sensitivity Analysis of Probabilistic Graphical Models*. University of California at Los Angeles, PhD Thesis.

Chan, H. and Darwiche, A. (2001). When do numbers really matter? In *Proceedings of the* $17^{th}$ *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 65–74. Morgan Kaufmann Publishers Inc.

Chandrasekaran, V., Srebro, N., and Harsha, P. (2012). Complexity of inference in graphical models. *arXiv preprint arXiv:1206.3240*.

Chater, N. and Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3(2):57–65.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Chomsky, N. (2000). Minimalist inquiries: The framework. Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik, In Roger Martin, David Michaels, and Juan Uriagereka (Eds.), 89–155, MIT Press.

Cichy, R. M., Heinzle, J., and Haynes, J.-D. (2012). Imagery and perception share cortical representations of content and location. *Cerebral Cortex*, 22(2):372–380.

Clark, C. W. and Dukas, R. (2003). The behavioral ecology of a cognitive constraint: limited attention. *Behavioral Ecology*, 14(2):151–156.

Coenen, A., Rehder, B., and Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, 79:102–133.

Cohen, R., Havlin, S., and Ben-Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. MIT Press Cambridge.

Cummins, D. D. (2014). The impact of disablers on predictive inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6):1638.

Dasgupta, I., Schulz, E., and Gershman, S. J. (2016). Where do hypotheses come from? Center for Brains, Minds and Machines (CBMM) Memo No. 056.

Dasgupta, I., Schulz, E., Goodman, N. D., and Gershman, S. J. (2017). Amortize hypothesis generation. In *Proceedings of the Cognitive Science Society*.

De Beeck, H. P. O., Haushofer, J., and Kanwisher, N. G. (2008). Interpreting fmri data: maps, modules and dimensions. *Nature Reviews Neuroscience*, 9(2):123.

De Neys, W., Schaeken, W., and d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, 31(4):581–595.

Deneve, S. (2008a). Bayesian spiking neurons I: Inference. *Neural Computation*, 20(1):91–117.

Deneve, S. (2008b). Bayesian spiking neurons II: Learning. *Neural Computation*, 20(1):118–145.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222.

Dijkstra, E. W., Feijen, W. H., and Van Gasteren, A. M. (1983). Derivation of a termination detection algorithm for distributed computations. *Information Processing Letters*, 16(5):217–219.

Dukas, R. (2004). Causes and consequences of limited attention. *Brain, Behavior and Evolution*, 63(4):197–210.

Dukas, R. and Kamil, A. C. (2001). Limited attention: the constraint underlying search image. *Behavioral Ecology*, 12(2):192–199.

Edwards, B. J., Burnett, R. C., and Keil, F. C. (2015). Effects of causal structure on decisions about where to intervene on causal systems. *Cognitive Science*, 39(8):1912–1924.

Ericsson, K. A. and Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2):211.

Fahlman, S. E. and Lebiere, C. (1989). The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems*, pp. 524—532.

Fernando, C. (2013). From blickets to synapses: Inferring temporal causal networks by observation. *Cognitive Science*, 37(8):1426–1470.

Fernbach, P. M., Darlow, A., and Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2):168–185.

Fernbach, P. M. and Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1):64–88.

Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres.

Galbiati, M., Delpini, D., and Battiston, S. (2013). The power to control. *Nature Physics*, 9(3):126–128.

Galilei, G. (1967/1632). *Dialogue Concerning The Two Chief World Systems*. University of California Berkeley Press.

Gao, J., Liu, Y.-Y., D'Souza, R. M., and Barabási, A.-L. (2014). Target control of complex networks. *Nature Communications*, 5, 5415.

Geiger, D. and Heckerman, D. (1991). Advances in probabilistic reasoning. In *Proceedings of the $7^{th}$ Conference on Uncertainty in Artificial Intelligence*, pages 118–126. Morgan Kaufmann Publishers Inc.

Geiger, D., Verma, T., and Pearl, J. (1989). d-separation: From theorems to algorithms. In *Proceedings of $5^{th}$ Workshop on Uncertainty in Artificial Intelligence*, pages 118–125.

George, D. and Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology*, 5(10):e1000532.

Gershman, S. and Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the Cognitive Science Society*.

Gershman, S. J. and Beck, J. M. (2017). Complex probabilistic inference: From cognition to neural computation. In A. Moustafa (Ed.) *Computational Models of Brain and Behavior.* Wiley-Blackwell.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, 24(1):1–24.

Glymour, C. (1987). Android Epistemology and the Frame Problem, *The Robot's Dilemma: The Frame Problem in AI.* pp. 65–75.

Glymour, C. (2003). Learning, prediction and causal bayes nets. *Trends in Cognitive Sciences*, 7(1):43–48.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological Review*, 111(1):3–32.

Greville, W. J. and Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, 139(4):756.

Griffiths, T., Levy, R., McKenzie, C. R., Steyvers, M., Tenenbaum, J., and Vul, E. (2009). Rational process models. In *Proceedings of the Cognitive Science Society.*

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229.

Griffiths, T. L., Vul, E., and Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4):263–268.

Grill-Spector, K. and Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27:649–677.

Hansen, P., Jaumard, B., Nguetse, G.-B. D., and De Aragao, M. P. (1995). Models and algorithms for probabilistic and bayesian logic. In *Proceedings of the* $14^{th}$ *International Joint Conference on Artificial Intelligence*, pages 1862–1868.

Hassabis, D., Kumaran, D., Vann, S. D., and Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731.

Horner, V. and Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens). *Animal Cognition*, 8(3):164–181.

Horvitz, E. J. (1990). *Computation and Action under Bounded Resources*. PhD Dissertation, Stanford University.

Hume, D. (1748/1975). *An Inquiry Concerning Human Understanding*. Oxford University Press.

Icard, T. F. and Goodman, N. D. (2015). A resource-rational approach to the causal frame problem. *Proc. of the 37th Annual Meeting of the Cognitive Science Society*.

Ishai, A., Ungerleider, L. G., and Haxby, J. V. (2000). Distributed neural systems for the generation of visual images. *Neuron*, 28(3):979–990.

Jern, A. and Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, 66(1):85–125.

Kant, I. (2013/1781). *Immanuel Kant's Critique of Pure Reason*. Read Books Ltd.

Kastner, S. and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, 23(1):315–341.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Korb, K. B., Hope, L. R., Nicholson, A. E., and Axnick, K. (2004). Varieties of causal intervention. In *Proceedings of Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 322–331. Springer.

Krynski, T. R. and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3):430.

Kshemkalyani, A. D. and Singhal, M. (2008). *Distributed Computing: Principles, Algorithms, and Systems*. Cambridge University Press.

Kuchtey, J., Fulton, S. A., Reba, S. M., Harding, C. V., and Boom, W. H. (2006). Interferon-$\alpha\beta$ mediates partial control of early pulmonary mycobacterium bovis bacillus calmette–guérin infection. *Immunology*, 118(1):39–49.

Lagnado, D. A. and Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3):451–60.

Lamport, L. (1978). Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H. G. (1990). Independence properties of directed markov fields. *Networks*, 20(5):491–505.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting Structured Data.* MIT Press.

Lin, C. T. (1974). Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201–208.

Litvak, S. and Ullman, S. (2009). Cortical circuitry implementing graphical models. *Neural Computation*, 21(11):3010–3056.

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2011). Controllability of complex networks. *Nature*, 473(7346):167–173.

Lochmann, T. and Deneve, S. (2011). Neural processing as causal inference. *Current Opinion in Neurobiology*, 21(5):774–781.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759.

Lynch, N. A. (1996). *Distributed Algorithms*. Morgan Kaufmann.

Lyons, D. E., Young, A. G., and Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, 104(50):19751–19756.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Mahoney, S. M. and Laskey, K. B. (1998). Constructing situation specific belief networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 370–378.

Mareschal, D. and Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science*, 11(2):149–186.

Marr, D. (1982). *Vision: A Computational Approach*.

Mattern, F. (1987). Algorithms for distributed termination detection. *Distributed Computing*, 2(3):161–175.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In: Morris MGM (Ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology*. pages 8–45.

McGuigan, N. and Whiten, A. (2009). Emulation and "overemulation" in the social learning of causally opaque versus causally transparent tool use by 23-and 30-month-olds. *Journal of Experimental Child Psychology*, 104(4):367–381.

McGuigan, N., Whiten, A., Flynn, E., and Horner, V. (2007). Imitation of causally opaque versus causally transparent tool use by 3-and 5-year-old children. *Cognitive Development*, 22(3):353–364.

Meder, B., Mayrhofer, R., and Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, 121(3):277.

Mittal, N., Venkatesan, S., and Peri, S. (2004). Message-optimal and latency-optimal termination detection algorithms for arbitrary topologies. In *Proceedings of International Symposium on Distributed Computing (DISC)*, pages 290–304. Springer.

Mittal, N., Venkatesan, S., and Peri, S. (2007). A family of optimal termination detection algorithms. *Distributed Computing*, 20(2):141–162.

Mohan, K. and Pearl, J. (2014). On the testability of models with missing data. In *Proceedings of Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 643–650.

Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:342–345.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496.

Nagell, K., Olguin, R. S., and Tomasello, M. (1993). Processes of social learning in the tool use of chimpanzees (pan troglodytes) and human children (homo sapiens). *Journal of Comparative Psychology*, 107(2):174.

Neapolitan, R. E. (2009). *Probabilistic Methods for Bioinformatics: With an Introduction to Bayesian Networks*. Morgan Kaufmann.

Newton, I. (1964/1687). *The Mathematical Principles of Natural Philosophy (Principia Mathematica)*. New York: Citadel Press.

Nielsen, M. (2006). Copying actions and copying outcomes: social learning through the second year. *Developmental Psychology*, 42(3):555.

Nielsen, M. (2012). Imitation, pretend play, and childhood: essential elements in the evolution of human culture? *Journal of Comparative Psychology*, 126(2):170.

Nielsen, M. and Tomaselli, K. (2010). Overimitation in kalahari bushman children and the origins of human cultural cognition. *Psychological Science*.

Oakes, L. M., Cashon, C., Casasola, M., and Rakison, D. (2011). *Infant Perception and Cognition: Recent Advances, Emerging Theories, and Future Directions*. Oxford University Press, USA.

O'Craven, K. M. and Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, 12(6):1013–1023.

Park, J. and Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the markov property in causal reasoning. *Cognitive Psychology*, 67(4):186–216.

Pashler, H., Johnston, J. C., and Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, 52(1):629–651.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the $7^{th}$ Conference of the Cognitive Science Society, 1985*, pages 329–334.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (1990). Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4(5-6):363–389.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pearl, J. and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

Pecevski, D., Buesing, L., and Maass, W. (2011). Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology*, 7(12):e1002294.

Polak, B. (2007). Handout on mixed strategies. ECON-159: Open Yale Courses.

Rao, R. P. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, 16(1):1–38.

Reddy, L. and Kanwisher, N. (2007). Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. *Current Biology*, 17(23):2067–2072.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, 72:54–107.

Rehder, B. (2016). Beyond markov: Accounting for independence violations in causal reasoning. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 1853–1858.

Rehder, B. and Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, 45:245—-260.

Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Russell, S., Binder, J., Koller, D., and Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables. In *International Joint Conference of Artificial Intelligence (IJCAI)*, volume 95, pages 1146–1152. Citeseer.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, 94(1-2):57–77.

Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12):883–893.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, 117(4):1144.

Savin, C. and Deneve, S. (2014). Spatio-temporal representations of uncertainty in spiking neural networks. In *Advances in Neural Information Processing Systems*.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: how children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, 35(1):303.

Shachter, R. D. (1988). Probabilistic inference and influence diagrams. *Operations Research*, 36(4):589–604.

Shachter, R. D. (1998). Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 480–487. Morgan Kaufmann Publishers Inc.

Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science*, 1(1):103–126.

Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. *Processes of Change in Brain and Cognitive Development: Attention and Performance*, 21:61–86.

Shultz, T. R. and Bale, A. C. (2006). Neural networks discover a near-identity relation to distinguish simple syntactic forms. *Minds and Machines*, 16(2):107–139.

Shultz, T. R., Mareschal, D., and Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning*, 16(1-2):57–86.

Shultz, T. R. and Rivest, F. (2001). Knowledge-based cascade-correlation: Using knowledge to speed learning. *Connection Science*, 13(1):43–72.

Shultz, T. R. and Takane, Y. (2007). Rule following and rule use in the balance-scale task. *Cognition*, 103(3):460–472.

Shultz, T. R. and Vogel, A. (2004). A connectionist model of the development of transitivity. In *Proceedings of the* $26^{th}$ *Annual Conference of the Cognitive Science Society*, pages 1243–1248.

Simon, H. A. (1957). *Models of Man*. Wiley.

Sloman, S. A. and Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, 66:223–247.

Spiers, H. J., Maguire, E. A., and Burgess, N. (2001). Hippocampal amnesia. *Neurocase*, 7(5):357–382.

Stuhlmüller, A., Taylor, J., and Goodman, N. (2013). Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056.

Tel, G. (2000). *Introduction to Distributed Algorithms*. Cambridge University Press.

Thöne, H., Güntzer, U., and Kießling, W. (1992). Towards precision of probabilistic bounds propagation. In *Proceedings of the* $8^{th}$ *Conference on Uncertainty in Artificial Intelligence*, pages 315–322. Morgan Kaufmann Publishers Inc.

Tian, J. (2008). Identifying dynamic sequential plans. *Proceedings of the* $24^{th}$ *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 554–561.

Tomasello, M. (2009). *The Cultural Origins of Human Cognition*. Harvard University Press.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, 24(5):295–300.

Uzgiris, I. C. (1981). Two functions of imitation during infancy. *International Journal of Behavioral Development*, 4(1):1–12.

Want, S. C. and Harris, P. L. (2002). Social learning: Compounding some problems and dissolving others. *Developmental Science*, 5(1):39–41.

Westermann, G., Sirois, S., Shultz, T. R., and Mareschal, D. (2006). Modeling developmental cognitive neuroscience. *Trends in Cognitive Sciences*, 10(5):227–232.

White, P. A. (1997). Naive ecology: Causal judgments about a simple ecosystem. *British Journal of Psychology*, 88(2):219–233.

White, P. A. (2000). Naive analysis of food web dynamics: A study of causal judgment about complex physical systems. *Cognitive Science*, 24(4):605–650.

Whiten, A., Custance, D. M., Gomez, J.-C., Teixidor, P., and Bard, K. A. (1996). Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes). *Journal of Comparative Psychology*, 110(1):3.

Whiten, A., Hinde, R. A., Laland, K. N., and Stringer, C. B. (2011). *Culture Evolves*. The Royal Society.

Younger, B. A. and Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development*, pages 858–867.

Younger, B. A. and Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development*, pages 803–815.